Fair and Faithful Processing of Referring Expressions in English

Vagrant Gautam

A dissertation submitted towards the degree

Doctor of Engineering (Dr.-Ing.)

of the Faculty of Mathematics and Computer Science

of Saarland University

Saarbrücken, 2025

Vagrant Gautam

Fair and Faithful Processing of Referring Expressions in English

DAY OF COLLOQUIUM:

January 9th, 2026

DEAN OF THE FACULTY:

Prof. Dr. Roland Speicher

EXAMINATION BOARD:

Chair – Prof. Dr. Ingmar Weber

Advisor, Reviewer – Prof. Dr. Dietrich Klakow

Reviewer – Prof. Dr. Rachel Rudinger

Reviewer – Prof. Dr. Vera Demberg

Academic Assistant – Dr. Koel Dutta Chowdhury

Abstract

Names ("*Vagrant*"), pronouns ("*they*") and definite descriptions ("*the birder*") are examples of referring expressions, linguistic forms that point to referents. The complexity of reference lies in how we can contextually map the same referring expression to different individuals, and the same individual to different referring expressions. Thus, despite recent advances in natural language processing (NLP), faithfully resolving and doing reference still presents a significant challenge for systems that deal exclusively with linguistic form. Beyond denotational meanings, referring expressions can also have gendered and racial connotations, and are therefore widely used to measure social biases and fairness in society and NLP systems. This typically involves several simplifications that hamper valid and ethical fairness research, including the assumption that names map one-to-one to their referents' race and gender, and that the grammatical gender of English pronouns maps one-to-one to their referents' gender.

In this thesis, I tackle these issues and make a number of contributions towards fair and faithful computational processing of English referring expressions. First, I provide theoretical arguments informed by other disciplines to critique the validity of using pronouns and names as a proxy for sociodemographic factors such as gender. I empirically show that in the task of coreference resolution, this assumption can misrepresent system performance and bias, with a novel method to measure stereotypical bias in this context. Pivoting to language modelling next, I show that large language models can—in simple settings—frequently overcome stereotypical biases to do pronominal reference correctly just as humans do. Finally, with a controlled evaluation to disentangle true reasoning about reference from shallow repetition of referring expressions, I show that today's large language models are not up to the task of faithful reasoning about reference. The arguments in this thesis are of wide relevance to researchers and practitioners who work on fairness, reasoning, and reference, more broadly.

Zusammenfassung

Namen ("*Vagrant*"), Pronomen ("*er*") und eindeutige Beschreibungen ("*der Vogelbeobachter*") sind Beispiele für referierende Ausdrücke, sprachliche Formen, die auf Referenten verweisen. Ihre Komplexität liegt darin, dass wir denselben Ausdruck kontextuell auf verschiedene Personen und dieselbe Person auf verschiedene Ausdrücke beziehen können. Trotz der Fortschritte im Bereich Sprachverarbeitung stellt die getreue Auflösung und Ausführung von Referenzen daher immer noch eine große Herausforderung für Systeme dar, die sich ausschließlich mit sprachlichen Formen befassen. Neben denotationalen Bedeutungen können referenzierende Ausdrücke auch soziale Konnotationen haben und werden daher häufig zur Messung sozialer Vorurteile und Fairness verwendet. Dies beinhaltet in der Regel mehrere Vereinfachungen, die eine valide und ethische Fairnessforschung erschweren, darunter die Annahme, dass Namen eins-zu-eins auf die Ethnie und das Geschlecht ihrer Bezugspersonen abgebildet werden und dass das grammatikalische Geschlecht der englischen Pronomen eins-zu-eins auf das Geschlecht ihrer Bezugspersonen abgebildet wird.

In dieser Arbeit leiste ich eine Reihe von Beiträgen zu einer fairen und getreuen Verarbeitung von englischen Referenzausdrücken. Zunächst führe ich theoretische Argumente aus anderen Disziplinen an, um die Gültigkeit der Verwendung von Pronomen und Namen als Stellvertreter für soziodemografische Faktoren zu kritisieren. Empirisch zeige ich, dass diese Annahme bei der Aufgabe der Coreference-Resolution die Systemleistung und -verzerrung falsch darstellen kann. In einem nächsten Schritt zeige ich, dass große Sprachmodelle - in einfachen Situationen - häufig stereotype Verzerrungen überwinden können, um pronominalen Verweis korrekt auszuführen, so wie es Menschen tun. Abschließend zeige ich anhand einer kontrollierten Evaluierung, dass die heutigen großen Sprachmodelle nicht in der Lage sind, treffsichere Aussagen über die Referenz zu treffen.

Acknowledgements

Let us be grateful to the people who bring us joy;
they are the charming gardeners who make our souls blossom

— Marcel Proust

This thesis may have a single author, but neither the work in it nor I in my current form would exist without the people who have sustained me and helped me grow as a person and scientist. This section is about them.

To *Dietrich Klakow*, my advisor, I am thankful for the profound academic freedom he gave me in this position, which allowed me to do exactly the kind of work I find interesting and important, and very little else. Thanks for the dead bugs too!

To the rest of my committee, *Rachel Rudinger*, *Vera Demberg*, and *Ingmar Weber*: Thank you for your engagement with my work. At the intersection of fairness, reasoning, linguistics, machine learning, and computational social science, you are really the dream committee to give me this academic stamp of approval.

A big thank you to everyone who published with me during my PhD: *Elia Ovalle*, *Arjun Subramonian*, *Gilbert Gee*, *Kai-Wei Chang*, *Miaoran Zhang*, *Paloma García-de-Herreros*, *Philipp Slusallek*, *Marius Mosbach*, *Mingyang Wang*, *Jesujoba O. Alabi*, *Xiaoyu Shen*, *Anne Lauscher*, *Os Keyes*, *Zeerak Talat*, *Tomás Vergara-Browne*, *Mor Geva*, *Julius Steuer*, *Eileen Bingert*, *Ray Johns*, *Andreas Waldis*, *Iryna Gurevych*, *Preethi Seshadri*, *Yizhou Sun*, *Elisa Forcada Rodríguez*, *Olatz Perez-de-Viñaspre*, and *Jon Ander Campos*. I have learned a lot from my collaborators, and I wish to highlight *Anne Lauscher*, *Arjun Subramonian*, *Zeerak Talat*, and *Vlad Niculae* (a collaborator in spirit but not on paper), without whom I wouldn't be half the scientist I am today. I also want to thank my students, *Jingyan Chen*, *Eileen Bingert*, *Sambit Bhaumik*, and *Elisa Forcada Rodríguez*, for trusting me, working with me, and letting me learn with you.

For the papers that made it into this thesis, I am grateful to *Lucy Li* for literature recommendations for Chapter 4, *Timm Dill* for several rounds of annotation in Chapters 5, 6 and 7, and many others for feedback - all our reviewers, members of the *Critical Media Lab Basel* and the *Interdisciplinary Institute for Societal Computing* for Chapter 4, *Rachel Rudinger* and *Benjamin Van Durme* for Chapter 5, and finally, *Aaron Mueller*, *Marius Mosbach*, *Vlad Niculae*, *Yanai Elazar* and our action editor *Hai Zhao* for Chapters 6 and 7. These papers were partially funded by the BMBF's (German Federal Ministry of Education and Research) SLIK project under the grant 01IS22015C.

Of course, this thesis wasn't borne solely of intellectualism and labour. I also owe a debt to the people who are part of my social networks in Germany who have given me support in many ways beyond discussing work: *Kate Arkhangelskaia*, *Flora Gessner*, *Julian*, *Naya*, *Noa*, *Merlin*, *Vincent*, *Aude Benenson*, *Anna*, *Toni Mattheis*, *Can*, *Orion Junkers*, *Valerie*, *Tate*, *Diana Davidson*, *Isabell Landwehr*, and *Amira Palisch*. Also, shout out to the man who runs *Berliner Gemüse Döner* in Saarbrücken, as well as the people at *Unique Cafe* and *Comame*, for fuelling my PhD with halloumi and coffee.

My PhD was also enriched, enabled and made so much more fun by the people at LSV that I got to see almost every day for lunch, coffee, and cake breaks. I want to mention *Miaoran Zhang*, *Marius Mosbach*, *Julius Steuer*, *Paloma García-de-Herreros*, *Zena Al-Khalili*, *Koel Dutta Chowdhury*, *Nicolas Louis*, and *Jesujoba O. Alabi* here. *Miaoran*, thank you for collaborating with me on my first first-author paper, which gave me the courage and confidence to keep going, and for our bets that got me out of the house and to the office before 9:30 even on cold, dark, and rainy mornings; *Marius*, thank you for teaching me the most important things I learned during my PhD (about beer, sportsball, and Saarland); *Julius*, you've been a great co-parent to our coffee machine and tour guide for Roman stuff and Saarland's countryside; *Paloma*, thanks for the spirit of adventure and cheer you bring to everything you do, for political solidarity I desperately needed, and for your delicious focaccia; *Zena*, thanks for being the best at bullying me to write this damn thesis, and for all the Syrian food and good conversations you've given me; *Koel*, thanks for holding up a mirror to myself and feeding me *gajar ka halwa*; *JJ*, thanks for letting me drag you to new places and for playing football with

me for the first time in my life; *Nico*, thanks for all the metal recommendations and for putting up with my bugs, both real and tech-related.

Thank you to the people rooting for me from back in Canada: *Michael Wortis*, *Madeleine Smiciklas* and the entire *Beale-Smiciklas* clan, *Kaschelle Thiessen*, *Danica Reid*, *Ashley Farris-Trimble*, *Luna Cavasso*, *Louise Sauwan*, *Mike Fuchs*, *Maite Taboada*, *Brad Bart*, *Simon Vandieken*, *Bran Eveland Cron*, *Robert Beda*, *Siobhan Ennis*, *Kelly Ma*, and my therapist, *Monica*, as well as *ammamma* and my dear friends cheering me on from other places: *Arjun Subramonian*, *EJ Mason*, *Fred Mailhot*, *Dave Howcroft*, *Sy Brand*, *Pranav A*, *Sabrina Mielke*, *Maureen Kosse*, *Kriti Tripathi*, and *Vlad Niculae*. Thanks also to *Rupak Sarkar* for helping schedule my defense! I will definitely run out of space if I start to name all the people on the internet whose intellectual engagement and support I've benefited from, so let me just say I hugely appreciate the people (especially linguists) I've met online through fora including *Queer in AI*, *Twitter*, *Margiebomx*, the *NLP Shitposters' Caucus*, the *Lavender Languages Summer School*, *Twitterlings*, *da c00l zone*, *BlueSky*, and so on.

Lastly, my most important thank-yous go to the people who've kept me going through the years and who are my family. I am so grateful for the threads that connect us. *Sabrina*, thank you just for being there; *Kriti*, don't forget, you'll always be the more useful doctor! *Michael*, thank you for making this vagrant feel like they had a home they could always go to. I aspire to be as generous as you. *Mady*, I am so lucky to know someone as perfect, supportive and strong as you are. Thanks for being my best friend.

Contents

1

Introduction

Referring expressions are linguistic forms that point to real-world referents (e.g., names, pronouns, and definite descriptions), putting them at the heart and soul of the largely arbitrary system of communication we call natural language. Due to their significance in natural language, referring expressions are also central to the computational field of natural language processing (NLP). For NLP systems that exclusively see linguistic forms without the connections we humans typically make with their corresponding real-world referents, referring expressions present a particularly interesting challenge.

Referring expressions have thus been used in several, vastly different areas of NLP since the field's inception. Coreference resolution, the identification of expressions in natural language that have the same referent, is one of the oldest tasks in computational linguistics (Sukthanker et al., 2020).[1] Referring expression generation, or selecting the most contextually and semantically appropriate referring expression in a given context, is similarly considered one of the most well-developed areas of natural language generation (Krahmer and Deemter, 2012). More recently, NLP researchers have begun to use another feature of referring expressions such as names and pronouns, i.e., that they often index sociodemographic features of relevance in societies, such as gender. Thus, referring expressions, when used as textual proxies for sociodemographic factors, allow us to measure biases in society—e.g., to study representation in the ACL Anthology via names (Vogel and Jurafsky, 2012)—as well as in NLP systems—e.g., to study biased

1 Algorithms to resolve pronominal anaphora and their evaluations have been studied for nearly 20 years before the author of this thesis was even born (Hobbs, 1978; Lappin and Leass, 1994)!

associations in word embeddings via the pronouns *he* and *she* (Bolukbasi et al., 2016). Given the wealth of prior study on referring expressions, what then is left for a thesis in 2025 to contribute to this area?

1.1 Motivation

I see two main threads that present novel challenges with processing referring expressions: **(1)** The rise of transformer-based large language models, and **(2)** our understanding of the relationship between language and sociodemographic categories.

Large language models are neural-network based probabilistic models of language that are trained using vast quantities of text data. This simple recipe, enabled by advancements in hardware, contrasts with earlier decades of work in NLP, where task-specific architectures, rule-based models, and approaches with structural biases reigned. Now, a single large language model can be used for a variety of complex tasks involving language, including writing entire essays, summarizing papers, and answering questions. In this new paradigm, referring expression generation is no longer one component of a larger system of moving pieces involved in generation, since a single model is now responsible for everything. This therefore changes how we should think about the problem of selecting contextually and semantically appropriate referring expressions.

Orthogonally, most work on using referring expressions as a proxy for sociodemographic characteristics involves simplifying assumptions that have gone unquestioned. For instance, the connection between the referring expression and the characteristic (e.g., pronouns and gender) is generally assumed to be one-to-one. Additionally, it is often assumed that all referring expressions are informative of the characteristic of interest (e.g., names and race). These assumptions are widely contested in sociolinguistics (Bucholtz and Hall, 2005; Conrod, 2018) and in onomastics (Hough, 2016), but much research in NLP fails to account for these criticisms, with serious implications for the accurate measurement of social biases in society and NLP systems. This highlights the need for an interdisciplinary re-contextualization of such work.

1.2 Research Objectives

This thesis addresses the gaps in the literature described above, asking how we can fairly and faithfully process referring expressions in English. The high-level research questions I seek to answer are:

1. How valid is it to use referring expressions like pronouns and names as a proxy for social categories of interest such as gender?
2. How can we measure stereotypical biases in how NLP systems resolve references?
3. Can large language models overcome their stereotypical biases and do reference correctly?
4. When large language models do reference correctly, are they really reasoning about reference or shallowly repeating referring expressions?

Research questions 1 and 2 have to do with fair processing of referring expressions. Research question 1 interrogates, from an interdisciplinary theoretical standpoint, current practices in NLP fairness research, where referring expressions are very commonly used as proxies for sociodemographic categories such as gender. Is such work still valid under the criticisms from outside of NLP? Research question 2 seeks to apply the theoretical arguments of research question 1 to an empirical context, to quantify the impact of conflating pronouns and social gender when measuring stereotypical biases. Work in this area has fallen into similar patterns of simplification, and also does not disentangle bad performance from stereotypically biased performance.

Moving to faithful processing of referring expressions, research question 3 builds on the previous question and asks whether language models can overcome their stereotypical biases about reference in context, and choose the correct referring expression for someone. Finally, research question 4 probes this further, asking whether good performance in this context is due to true "reasoning" about reference, or just shallow copying, towards accurately representing language model capabilities when it comes to reference.

1.3 Contributions

This thesis takes a critical look at current practices in studying the processing of English referring expressions, by systematically addressing the research questions outlined above. Chapters 4 and 5 present theoretical limitations of using referring expressions as a proxy for categories of interest (RQ1), as well as recommendations for how to mitigate these limitations; specifically, Chapter 4 focuses on names as a proxy for sociodemographic categories such as gender and race, while Chapter 5 examines the use of third-person pronouns in English as a proxy for social gender. The latter chapter also tests these theoretical arguments empirically, in the context of measuring stereotypical biases in systems for coreference resolution (RQ2). To this end, this chapter also contributes a new dataset and an evaluation procedure that disentangles bad performance from stereotypically biased performance, two factors which have been conflated in prior work.

Chapters 6 and 7 pivot to how language models use referring expressions to do reference successfully, overcoming any stereotypical biases they might have (RQ3). This is measured with a new task, called pronoun fidelity, and a new dataset to measure it. While Chapter 6 considers a simple version of pronoun fidelity for just one referent, Chapter 7 attempts to disentangle whether good performance happens due to true "reasoning" about reference, or simply shallow repetition of pronouns (RQ4). We accomplish this by using a setting with two distinct referents, making the choice of the appropriate referring expression more challenging for language models.

The main content of this thesis is based on the following three publications:

Vagrant Gautam, Arjun Subramonian, Anne Lauscher, and Os Keyes (Aug. 2024c). "Stop! In the Name of Flaws: Disentangling Personal Names and Sociodemographic Attributes in NLP." In: *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Ed. by Agnieszka Faleska, Christine Basta, Marta Costa-jussà, Seraphina Goldfarb-Tarrant, and Debora Nozza. Bangkok, Thailand: Asso-

ciation for Computational Linguistics, pp. 323–337. DOI: 10.18653/v1/2024.gebnlp-1.20. URL: https://aclanthology.org/2024.gebnlp-1.20/

Vagrant Gautam, Julius Steuer, Eileen Bingert, Ray Johns, Anne Lauscher, and Dietrich Klakow (Nov. 2024b). "WinoPron: Revisiting English Winogender Schemas for Consistency, Coverage, and Grammatical Case." In: *Proceedings of The Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*. Ed. by Maciej Ogrodniczuk, Anna Nedoluzhko, Massimo Poesio, Sameer Pradhan, and Vincent Ng. Miami: Association for Computational Linguistics, pp. 52–66. DOI: 10.18653/v1/2024.crac-1.6. URL: https://aclanthology.org/2024.crac-1.6/

Vagrant Gautam, Eileen Bingert, Dawei Zhu, Anne Lauscher, and Dietrich Klakow (Dec. 2024a). "Robust Pronoun Fidelity with English LLMs: Are they Reasoning, Repeating, or Just Biased?" In: *Transactions of the Association for Computational Linguistics* 12, pp. 1755–1779. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00719. URL: https://doi.org/10.1162/tacl_a_00719

1.4 Outline

Although my thesis focuses on fair and faithful processing of English referring expressions, this is but one part of my broader research program on trustworthy natural language processing. I begin in Chapter 2 with an overview of nine other papers and pre-prints completed during the course of my PhD towards this goal. Then I provide brief background on reference, fairness, and faithfulness for the rest of this thesis in Chapter 3. Chapters 4-7 represent the main content of this thesis, followed by a broad outlook on future work in Chapter 8, before I conclude.

1.5 Significance

The arguments presented in this thesis reveal significant problems with current use of referring expressions as proxies for social categories of interest, current practices when evaluating stereotypical biases, and evaluations of and claims about language model "reasoning." To address these problems, this thesis presents new recommendations, datasets and metrics that future work can build on. The ubiquity of referring expressions in different subareas of natural language processing makes these insights and resources relevant to researchers and practitioners who work on fairness, reasoning, coreference resolution, natural language generation, and computational linguistics.

2

Additional Papers

My broader research program has the goal of advancing the next generation of trustworthy natural language processing systems and research. Under this broad umbrella, my three focus areas are: Fairness, faithfulness, and meta-evaluation, in contexts beyond referring expressions and English. To this end, my PhD work has resulted in nine papers and pre-prints in addition to the three publications that are the primary focus of this thesis. In this chapter, I provide an overview of my other work, before narrowing my focus to fair and faithful processing of English referring expressions.

2.1 Fairness in NLP

I see fairness as a critical goal for NLP, and I define it as having NLP systems that work for everyone and can handle sociolinguistic variation of many kinds. Within this thesis, I focus on studying gender fairness in systems and societies, which is frequently quantified through the use of referring expressions. Specifically, I question the use of names and pronouns as proxies for gender in Chapters 4 and 5, on theoretical and empirical grounds. Additionally, I study stereotypical biases and extrinsic harms in Chapters 5, 6 and 7.

My thesis work is reflective of a broader tendency in fairness literature to focus on English, single-axis stereotypical biases, and demographic characteristics such as gender and race. Despite this, I have a wide view of fairness that includes inter-language

variation as well. Other work from my PhD reflects this, and includes a critical review of "intersectionality" in NLP and ML, an evaluation of multilingual country and gender biases, and a study of language gaps in multilingual in-context learning.

Intersectionality. The critical framework of "intersectionality" examines how social inequalities persist structurally because of power. Given the goal of AI "fairness," intersectionality as an analytical framework is thus pivotal to operationalizing fairness. Indeed, intersectionality is often invoked in studies of AI fairness and bias, particularly when studying multiple intersecting demographic categories. We critically review this literature (making this also something of a meta-evaluation paper), with deductive and inductive coding to map how the central tenets of intersectionality operate within the paradigm of AI fairness, and to uncover gaps in the conceptualization and operationalization of intersectionality. We find that researchers overwhelmingly reduce intersectionality to optimizing for fairness metrics over intersecting demographic subgroups. They also fail to discuss their social context and when mentioning power, they mostly situate it only within the AI pipeline, despite intersectionality explicitly necessitating both inquiry and praxis. Finally, we outline and assess the implications of these gaps for critical inquiry and praxis, and provide actionable recommendations for AI fairness researchers to engage with intersectionality in their work by grounding it in AI epistemology. My contributions to this paper were: Reviewing the literature on intersectionality, collaboratively coming up with guiding questions for our inductive and deductive coding, reading and annotating papers for the critical review, analyzing our findings, and writing.

Anaelia Ovalle, Arjun Subramonian, Vagrant Gautam, Gilbert Gee, and Kai-Wei Chang (2023b). "Factoring the Matrix of Domination: A Critical Review and Reimagination of Intersectionality in AI Fairness." In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '23. Montréal, QC, Canada: Association for Computing Machinery, pp. 496–511. DOI: 10.1145/3600211.3604705. URL: https://doi.org/10.1145/3600211.3604705

Multilingual country-gender biases. Directly addressing the limitations of prior work on stereotypical biases in NLP (single-axis bias, most often gender, intrinsic bias, and a focus on English), we contribute the first study of multilingual intersecting country and gender biases, with a focus on occupation recommendations generated by large language models. We construct a benchmark of prompts in English, Spanish and German, where the user requests job recommendations for a recently laid-off friend, naturalistically incorporating the friend's gender (indexed via four pronoun sets), country of origin (25 options), and current country (five options). Then, we evaluate a suite of 5 LLAMA-based models on this benchmark, finding that LLMs encode significant gender and country biases. Notably, we find that even when models show parity for gender or country individually, intersectional[1] occupational biases based on both country and gender persist. We also show that the prompting language significantly affects bias (with Spanish showing the least bias, possibly due to pronoun-dropping in Spanish pre-training data), and instruction-tuned models consistently demonstrate the lowest and most stable levels of bias. Our findings highlight the need for fairness researchers to use intersectional and multilingual lenses in their work. My contributions to this work were formal supervision, analysis, and writing.

Elisa Forcada Rodríguez, Olatz Perez-de-Vinaspre, Jon Ander Campos, Dietrich Klakow, and Vagrant Gautam (Aug. 2025). "Colombian Waitresses y Jueces canadienses: Gender and Country Biases in Occupation Recommendations from LLMs." In: *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Ed. by Agnieszka Faleska, Christine Basta, Marta Costa-jussà, Karolina Staczak, and Debora Nozza. Vienna, Austria: Association for Computational Linguistics, pp. 182–194. ISBN: 979-8-89176-277-0. DOI: 10.18653/v1/2025.gebnlp-1.18. URL: https://aclanthology.org/2025.gebnlp-1.18/

Multilingual in-context learning. In-context learning is a popular inference strategy where large language models solve a task using only a few labeled demonstrations without needing any parameter updates. Although it is widely thought of as a way

1 Note that this is not intersectional in the same sense as the previous paper.

to improve language model performance on a variety of tasks compared to zero-shot learning, this view is primarily based on extensive studies on *English* in-context learning. In contrast, multilingual in-context learning remains under-explored, and we lack an in-depth understanding of the role of demonstrations in this context. To address this gap, we conduct a multidimensional analysis of multilingual in-context learning, experimenting with 5 models from different model families, 9 datasets covering classification and generation tasks, and 56 typologically diverse written languages. Our results reveal that the effectiveness of demonstrations varies significantly across models, tasks, and languages, and few-shot learning sometimes even worsens performance compared to zero-shot learning on a variety of non-English languages. However, strong instruction-following models including LLAMA 2-CHAT, GPT-3.5, and GPT-4 are largely insensitive to the quality of demonstrations. Here, a carefully crafted template often eliminates the benefits of demonstrations for some tasks and languages altogether. These findings show that the importance of demonstrations might be overestimated. Our work highlights the need for granular evaluation across multiple axes towards a better understanding of in-context learning. I contributed to this work by playing an advisory role throughout the project, analyzing data and results, and writing.

Miaoran Zhang, Vagrant Gautam, Mingyang Wang, Jesujoba Alabi, Xiaoyu Shen, Dietrich Klakow, and Marius Mosbach (Aug. 2024). "The Impact of Demonstrations on Multilingual In-Context Learning: A Multidimensional Analysis." In: *Findings of the Association for Computational Linguistics: ACL 2024*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, pp. 7342–7371. DOI: 10.18653/v1/2024.findings-acl.438. URL: https://aclanthology.org/2024.findings-acl.438/

2.2 Faithfulness in NLP

A second important goal for NLP is faithfulness, of which I see two related components: Having NLP systems that are faithful to facts and input, as well as having a faithful

mechanistic understanding of how they work. Both qualities are particularly consequential given how widely large language models are used today. In this thesis, I focus on the former criterion, i.e., faithfulness to the use of referring expressions in input, i.e., pronoun fidelity with one referent (Chapter 6) and with two referents (Chapter 7).

I have also previously worked on faithfulness to context in question answering systems. In addition, I see it as a critical goal for the future of NLP to have a faithful understanding of the inner workings of NLP systems. Towards achieving this, I have proposed a new method for probing that unites behavioural and model-internal perspectives on toxic generations, and shown that cross-modal transfer methods for LLMs are unfaithful to their purported reasons for working.

Faithful question answering. If a question cannot be answered with the available information, robust systems for question answering (QA) should know *not* to answer. In the context of extractive QA, one way to build models that do this is to train them with both answerable questions (i.e., questions that can be answered with a given text) as well as unanswerable ones (i.e., questions that cannot be answered with a given text, to provide a negative signal). As many datasets of answerable questions exist, the focus of prior work has been to create unanswerable questions, either by employing annotators or through automated LLM-based methods for unanswerable question generation. We propose a simpler data augmentation method for unanswerable question generation in English: Performing antonym and entity swaps on answerable questions. Our method is built on a pipeline of traditional computational linguistics approaches like part-of-speech tagging, WordNets and named entity recognition, and our results show that the complexity of existing automated approaches is not justified. Compared to the prior state-of-the-art, data generated with our training-free and lightweight strategy has higher human-judged relatedness and readability, and results in better models (+1.6 F_1 points on SQuAD 2.0 data with BERT-LARGE). We quantify the raw benefits of our approach compared to no augmentation across multiple encoder models, using different amounts of generated data, and also on TydiQA-MinSpan data (+9.3 F_1 points with BERT-LARGE). In sum, our results establish swaps as a simple but strong baseline

for future work. I led this work, including conceptualization, building the pipeline to generate unanswerable questions, performing analysis, and writing the paper.

Vagrant Gautam, Miaoran Zhang, and Dietrich Klakow (Dec. 2023). "A Lightweight Method to Generate Unanswerable Questions in English." In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 7349–7360. DOI: 10.18653/v1/2023.findings-emnlp.491. URL: https://aclanthology.org/2023.findings-emnlp.491/

Uniting model-internal and behavioural toxicity. Prior work studying language model toxicity has taken either an exclusively behavioural perspective, i.e., how the toxicity of input prompts affects the toxicity of language model generations, or focused on how toxicity is encoded within model internals (e.g., through layer-wise probing). To unify these two distinct perspectives, we introduce aligned probing, a novel interpretability framework that aligns the behaviour of language models (based on their outputs) and their internal representations (internals). Using this framework, we examine the flow of toxicity from input, through model internals, to the output, in over 20 OLMo, LLAMA, and MISTRAL models. Our results show that LMs strongly encode information about the toxicity level of inputs and subsequent outputs, particularly in lower layers. Focusing on how unique LMs differ offers both correlative and causal evidence that when models "know" more about the input toxicity (i.e., more strongly encode this information in their representations), they generate less toxic output. We also highlight the heterogeneity of toxicity, as model behavior and internals vary across different types of toxicity, such as threats. Finally, with four case studies analyzing detoxification, multi-prompt evaluations, model quantization, and pre-training dynamics, we underline the practical impact of aligned probing. Our findings contribute to a more holistic understanding of LMs, both within and beyond the context of toxicity. I played an advisory role in this project, helping with conceptualization and paper writing.

Andreas Waldis, Vagrant Gautam, Anne Lauscher, Dietrich Klakow, and Iryna Gurevych (2025). *Aligned Probing: Relating Toxic Behavior and Model Internals*. arXiv: 2503.13390 [cs.CL]. URL: https://arxiv.org/abs/2503.13390

Cross-modal transfer of LLMs. Although language models are, as the name suggests, trained and typically used for tasks involving natural language, recent work has shown that adapting them to new non-language modalities shows promising results and better performance than training dedicated neural networks from scratch. These new modalities include partial differential equations, satellite image time series, electrocardiogram recordings, and more. ORCA is one such recent technique for cross-modal transfer of pre-trained transformer models, which consists primarily of training an embedder, and subsequently fine-tuning both the embedder and model on task data. Despite its high performance on a variety of downstream tasks, we do not understand precisely how each of these components contributes to ORCA's success. Therefore, we run a series of ablation studies and find that embedder training does not help 2D (i.e., matrix-style input) tasks at all, contrary to what the original paper posits. In 1D (i.e., sequence-style) tasks, some amount of embedder training is necessary but more is not better. In 4 out of 6 datasets we experiment with, it is model fine-tuning that makes the biggest difference. Through our ablations and baselines, we contribute a better understanding of the individual components of ORCA. I co-led this project, conceptualizing the research, visualizing and analyzing the results, and writing the paper.

Paloma García-de-Herreros, Vagrant Gautam, Philipp Slusallek, Dietrich Klakow, and Marius Mosbach (June 2024). "What explains the success of cross-modal fine-tuning with ORCA?" In: *Proceedings of the Fifth Workshop on Insights from Negative Results in NLP*. Ed. by Shabnam Tafreshi, Arjun Akula, João Sedoc, Aleksandr Drozd, Anna Rogers, and Anna Rumshisky. Mexico City, Mexico: Association for Computational Linguistics, pp. 8–16. DOI: 10.18653/v1/2024.insights-1.2. URL: https://aclanthology.org/2024.insights-1.2/

2.3 Meta-Evaluation

One of the challenges of NLP research in 2025 is that people use NLP systems in a variety of creative ways that we may not be directly evaluating. We come up with "tasks" that are meant to formalize these human uses, we create datasets of examples, and metrics to help us automatically evaluate systems at scale. However, every step of this process can introduce gaps between what we want to measure and what we are actually measuring. This is why I believe meta-evaluation to be of critical importance in NLP, also to reduce the gaps we introduce during conceptualization and operationalization, and to have valid and reliable research practices when we work with NLP systems and do research on them.

The work described in this thesis makes important contributions to this area, e.g., Chapter 7 shows that system reasoning about pronouns can be overestimated without the more realistic multi-person setting, Chapter 5 demonstrates empirical issues with a widespread conflation of pronoun forms with social gender, and Chapter 4 provides concrete guidelines towards valid use of names in sociodemographic research.

In other work during my PhD, I have interrogated the use and impact of other trustworthy NLP concepts, such as misgendering, democracy, interpretability, and intersectionality. As such concepts have been studied for much longer and with a variety of methods in other disciplines, I embrace interdisciplinarity and methodological pluralism when I study them in the NLP context, using definitions from relevant disciplines and flexibly choosing the best methods (whether quantitative, qualitative, or a mixture).

Meta-evaluation of misgendering. Misgendering, a harm that is particularly relevant to trans individuals, has only recently begun to be studied in NLP fairness, and there is a lack of consensus on how to evaluate it. Numerous methods have been proposed to measure LLM misgendering, including probability-based evaluations (e.g., automatically with templatic sentences) and generation-based evaluations (e.g., with automatic heuristics or human validation). However, it has gone unexamined whether these evalua-

tion methods have convergent validity, that is, whether their results align. Therefore, we conduct a systematic meta-evaluation of these methods across three existing datasets for LLM misgendering (including the RUFF dataset I introduce in Chapter 6). We propose a method to transform each dataset to enable parallel probability- and generation-based evaluation. Then, by automatically evaluating a suite of 6 models from 3 families, we find that these methods can disagree with each other at the instance, dataset, and model levels, conflicting on 20.2% of evaluation instances. Finally, with a human evaluation of 2400 LLM generations, we show that misgendering behaviour is complex and goes far beyond pronouns, which automatic evaluations are not currently designed to capture, suggesting essential disagreement with human evaluations. Based on our findings, we provide recommendations for future evaluations of LLM misgendering. Our results are also more widely relevant, as they call into question broader methodological conventions in LLM evaluation, which often assume that different evaluation methods agree. My contributions to this project were research conceptualization, data annotation, analysis and visualizations, and paper writing.

Arjun Subramonian, Vagrant Gautam, Preethi Seshadri, Dietrich Klakow, Kai-Wei Chang, and Yizhou Sun (2025). "Agree to Disagree? A Meta-Evaluation of LLM Misgendering." In: *Second Conference on Language Modeling*. URL: https://openreview.net/forum?id=vgmiRvpCLA

Understanding "democratization". Recent improvements in NLP and machine learning and increased mainstream adoption have led to researchers frequently discussing the "democratization" of artificial intelligence. In this paper, we seek to clarify how democratization is understood in NLP and ML publications, through large-scale mixed-methods analyses of papers using the keyword "democra*" published in NLP and adjacent venues. We find that democratization is most frequently used to convey (ease of) access to or use of technologies, without meaningfully engaging with theories of democratization, while research using other invocations of "democra*" tends to be grounded in theories of deliberation and debate. Overall, our conceptual analysis shows that democratization is mostly used as a buzzword instead of drawing from the

3000+ years of thought on democracy and democratization. Based on our findings, we call for researchers to enrich their use of the term democratization with appropriate theory, towards democratic technologies beyond superficial access. I co-led this project, including conceptualization, data annotation, analysis and visualizations, and writing.

Arjun Subramonian, Vagrant Gautam, Dietrich Klakow, and Zeerak Talat (Nov. 2024). "Understanding Democratization in NLP and ML Research." In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, pp. 3151–3166. DOI: 10.18653/v1/2024.emnlp-main.184. URL: https://aclanthology.org/2024.emnlp-main.184/

Quantifying the impact of interpretability and analysis work. Interpretability and analysis (IA) research is a growing subfield within NLP with the goal of developing a deeper understanding of the behaviour or inner workings of NLP systems and methods. Such work purports to build trust in NLP systems by contributing a mechanistic understanding of them, explaining their predictions, and improving their faithfulness. Despite growing interest in the subfield, a criticism of this work is that it lacks actionable insights and therefore has little impact on NLP. In this paper, we seek to quantify the impact of IA research on the broader field of NLP. We approach this with a mixed-methods analysis of: **(1)** A citation graph of 185K+ papers built from all papers published at ACL and EMNLP conferences from 2018 to 2023, and their references and citations, and **(2)** a survey of 138 members of the NLP community. Our quantitative results show that IA work is well-cited outside of IA, and central in the NLP citation graph. Through qualitative analysis of survey responses and manual annotation of 556 papers, we find that NLP researchers build on findings from IA work and perceive it as important for progress in NLP and multiple subfields. Researchers also rely on its findings and terminology for their own work. Many novel methods are proposed based on IA findings and highly influenced by them, but highly influential non-IA work cites IA findings without being driven by them. We end by summarizing what is missing in IA work today and provide a call to action to make IA research more impactful with

big-picture thinking, actionable insights, human-centered evaluations, and standardized methods. In this project, I led the manual annotation effort, and helped with analysis and writing.

Marius Mosbach, Vagrant Gautam, Tomás Vergara Browne, Dietrich Klakow, and Mor Geva (Nov. 2024). "From Insights to Actions: The Impact of Interpretability and Analysis Research on NLP." In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, pp. 3078–3105. DOI: 10.18653/v1/2024.emnlp-main.181. URL: https://aclanthology.org/2024.emnlp-main.181/

2.4 Conclusion

With these nine additional papers, I have contributed broadly to research on trustworthy natural language processing during my PhD. I have addressed aspects of fairness (intersectional and multilingual fairness, as well as democracy and misgendering in NLP) and faithfulness (interpretability and analysis in the context of question answering, toxicity, and cross-modal transfer) that go far beyond my focus on English referring expressions in the rest of this thesis.

3

Background

This chapter briefly introduces the definitions, concepts, and techniques central to the research in this thesis. I begin with reference, fairness, and faithfulness, the concepts mentioned in the title of this thesis. Finally, I provide background on the two main NLP tasks that I consider: Coreference resolution and language modelling.

3.1 Reference

Reference is notoriously hard to pin down, and there are different definitions among philosophers, linguists, and psychologists (Russell, 1905; Strawson, 1950; Donnellan, 1966; Gundel and Abbott, 2019). Thankfully, they all agree on the prototypical form of reference, i.e., an expression that uniquely denotes a single specific individual. Strawson's (1950) examples of these are: Names, personal pronouns, and definite descriptions.

Another important theoretical distinction is whether reference is treated as a semantic phenomenon, i.e., something that expressions do, or a pragmatic one, i.e., something that language users do with expressions (Abbott, 2010). One way of viewing reference as a pragmatic phenomenon is to frame it as a speech act, which helps us account for when one mistakes a person in the distance for one's friend (Gundel and Abbott, 2019). While this can be very useful, I do not wish to equate NLP systems with language users that have pragmatic intentions, and I thus follow common practice in

NLP with my focus on referring expressions as *semantic* relations between linguistic forms and people.

In this thesis, I focus on the three prototypical classes of referring expressions, which I elaborate on below. While I continue to touch on some theoretical views on reference where relevant, I use theory as a starting point in considering NLP research and practice, which tends to be mostly applied and empirical.

3.1.1 Personal Names

Personal names are proper nouns used to identify individual people (Hough, 2016). They are hotly contested in theoretical discussions of reference, with arguments that they are purely denotational (i.e., used exclusively to denote their referent), and that they also have connotations and lexical meaning (Abbott, 2010; Hough, 2016). Their denotational meaning is transparent in examples such as the following:

(1) **Justin Trudeau** is **Pierre Trudeau**'s son.

Most work on names in reference tends to take famous people as examples as above, but in this thesis, we concern ourselves with names in a broader sense. As Pelczar and Rainsbury (1998) observe, referring expressions like *John* and *he* seem to appeal to a "natural" feature, that of a "salient male," independently of any context. It is precisely this kind of gendered association that is exploited in NLP research, where personal names are sometimes used not just as traditional referring expressions that map to specific individuals, but rather as linguistic forms that index a particular gender (Karimi et al., 2016; Asr et al., 2021, *inter alia*).

(2) a. **Madeleine** is petting a dog.
b. **Julius** is petting a cat.
c. **Rain** has a pet snake.

In Example (2), *Madeleine* is often used to index female-ness, while *Julius* is male-typed and *Rain* may not even be recognized as a name, regardless of the real-life gender of the referents, the bearers of these names. In Chapter 4, we highlight failures that can come from this dual character of personal names as designators that also have indexical characteristics (Pelczar and Rainsbury, 1998; Kosse, 2021).

3.1.2 Pronouns

Pronouns are expressions that can stand in for a noun or noun phrase. In this thesis, we focus on personal pronouns that are linked to a referent in the context, an antecedent, as shown below. This referential use of personal pronouns is broadly known as anaphora.[1]

(3) a. Michael$_i$ said that **he**$_i$ had already made rhubarb ice cream.

b. [The Scottish poet]$_i$ had written about **their**$_i$ cats.

The pronouns I consider in this thesis are in fact an even narrower subset of referential personal pronouns, namely, third-person singular animate pronouns in English. These are sometimes treated as a closed class, consisting of *he/him/his*, *she/her/her*, and sometimes singular *they/them/their*, depending on which linguists you ask, all of which I consider in this thesis. There is also some evidence that they should instead be treated as an open class, given corpus linguistic and historical evidence of diverse and novel pronoun phenomena (McGaughey, 2020; Lauscher et al., 2022). As one example of how NLP systems handle this, I include *xe/xem/xyr* pronouns in my analyses as well.

What is particularly interesting about this subset of English pronouns is that they mark grammatical gender, which is thought to directly index the referent's social gender. Once again, as with personal names, this assumption is used in NLP to create one-to-one mappings between personal pronouns and social gender. However, as work

1 Confusingly, this is distinct from the use of "anaphor" in binding theory, which applies only to reflexive (*himself*) and reciprocal (*each other*) pronouns, neither of which I consider in this thesis. Another phenomenon which is highly relevant for binding theory but that I do not touch on is variable pronouns, e.g., "Every lawyer made their case successfully" (Han and Moulton, 2022).

in sociolinguistics and semiotics has pointed out (Ochs, 1992; Conrod, 2018), the way people use pronouns for other people is far more complex, with elements of performativity (Butler, 1988), play (Rudes and Healy, 1979), and politeness (Tovar, 2025). Inspired by this line of work in linguistics, Chapter 5 attempts to sever the assumed one-to-one connection between pronouns and social gender in the NLP context, cf. Bolukbasi et al. (2016) and Rudinger et al. (2018), *inter alia*. I do this by examining pronominal biases in their own right, independent of social gender biases. Using this newer definition of pronominal bias, I examine how NLP systems process referential pronouns; Chapter 5 examines how they *resolve* pronominal reference, while Chapters 6 and 7 consider how they *do* pronominal reference.

3.1.3 Definite Descriptions

Definite descriptions are noun phrases with *the* as the determiner, which typically refer to a specific individual, as shown in the examples below (Abbott, 2006).

(4) a. **The footballer** also liked to bake and do CrossFit.
b. **The phonologist** and **the artist** wanted to move to Glasgow together.

In this thesis, we sidestep theoretical debates on reference in this context (Donnellan, 1966; Prince, 1992, *inter alia*) as we only use definite descriptions to aid our analysis of resolving and doing *pronominal* reference in NLP. In particular, we focus on definite descriptions of occupations (*the surgeon*, *the nurse*, etc.) to elicit reasoning and gendered stereotypes in coreference resolution and language modelling with pronouns. Such stereotypes have also been shown in coreference resolution by humans, with eye-tracking (Pyykkönen et al., 2010), reading times (Kennison and Trofe, 2003), event-related brain potentials (Osterhout et al., 1997), and production studies (Boyce et al., 2018; Boyce et al., 2019; Morehouse et al., 2022). Definite descriptions also have a long history of being used to probe stereotypes in NLP, as in Rudinger et al. (2018) and Zhao et al. (2018).

3.2 Fairness in NLP

Work on fairness in NLP is about measuring and mitigating biases, stereotyping, discrimination, and other harms via language (Gallegos et al., 2024). This includes studying these phenomena in NLP systems as well as in society. The former involves studying fairness in specific NLP tasks, such as coreference resolution (Rudinger et al., 2018; Cao and Daumé III, 2021) and natural language generation (Kotek et al., 2023; Ovalle et al., 2023a), as in this thesis; the latter involves analyzing society through language, studying biases in academia (Vogel and Jurafsky, 2012; Mohammad, 2020) and in media representation (Asr et al., 2021; Bamman et al., 2024), among others.

The overarching goal of such work is "fairness," which is variously defined as parity or as equity, usually between social groups (defined along gendered, racial, or linguistic lines), and sometimes also at the level of the individual (Gallegos et al., 2024). Beyond this diversity in how one actually defines fairness as a goal and measures it (Jacobs and Wallach, 2021), some critical work has also noted a lack of consensus on what behaviours qualify as harms, to whom, and why (Selbst et al., 2019; Blodgett et al., 2020), problems with how stereotypes are defined and measured (Blodgett et al., 2021), and how definitions of fairness rest on contested classifications (Field et al., 2021; Weinberg, 2022). These critiques inform Chapters 4 and 5, which critically question how names and pronouns are (and should be) used in fairness research.

In addition to these more theoretical arguments, empirical research has also questioned how fairness is *measured*. Broadly, measurements of fairness can be classified as either **intrinsic** or **extrinsic**. Intrinsic metrics look at system-internal associations, such as between *surgeons* and the pronoun *he*, while extrinsic metrics evaluate downstream harms in applications of these systems, such as errors and differences in quality of service. Empirical meta-evaluations of fairness include work showing that templatic probability-based evaluations of intrinsic biases are brittle (Seshadri et al., 2022; Goldfarb-Tarrant et al., 2023; Selvam et al., 2023), and that intrinsic biases do not correlate with extrinsic biases (Goldfarb-Tarrant et al., 2021; Cao et al., 2022). My work in Chapter 5 similarly demonstrates the empirical problems that arise from treating

all pronouns of a certain grammatical gender as equivalent when it comes to measuring social gender biases, Chapter 6 proposes an extrinsic evaluation of the harm of misgendering, and Chapter 7 further investigates the robustness of this measure.

3.3 Faithfulness in NLP

In contrast to fairness, faithfulness is not a clearly-defined subfield of NLP with a dedicated research community. Faithfulness is instead a desideratum of several domains of NLP, including interpretability/explainability, machine translation and summarization, where it means different things. When interpreting and explaining NLP systems, we seek faithful methods that accurately represent the reasoning process within the system (Jacovi and Goldberg, 2020). Faithfulness is an important consideration regardless of whether we use attention mechanisms (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019), saliency methods (Bastings and Filippova, 2020), or system-generated reasoning chains in natural language (Turpin et al., 2023). On the other hand, in the context of machine translation and summarization, faithfulness refers to the property of creating a target text that is consistent with the source, in content (Maynez et al., 2020), style (Genzel et al., 2010), and the experience of reading a text (Jones and Irvine, 2013).

In this thesis, I consider the latter kind of faithfulness and not the former, i.e., faithful use of referring expressions means using them to point to individuals in a manner that is consistent with provided information. In Chapter 5, I use this definition of faithfulness to propose a novel metric for isolating bias in coreference resolution. Measuring faithful reuse of a specified pronoun for someone introduced with a definite description is the focus on Chapter 6, and Chapter 7 probes whether successful reuse is due to faithful reasoning, or simply shallow repetition. The work in this thesis does not attempt to interpret or explain system-internal processing of referring expressions, although this is a fruitful area for future work, as I describe in Chapter 8.

3.4 Coreference Resolution

Coreference resolution is the NLP task of identifying which surface forms in text refer to the same entity, i.e., which surface forms corefer. In Example (5), we see how names, definite descriptions, and personal pronouns can all refer to the same entity.

(5) [The machine learning professor]$_i$, Vlad$_i$, wanted [his$_i$ students]$_j$ to test all of their$_j$ code.

We also see how text spans referring to an entity can overlap with each other (note the two indices within the span *his students*). Due to the complexity of the task, coreference resolution has been formulated in a number of ways: **Span-based approaches** treat the task as a two-step problem of detecting spans that are likely to be mentions, and subsequently linking them (Lee et al., 2017; Joshi et al., 2020; Otmazgin et al., 2023). The same general steps are swapped in order in **word-level approaches**, where coreference resolution is formulated as a two-step problem of computing coreference links between words, and then reconstructing the spans (Dobrovolskii, 2021; D'Oosterlinck et al., 2023). In contrast to the above approaches, **autoregressive approaches** tailored for coreference resolution involve multiple forward passes to build a coreference structure (Bohnet et al., 2023; Liu et al., 2022). Finally, it is also possible to simply prompt generative language models in natural language to perform coreference resolution (Xu et al., 2023; Gan et al., 2024). In Chapter 5, we use all of these approaches except for autoregressive approaches, which are particularly computationally expensive.

The complexity of coreference structures has also resulted in a range of specialized metrics and multi-component evaluations, surveyed in Liu et al. (2023). However, this thesis examines a specialized and simple version of coreference resolution inspired by Winograd Schemas (Levesque et al., 2012) and Winogender Schemas (Rudinger et al., 2018). These schemas, explained in detail in Chapter 5, always contain one pronoun and two non-overlapping candidate antecedents, only one of which is correct in context. In Liu et al. (2023), this specialized version is known as a "gold-two-mention problem,"

and is typically evaluated with more widely used metrics such as accuracy, precision, recall and F_1, as described below.

3.4.1 Evaluation

The following sentence is an example from Winogender Schemas, to illustrate how scoring works with each of the performance metrics we consider:

(6) **The cashier**$_i$ told **the customer**$_j$ that **her**$_j$ card was declined.

In the above example, there are three entity mentions–*the cashier*, *the customer*, and *her*. We are interested in all expressions that co-refer with the pronoun *her*, which we can extract and normalize regardless of the coreference resolution approach that is used. Based on the scenario described about a card being declined, it is only *the customer* whose card could be declined, and not *the cashier*. Thus, to achieve a correct score on this example, a coreference resolution system needs to map *her* to *the customer*, and it must also not map *her* to *the cashier*. These are, respectively, a true positive (TP) and a true negative (TN). Any time that a coreference resolution system maps *her* to *the cashier* in this context, this is a false positive (FP), and similarly, when it does not map *her* to *the customer*, this is a false negative (FN). Given these definitions, we compute:

$$\textbf{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

$$\textbf{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\textbf{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\textbf{F}_1 = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

3.5 Language Modelling

Language modelling involves creating computational models of natural language and has long been a pursuit of NLP. The history of language modelling includes rule-based models (e.g., formal grammars), statistical models estimated from corpora (e.g., n-gram models), and neural network-based approaches. In this thesis, we focus on the latest instance of the latter, transformer-based language models (Vaswani et al., 2017), as they have been the predominant style of language model since 2018. Specifically we consider three types: Encoder-only, decoder-only and encoder-decoder models.

Although language models can vary a lot architecturally, fundamentally, they are all probabilistic models of language. This means that a good language model tends to assign higher probabilities to valid sequences in the language, and lower probabilities to invalid ones. This fact is used in **probability-based evaluations** with minimal pair data, where one item of the pair is valid and the other is invalid. The language model gets the pair right if it assigns a higher probability to the valid item, and is wrong otherwise. These binary scores can be aggregated across a full dataset of such minimal pairs or minimal sets to give a final percentage accuracy on the entire dataset. Probability-based evaluations on minimal pair/set data is used to target performance on syntax (Warstadt et al., 2020), world knowledge (Ivanova et al., 2024), bias (Kurita et al., 2019), and more. In the context of this thesis, I use probability-based evaluations with minimal sets that vary in whether or not reference is performed correctly.

Language models can also be used to generate text, by sampling from the probability distribution they model. **Generation-based evaluations** involve using language models to generate text and then either automatically or manually evaluating the generations. Increasingly, with the advent of models that are specialized to follow instructions or natural language prompts, such evaluations involve prompting the model to do a certain task. Although prompting is brittle (Sclar et al., 2024; Mizrahi et al., 2024) and is not a substitute for directly measuring language model probabilities (Hu and Levy, 2023; Hu and Frank, 2024), I also use generation-based evaluations in this thesis for some models that are tuned to be used like this. Generation-based evaluations also allow

us to use "reasoning" techniques, such as chain-of-thought (Wei et al., 2022), where a language model generates additional text that it can condition its final answer on.

3.5.1 Encoder-Only Language Models

Encoder-only models such as BERT (Devlin et al., 2019), ALBERT (Lan et al., 2020), and RoBERTa (Liu et al., 2019) consist of only the encoder component of the original transformer architecture. All of these models are pre-trained with a masked languaged modelling objective, i.e., prediction of tokens in a sequence that have been randomly replaced with special [MASK] tokens. RoBERTa is additionally pre-trained on next sentence prediction, i.e., classifying whether or not one sentence should follow another in a text, to encourage reasoning about sentence-level relationships. This training objective was later found to be ineffective and was replaced with a sentence order prediction loss when training ALBERT. Encoder-only models are typically used with probability-based evaluations, as described below. Although generation-based evaluation of encoder-only models has been shown to be possible (Wang and Cho, 2019), this is not a method that is used in practice, including in this thesis.

Probability-based evaluation. For probability-based evaluations with encoder-only models, we compute pseudo log likelihood scores on minimal pairs or minimal sets of sequences (Salazar et al., 2020; Kauf and Ivanova, 2023). This method relies on these models' masked language modelling abilities, and we use it in Chapters 6 and 7. Formally, assume a sequence of tokens $S := w_1, w_2, w_3, ..., w_N$, of length N. We make N copies of S, where each S_i is identical to S except that w_i is replaced with [MASK]. We feed each S_i to an encoder-only language model M, allowing us to obtain $\log P_M([\text{MASK}] = w_i | S_i)$. Our final pseudo log probability score is obtained by summing together the log probabilities for each S_i:

$$\sum_{i}^{N} \log P_M([\text{MASK}] = w_i | S_i)$$

3.5.2 Decoder-Only Language Models

In contrast to encoder-only models, decoder-only models such as OPT (Zhang et al., 2022), PYTHIA (Biderman et al., 2023), and LLAMA (Touvron et al., 2023) only use the decoder component of the transformer. These models are pre-trained on large corpora to predict the next token (Radford et al., 2018), because of which they are also known as autoregressive language models. In addition to these **base models**, we also use **chat models** that have been post-trained to optimize them for dialogue. Approaches for post-training can vary significantly, but LLAMA 2-CHAT, the only chat model we consider in this thesis, is post-trained with supervised instruction tuning and reinforcement learning from human feedback (Touvron et al., 2023). We use decoder-only models with probability- and generation-based evaluations in Chapters 6 and 7.

Probability-based evaluation. For probability-based evaluations with decoder-only base models, we compute log likelihood scores on minimal pairs or sets of sequences, as in previous work (Warstadt et al., 2020; Kauf and Ivanova, 2023). This method follows directly from the pre-training objective of next word prediction. Formally, assume a sequence of tokens $S := w_1, w_2, w_3, ..., w_N$, of length N. We feed the sequence S to a decoder-only language model M, and for each token w_i, we obtain the log probability of predicting it given the left context, i.e., $\log P_M(w_i|w_1, ..., w_{i-1})$. Our final log probability score is obtained by summing together all of these log probabilities:

$$\sum_{i}^{N} \log P_M(w_i|w_1, ...w_{i-1})$$

Generation-based evaluation. As chat models are optimized to follow instructions in natural language, we use generation-based evaluations with them. This involves providing a model with a natural language prompt, and then using it to generate some number N of tokens conditioned on this prompt. The output of the model then needs to be preprocessed and normalized to extract the final answer. N is generally determined through experimentation, and we generally report a range of performance on multiple

prompts, as recent work has shown that prompting is brittle and can lead to large variance in results (Sclar et al., 2024; Mizrahi et al., 2024).

3.5.3 Encoder-Decoder Language Models

Encoder-decoder language models (also called sequence-to-sequence or seq2seq models) closely follow the original transformer architecture in having both an encoder and a decoder component (Raffel et al., 2020). The encoder thus fully processes an input sequence before beginning to causally decode the output sequence. A variety of tasks (translation, question answering, and classification) can be cast to this input-output or "text-to-text" format, which is used to train a generalist T5 model (Raffel et al., 2020). The pre-training objective is a variant of masked language modelling where spans of tokens are masked, rather than individual tokens. Thus, both probability-based evaluations (i.e., by masking and predicting spans), as well as generation-based evaluations (via prompting) are possible with this base model. In this thesis, however, we focus on a variant of T5 known as FLAN-T5 (Chung et al., 2024), which is post-trained to follow instructions through supervised fine-tuning. We thus use generation-based evaluations for this model in Chapters 5, 6, and 7, just as we do with decoder-only chat models.

Generation-based evaluation. As FLAN-T5 is trained to follow instructions in natural language, we use generation-based evaluations exactly as described for decoder-only chat models above.

4

Disentangling Personal Names and Sociodemographic Attributes

Personal names are an example of a cross-culturally universal type of referring expression that differentiates individuals, allowing us to refer to them individually. Simultaneously, and somewhat paradoxically, personal names also categorize people based on sociodemographic characteristics that are important in a given society. Based on this, the natural language processing community has associated personal names with sociodemographic characteristics like gender and race in a variety of tasks. However, these works show varying degrees of critical engagement with the established methodological problems of using names as a proxy for categories of interest. To guide future work in NLP that uses names and sociodemographic characteristics, we provide an overview of relevant research: First, we present an interdisciplinary background on names and naming. We then survey the issues inherent to associating names with sociodemographic attributes, covering problems of validity (e.g., systematic error, construct validity), as well as ethical concerns (e.g., harms, differential impact, cultural insensitivity). Finally, we provide guiding questions along with normative recommendations to avoid validity and ethical pitfalls, towards more fairly handling names and sociodemographic characteristics in natural language processing.

The content in this chapter is based on:
Vagrant Gautam, Arjun Subramonian, Anne Lauscher, and Os Keyes (Aug. 2024c). "Stop! In the Name of Flaws: Disentangling Personal Names and Sociodemographic

Attributes in NLP." In: *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Ed. by Agnieszka Faleska, Christine Basta, Marta Costa-jussà, Seraphina Goldfarb-Tarrant, and Debora Nozza. Bangkok, Thailand: Association for Computational Linguistics, pp. 323–337. DOI: 10.18653/v1/2024.gebnlp-1.20. URL: https://aclanthology.org/2024.gebnlp-1.20/

Vagrant Gautam conceptualized the research and led the interdisciplinary literature review and paper writing. Arjun Subramonian helped with reviewing the literature and writing the paper. Os Keyes provided a philosopher's perspective and, along with Anne Lauscher, advised and gave feedback on the project.

4.1 Introduction

A person's identity is a complex and paradoxical thing - it simultaneously identifies someone's uniqueness, allowing us to refer to them specifically, and categorizes them, identifying what they have in common with others (Strawson, 1950; Strauss, 2017). A perfect example of this phenomenon is a person's *name*. Personal names are proper nouns used to refer to individuals. They play an important distinguishing role in our lives, as they let us refer to people directly in language, uniquely represent them mentally, and underscore their significance as individuals (Jeshion, 2009). For these reasons, personal names are a linguistic universal, i.e., they appear across languages and cultures, although naming customs vary across the world (Hough, 2016).

But alongside differentiating people, names also categorize them in their society. Names assigned to people often index, or point to, aspects of identity that are important in the context of their society, including sex, religion, tribe, stage of life, etc. Personal names are thus rich resources to understand the social organization of communities, and have been studied across anthropology (Alford, 1987; Hough, 2016), sociology (Marx, 1999; Pilcher, 2017), linguistics (Allerton, 1987; Anderson, 2003), and onomastics (Alvarez-Altman et al., 1987; Adams, 2009).

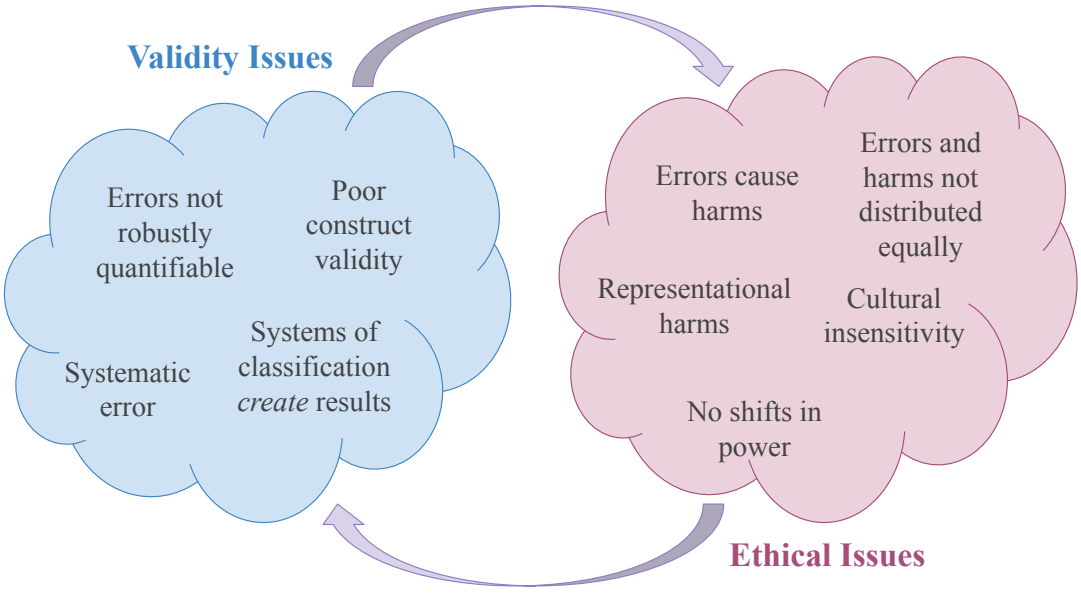


Figure 4.1: Overview of the methodological issues (concerning validity and ethics) of the use of personal names and sociodemographic characteristics in NLP.

In natural language processing (NLP) as well, personal names have a long history of use—NLP researchers have worked on identifying and disambiguating uses of personal names (Mann and Yarowsky, 2003; Minkov et al., 2005; Färber and Ao, 2022) and have examined name translation (Sennrich et al., 2016; Wang et al., 2022; Sandoval et al., 2023) and name transliteration (Li et al., 2007; Benites et al., 2020; Sälevä and Lignos, 2024). Increasingly, NLP researchers also use personal names along with sociodemographic characteristics for passive analysis of media and scholarly content (Vogel and Jurafsky, 2012; Knowles et al., 2016; Mohammad, 2020; Asr et al., 2021), or to examine model biases and harms (Maudslay et al., 2019; Romanov et al., 2019; Webster et al., 2021). However, these papers engage to varying degrees with concerns that have been raised outside of NLP about the methodological validity and ethics of associating names with sociodemographic characteristics. We argue that neglecting these issues is a significant barrier to valid and respectful research, as well as fair, inclusive NLP systems.

Thus, we contribute an overview of the issues with associating names with sociodemographic attributes (focused on gender and race, two popular categories used in

NLP research), as shown in Figure 4.1. We begin with background on names in other fields (Section 4.2) and in NLP (Section 4.3), and lay out the problems with validity (Section 4.4) and ethical concerns (Section 4.5) raised when associating personal names with sociodemographic characteristics. Finally, we present guiding questions along with normative recommendations (Section 4.6) to promote fairer use of names in future work in the field.

4.2 Background: Names and Naming

Names are generally regarded as social phenomena that serve two central functions that are sometimes in conflict: Differentiation and categorization of individuals (Alford, 1987). Differentiation is important psychologically and semantically for us to be able to directly refer to and mentally represent individuals, and names also serve to underscore their referent's significance as an individual (Jeshion, 2009). Categorization, on the other hand, is important for the social organization of communities, and naming conventions tend to reflect factors that are important to a community at a given point in time, e.g., gender, religion, descent, transition to adulthood, and so on (Hough, 2016). For instance, the practice of naming someone after their father or grandfather—patronymic naming—was once common across Europe, and was popular in Sweden until the nineteenth century (e.g., *Samuelsson*) and continues into Iceland today (e.g., *Gunnarsdóttir*) (Hough, 2016). This example shows how names and naming can *only* be understood in a specific (geographic, cultural, temporal) context, and even then includes a lot of variation. As folk assumptions about names tend to overlook the wide variation in names and naming (McKenzie, 2010), we present an overview of naming as it relates to sociodemographic characteristics below.

Variation in societal conventions. The markers considered important to index in a name vary widely across cultures. For example, almost all European naming systems and indeed most societies across the world tend to assign sex-typed names (Hough, 2016), while South Indian naming conventions often index caste (Meganathan, 2009). However,

convention does not mean that every single individual is assigned a name that neatly follows that convention, as shown by the long history of gender-ambiguous names in the USA (Barry and Harper, 1982). Additionally, gendered associations for specific names change over time (Barry and Harper, 1993), as do naming conventions in societies. For example, it is becoming increasingly popular to assign non-gendered names in the USA and in Israel (Hough, 2016; Obasi et al., 2019). Apart from conventions, names are not static and unchanging from birth, with many people changing their names due to partnerships, adoption, transition to a different life stage or gender, and so on (Hough, 2016; Obasi et al., 2019; AIATSIS, 2022).

Assimilation and resistance to convention. Trends in big-picture naming conventions are complicated by factions of society who want to resist imposed classification. Increasingly heterogeneous societies are a natural setting for such tensions; cross-cultural associations with sociodemographic characteristics can differ and sometimes clash, complicating naming, e.g., names like *Nicola* and *Andrea* tend to be assigned to boys in Italy but to girls in Germany. As Germany is a society with highly regulated naming practices, inclusion of these names necessitated a court judgment (Hough, 2016). Immigrant families thus have to juggle the delicate balance of asserting their identity but avoiding name-based stigma and discrimination in the new culture. Their naming practices have therefore been studied as an indicator of attitudes towards assimilation or its rejection, showing how names are not a transparent indicator of race (Sue and Telles, 2007; Becker, 2009). Even among adults, imperialism and colonialism are forces that affect naming. Indigenous individuals have been forced to adopt Western names in settler colonial and postcolonial societies, e.g., the USA, Canada, Australia (AIATSIS, 2022). Similarly, Chinese individuals around the world adopt Western names in conversational (Li, 1997) and professional settings (Chan, 2016). Among trans and gender-nonconforming adults, many choose a new name to reflect and express their gender, walking the tightrope between normativity and self-assertion (Konnelly, 2021); Obasi et al. (2019) find that 50% of gender-nonconforming respondents who change their name pick a gender-neutral name. Indeed, two of the gender-nonconforming

authors of this paper have changed their names to gender-neutral names. Beyond transgender people, new names and pseudonyms are also often self-selected to assert agency in identity creation, e.g., *bell hooks*, *Sojourner Truth*, and *Malcolm X* (Baker and Green, 2021).

Quantitative aspects of naming. As naming involves a trade-off between differentiation and categorization, names often recur, a quantitative assumption that a lot of sociological, anthropological and NLP classification relies on (Alford, 1987). However, the distributions of names and people can be very different. Weitman (1981) finds that in 100 years of first names from Israel's Population Registry, the most frequent names (101+ occurrences) account for the majority of the population of a society (91%), but this corresponds to just a tiny minority of all assigned *names* (2.93%). These numbers could vary widely depending on the society, as, for example, the Chuukese people of Micronesia have a tradition of giving entirely unique names to children (Alford, 1987). Hence, it is important to distinguish when names are the object of study and when people are, to contextualize any results that involve the analysis of names.

4.3 Names and Sociodemographic Characteristics in NLP

Here, we present a non-comprehensive list of papers to illustrate some common uses of names and sociodemographic characteristics in NLP. Most papers that deal with names and sociodemographic attributes can be classified into one of two categories: Papers that attempt to infer the sociodemographic attributes of real people via their names, and papers that use names to exemplify certain sociodemographic groups, following their naming practices.

NLP tasks and problems. Numerous NLP works have developed algorithms to infer sociodemographic attributes from names (Chang et al., 2010; Liu and Ruths, 2013;

Knowles et al., 2016), e.g., for passive analysis of social media content. Another line of NLP papers has relied on names to quantify gender disparities in academic publishing (Vogel and Jurafsky, 2012; Mohammad, 2020) or media representation (Asr et al., 2021). Some NLP works have identified preserving dominant gender associations as an important criterion for transliteration and translation (Li et al., 2007; Wang et al., 2022). Names are also used to investigate social biases in NLP systems and language models (Kotek et al., 2023; An et al., 2023; Ibaraki et al., 2024). For example, De-Arteaga et al. (2019) study how first names, which they consider "explicit gender indicators," affect the gender bias of occupation prediction from biographies. Similarly, Jeoung et al. (2023) assess the causal impact of first names, which they posit "may serve as proxies for (intersectional) socio-demographic representations," on the commonsense reasoning performance of language models. Smith and Williams (2021) measure racial biases as well, evaluating generative dialogue models by having "one conversational partner [. . .] state a name commonly associated with a certain gender and/or race/ethnicity." In this line of research, it is commonplace to use skewed reference populations such as USA census data (U.S. Census, 2020) and Social Security Administration baby names (U.S. Social Security Administration, 2023) for gender assocations (Lockhart et al., 2023).

Engagement with pitfalls. In these works, researchers engage to varying degrees with the established methodological and ethical problems of associating names with sociodemographic characteristics. Some NLP papers make unfounded assumptions about names, e.g., Vogel and Jurafsky (2012) posit that certain names are "unambiguous" with respect to gender across languages, and Wang et al. (2022) claim that there exist "names with obvious gender." Other papers are more critically reflective, acknowledging the limitations of their work: Knowles et al. (2016) state that their classifier to predict gender from names is biased towards the USA and assumes gender is binary, but leaves these issues "to be addressed in future work." Mohammad (2020) acknowledges that inferring gender from names can yield misgendering because "names do not capture gender fluidity or contextual gender," but suggest a trade-off with "the benefits of NLP techniques and social category detection." Encouragingly, some recent papers opt for

more inclusive study designs after engaging deeply with the pitfalls of using names and sociodemographic characteristics (Sandoval et al., 2023; Saunders and Olsen, 2023; Lassen et al., 2023).

4.4 Validity Issues

In this section, we present issues of validity when associating names with sociodemographic categories, or using names to infer them for real people. Issues of validity mean that results with these operationalizations may neither be indicative of what we actually want to measure, nor of reality.

Inference errors are not quantifiable without asking humans. The accuracy of using names to infer sociodemographic characteristics of real people cannot be quantified without ground truth data, which for people's identities, can *only* be obtained by asking them. Multiple studies thus empirically analyze the error rates of name-based gender and race inference systems as compared to gold data in different contexts (Karimi et al., 2016; Kozlowski et al., 2022; Van Buskirk et al., 2023; Lockhart et al., 2023).[1] For example, Lockhart et al. (2023) evaluate gender and race inference systems using self-reported data from nearly 20,000 individuals. Importantly, their self-reported data does not directly transfer to other contexts, as their respondents are authors of English language social science journal articles who are mostly located in the USA. Using this data as reference data for a system with users located primarily in India, or for USA authors in a different century, makes little sense. In new environments, it is simply not possible to reasonably estimate the bounds of error of a name-based analysis, and results without a corresponding analysis of self-reported data should not be taken seriously.

1 All of these studies look at imputing an individual's gender, but the gold labels they compare to are, confusingly, not always self-reported gender! Some use gender assigned by annotators as the ground truth, which would be fine if they were comparing to *perceptions* of an individual based on their name, but these studies do not, raising further questions about their methodological validity.

Popular design choices for handling uninformative names lead to systematic error and selection bias. Names that are uninformative of a sociodemographic characteristic present an issue for computational analyses that rely on a connection between them. This is an issue that comes up frequently: In the context of gender, names like *Alex* have no unique gendered association in the USA and Canada; with race, names assigned by Black and white parents overlap in the USA (Lockhart et al., 2023), and religious names are used around the world (Curtis, 2005; Olúwáfei, 2014); at the intersection of gender and race, many Chinese names are not gender-associated when Romanized, and infrequent names are also not informative. In any type of computational analysis with names, two common design choices for handling uninformative names are to assign the majority class label anyway, or, alternatively, to just exclude them. Assigning the majority class (i.e., classifying *all* people named *Miaoran* as female if a gender prediction tool predicts the name to be "60% female") results in systematic error (Kirkup and Frenkel, 2006). This choice is common when inferring sociodemographic attributes from names. When using names as examples of sociodemographic groups, on the other hand, it is more common to simply exclude uninformative names from the analysis, even when this makes up the majority of all names. This completely alters the makeup of the data and therefore the results (Mihaljevi et al., 2019), resulting in selection bias. Both choices affect internal validity, i.e., gaps in the translation from measurements to overall conclusions (Liao et al., 2021), leading to less robust and trustworthy results.

Poor construct validity. Construct validity asks how well an abstract concept can be measured through some indicator (Messick, 1995); in our case, the question is: How valid is it to assign sociodemographic categories via names?[2] The answer to this depends on what aspects of the sociodemographic category we are interested in: Identity, socialization, expression, perception—all of which could differ and are frequently conflated (Keyes et al., 2021). As discussed previously, many names are

2 While we focus on the construct validity of names in this section, we note that poor construct validity also applies to the sociodemographic categories themselves (Benthall and Haynes, 2019; Hanna et al., 2020) and to abstract concepts such as "bias" and "fairness," which show up frequently in the study of names and sociodemographic categories in NLP (Blodgett et al., 2020; Jacobs and Wallach, 2021).

simply not informative of certain sociodemographic *identities* in given contexts and with homogeneous populations; Lockhart et al. (2023) find that overall error rates of name-based gender and race imputation tools range from 4.6% to 86% overall, and up to 100% for particular subgroups, depending on the tool. However, when it comes to the *perception* of names as indexing a sociodemographic category, some names may have stronger construct validity, an assumption used by Sandoval et al. (2023) in their examination of names assigned at birth that are strongly associated with the baby's sex and the parents' race/ethnicity. On the other hand, Mohammad (2020) uses names to operationalize both identity (to investigate trends in authorship) and perception (to investigate trends in citation) in a bibliometric analysis of the ACL Anthology, even though these need not match, and many underrepresented names are uninformative of identity as well as perception (Van Buskirk et al., 2023). As names do not neatly line up with sociodemographic identities, perceptions, or experiences in a context-independent way, it is critical to investigate construct validity of names in any setting where they are used, and use this to inform study design.

Systems of classification *create* results. Although classification is inherently human, classification systems are produced by culture and politics, and end up *creating* a view of the world (Bowker and Star, 2000). In computing, researchers have power and our positionality shapes how we view and operationalize categories of classification such as race and gender (Scheuerman et al., 2020b; Scheuerman and Brubaker, 2024). However, many such categories are unstable and contested (Keyes et al., 2021; Mickel, 2024). For instance, it has been shown that different ways of operationalizing race can result in entirely different conclusions (Steidl and Werum, 2019; Benthall and Haynes, 2019; Hanna et al., 2020). Individuals and groups thus cannot be treated as monoliths that can be characterized one-dimensionally via names.

4.5 Ethical Issues

The issues we have examined so far impact the scientific validity of claims made using personal names and sociodemographic categories. Many of these problems arise from assumptions that can also be criticized on ethical grounds, as we show.

Errors cause harms. Harms can be broadly described as a setback in the interests or progress of people due to, e.g., the outcomes of an automatic process (Feinberg, 1984). Group-level harms are experienced collectively by people in a sociodemographic group, while individual harms (which might result from group membership) are experienced at the person-person or person-technology level. Inferring gender from names frequently misgenders trans people and erases non-binary people (Keyes, 2018). This perpetrates group-level erasure, as well as individual harms including damaging autonomy and dignity (Mcnamarah, 2020), inflicting psychological harms (Dev et al., 2021), and a failure to show recognition respect to people (Darwall, 1977). Certain types of name-based classification (e.g., of persecuted ethnic or religious groups) can threaten individual safety, and when NLP infrastructure is used for surveillance and targeting, this also threatens the safety of entire groups of people (Wadhawan, 2022). Regulation efforts such as the AI Act (Commission, 2021) in the EU try to mitigate this, but this does not apply to authoritarian regimes' use of such technology (Briglia, 2021). NLP systems reinforce group-level structural discrimination in other ways as well; name-based studies of racial disparities in academia have been shown to systematically discount the intellectual contributions of Black researchers (Kozlowski et al., 2022).

Errors and harms are not distributed equally. In their work on name-based gender classification, Van Buskirk et al. (2023) note that for names with no available data, assigning the majority class (in their case, male) maximizes accuracy, but results in 0% error for the male class and 100% error for any other classes. For non-binary people, who are generally excluded from gender classification by design, the error rate is also almost always 100%. As for name-based race/ethnicity classifiers, Lockhart

et al. (2023) show that people who self-identify as Filipino, Black, or Middle Eastern and North African, are misrecognized 55-75% of the time, as compared to those who identify as white, Chinese, or Korean, who are mislabelled less than 10% of the time. As described above, misrecognition errors cause harms, which are then disproportionately experienced by these individuals. We echo the conclusions of Mihaljevi et al. (2019) and Lockhart et al. (2023), i.e., that inclusive analyses are only possible when names are no longer used as a proxy to infer individuals' gender or race/ethnicity.

Representational harms. The erasure of identities and the flattening of variation in naming customs leads to representational harms, which include the reinforcement of essentialist categories and power structures (Chien and Danks, 2024). These harms primarily affect sociodemographic groups, e.g., non-binary people, who are often incorrectly and unjustly treated as a novel social phenomenon. Groups of people with a certain name are often subject to a different type of representational harm, i.e., stereotyping. For instance, the name *Kevin* is associated with lower socioeconomic class in Germany (Kaiser, 2010). This stereotype, if encoded in an NLP system, could lead to quality-of-service differentials, as class is a sociodemographic characteristic that correlates with lower NLP performance in other contexts (Curry et al., 2024).

Cultural insensitivity. Conceptualizations of names and sociodemographic characteristics in NLP are often Western-centric, with folk assumptions about what names look like and the application of USA racial categories and naming preferences to areas outside the USA, where they are unintelligible (Field et al., 2021). Non-Western naming practices are only sometimes described in papers where there is a specific language of study that is not English, e.g., name tagging in Arabic (Shaalan and Raza, 2007) and Uyghur (Abudukelimu et al., 2018). Even within English, there is little recognition of, e.g., English common nouns used as names in China (*Billboard*, *Shooting*, *Pray*, etc.; Chan, 2016), names containing spelling variations (AIATSIS, 2022), and names that overlap in different cultures but have different associations, e.g., *Jan* in the USA compared to *Jan* in Germany. Beyond names, even gender, race, and other sociode-

Theme	Guiding questions
Names vs. people	What are you aiming to study–names? Or people, via their names? What aspects of names are you interested in? What aspects of people are you interested in?
Context	What is your context? Is processing names with NLP systems necessary to answer your questions?
Harms and power	What kinds of harms apply? How can you mitigate them? Are you describing or prescribing? How does your work reify/redistribute power?
Refusal	Is it still worth it?

Table 4.1: Our list of guiding questions for the use of names and sociodemographic categories in NLP, grouped by theme. See paragraphs in Section 4.6 for detailed recommendations.

mographic categories of relevance are different across cultures. Many cultures have definitions of gender that go beyond the binary. Enforcing binary gender can thus be seen as an example of what Lugones (2016) calls the "coloniality" of gender, which also results in epistemic violence, i.e., inhibiting people from producing knowledge, or silencing and discrediting their knowledge (Chilisa, 2019).

Power is centralized. Names are a site for enforcing institutional power, as seen in "real name" policies (Haimson and Hoffmann, 2016), the (non-consensual) permanence of names in data infrastructure including Google Scholar (Speer, 2021), governmental name regulation (Te Tari Taiwhenua, 2021), and the "collective delusion" of legal names, at least in the USA (Baker and Green, 2021). Names are also regulated socially through norms and expectations, many of which end up baked into our NLP systems. We exercise power as NLP researchers and practitioners via our assumptions, which may reify sociodemographic categories, codify (or dismantle) associations between

names and these categories, and create infrastructure that harms people at scale through surveillance or mislabelling. Knowles et al. (2016) open-sourced their name-based gender inference tool, and Vogel and Jurafsky (2012) published (binary) gender labels with names of authors of NLP papers, which continue to be used in research (Mohammad, 2020; Van Buskirk et al., 2023). This data reflects folk assumptions about gender, i.e., that it is binary, immutable and in perfect correspondence with names (Keyes, 2018; Cao and Daumé III, 2021). These datasets also deadname and misgender scientists from the NLP community, some of whom have spoken about its harms (Mielke, 2024). Transgender people can only be counted in such a system if they conform to normative expectations (Johnson, 2016; Konnelly, 2021), and if not, the burden is disproportionately on them to seek redress. Even Asr et al. (2021)—a system relying on name-based gender inference that considers gender beyond the binary and does not publicly misgender individuals—does not shift power, as workarounds are a patch rather than built-in to the method; gender inference still relies on APIs that use binary gender, and mistakes (typically, famous non-binary people) are manually corrected. As all these examples show how power remains centralized, we echo previous calls to reimagine and reconfigure power relations in service of user autonomy (Keyes et al., 2019; Blodgett et al., 2020; Hanna and Park, 2020).

4.6 Guiding Questions and Recommendations

In the previous sections, we have reviewed the myriad of issues surrounding the accuracy, validity and ethical use of names along with sociodemographic characteristics, and noted that all these issues arise from the same assumptions and inform each other. In addition, we have shown that these problems apply overwhelmingly to those who are not cisgender, white, normatively named in a Western context, and well-represented in publicly available data. Thus, work that uses names to operationalize people's sociodemographic categories most misrepresents and further marginalizes those who are already at the margins. We take the normative position that this is not acceptable collateral damage, even (and especially!) in the name of ostensible fairness. Thus, we

come up with guiding questions and recommendations for NLP practitioners who are considering the use of names as they relate to sociodemographic categories. These are summarized in Table 4.1.

What are you aiming to study–names? Or people, via their names? It is acceptable to investigate what concepts NLP models associate with names, e.g., *Madeleine* with *kindness*. It is even acceptable to demonstrate that NLP models associate *Marius* with the pronoun *he* or with being male, and that these associations mirror common human associations (Caliskan et al., 2017; Crabtree et al., 2023). It is marginally acceptable to associate names with sociodemographic characteristics using imaginary people, e.g., drawing insights about gender bias more broadly based on how NLP models handle synthetic names of people assumed to be exclusively female; while doing so does not compromise people's autonomy and dignity, it does further entrench hegemonic folk theories of names and people's identities, which has cultural harms. Finally, it is unacceptable to present results about real people based solely on the assumption that their names provide a reliable signal about their identities, e.g., NLP papers authored by people named *Madeleine* and *Marius* cannot on their own provide trustworthy insights into gender and racial representation in the field, unless those specific individuals are asked about their gender.

What aspects of names are you interested in? Names are rich objects of study with variation in form, length, training data frequency, tokenization, associations, the strengths of these associations, and more, some of which have already been explored in prior work in NLP (Shwartz et al., 2020; Wolfe and Caliskan, 2021; Sandoval et al., 2023). Once you have decided what aspects to study, they must be operationalized and measured carefully, with attention to the context of the study or eventual system deployment. This includes the scope of what counts as a "name." For instance, considering the use of English common nouns as names (e.g., *Cloud*) is particularly important when working with data from or systems deployed in China, where this naming practice is common (Chan, 2016). Ensure that pre-processing choices are contextualized and

do not distort results, that names are understood within context, and that error can be quantified robustly in the given context. Thus, when measuring training data frequency of names, counting *Cloud* tokens as names must consider when it is used as a name and when it is used simply as a noun. Error could be quantified through manual analysis on a subset of the data.

What aspects of people are you interested in? People's identity and perceptions of them can differ, and these shape their experiences in various ways. Therefore, it is first necessary to decide which aspects are relevant for a study. *Attempting to infer someone's identity using names is simply unacceptable due to the range of methodological and ethical concerns we list in this paper.* We echo onomastic advice from nearly 40 years ago (Weitman, 1981), i.e., that "inferences from names must be to the givers of these names, not to their bearers. What is more, inferences must always be to sociological formations (such as social classes, ethnic groups, historical generations, and the like), not to individual name-givers." In addition to studying formations of name-givers, it can also be acceptable to study perceptions of identity based on names. For instance, numerous sociology papers have investigated racial and ethnic perceptions, as well as occupational stereotypes, based on names (King et al., 2006; Gaddis, 2017a; Gaddis, 2017b). Again, we emphasize that perceptions based on names are also highly contextual and non-universal.

What is your context? It is essential to understand the geographical, temporal, and cultural context of data with names, and document this information for datasets, e.g., with datasheets (Gebru et al., 2021). What is the geographic, temporal, cultural, and political context of the name data, name-bearers, models, and sociodemographic categories you use? Who are the people who will be impacted by your work, and what is their context? What do you know about the naming practices in these contexts and the hetereogeneity in these practices? Are you quantifying error with self-reported data? We posit that it is unacceptable to use names without deeply engaging with context

in these senses, and stress that ascribing contemporary Western identity categories to historical peoples without acknowledging the difference in contexts is reductive.

Is processing names with NLP systems necessary to answer your questions? For questions about human identity and perception based on names, NLP may not be the only or best method available. We warn against technical solutionism (Green, 2021); researchers should reflect on whether their questions could be approached with interviews, case studies, and so on (Cameron, 2004). Qualitative methods can provide deeper, richer evidence while respecting people's autonomy, dignity and context. If your questions are instead about NLP systems, then processing names with them is certainly necessary, but we note that methodological pluralism and interdisciplinarity can enrich our practice as NLP researchers and practitioners regardless (Wahle et al., 2023).

What kinds of harms apply? How can you mitigate them? Our paper provides a starting point for harms that are relevant to the use of names and sociodemographic characteristics in NLP, and we encourage transparency about methodological and ethical problems (Bietti, 2019; Hao, 2019). It is unacceptable to sideline these problems in the name of "social good" (Green, 2019; Greene et al., 2019; Bennett and Keyes, 2020). Rather than treating entire segments of the population as limitations of or future work for one's research, we encourage changing the methods themselves, as Lauscher et al. (2022) do with neopronouns. We recommend firmly grounding work in the ethical principles of autonomy, justice, and beneficence for people (Floridi and Cowls, 2019), which we note are sadly under-represented in machine learning research (Birhane et al., 2022).

Are you describing or prescribing? Despite being distinct from each other, descriptions of social phenomena are often conflated with normative behaviour (i.e., assumptions and assertions that create and reinforce norms) in NLP (Vida et al., 2023). This is the subtle but significant difference between showing that sociodemographic name associations in language models mirror the judgements of some group of humans,

versus stating that model associations *should* mirror the judgements of some group of humans. The latter "cannot avoid creating and reinforcing norms" (Talat et al., 2022). Therefore, researchers should clearly distinguish descriptive and normative behaviours in the design, execution, and presentation of their experiments (Vida et al., 2023). System designers do have to make decisions about how systems *should* behave, i.e., they need to choose to perpetuate harmful structures in service of usability or to impose their own values on users and stakeholders when they take an advocacy position. This is an ethical dilemma in design that participatory methods and feminist epistemologies are uniquely positioned to help with (Bardzell, 2010).

How does your work reify or redistribute power? Central to NLP and computer science at large are scale thinking (Hanna and Park, 2020), quantitative methodologies (Birhane et al., 2022), and the illusion of objectivity (Waseem et al., 2021). All these values serve to reify existing hierarchies and power structures. We must first recognize our own power as NLP researchers and practitioners, and how our work can reinforce infrastructure for (mis)classifying real people, and enable surveillance and harms at scale. We recommend a counterpower stance (Keyes et al., 2019), situated knowledges (Haraway, 1988), and methods informed by a politic, e.g., intersectionality, a critical framework that centers justice, power, and reflexivity, and mandates praxis with teeth (Collins, 2019; Erete et al., 2018; Ovalle et al., 2023b). Particularly for those of us who are interested in using NLP for social good, we should constantly be asking: "Social good for whom?" The differential impact on people matters, and as researchers and practitioners, we have a responsibility to attend to it and resist the othering perpetuated by classification systems.

Is it still worth it? After considering all these guiding questions, we remind the reader that refusal is possible (Honeywell, 2016; Tatman, 2020; Lockhart et al., 2023; Mihaljevi et al., 2019), and indeed an important part of the history of science (Williams, 1924; United States National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1978; Weindling, 2001).

4.7 Related Work

Several papers study and critically interrogate the inference and use of sociodemographic information in computing (Larson, 2017; Keyes, 2018; Benthall and Haynes, 2019; Hanna et al., 2020; Keyes et al., 2021; Field et al., 2021; Devinney et al., 2022), many of which touch upon names but do not address them in detail. The work that deals with names in particular are all outside of NLP: Karimi et al. (2016), Keyes (2017), Tzioumis (2018), Mihaljevi et al. (2019), Scheuerman et al. (2019), Lockhart et al. (2023), and Van Buskirk et al. (2023). These papers have different scopes and take a variety of positions with regards to the ethics of name-based inference, some of which we find insufficiently radical. Finally, our recommendations echo those from prior work (particularly in the fields of human-computer interaction and science and technology studies), but are contextualized for names in NLP. Among others, we take inspiration from Keyes et al. (2019), Hanna and Park (2020), Blodgett et al. (2020), Scheuerman et al. (2020a), and Green (2021).

4.8 Limitations

Our background on names and naming is limited, and meant only as a brief introduction to onomastics and related fields that use names and sociodemographic characteristics; we refer the interested reader to our references for deeper discussion of onomastic variation. Additionally, we know that problematic and decontextualized assumptions about names are rife within NLP based on our background as authors within or adjacent to the field, as well as writing in other fields about methods that are also popular in NLP. However, as we do not undertake a comprehensive, critical survey of NLP papers that use names and sociodemographic characteristics, we cannot empirically quantify the extent to which the problems we outline plague NLP research, and we leave a more systematic study of this to future work.

4.9 Conclusion

In this chapter, we answered the first research question of this thesis, showing that personal names and not a reliable proxy for sociodemographic characteristics of interest, such as race and gender. We did this by presenting an overview of names and naming practices around the world, and as as discussed in disciplines outside of NLP. We showed how NLP uses of names and sociodemographic characteristics has issues of validity (e.g., selection bias and construct validity) and ethical concerns (e.g., harms, cultural insensitivity). Finally, we presented a list of guiding questions and normative suggestions towards addressing these concerns for fairer future work involving names in NLP. Next, we turn our attention to personal pronouns, asking whether they fare better as a proxy for social gender.

5

Revisiting the Relationship between Pronouns and Social Gender

When referring expressions have the same referent, they are said to co-refer, and the task of identifying such co-referring expressions in natural language is called coreference resolution. Several studies have shown that biases can cause both humans and machines to incorrectly resolve coreferences; one such influential study is Winogender Schemas (Rudinger et al., 2018), which uses pronouns as a proxy for gender to evaluate gender bias in coreference resolution. However, a closer look at their data reveals an untested assumption that different pronominal forms are equivalent, violations of template constraints, and typographical errors. As these issues compromise the dataset's use for reliable evaluation, we fix them and contribute a new dataset: WINOPRON. Using WINOPRON, we evaluate two state-of-the-art supervised coreference resolution systems, SPANBERT, and five sizes of FLAN-T5, and demonstrate that accusative pronouns are harder to resolve for all models. We also propose a new method to evaluate pronominal bias in coreference resolution that goes beyond the binary. With this method, we also show that bias characteristics vary not just across pronoun sets (e.g., *he* vs. *she*), but also across surface forms of those sets (e.g., *him* vs. *his*). Our analyses show that pronouns are a noisy stand-in for gender, and that different forms indicating the same grammatical gender do not show consistent performance and bias characteristics.

The content in this chapter is based on:

Vagrant Gautam, Julius Steuer, Eileen Bingert, Ray Johns, Anne Lauscher, and Dietrich Klakow (Nov. 2024b). "WinoPron: Revisiting English Winogender Schemas for Consistency, Coverage, and Grammatical Case." In: *Proceedings of The Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*. Ed. by Maciej Ogrodniczuk, Anna Nedoluzhko, Massimo Poesio, Sameer Pradhan, and Vincent Ng. Miami: Association for Computational Linguistics, pp. 52–66. DOI: 10.18653/v1/2024.crac-1.6. URL: https://aclanthology.org/2024.crac-1.6/

Vagrant Gautam conceptualized the research, created the dataset together with Eileen Bingert, and led the paper writing. Dataset validation was performed by Vagrant Gautam, Eileen Bingert, and Ray Johns. Vagrant Gautam and Julius Steuer performed experiments with coreference resolution systems and large language models, and analyzed the results. Anne Lauscher and Dietrich Klakow advised and provided feedback.

5.1 Introduction

To refer to someone, we can choose from a variety of linguistic expressions, including definite descriptions (*the birder*), pronouns (*xe*), and names (*Vagrant*), under certain cognitive and pragmatic constraints (Prince, 1981; Prince, 1992; Gundel et al., 1993). Among these options, third-person pronouns (*he*, *she*, singular *they*, etc.) are unique in English as they explicitly mark grammatical gender, which also often indexes social gender of the person being referred to. When referential gender (e.g., the feminine *she*) conflicts with the gendered expectations of a co-referring expression (e.g., the conceptually masculine *surgeon*), people have difficulties mapping these pronouns to the right referent (Kennison and Trofe, 2003; Bjorkman, 2017; Ackerman, 2019; Morehouse et al., 2022).

Rudinger et al. (2018) demonstrated that these biases also emerge within natural language processing systems, by introducing Winogender Schemas, a challenge dataset to evaluate occupational gender bias in coreference resolution systems. The dataset has become popular due to its careful construction; it has been translated to other

languages (Hansson et al., 2021; Stanovsky et al., 2019) and used in framings beyond coreference resolution, e.g., to evaluate natural language inferences (Poliak et al., 2018) and intrinsic bias in language models (Kurita et al., 2019).

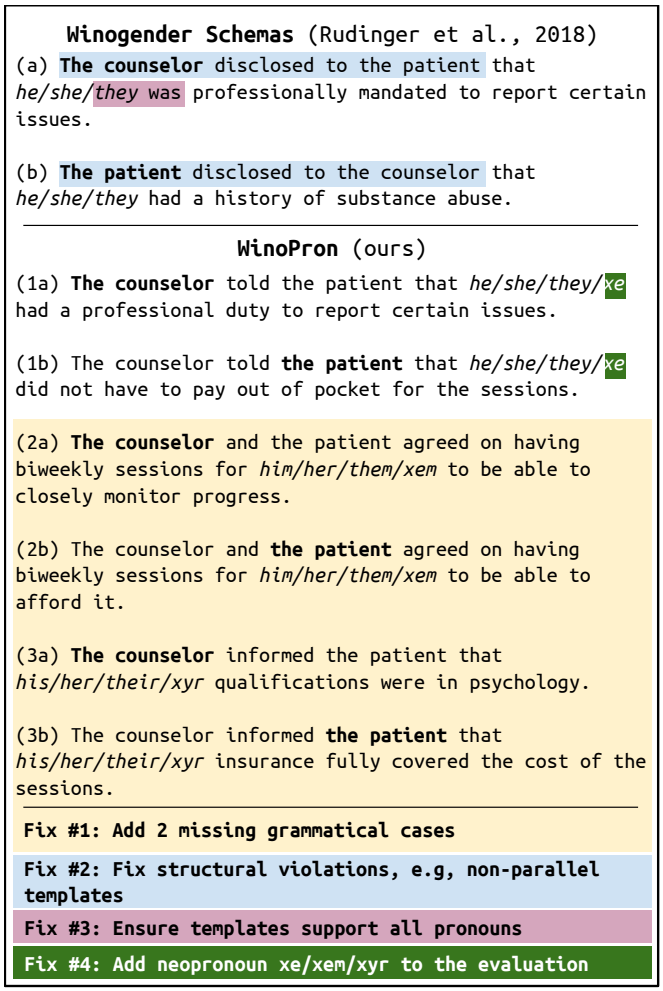


Figure 5.1: Problems with Winogender Schemas that we fix in our new coreference resolution dataset, WINOPRON. Correct antecedents appear in **bold**.

However, a closer look at the dataset reveals weaknesses that compromise its use for reliable evaluation (see Figure 5.1), which we hypothesize would affect both performance and bias evaluation. These include: Treating different pronominal forms as equivalent, varying more than just pronouns in templates that should otherwise be identical, and not conforming to structural requirements. In this chapter, we identify issues with the original dataset and fix them to create a new dataset we call WINOPRON (Section 5.3). We also propose a novel method to evaluate pronominal bias in

coreference resolution that goes beyond the binary and focuses on linguistic rather than social gender (Cao and Daumé III, 2021). With this dataset and method, we empirically test the assumption that different pronouns with the same grammatical gender behave similarly with regards to coreference resolution system performance (Section 5.5) as well as bias (Section 5.6), and find that they do not.

Our fixes reveal that grammatical case, which we balance for in WINOPRON, does indeed affect both performance and bias results, even though prior work does not distinguish between pronouns of different grammatical cases with the same grammatical gender; accusative pronouns are harder to resolve than nominative or possessive pronouns, and system pronominal bias is not always consistent across different grammatical cases of the same pronoun set. We find that singular *they* and the neopronoun *xe* are extremely hard for supervised coreference resolution systems to resolve, but surprisingly easy for FLAN-T5 models of a certain size. Since pronoun form has such a profound effect on performance and bias, this chapter highlights the need for future studies of stereotypical bias in coreference resolution to control this aspect of their data. We release our data and code at github.com/uds-lsv/winopron.

5.2 Background: Winogender Schemas

Winogender Schemas (Rudinger et al., 2018) are a widely-used dataset consisting of paired sentence templates in English, with slots for two human entities (an occupation and a participant), and a third person singular pronoun.

(1) Winogender Schemas for *cashier*, *customer* and possessive pronouns

a. The cashier$_i$ told **the customer**$_j$ that {**his**$_j$ / **her**$_j$ / **their**$_j$} card was declined.

b. **The cashier**$_i$ told the customer$_j$ that {**his**$_i$ / **her**$_i$ / **their**$_i$} shift ended soon.

As Example (1) shows, the second part of each template disambiguates which of the two entities the pronoun uniquely refers to, similar to Winograd schemas (Levesque et al.,

2012). Changing the pronoun (e.g., from *his* to *her*) maintains the coreference, allowing us to measure whether coreference resolution systems are worse at resolving certain pronouns to certain entities. Beyond analyzing differences in system performance with masculine, feminine and neutral pronouns, Rudinger et al. (2018) also map the grammatical gender of pronouns to the *social* gender they are thought to index, to show that real-world gender bias affects coreference resolution performance. In other words, masculine *he* is mapped to male, and feminine *she* is mapped to female for comparison to (binary) labour statistics in the USA.

The original list of entities consists of 60 occupations, each paired with a contextually appropriate participant, e.g., *accountant* is paired with *taxpayer*. For each occupation-participant pair, two templates with pronouns are created, such that one resolves to the occupation, and the other resolves to the participant. The template pairs are designed to be parallel until the pronoun, such that only the ending can be used to disambiguate how to resolve the pronoun. In total, these 120 unique templates are each instantiated with three pronoun sets (*he*, *she*, and singular *they*), for a total of 120 x 3 = 360 sentences for evaluation.

5.3 WinoPron Dataset

Although Winogender Schemas are established in the coreference resolution literature, we find issues with the dataset that compromise its use for reliable evaluation (see Figure 5.1 for examples). We first motivate these issues and our fixes, and then describe how we create and systematically validate our new dataset, WINOPRON.

We mostly reuse the occupation-participant pairs from Winogender Schemas (see Appendix A.1 for the full list of pairings), but add 240 new templates to cover missing grammatical cases, for a total of 360 templates. We also include a neopronoun set (*xe/xem/xyr*), giving us 360 templates x 4 pronoun sets = 1,440 sentences for evaluation.

Grammatical case	Winogender Schemas	WINOPRON
Nominative (*he*, *she*, *they*, *xe*, etc.)	89	120
Accusative (*him*, *her*, *them*, *xem*, etc.)	4	120
Possessive (*his*, *her*, *their*, *xyr*, etc.)	27	120
Total	120	360

Table 5.1: Number of templates per grammatical case in Winogender Schemas and WINOPRON.

5.3.1 Issues in Winogender Schemas and Solutions in WinoPron

Here, we detail the issues with Winogender Schemas that we fix in WINOPRON.

Support for 3 grammatical cases. We hypothesize that systems have different performance and bias characteristics with pronouns in different grammatical cases (here, we mean the surface form of the pronoun). However, as Table 5.1 shows, there is a variable number of pronouns per grammatical case in Winogender Schemas, and all of them are treated as equivalent. To enable more granular evaluation, we balance this distribution in WINOPRON.

Consistency fixes. Winograd-like schemas have strict structural constraints so that models cannot inflate performance through heuristics. However, when analyzing Winogender Schemas, we found constraint violations, e.g., non-parallel paired templates, which could be gamed by always guessing the first entity in the template. We fixed these along with typographical errors to ensure robust and reliable evaluation.

Support for all English pronouns. For a controlled evaluation comparing pronouns, it is common to use templates that only vary the pronoun. However, 17% of Winogender Schemas must be modified to work with singular *they* due to its different verbal agreement ("he was" but "they were"). To ensure a fair comparison across differ-

ent pronouns, we modify these templates to work with any pronouns. This requires only using past tense constructions, and avoiding the verb *be*.

Single-entity versions. When evaluating large language models on coreference resolution when they have not explicitly been trained for it, poor performance could mean that the model simply cannot perform the task (with a given prompt). In its current form, Winogender Schemas do not allow us to disentangle *why* bad model performance is bad. In WINOPRON, we create single-entity sentences that are parallel to the traditional, more complex double-entity sentences, for a simple setting to test this, and a useful baseline for all systems.

5.3.2 Data Creation

Two authors with linguistic training iteratively created sentence templates until we reached consensus on their grammaticality and correct, unique coreferences. We found template construction to be particularly challenging and time-consuming, due to ambiguity and verbal constraints.

Ambiguity. Our biggest source of ambiguity during template creation was singular *they*, as *they* is also a third person plural pronoun. For example, if an *advisor* and *student* were meeting to discuss *their* future, this could potentially refer to their future *together*. This problem applied across grammatical cases. In addition, possessive sentences were potentially ambiguous across all pronoun series; when discussing a *doctor* and a *patient* and someone's diagnosis, this could be the *doctor*'s diagnosis (i.e., the diagnosis made by the doctor), or the *patient*'s diagnosis (i.e., the diagnosis the patient received). All ambiguous templates were discarded and subsequently reworked.

Verbal constraints. The structural constraint of template pairs being identical until the pronoun led to some difficulties in finding appropriate (logically and semantically plausible) endings for the two sentences, particularly with accusative pronouns. With

nominative pronouns, we had to ensure we used verbs in the past tense and avoid was/were, so that our templates could be used with both *he/she/xe* and singular *they*. It was also sometimes difficult to create single-entity sentences that were semantically close to the double-entity versions because the latter only made sense with two entities.

5.3.3 Data Validation

As WINOPRON templates have structural constraints that can be programmatically validated, we wrote automatic checks for these. In addition, we performed human annotation of the sentences for grammaticality, and unique, correct coreferences.

Automatic checks. We automatically checked our data for completeness first, i.e., that every occupation-participant pair had sentence templates for nominative, accusative, and possessive pronouns. We then automatically checked structural constraints, e.g., that a pair of templates must always be identical until the pronoun slot, and that no additional pronouns appeared in the sentence.

Human annotation. Both authors who created the schemas systematically annotated them, rating 100% of the final instances as grammatical and 100% of them as having unique, correct coreferences. We confirmed the uniqueness of coreferences by marking each data instance as coreferring with the appropriate antecedent and also not coreferring with the other antecedent. An additional annotator independently verified the final templates, rating 100% of them as grammatical, and 98.2% as having unique, correct coreferences.

5.4 Models

We select a range of systems with different architectures and levels of training for coreference resolution, for evaluation of performance, consistency, and bias using WINOPRON.

LINGMESS. LINGMESS is a state-of-the-art, linguistically motivated, mixture-of-experts system for coreference resolution (Otmazgin et al., 2023). Expert scorers—each specializing in different linguistic features such as syntax, semantics, and discourse—independently evaluate coreference links, and their scores are combined to make the final resolution decisions. This approach allows the model to consider a wider range of linguistic signals than traditional single-scorer models.

CAW-COREF. CAW-COREF (D'Oosterlinck et al., 2023) is a state-of-the-art word-level coreference resolution system based on an encoder-only model that builds on Dobrovolskii's (2021) encoder-only architecture. After an initial forward pass through an encoder model, an antecedent scoring matrix is constructed from contextualized word embeddings; for each word, the top k antecedents are scored again by a feedforward neural network. Both scores are then used to identify the most likely antecedents of each word.

SPANBERT. SPANBERT is an encoder-only language model pre-trained with a span prediction objective and further enhanced for coreference resolution with fine-tuning data (Joshi et al., 2020). This pre-training method improves model performance on tasks that rely on spans of text, e.g., question answering, named entity recognition, and coreference resolution. We use both available model sizes (base and large) for evaluation.

FLAN-T5. FLAN-T5 is an encoder-decoder language model that is instruction-tuned, but not explicitly trained for coreference resolution (Chung et al., 2024). We

evaluate on five model sizes (small, base, large, xl, and xxl), with 10 prompts from the FLAN collection (Longpre et al., 2023). See Appendix B.2 for more details on prompting.

5.5 Performance and Consistency

To demonstrate the effects of our changes, we evaluate performance and consistency metrics with all models on WINOPRON.

5.5.1 Performance Results

We first show how our changes affect overall performance between Winogender Schemas and WINOPRON. Then we use WINOPRON to investigate differences across case (which we have balanced for) and pronoun sets (which can now be evenly compared). Detailed results with single- and double-entity sentences and additional metrics (F_1, precision, recall) are provided in Appendix C.

WINOPRON is harder than Winogender Schemas. As Table 5.2 shows, all the systems we evaluate perform worse on WINOPRON, with F_1 dropping on average by 10 percentage points compared to Winogender Schemas. Patterns of performance across models are similar between Winogender Schemas and WINOPRON, with similar scaling behaviour for both SPANBERT and FLAN-T5. Notably, scale seems to supercede supervision, as the largest FLAN-T5 models perform the best overall. Smaller FLAN-T5 models perform at chance level, which is likely a reflection of the "demand gap" induced through prompting (Hu and Frank, 2024).

Accusative pronouns are harder. When model accuracy is split by grammatical case and pronoun series, we see that *all* models struggle with accusative pronouns. In general, systems perform best at resolving nominative pronouns, with a slight decrease

System	Winogender Schemas	WINOPRON	ΔF_1
LINGMESS	85.5	64.4	-21.1
CAW-COREF	81.3	67.3	-14.0
SPANBERT-BASE	71.8	61.6	-10.2
SPANBERT-LARGE	82.0	70.1	-11.9
FLAN-T5-SMALL	52.2	51.6	-0.6
FLAN-T5-BASE	66.6	62.4	-4.2
FLAN-T5-LARGE	89.2	78.0	-11.2
FLAN-T5-XL	97.4	89.0	-8.4
FLAN-T5-XXL	97.5	88.8	-8.7

Table 5.2: Overall performance (F_1) of coreference resolution systems on Winogender Schemas and WINOPRON. WINOPRON is harder for all systems.

for possessive pronouns and a large drop for accusative pronouns, as seen in Figure 5.2. This finding holds even for the best performing models on WINOPRON, FLAN-T5-XL and FLAN-T5-XXL, where accuracy with accusative pronouns (81.9% and 78.6%) is much lower than with nominative (94.3% and 96.3%) or possessive (89.3% and 90.0%) pronouns. We hypothesize that the performance gap for accusative pronouns is partially an effect of frequency; *him* tokens appear roughly half as often in large pre-training corpora as *he* and *his* tokens (Elazar et al., 2024).

Performance with singular *they* and neopronouns is bimodal. For the supervised coreference resolution systems (LINGMESS and CAW-COREF), performance with singular *they* is close to chance, and performance with the neopronoun *xe* is far below chance, despite good performance with *he/him/his* and *she/her/her*. SPANBERT performance also shows a gap between singular *they* and neopronoun performance compared to data-rich pronouns, although the gap is much smaller. These findings mirror those of Cao and Daumé III (2020) and Lauscher et al. (2022).

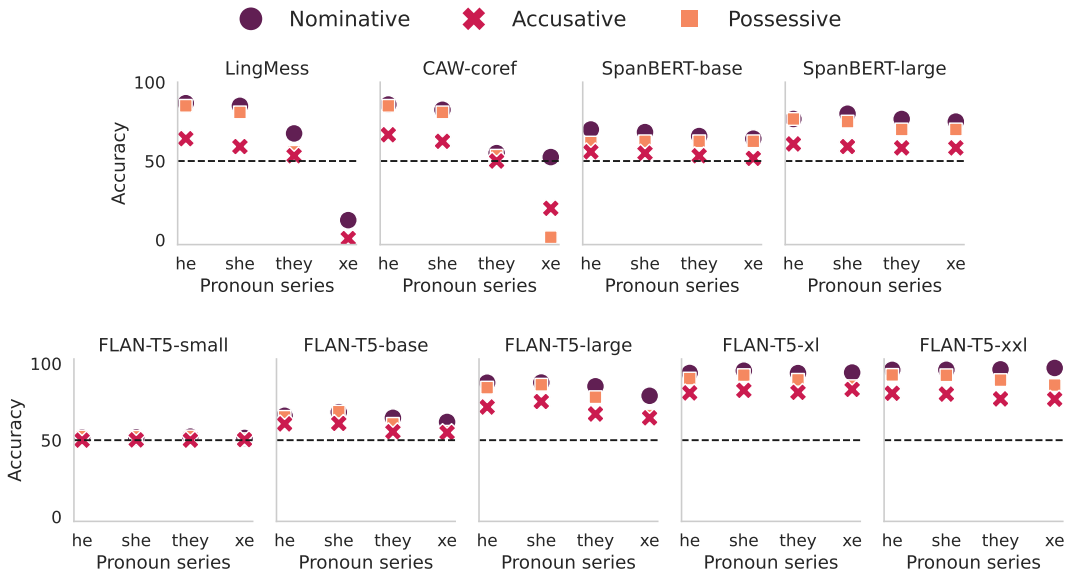


Figure 5.2: Accuracy on WINOPRON by case and pronoun series with supervised coreference resolution systems (CAW-COREF and LINGMESS), and language models fine-tuned for coreference resolution (SPANBERT) and prompted zero-shot (FLAN-T5), compared to random performance (50%). Accusative pronoun performance is worse than other grammatical cases, and singular *they* and the neopronoun *xe* are challenging for several models.

5.5.2 Consistency Results

Next, we evaluate system consistency on groups of closely related instances in WINOPRON, in order to dissect performance results and examine if systems are really right for the right reasons. We follow Ravichander et al. (2022) in operationalizing consistency by taking the score of the lowest-performing instance in the group as the group's score. We consider two groups: (a) *Pronoun consistency*, and (b) *disambiguation consistency*, inspired by Abdou et al.'s (2020) pair accuracy on Winograd Schemas. In both cases, we report the percentage of groups for which a model performs consistently.

(2) Group to measure pronoun consistency

a. **The counselor**$_i$ informed the patient$_j$ that **his**$_i$ qualifications were in psychology.

b. **The counselor**$_i$ informed the patient$_j$ that **her**$_i$ qualifications were in psychology.

c. **The counselor**$_i$ informed the patient$_j$ that **their**$_i$ qualifications were in psychology.

d. **The counselor**$_i$ informed the patient$_j$ that **xyr**$_i$ qualifications were in psychology.

(3) Group to measure disambiguation consistency

a. **The counselor**$_i$ informed the patient$_j$ that **xyr** $_i$ qualifications were in psychology.

b. **The counselor**$_i$ informed the patient$_j$ that **xyr** $_j$ insurance covered the cost of the sessions.

As shown in Example (2), pronoun consistency measures model robustness across pronoun sets, i.e., if a model fails with even one pronoun set on a given template, then its score for that template is zero. As we consider four pronoun sets, chance is $50\%^4$, or 6.25%. Disambiguation consistency, shown in Example (3), measures a system's ability to resolve a fixed pronoun to competing antecedents in paired templates, i.e., to disambiguate them. Chance is thus $50\%^2$, or 25%.

SPANBERT-LARGE is more robust to pronoun variation. As Table 5.3 shows, LINGMESS and the small and base sizes of FLAN-T5 score below chance, the former due to near-zero performance on *xe/xem/xyr*, and the latter due to poor performance overall. Interestingly, SPANBERT-LARGE is more consistent (60.0%) than FLAN-T5-XL (55.3%) and FLAN-T5-XXL (43.9%). This indicates that despite its lower overall performance in Section 5.5.1, SPANBERT-LARGE is more robust to pronominal variation.

The best model can only disambiguate half of the sentence pairs. Following from its high overall performance, FLAN-T5-XL has the highest disambiguation consistency score at 51.4%, just over half the template pairs we evaluate. In contrast,

Model	Pronoun Cons.	Disambiguation Cons.
LingMess	*4.2*	33.3
CAW-coref	18.3	34.7
SpanBERT-base	50.0	*24.3*
SpanBERT-large	**60.0**	41.2
FLAN-T5-small	*3.9*	*0.0*
FLAN-T5-base	*0.8*	*0.0*
FLAN-T5-large	14.4	*5.4*
FLAN-T5-xl	55.3	**51.4**
FLAN-T5-xxl	43.9	43.3

Table 5.3: Consistency results on WinoPron. Chance is 6.25% for pronoun consistency and 25% for disambiguation consistency. *Red, italicized numbers* are worse than chance.

SpanBERT-base has disambiguation consistency below chance (24.3%). Given its reasonable overall performance, this result could plausibly stem from model bias, i.e., over-resolving a pronoun to a particular antecedent, disregarding the disambiguating context. We thus investigate bias in more detail next.

5.6 Pronominal Bias

So far, we have focused on coreference resolution performance and consistency and found that accusative forms and less frequent pronoun sets are harder, and models are mostly non-robust to pronominal variation and antecedent disambiguation. However, we have not established the extent to which models over-resolve a pronoun to a particular antecedent due to pronominal biases, or because they simply cannot perform the task—due to the prompting method, insufficient model capacity, or even artifacts of the template creation, such as implicit causality (Brown and Fish, 1983; Sieker et al.,

2023; Kankowski et al., 2025). Thus, one of our aims in this section is to disentangle performance and bias.

Our second goal in this section is to focus on pronominal, rather than gender biases, such that we can evaluate coreference resolution with singular *they* and the neopronoun *xe*, as well as to avoid conflating grammatical and social gender (see Cao and Daumé III (2021) for a critical discussion). Addressing these goals, we propose a new method for evaluating pronominal bias in coreference resolution. We then apply our method to investigate bias in SPANBERT models on WINOPRON.

5.6.1 Evaluating Pronominal Bias

When proposing a new method to evaluate pronominal bias in coreference resolution systems, our primary goal is to disentangle performance and bias. In other words, we should have reason to believe that the templates can be disambiguated and the model can perform the task, and that the reason it gets an instance wrong with a particular pronoun is specifically due to pronominal bias. Additionally, we would like our method to work with an arbitrary set of pronouns of interest, and multiple surface forms of those pronouns.

Measuring performance. We first **(1)** isolate template pairs where the system attempts the task of coreference resolution as intended, i.e., the system resolves each pronoun to the occupation or participant (regardless of correctness). Next, we **(2)** focus on the template pairs that the model can *correctly* disambiguate with at least one pronoun set, p_a. We deem the model capable of performing coreference resolution on this set of template pairs if it can resolve them with at least one pronoun set.

Measuring bias. Of the template pairs that a model can successfully disambiguate with at least one pronoun p_a, we then **(3)** focus on cases where the model fails to disambiguate the exact same template pair with a different pronoun $p_b \neq p_a$, as this is likely due to pronominal bias. If the model over-resolves p_b to the occupation, we posit

that the model has a *positive bias* between p_b and that occupation. On the other hand, if it over-resolves p_b to the participant, the model is biased against associating p_b with the occupation, i.e., it has a *negative bias*.

Comparing results. With sets of positively and negatively biased occupations for each pronoun form, we want to quantify how many of a model's reasonable attempts to resolve a pronoun gave biased outputs. We thus compute the percentage of templates that result in bias (see Measuring Bias) of the total templates that a model attempts to resolve with that pronoun, given that it can correctly solve it with at least one pronoun (see Measuring Performance). This gives us a quantitative measure of "how biased" a model is which also controls for whether a model is attempting the task and can perform the task with another pronoun. In addition, we can quantify whether two models or two surface forms of a pronoun set show similar resolution biases by computing the Jaccard index (Jaccard, 1912), i.e., the size of the intersection of the biased occupation sets divided by the size of their union.

5.6.2 Results

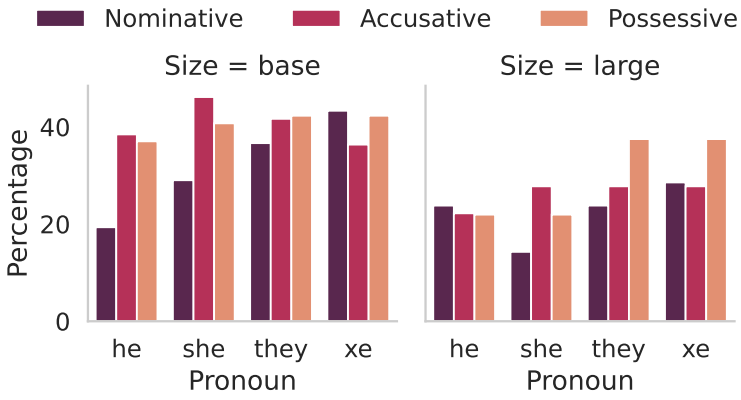


Figure 5.3: Percentage of model-attempted templates that show bias, for SPANBERT-BASE and SPANBERT-LARGE.

Nominative case		Accusative case		Possessive case	
Positive	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**
he		**him**		**his**	
engineer	*receptionist*	–	*dietitian*	*practitioner*	*hairdresser*
painter	*secretary*		*secretary*	*chef*	*secretary*
she		**her**		**her**	
hairdresser	*accountant*	*cashier*	*firefighter*	*practitioner*	*accountant*
painter	*plumber*		*mechanic*	*painter*	*surgeon*
they		**them**		**their**	
–	*accountant*	–	*cashier*	*advisor*	*accountant*
	plumber		*dietitian*	*baker*	*surgeon*
xe		**xem**		**xyr**	
–	*hairdresser*	–	*mechanic*	*advisor*	*engineer*
	engineer		*cashier*	*baker*	*supervisor*

Table 5.4: A sample of SPANBERT-LARGE's biases when resolving pronouns to occupations. Positive bias means that the model over-resolves the pronoun to that occupation. Negative bias means the model under-resolves the pronoun to the occupation.

We apply our method to SPANBERT-BASE and SPANBERT-LARGE and collect all instances of positive and negative bias between a pronoun form and an occupation. Aggregated bias results for both models are shown in Figure 5.3, and Table 5.4 shows a sample of biased occupations for SPANBERT-LARGE.

SPANBERT-BASE is more biased than SPANBERT-LARGE. As Figure 5.3 shows, a larger percentage of SPANBERT-BASE's attempted and resolvable templates show biased behaviour when compared to SPANBERT-LARGE. This pattern holds even when examining positive and negative biases separately. However, there are more negatively biased occupations than positively biased ones for both models.

Grammatical case	he	she	they	xe
Nominative	0.14	0.15	0.17	0.32
Accusative	0.12	0.10	0.25	0.29
Possessive	0.12	0.18	0.24	0.24

Table 5.5: Similarity of biased occupations between SPANBERT-BASE and SPANBERT-LARGE, quantified with the Jaccard index (0.0 -1.0; higher is more similar).

Bias is qualitatively different across model sizes. In addition to being quantitatively different, we find that despite being trained and fine-tuned on the same data, there is low overlap between the occupational biases acquired by SPANBERT-BASE and SPANBERT-LARGE (see Table 5.5). For instance, the former positively associates *she* with *machinist*, while the latter positively associates *she* with *hairdresser* and *painter*. Only *they/them/their* and *xe/xem/xyr* have slightly higher overlap, mostly due to negative bias, as these models under-resolve these particular pronouns to all occupations.

Grammatical case pairing	SPANBERT-BASE				SPANBERT-LARGE			
	he	she	they	xe	he	she	they	xe
Nominative-Accusative	0.10	0.00	0.00	0.00	0.10	0.00	0.07	0.06
Accusative-Possessive	0.07	0.13	0.14	0.07	0.22	0.00	0.06	0.06
Nominative-Possessive	0.07	0.11	0.10	0.09	0.17	0.29	0.15	0.19

Table 5.6: Similarity of biased occupations across pairings of grammatical case of a pronoun set, quantified with the Jaccard index (0.0 -1.0; higher is more similar).

Bias does not match qualitatively across grammatical case. In other words, positive bias with *she* for an occupation does not entail positive bias with *her*. We quantify this systematically by computing Jaccard indices in Table 5.6, where we find that most pairings of grammatical case have very low overlap in their biases. In fact, even contradictory associations are possible; SPANBERT-BASE has a positive bias between *manager* and *them*, but a negative bias betweeen *manager* and *their*.

Only nominative and possessive occupational biases in SPANBERT-LARGE appear to somewhat consistently overlap with each other. Although some of these instances (e.g., negative bias for *secretary* with *he*, *him*, and *his*) align with social stereotypes (Haines et al., 2016), the overall pattern provides evidence that grammatical case in pronouns has biases that should be examined in their own right.

Bias is not additive. Even though SPANBERT-LARGE has positive bias for *baker* and *her*, *their* as well as *xyr*, this does not imply that the model must have a negative bias between *baker* and *his*; it does not. This further highlights the need for evaluation that goes beyond binary, oppositional operationalizations of gender via pronouns.

5.7 Discussion

By systematically identifying and fixing issues with Winogender Schemas (Rudinger et al., 2018), we create a new dataset, WINOPRON, and introduce a novel method to evaluate pronominal biases in coreference resolution. We find that: **(1)** Different grammatical cases of pronouns show vastly different performance and bias characteristics, **(2)** pronominal biases are rich and varied, of which *he* and *she* are only the tip of the iceberg, and **(3)** model biases are complex and do not necessarily match our intuitions about them. Based on our findings, we make some recommendations for researchers who study coreference resolution and those who study bias and fairness via pronouns.

First, grammatical case is a dimension of pronominal performance and bias that warrants more study (Munro and Morrison, 2020). In particular, we hope that future work further investigates *why* accusative pronouns are harder. The patterns we demonstrate (both for performance and bias) could arise from a number of sources beyond mere frequency, including quirks of our dataset, or the distribution of semantic roles in training data for coreference resolution systems.

Second, we echo prior calls for fairness researchers to attend to the differences between social gender and terms that index it (Cao and Daumé III, 2021), to include more diversity in pronouns (Baumler and Rudinger, 2022; Lauscher et al., 2022; Hossain

et al., 2023), and to move towards richer operationalizations of gender (Devinney et al., 2022; Ovalle et al., 2023a) and bias (Blodgett et al., 2020), particularly since we find evidence that pronouns of a particular grammatical gender are inconsistent proxies for social gender across different grammatical cases. Specifically, future work on bias in coreference resolution should treat pronominal bias as distinct from (social) gender bias, defend how and why pronouns are mapped to social gender, and move beyond binary, oppositional methods of evaluation, as we do.

Lastly, as our work is a case study in how careful data curation and operationalization affects claims about system performance and bias, we emphasize the need for thoughtful data work (Sambasivan et al., 2021), and encourage the use of automatic checks when feasible, as in our work.

5.8 Related Work

Besides Rudinger et al. (2018), there are a number of papers that tackle gender bias in coreference resolution, all of which differ from ours. Similar to Winogender Schemas, WinoBias (Zhao et al., 2018) proposes Winograd-like schemas that focus on occupations to evaluate gender bias in coreference resolution. However, WinoBias only covers *he* and *she*, rather than our coverage of all English pronoun sets by design. In addition, like Winogender, WinoBias also treats pronouns in all grammatical cases the same way. WinoNB schemas (Baumler and Rudinger, 2022) evaluate how coreference resolution systems handle singular *they* and plural *they* with similar schemas. Beyond these constructed schemas, there also exist datasets of challenging sentences found "in the wild," such as BUG (Levy et al., 2021), GAP (Webster et al., 2018), and GICOREF (Cao and Daumé III, 2021). However, as these natural datasets are not carefully constructed like Winograd-like schemas, pronouns cannot be swapped in dataset instances and still be assumed to be grammatical or coherent.

Our work is also one among several papers that investigate datasets for problems including low quality or noisy data (Elazar et al., 2024; Abela et al., 2024), artifacts (Shwartz et al., 2020; Herlihy and Rudinger, 2021; Elazar et al., 2021; Dutta

Chowdhury et al., 2022), contamination (Balloccu et al., 2024; Deng et al., 2024), and issues with conceptualization and operationalization of bias (Blodgett et al., 2021; Selvam et al., 2023; Nighojkar et al., 2023; Subramonian et al., 2023). We cover many of these areas, but do not control for dataset artifacts, which we explain in our Limitations section.

5.9 Limitations

As in Winogender Schemas, our schemas are not "Google-proof" and could conceivably be solved with heuristics, including word co-occurrences, which is a primary concern when creating and evaluating *Winograd* schemas (Levesque et al., 2012; Amsili and Seminck, 2017; Elazar et al., 2021). The fact that we do not control for this means that our dataset gives *generous* estimates of system performance, particularly for strong language models like FLAN-T5, but it also means that this dataset is inappropriate to test reasoning, a challenge we take on in Chapter 7. Our dataset construction instead controls for simple system heuristics that are relevant for coreference resolution, such as always picking the first entity in the sentence, or always picking the second.

We take steps to prevent data contamination (Jacovi et al., 2023), including not releasing our data in plain text, and not evaluating with language models behind closed APIs that do not guarantee that our data will not be used to train future models (Balloccu et al., 2024). However, as we cannot guarantee a complete absence of data leakage unless we never release the dataset, we encourage caution in interpreting results on WINOPRON with models trained on data after August 2024.

Finally, we note that as our evaluation set only contains one set of templates per occupation-participant pair, our results represent a point in the distribution of bias related to that occupation. When results are aggregated, our dataset gives us a big-picture overview of performance and bias in coreference resolution. We thus echo Rudinger et al.'s (2018) view of Winogender Schemas as having "high positive predictive value and low negative predictive value" for bias. In other words, they may demonstrate evidence of pronominal bias in systems, but not prove its absence. In the case of large language

models in particular, using a small number of templates for templatic evaluation is known to be brittle even to small, meaning-preserving changes to the template (Seshadri et al., 2022; Selvam et al., 2023). Our dataset's small size is a result of us requiring a tightly controlled and structured dataset to evaluate how coreference resolution varies. Thus, it may differ from realistic examples (which would have other differences that confound bias results). We wish to emphasize that in addition to controlled datasets like ours, realistic evaluation is also necessary for holistically evaluating performance, robustness and bias in coreference resolution.

5.10 Conclusion

This chapter addressed research questions 1 and 2 in this thesis, by questioning the connection between pronouns and gender, and revisiting measurements of bias in coreference resolution. In particular, we demonstrated a number of issues with the well-known Winogender Schemas dataset, which we fixed in our new, expanded WINOPRON dataset. In addition, we propose a novel way to evaluate pronominal bias in coreference resolution that goes beyond the binary and focuses on grammatical gender. With our new dataset, we evaluate both supervised coreference resolution systems and language models, and find that the grammatical case of pronouns affects model performance and bias, and that bias varies widely across models, pronoun sets and even grammatical cases of the same pronoun set, which have been treated as equivalent in prior work. This chapter demonstrates that measurements of bias and robustness are only as good as the datasets and metrics we use to measure them, and we call for careful attention when developing future resources for evaluating bias, with attention to grammatical case, more careful operationalizations of bias, and greater diversity in the pronouns we consider. In the next two chapters, we investigate whether stereotypical biases such as the ones we found in this chapter can be overcome with additional information, shifting our focus from reference *resolution* to actually *doing* reference.

6

Pronoun Fidelity of English LLMs

As we have seen in the last chapter, social gender stereotypes about occupations and their corresponding reflections in language cause biases when resolving reference for both humans and NLP systems, although they manifest differently. In this chapter, we ask whether large language models can *overcome* their stereotypical biases and faithfully reuse pronouns for someone, as humans do. To measure this, we introduce the task of pronoun fidelity: Given a context introducing an entity with a co-referring definite description and pronoun, the task is to reuse the correct pronoun later. With a new dataset to evaluate pronoun fidelity in English, we evaluate 37 model variants from nine popular families, across architectures (encoder-only, decoder-only and encoder-decoder) and scales (11M-70B parameters). We find that when an individual of a certain occupation is introduced with a pronoun, models can mostly overcome their stereotypical biases and faithfully reuse this pronoun in the next sentence. However, they show significantly worse performance with *she/her/her*, singular *they* and neopronouns, compared to *he/him/his*, showing that doing fair reference remains an open problem.

The content in this chapter is based on:
Vagrant Gautam, Eileen Bingert, Dawei Zhu, Anne Lauscher, and Dietrich Klakow (Dec. 2024a). "Robust Pronoun Fidelity with English LLMs: Are they Reasoning, Repeating, or Just Biased?" In: *Transactions of the Association for Computational Linguistics* 12, pp. 1755–1779. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00719. URL: https://doi.org/10.1162/tacl_a_00719

Vagrant Gautam conceptualized the research, created the dataset together with Eileen Bingert, analyzed the results, and led the paper writing. Dataset validation was performed by Vagrant Gautam and Eileen Bingert. Vagrant Gautam and Dawei Zhu performed experiments, and Anne Lauscher and Dietrich Klakow advised and provided feedback.

6.1 Introduction

Third-person pronouns (*he*, *she*, *they*, etc.) are words that construct individuals' identities in conversations (Silverstein, 1985). In English, these pronouns mark referential gender for the entity they refer to, which can index social gender and potentially clash with stereotypical gendered associations of the referent. As we saw in the previous chapter, this can create problems for NLP systems in resolving reference. Correctly *using* the pronouns an individual identifies with is also important to avoid misgendering (including through incorrect pronoun use), which can in the best case be a social faux pas (Stryker, 2017) and in the worst case, cause psychological distress and exacerbate suicidal thoughts, particularly among transgender individuals (McLemore, 2018).

Accordingly, it is important for large language models (LLMs) to use pronouns faithfully and without causing harm. To this end, many studies have explored how LLMs handle pronouns, showing that they stereotypically associate pronouns and occupations (Kurita et al., 2019), reason about co-referring pronouns and entities better when they conform to stereotypes (Tal et al., 2022), fail when exposed to novel pronoun phenomena such as neopronouns (Lauscher et al., 2023), and cannot consistently reuse neopronouns during generation (Ovalle et al., 2023a). These shortcomings create differences in quality of service and cause representational harm, amplifying discrimination against certain pronoun users (Blodgett et al., 2020; Dev et al., 2021).

In such work on LLM pronoun use, a question that has gone unexamined thus far is: *Can models faithfully reuse pronouns*? To answer this question, we propose *pronoun fidelity* (Section 6.2), a new task to investigate realistic model reasoning about pronouns, and we introduce RUFF (Section 6.3), a dataset to evaluate this task. With this dataset,

we present an analysis of pronoun fidelity across 37 variants from nine popular language model families covering architectures and scales, to investigate whether models can overcome their stereotypical biases in the appropriate context. To do so, we first establish a "bias baseline" by collecting model pronoun predictions for occupations in the absence of any context (Section 6.5). Then, we evaluate whether models can overcome their biased predictions when explicitly shown what pronoun to use in context (Section 6.6). All models turn out to be good at this task, performing well above chance but below human performance, which is perfect in this simple setting. However, there are significant disparities across pronoun sets and architectures; models are statistically significantly better at reusing *he/him/his* pronouns compared to *she/her/her*, which in turn shows better reuse than singular *they* and the neopronoun *xe/xem/xyr*. Additionally, encoder-only models show much better performance than decoder-only models of the same scale, despite the latter being the preferred architecture of today's LLMs, as shown in Figure 6.1. Overall, our results show that even in this simple and realistic setting to evaluate downstream impacts of stereotypical biases, LLMs show significant quality of service differentials with how they do reference.

6.2 Pronoun Fidelity Task

Using multiple referring expressions (definite descriptions, names, and pronouns) to discuss a single individual is a well-studied phenomenon in discourse (Grosz et al., 1995). Building on this, we formalize our task of pronoun fidelity: Given a context in which an entity is introduced with a co-referring definite description and pronoun, the task is to reconstruct the pronoun later in a sentence about the entity.

Introduction: *The accountant had just eaten a big meal so her stomach was full.*

Task sentence: *The accountant was asked about ___ charges for preparing tax returns.*

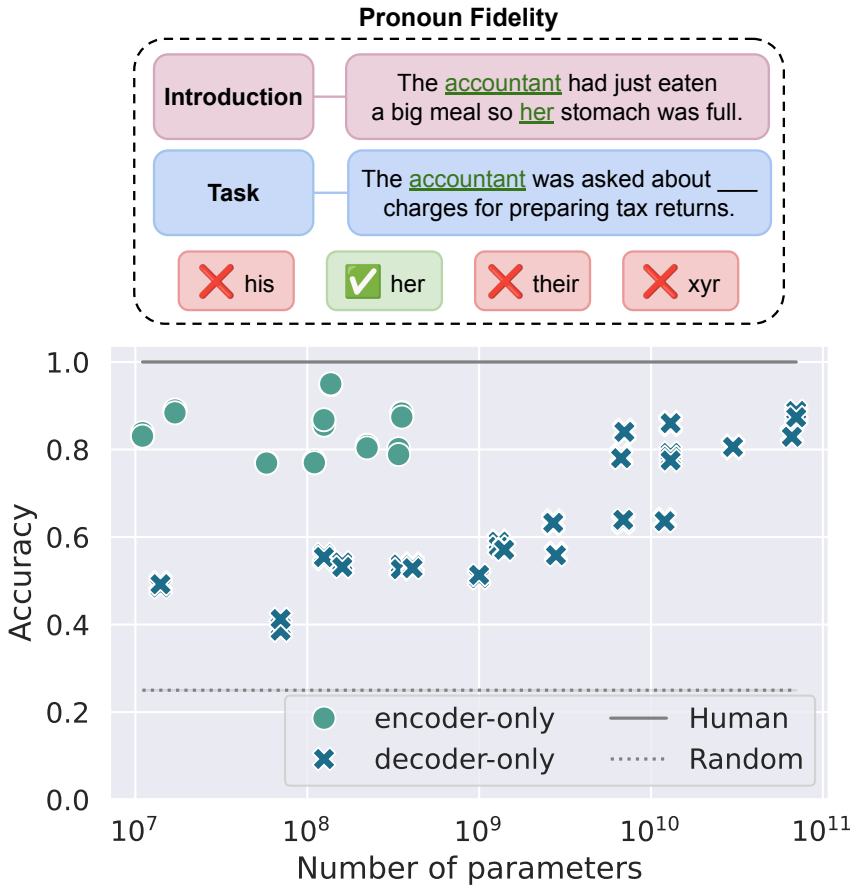


Figure 6.1: We evaluate model accuracy at using the correct pronoun for an entity when provided with an explicit introduction. All models perform above chance, but below human performance. Plotting scaling behaviour split by architecture shows that encoder-only models are better than decoder-only models of the same scale, and comparable to decoder-only models orders of magnitude larger.

More formally, an introduction sentence $i(e_a, p_a)$ establishes a coreference between an entity e_a and a pronoun p_a, creating a context. Then, a task sentence $t(e_a, p)$ contains an unambiguous coreference between the entity e_a from the introduction and a pronoun slot p which must be filled. The task is to maximize the probability P

$$P[\, t(e_a, p = p_a) \mid i(e_a, p_a) \,], \tag{6.1}$$

of reconstructing the correct pronoun p_a in the sentence $t(e_a, p)$, given the context.

6.3 RUFF Dataset

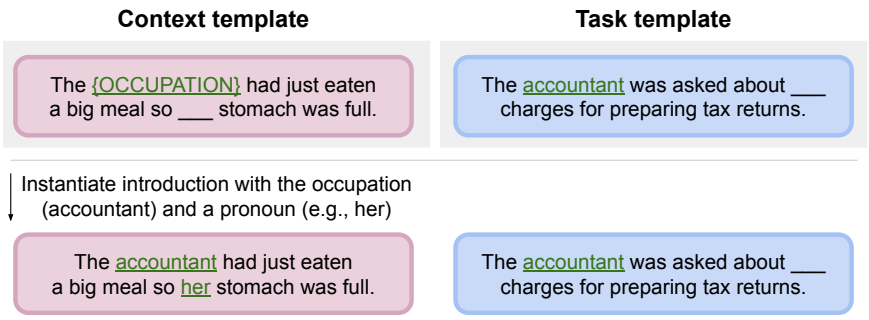


Figure 6.2: Template assembly for RUFF: Occupation-specific task templates are matched with generic context templates that are instantiated with a pronoun set to introduce the person and pronouns. This creates realistic but controlled narratives that allow us to measure robust pronoun fidelity.

To evaluate pronoun fidelity at scale, we create RUFF, an evaluation dataset of narratives about people with occupations that have been studied extensively in the context of gendered stereotypes in NLP (Rudinger et al., 2018). Our narratives cover the 60 occupations studied in Winogender schemas; see Appendix A.1 for a full list. As for pronouns, we consider four third-person pronouns in three grammatical cases (nominative, accusative and possessive dependent). In addition to the English masculine (*he/him/his*) and feminine (*she/her/her*) pronouns, we heed Lauscher et al.'s (2022) call for more inclusive NLP research by examining two more pronoun sets that are less well-studied in NLP: Singular *they* (*they/them/their*), the pronoun of choice of over 75% of respondents to the Gender Census (Lodge, 2023), and *xe/xem/xyr*, the most popular neopronoun according to the same census. In total, RUFF contains 7,200 data instances, each consisting of an introduction sentence and a task sentence. The instances are constructed with a 3-step pipeline: Template creation (Section 6.3.1), template assembly (Section 6.3.2), and data validation (Section 6.3.3).

6.3.1 Template Creation

Below, we describe how we create occupation-specific task templates and generic context templates for introductions.

Task templates. We reuse the 180 single-entity sentences about occupations from Chapter 5 as task sentence templates, as they show unique, unambiguous coreference between the occupation and the pronoun.

Introduction templates. The ideal introduction template would be: **(1) Flexible** across different occupations for a controlled setting to test pronoun fidelity; **(2) cohesive** in a multi-sentence narrative including the task template about an occupation; and **(3) neutral**, not dramatically affecting the prediction of a certain pronoun. Templates such as *He is an accountant* are well-established for testing word embedding associations (Caliskan et al., 2017; May et al., 2019). They are flexible and neutral (they are even referred to as "semantically bleached" templates in the literature), but it is unnatural to introduce an entity for the first time by beginning with a pronoun (Grosz et al., 1995). Natural corpora like Levy et al. (2021) have the most potential for creating cohesive narratives, but contain occupation-specific sentences that are inflexible and sometimes also non-neutral, e.g., ungrammatical with singular *they*.

For a setting that satisfies all three criteria, we create context templates with generic themes, e.g., universal human emotions and sensations (*hungry*/*full*, *tired*/*energetic*, *unhappy*/*happy*, etc.). The generic themes make them flexible for use across all occupations. Our templates are created to be grammatical with all pronoun sets we consider, which satisfies neutrality. Additionally, our use of both positive and negative versions of templates (i.e., *happy* and *unhapppy*) as well as our variety of templates allows us to mitigate potential implicit biases when aggregated (Alnegheimish et al., 2022).

6.3.2 Template Assembly

Figure 6.2 shows how we instantiate and combine templates to assemble our data instances: First, we select an occupation (e_a) and one of its task templates. We pick a pronoun (p_a) to use as ground truth and instantiate a random context template with the selected occupation and pronoun, creating an introduction sentence which will then be paired with the appropriate task sentence for that occupation and grammatical case. Instantiating 10 templates with 4 different pronoun sets and pairing them with task templates for 60 occupations across 3 grammatical cases gives us a total of 7,200 unique instances for this task.

Data type	**Number of instances**
Task sentences with no context (used in Section 6.5)	180
With introductory context (used in Section 6.6)	3 x 2,160 (of 7,200)

Table 6.1: Number of dataset instances with and without an introductory context. We subsample 3 sets of 2,160 sentences of the total number of instances we created.

Our stackable dataset design gives us a controlled setting to evaluate the effect of context on model predictions of reference. We subsample the data with three random seeds for the rest of our evaluation, ensuring that all occupations, cases, pronoun declensions and distractor pronouns are equally represented in each subsampled set of 2,160 sentences (see Table 6.1).

6.3.3 Data Validation

We validate all task and context templates, as well as a sample of pronoun fidelity instances. Annotator information is shown in Appendix A.2 and all annotator instructions are provided in Appendix A.3.

Templates. Two authors with linguistic training iteratively created and validated sentence templates for grammaticality and correct coreferences until consensus was reached. An additional annotator independently rated 100% of the sentences as grammatical and with the correct coreferences.

Pronoun fidelity task. To verify that the pronoun fidelity task is easy and unambiguous for humans, and to create a ceiling for model evaluation, we also validate a subset of 100 pronoun fidelity instances. One author and one annotator had to fill in the pronoun, and each performed with 100% accuracy.

6.4 Experimental Setup

Model	Sizes	Architecture
Evaluated with (Pseudo) Log Likelihoods		
ALBERT-V2	base (11M), large (17M), xlarge (58M), xxlarge (223M)	Encoder-only
BERT	base (110M), large (340M)	Encoder-only
ROBERTA	base (125M), large (355M)	Encoder-only
MOSAICBERT	137M	Encoder-only
OPT	125M, 350M, 1.3B, 2.7B, 6.7B, 13B, 30B, 66B	Decoder-only
PYTHIA	14M, 70M, 160M, 410M, 1B, 1.4B, 2.8B, 6.9B, 12B	Decoder-only
LLAMA-2	7B, 13B, 70B	Decoder-only
Evaluated with prompting		
FLAN-T5	small (77M), base (248M), large (783M), xl (2.85B), xxl (11.3B)	Encoder-decoder
LLAMA-2-CHAT	7B, 13B, 70B	Decoder-only

Table 6.2: Models we experiment with across a range of sizes (11M-70B parameters) and architectures.

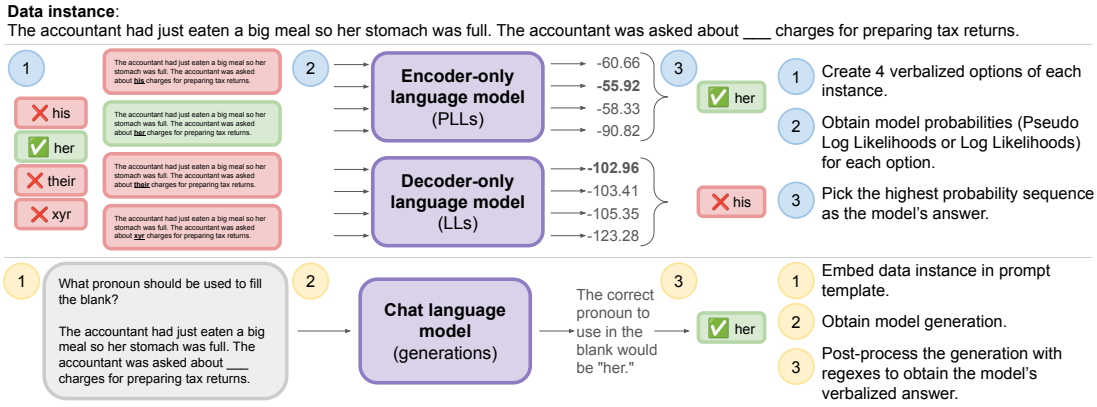


Figure 6.3: Model evaluation overview: Pseudo log likelihoods (PLLs) and log likelihoods (LLs) of verbalized instances are used for encoder-only and decoder-only models; generations are used for chat models.

Here we list our models, evaluation methods, and metrics. Further details are provided in Appendix B.1.

6.4.1 Models

We experiment with 37 transformer-based language model variants from nine popular model families (see Table 6.2), which we chose to evaluate the effects of architecture and scaling. Our encoder-only models are from the BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ALBERT-v2 (Lan et al., 2020) and MosaicBERT (Portes et al., 2023) model families, as the first three remain well-used in NLP, and the last is trained on much more data. As for our decoder-only models, we select the popular Llama-2 (Touvron et al., 2023) model family, as well as OPT (Zhang et al., 2022) and Pythia (Biderman et al., 2023) for their large range of model sizes. In addition, we experiment with popular chat models that are further trained with instruction-tuning and reinforcement learning, to evaluate task performance with prompting; specifically, we use decoder-only Llama-2-chat models (Touvron et al., 2023) and encoder-decoder FLAN-T5 models (Chung et al., 2024).

6.4.2 Obtaining Predictions

Figure 6.3 shows an overview of our evaluation methods. Decoder-only and encoder-only models are evaluated comparably in a forced choice setting: Following Hu and Levy (2023), we take direct measurements of probabilities as a proxy for models' metalinguistic judgements. Generations are obtained from chat models and post-processed to obtain unique pronouns, if any.

Encoder-only and decoder-only models. We verbalize four versions of each data instance, i.e., we fill in the blank with each of the four pronouns we consider, creating four options. We then obtain model probabilities for each of these four options, and select the highest probability option as the model's choice. Specifically, we use log likelihoods for decoder-only models and pseudo log likelihoods for encoder-only models, following prior work (Salazar et al., 2020; Kauf and Ivanova, 2023). We do not use masked token prediction due to tokenization issues with neopronouns (Ovalle et al., 2024); briefly, we want *xe* to be tokenized "normally" (which is often as two tokens) rather than a single UNK token.

Chat models. Following common practice, we evaluate chat models (FLAN-T5 and LLAMA-2-CHAT) using vanilla and chain-of-thought prompting. As Sclar et al. (2024) recommends, we show the range of expected performance with 10 different prompts, inspired by the prompts to elicit coreferences in the FLAN collection (Longpre et al., 2023). See Appendix B for more methodological details.

6.4.3 Metrics

As every instance of the pronoun fidelity task has a unique correct answer, we report accuracy averaged over the three randomly sampled subsets of our dataset. We show the standard deviation with error bars or shading. Where possible, we perform significance

testing with a Welch's t-test and a threshold of 0.05. We use human performance (which is 100%) as our ceiling, and compare models to a baseline of randomly selecting 1 of the 4 pronouns (i.e., 25%).

6.5 Model Predictions with No Context

We begin by creating a "bias baseline," i.e., obtaining pronoun predictions from models on our task sentences in the absence of any context. In Section 6.6, we will examine whether models can overcome this bias with reasoning when provided with context establishing a single correct answer.

Example: *The accountant was asked about ___ charges for preparing tax returns.*
No single answer (among *his*, *her*, *their*, *xyr*)

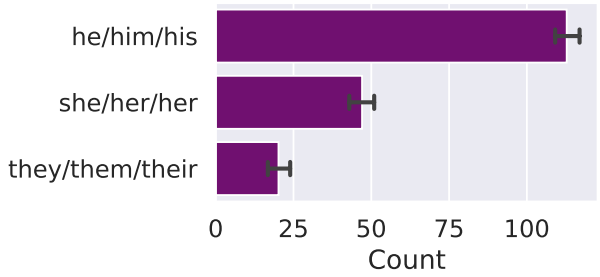


Figure 6.4: Counts of pronoun predictions from all models, in the absence of context. The pronouns *xe/xem/xyr* never appear as the highest-likelihood option for any model. Error bars indicate standard deviation across models.

As we cannot evaluate accuracy on a task with no single correct answer, we show the counts of model predictions of different pronoun declensions in Figure 6.4, averaged over all models. The results show an overall bias towards predicting *he/him/his* pronouns in the absence of any context, followed by *she/her/her*. Nevertheless, this is only a measurement of intrinsic bias, or stereotypical associations between pronouns and occupations, and models may still be able to overcome these biases in context.

Figure 6.5: Counts of pronoun predictions from all models, in the absence of context. The pronouns *xe/xem/xyr* are not plotted as no model assigns it the highest likelihood. The random baseline shows counts if each pronoun set was chosen equally often.

To further analyze how pronoun predictions differ across individual models, we plot per-model counts in Figure 6.5. Even though our task sentences are designed such that any pronoun set can be used grammatically, all models predict *he/him/his* more frequently than *she/her/her*, which is in turn predicted more frequently than *they/them/their*. No model ever assigns *xe/xem/xyr* pronouns the highest likelihood.

Obtaining pronoun predictions without context as we do in this section is a popular method to measure model bias, with numerous papers (Caliskan et al., 2017; May et al., 2019; Kurita et al., 2019, *inter alia*) showing that associations between occupations and pronouns are based on social gender stereotypes, e.g., *doctor-he* and *nurse-she*. More recent work has shown that model pronoun predictions might be a statistical accident of the chosen templates (Seshadri et al., 2022; Selvam et al., 2023), and that intrinsic biases may not correlate with downstream harms (Goldfarb-Tarrant et al., 2021). In order to test for such extrinsic behaviours, the rest of this chapter examines whether models can override their intrinsic statistical biases on these same templates when provided the right pronoun to use.

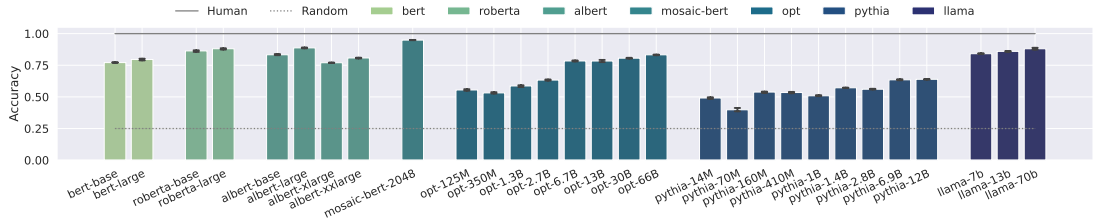


Figure 6.6: Pronoun fidelity by model with an introductory context. Accuracy is averaged across occupations, pronouns and grammatical cases, and is above chance (0.25) but below human performance (1.0).

6.6 Injecting an Introductory Context

When models are provided with an introductory sentence explicitly establishing the pronoun to use for an entity, can they use that pronoun to refer to the same entity in the immediate next sentence?

> **Example:** *The accountant had just eaten a big meal so her stomach was full. The accountant was asked about ___ charges for preparing tax returns.*
> **Correct answer:** her

As Figure 6.6 shows, **all models perform better than chance at pronoun fidelity with a simple introduction** (up to 0.95 with MOSAICBERT), but not as well as humans, who achieve perfect performance. We also see improvements with increasing model scale, with the exception of ALBERT-V2, as in Tay et al. (2023).

Which pronouns are harder? Even in the simplest case of the pronoun fidelity task, patterns emerge when split by pronoun, as shown in Figure 6.7. Overall model **accuracy on *he/him/his* is significantly higher than *she/her/her*, which in turn is significantly higher than both *they/them/their* and *xe/xem/xyr***, in line with previous findings that language technology has gaps when it comes to neopronouns (Lauscher et al., 2023). Models show intriguing patterns with these last two pronoun sets. Most encoder-only models appear to handle the neopronoun better than singular *they* (e.g., BERT-LARGE

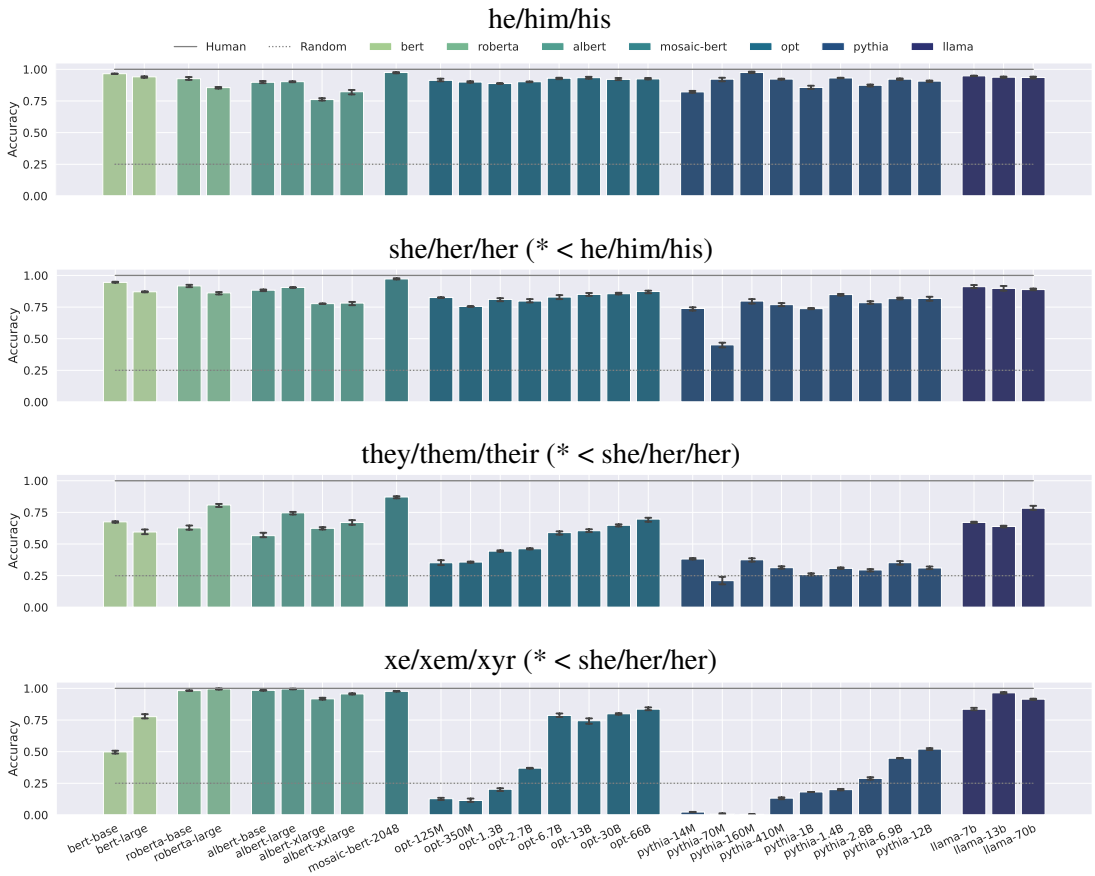


Figure 6.7: Pronoun fidelity by model with an introductory context, split by pronoun series. Model accuracy is compared to chance (0.25) and human performance (1.0) and * denotes statistical significance.

has an accuracy of 0.78 on *xe/xem/xyr* compared to 0.60 on *they/them/their*), which warrants further investigation. Decoder-only models smaller than 6.7B parameters struggle with the neopronoun, with every OPT and PYTHIA model smaller than 2.7B parameters performing below chance, and in some cases (e.g., PYTHIA-14M, PYTHIA-70M and PYTHIA-160M) even performing close to 0.0. Beyond this scale, however, models perform better on *xe/xem/xyr* than on singular *they*, with LLAMA-13B achieving 0.96 accuracy on the neopronoun. These differences are statistically significant. As the training data for individual model families is the same, this might suggest that decoder-only models generalize to novel pronouns starting at the scale of 6.7B parameters. In either case, our observations could also explain the poor performance that some previous studies of neopronouns find, as the largest model that Hossain et al. (2023) experiment

with, for instance, is OPT-6.7B. The lower performance of bigger models with singular *they* could also be a reflection of human processing difficulties with definite, specific singular *they*, as has been observed in linguistics (Conrod, 2019).

Which architectures are better? As Figure 6.1 shows, encoder-only models are much better than decoder-only models of the same scale, and their performance is comparable to or better than decoder-only models that are orders of magnitude larger; RoBERTa-base (125M) is 0.86 accurate compared to OPT-125M's 0.55, and exceeds OPT-66B's 0.83 despite being more than 500 times smaller. Possible reasons for this include that bidirectional attention in encoder-only models results in a stronger encoding of co-referring expressions. These results are striking in light of the fact that decoder-only models are the most popular architecture of today's best large language models, despite showing poorer performance on something as fundamental as reusing a previously-specified pronoun in this simple setting.

6.7 Discussion

Our results highlight the need for more extrinsic evaluations and raise questions about reasoning and the effects of architecture, which we elaborate on below:

The importance of extrinsic evaluations. Although work on fairness has historically focused on intrinsic evaluations of stereotypes, our experiments add to the growing body of work showing that extrinsic evaluations contradict results from intrinsic evaluations, which are often presumed to be directly connected. In fact, as Nissim et al. (2020) argues, intrinsic evaluations even sometimes presume their conclusions due to certain choices during evaluation and implementation. Our work shows that even when language models have intrinsic pronominal associations for certain occupations, this does not preclude them from reasoning about a different set of pronouns when necessary. In the case of *xe/xem/xyr* pronouns, in particular, the evidence is particularly

compelling: No model has intrinsic stereotypical associations with this neopronoun set, yet most models are able to successfully reuse it more than random chance.

Are we really seeing pronominal reasoning? Although we see good performance on pronoun fidelity with encoder-only models and large decoder-only models, the question remains whether this good performance is due to really "reasoning" about co-referring expressions the way humans do, i.e., learning a mapping between a person and a pronoun set, and then applying it in future contexts. A language model could conceivably also simply be "repeating" the pronoun without reasoning at all, with important implications for reference if they cannot truly reason to overcome stereotypical biases. This question is explored in more detail in the following chapter.

***How* exactly does architecture influence performance?** Referring to ALBERT-V2, BERT, ROBERTA and MOSAICBERT as encoder-only models, and OPT, PYTHIA and LLAMA-2 as decoder-only models conflates several specific differences between them. In general, the models we refer to as encoder-only have bidirectional attention, i.e., can attend to left and right tokens, and are often trained with masked language modelling. In contrast, the models we refer to as decoder-only have unidirectional attention, i.e., can attend only to left tokens, and are typically trained for the task of next token prediction. In addition, they are evaluated differently in our setup, with pseudo log likelihoods for the former and log likelihoods for the latter. Further investigation is needed to disentangle these different factors, each of which could impact our results.

6.8 Related Work

Misgendering. Hossain et al. (2023) and Ovalle et al. (2023a) both study misgendering when models are prompted with a pronoun series to use for an individual. Hossain et al. (2023) evaluates misgendering with stereotypically gendered names and explicit pronoun declarations (e.g., by using *Alex's pronouns are they/them/theirs.* as an introductory sentence), while we study occupational biases via definite descriptions that

are introduced in a naturalistic way. Ovalle et al. (2023a) evaluates misgendering in a variety of context but only focuses on open-ended generation with decoder-only models, in contrast to our probability-based evaluations with a range of architectures. Our evaluations are thus complementary, as Subramonian et al. (2025) has shown that probability- and generation-based evaluations measure different capabilities and disagree with each other on 20% of instances.

Faithful pronoun use. In contrast to our examination of within-language pronoun use, faithful pronoun use in context has been studied extensively in machine translation (Müller et al., 2018; Voita et al., 2018; Fernandes et al., 2023), where there is also a ground truth. Similar to our work, Sharma et al. (2022) injects context with an explicit coreference to encourage faithful pronoun translation.

Pronouns and occupational bias. Stereotypical associations between pronouns and occupations have been studied in masked token prediction (Kurita et al., 2019; Vassimon Manela et al., 2021; Tal et al., 2022) and embeddings (Bolukbasi et al., 2016; Zhao et al., 2019), but these studies typically use brittle methodology (Gonen and Goldberg, 2019; Seshadri et al., 2022) and measure intrinsic bias, which may not translate to extrinsic bias or harms (Goldfarb-Tarrant et al., 2021). Unlike these works, we evaluate extrinsic bias through our focus on natural pronoun use in context.

6.9 Limitations

Our task as it is defined in Section 6.2 is much broader than the scope of our dataset. We focus on occupations due to the wide attention they have received in prior literature, but we continue a long tradition of ignoring biases relating to the participants, e.g., *child*, *taxpayer*, etc. In addition, pronoun fidelity is only one dimension of inclusive language model behaviour and faithful reference, and indeed only one way in which misgendering occurs in language, even in morphologically poor languages like English.

We take steps to prevent data contamination following Jacovi et al. (2023), including not releasing our data in plain text, and not evaluating with models behind closed APIs that do not guarantee that our data will not be used to train future models. However, as we cannot guarantee a complete absence of data leakage unless we never release the dataset, we encourage caution in interpreting results on RUFF with models trained on data after March 2024.

6.10 Conclusion

In this chapter, we introduced the task of pronoun fidelity to evaluate whether language models can overcome their stereotypical occupational biases and do pronominal reference correctly. With RUFF, a dataset we designed to measure pronoun fidelity, we evaluated several models of different architectures. Our results show that overall, models are indeed able to overcome stereotypical biases to reuse a pronoun that was shown to them earlier, but encoder-only and decoder-only models differ dramatically in their performance at the same scale. Models show significant performance disparities with neopronouns, singular *they* and *she/her/her*, compared to *he/him/his*, in a setting that is simple for humans. Overall, our results highlight the importance of extrinsic evaluations of whether language models perpetuate or amplify discrimination against users of certain pronouns. One question about our results that remains unanswered is whether we are seeing true pronominal reference, or simply shallow repetition of referring expressions. We turn to this question next.

7

Pronoun Fidelity with Multiple Referents

When models do faithful reference, are they really reasoning, or are they simply shallowly repeating previous tokens? To disentangle reasoning from repetition in the context of faithful pronoun use, we introduce *robust pronoun fidelity*, the task of reusing a pronoun for an individual in a narrative about two people. Concretely, given a context introducing an entity with a co-referring definite description and pronoun, the task is to reuse the correct pronoun for that entity later, independent of intervening sentences that discuss a second entity with a different pronoun. To measure robust pronoun fidelity in English, we augment the RUFF dataset from the last chapter with carefully-designed non-adversarial "distractors," resulting in over 5 million instances. We use it to evaluate the same 37 model variants from nine popular families, across architectures (encoder-only, decoder-only and encoder-decoder) and scales (11M-70B parameters), and find that the pronoun fidelity of these models is not robust, despite this being a simple, naturalistic setting where humans achieve 99.8% accuracy. Models are easily distracted by sentences discussing other people, and even one sentence with a distractor pronoun causes accuracy to drop on average by 34 percentage points. Surprisingly, encoder-only models are better at this task than decoder-only models that are orders of magnitude larger. We encourage researchers to bridge the gaps we find and to carefully evaluate reasoning in settings beyond reference where superficial repetition might inflate perceptions of model performance.

The content in this chapter is based on:

Vagrant Gautam, Eileen Bingert, Dawei Zhu, Anne Lauscher, and Dietrich Klakow (Dec. 2024a). "Robust Pronoun Fidelity with English LLMs: Are they Reasoning, Repeating, or Just Biased?" In: *Transactions of the Association for Computational Linguistics* 12, pp. 1755–1779. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00719. URL: https://doi.org/10.1162/tacl_a_00719

Vagrant Gautam conceptualized the research, created the dataset together with Eileen Bingert, analyzed the results, and led the paper writing. Dataset validation was performed by Vagrant Gautam and Eileen Bingert. Vagrant Gautam and Dawei Zhu performed experiments, and Anne Lauscher and Dietrich Klakow advised and provided feedback.

7.1 Introduction

As social beings, we humans interact with a large number of individuals whom we subsequently need to refer to in language, for which there is a variety of referring expressions that we can choose from and must disambiguate in context. The process of both producing and perceiving these referring expressions requires sophisticated reasoning, particularly since (in different contexts) the same surface-level expression can refer to a different individual, and we can refer to the same individual with multiple referring expressions. This complexity makes for a fascinating testbed for LLMs.

In the last chapter, we saw that LLMs show pronoun fidelity in a simple setting, but were unable to say whether this was due to real "reasoning" as in humans, or shallow repetition of previous tokens. To investigate this, we design a more complex setting with multi-person reference, in order to measure *robust pronoun fidelity* (Section 7.2). Going beyond the previous chapter's focus on a single individual at a time, we augment RUFF (Section 7.3) to create a large-scale dataset of over 5 million instances of narratives with two referents, using two different sets of pronouns. With this dataset, we present an analysis of robust pronoun fidelity across the same 37 language model variants across architectures and scales, to investigate whether models are reasoning or repeating.

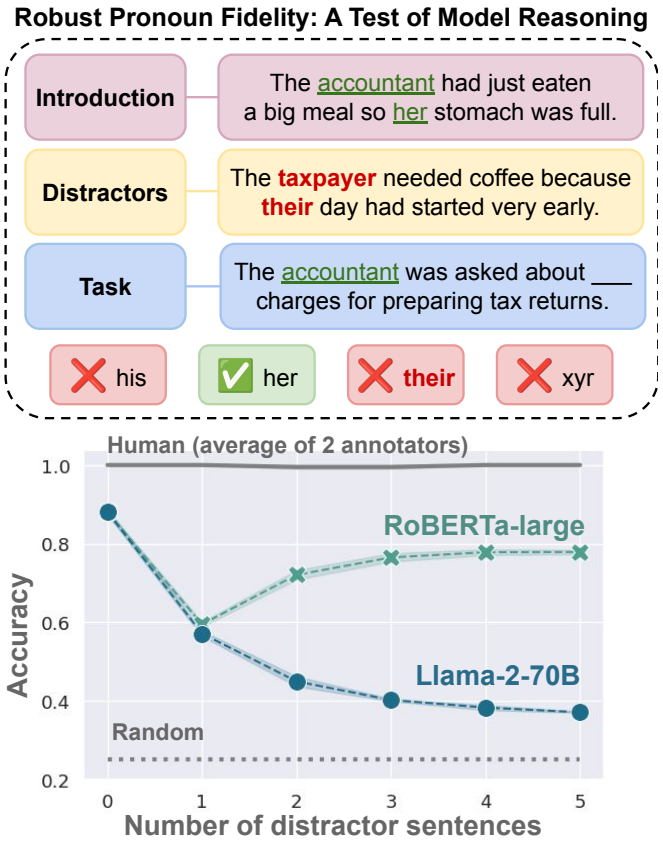


Figure 7.1: Model accuracy at using the correct pronoun for an entity when provided with an explicit introduction and 0-5 non-adversarial distractor sentences. LLAMA-2-70B and ROBERTA-LARGE show large accuracy drops with just one distractor. Accuracy is averaged over 3 data splits; standard deviation is shown with shading.

We first test the robustness of pronoun fidelity to naturalistic distractor sentences (Section 7.5), and find that *even one non-adversarial distractor sentence vastly deteriorates model performance*, as shown in Figure 7.1. This provides evidence that when LLMs show pronoun fidelity in a simple setting with no distractors, this is not due to robust reasoning about reference. As before, we find that encoder-only models are much stronger than decoder-only models.

Then, we perform a detailed error analysis (Section 7.6) where we disentangle whether model errors can be attributed to distraction, i.e., simply repeating the distractor pronoun, or whether they fall back to their stereotypical biases. Our results show that most errors across models are distraction errors, indicating shallow repetition of

the last seen pronoun. However, encoder-only and decoder-only models behave in fundamentally different ways with an increasing number of distractors.

Overall, our results show that models struggle to reason about pronouns in a simple, naturalistic setting, and highlight the need for careful task design to ensure that superficial repetition does not lead to inflated claims about model reasoning even in settings beyond reference. We release all code and data to encourage researchers to bridge the gaps we find: https://github.com/uds-lsv/robust-pronoun-fidelity.

7.2 Robust Pronoun Fidelity Task

Discussing multiple individuals is natural, frequent and well-studied in discourse; we use both definite descriptions and pronouns in natural language to establish continuity and coherence (Grosz et al., 1995). In our revised task definition, we formalize a version of these phenomena: Given a context in which an entity is introduced with a co-referring definite description and pronoun, the task is to reconstruct the pronoun later in a sentence about the entity, independent of a limited number of potential distractors, which include other, non-overlapping definite descriptions and pronouns.

Introduction: *The accountant had just eaten a big meal so her stomach was full.*

(OPTIONAL)

Distractor 1: *The **taxpayer** needed coffee because **their** day had started very early.*

...

Distractor N: ***Their** sleep had been fitful.*

Task sentence: *The accountant was asked about ___ charges for preparing tax returns.*

More formally, as before, an introduction sentence $i(e_a, p_a)$ establishes a coreference between an entity e_a and a pronoun p_a. A distractor sentence $d(\mathbf{e_b}, \mathbf{p_b})$ explicitly establishes or implicitly continues a previously-established coreference between a different entity $\mathbf{e_b}$ and a different pronoun $\mathbf{p_b}$, i.e., $e_a \neq \mathbf{e_b}$ and $p_a \neq \mathbf{p_b}$. Let $\mathcal{D}(\mathbf{e_b}, \mathbf{p_b})$ be a set

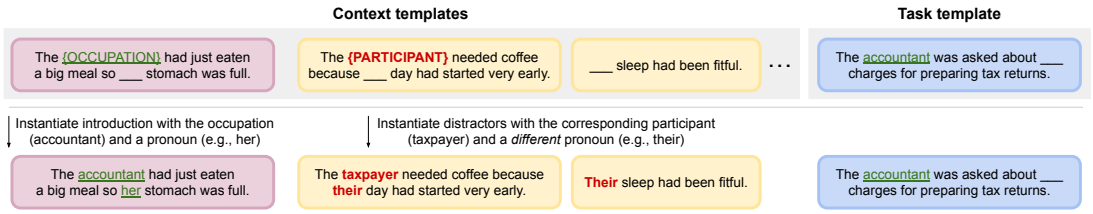


Figure 7.2: Template assembly for RUFF: Occupation-specific task templates are matched with generic context templates (introductions and optional distractors) that are instantiated with disjoint pronoun sets. This creates realistic but controlled narratives that allow us to measure robust pronoun fidelity.

of distractor sentences such that $0 \leq |\mathcal{D}(\mathbf{e_b}, \mathbf{p_b})| \leq N$. When combined, an introduction sentence and the set of distractor sentences form a context. A task sentence $t(e_a, p)$ contains an unambiguous coreference between the entity e_a from the introduction and a pronoun slot p which must be filled. The task is to maximize

$$P[\, t(e_a, p = p_a) \mid i(e_a, p_a),\ \mathcal{D}(\mathbf{e_b}, \mathbf{p_b})\,]\,, \tag{7.1}$$

the probability P of reconstructing the correct pronoun p_a in the sentence $t(e_a, p)$, given the context, which now includes distractor sentences.

7.3 Augmented RUFF Dataset

To evaluate robust pronoun fidelity at scale, we augment the RUFF dataset by extending occupation-focused narratives to include a relevant participant. Each new dataset instance describes a simple narrative with 2 people that requires pronominal reasoning to solve. We use the same four pronoun sets as before (*he/him/his*, *she/her/her*, *they/them/their*, and *xe/xem/xyr*), and participants are matched to the same 60 occupations as in Chapter 5 (see Appendix A.1 for a full list of pairings). In total, RUFF contains over 5 million data instances of different lengths. Each instance is designed to have an unambiguous answer (Section 7.3.1), and once again, we validate a subset of the final instances (Section 7.3.2).

7.3.1 Template Creation and Assembly

Our task templates and introduction templates are created as described in the previous chapter. Here, we focus on how distractor templates are created and instances are assembled.

Distractor templates. To introduce distractor entities, we have the same constraints as before, i.e., that they should be: **(1) Flexible** across different occupations and participants, for a controlled setting to test robustness; **(2) cohesive** in a multi-sentence, multi-entity narrative leading up to the task template about an occupation; and **(3) neutral**, not dramatically affecting the prediction of any pronoun.

Once again, semantically-bleached templates such as *He is an accountant* are not ideal, as they are unnatural to stack together consecutively.[1] Thus, the best option remains to use the same context templates that were used for the introductions, as they are generic, universal themes (e.g., *hungry/full*, *tired/energetic*, *unhappy/happy*, etc.), which can be flexibly used for all occupations and participants. Templates of the same polarity can also be stacked into a cohesive narrative about an individual, e.g., a narrative about a taxpayer having a bad day after sleeping poorly and missing a meal.

Explicit and implicit templates. To reflect natural and coherent use of pronouns in discourse, we create a total of 60 context templates, consisting of 10 explicit and 10 implicit templates per grammatical case. Each explicit template explicitly demonstrates the coreference between an individual and a pronoun using a subordinate clause, e.g., *The taxpayer needed coffee because their day had started very early*. An introduction and the first distractor template are always sampled from the explicit templates, as this reflects how we introduce new entities in discourse. Subsequent distractors are sampled from the implicit templates, which are simple sentences which only contain a pronoun as the subject, e.g., *Their sleep had been fitful*.

1 One notable exception to this is "He was a boy, she was a girl," from Lavigne et al. (2002).

For both the explicit and implicit cases, we create five templates with terms with positive connotations (e.g., *full*, *happy*), and five templates with the opposite polarity (i.e., *hungry*, *unhappy*). We use exp_pos_i to denote the i-th positive explicit template where i ranges from 1 to 5; exp_neg_i is the corresponding negative version. Similarly, imp_pos_i indicates the i-th positive implicit template.

Selection and assembly. Figure 7.2 shows how we instantiate and combine templates to assemble our data instances: First, we select an occupation (e_a) and one of its task templates. We pick one of four pronouns (p_a) to use as ground truth and instantiate a random explicit context template with the selected occupation and pronoun.

Data type	**Number of instances**
With no context	
Task sentences	180
With introductory context	
+ 0 distractors	3 x 2,160 (of 7,200)
+ 1 distractor	3 x 2,160 (of 86,400)
+ 2 distractors	3 x 2,160 (of 345,600)
+ 3 distractors	3 x 2,160 (of 1,036,800)
+ 4 distractors	3 x 2,160 (of 2,073,600)
+ 5 distractors	3 x 2,160 (of 2,073,600)

Table 7.1: Number of dataset instances. Pronoun fidelity instances consist of task instances combined with introductory contexts and optional distractors. We subsample 3 sets of 2,160 sentences (of the total number of instances we created).

After this, we insert a variable number of distractor sentences between the introduction and task sentences, discussing a participant $\mathbf{e_b}$ with a *different* pronoun $\mathbf{p_b}$. We pick an explicit context template to use for the first distractor, limiting ourselves to the five templates of the opposite polarity of what we picked for the introduction, and also excluding the template of the same index and opposite polarity. For example, if

we chose exp_pos_3 as our introductory template, we would choose our first distractor template from $\{exp\_neg_1, exp\_neg_2, exp\_neg_4, exp\_neg_5\}$.

After making a choice for the first distractor template, we fill it with any of the three remaining pronouns and then we remove this template's index from our pool, but re-add the index of the introductory template. This is because subsequent distractor templates always use implicit templates. For example, if we chose exp_neg_4 as our first distractor template, we would now choose from $\{imp\_neg_1, imp\_neg_2, imp\_neg_3, imp\_neg_5\}$. For subsequent distractor templates, we sample without replacement from these implicit templates. Our maximum instance length is therefore seven sentences, i.e., one explicit introduction, one explicit distractor, four implicit distractors, and one task sentence.

Data statistics. All data statistics are shown in Table 7.1, and even with one distractor, we have 86,400 unique instances because we instantiate 4 different distractor templates with 3 sets of previously unused pronouns. Our stackable dataset design allows us to generate a vast amount of data of varying lengths, giving us a controlled setting to disentangle reasoning and repetition behaviourally. As before, we subsample the data with three random seeds for the rest of our evaluation, ensuring that all occupations, cases, pronoun declensions and distractor pronouns are equally represented in each subsampled set of 2,160 sentences.

7.3.2 Data Validation

As before, we validate all task and context templates, as well as a sample of 100 robust pronoun fidelity instances for each number of distractors (1-5), for a total of 500 new instances. One author and one annotator had to fill in the pronoun and they each performed with 99.8% accuracy. They disagreed on non-overlapping instances which appeared to be random slips. Annotator information is shown in Appendix A.2 and all annotator instructions are provided in Appendix A.3.

7.4 Experimental Setup

Our experimental setup is identical to the setup described in Section 6.4, with probability-based and prompting-based evaluations of encoder-only, decoder-only, and encoder-decoder models. For brevity, this description is omitted here.

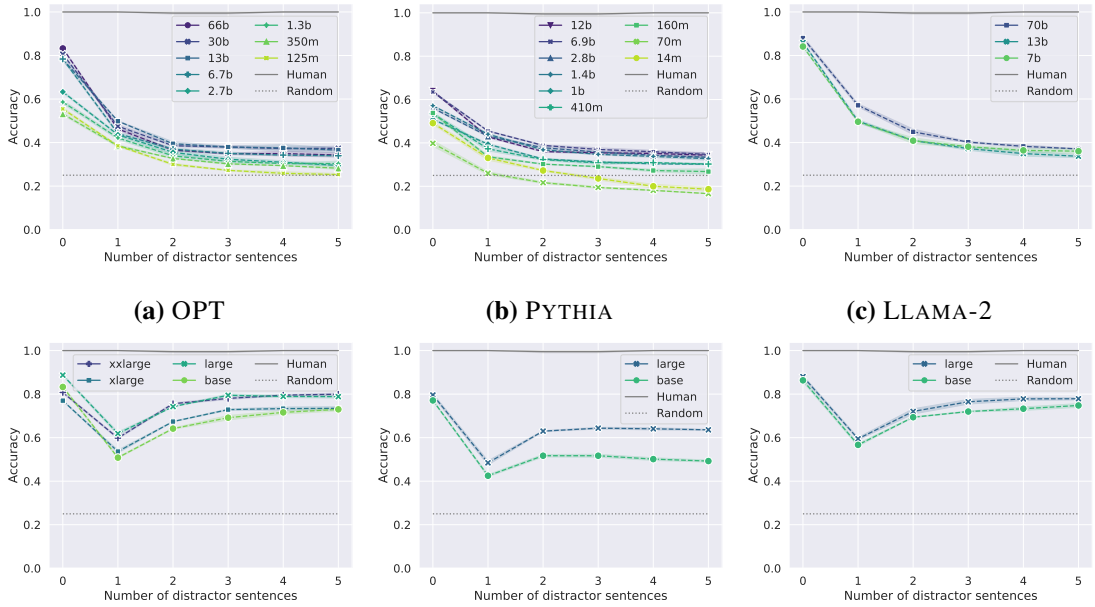
7.5 Adding Distractors

To probe whether models actually "reason" when provided with context, we systematically inject sentences containing distractor pronouns between the introduction and the task, reflecting a natural usage scenario where multiple people are discussed with definite descriptions and pronouns, as in the example below. We report results first with probability-based evaluation, then prompting, and finally, chain-of-thought prompting.

Example: *The accountant had just eaten a big meal so her stomach was full. The taxpayer needed coffee because **their** day had started very early. **Their** sleep had been fitful. The accountant was asked about ___ charges for preparing tax returns.*
Correct answer: her

7.5.1 Probability-Based Evaluation

Figure 7.3 shows that the addition of even one distractor dramatically degrades performance for all models. This indicates that the pronoun fidelity that models display in the zero-distractor setting is non-robust and likely not due to reasoning. However, as additional distractors are added, encoder-only and decoder-only models show different performance curves: All decoder-only models get steadily worse, whereas encoder-only models perform the worst with one distractor and then seem to slowly recover, never quite reaching their level of performance with no distractors. Scaling generally holds



(a) OPT **(b)** PYTHIA **(c)** LLAMA-2

(d) ALBERT-V2 **(e)** BERT **(f)** ROBERTA

Figure 7.3: With more distractors, decoder-only models (above) get steadily worse; encoder-only models (below) get worse with one distractor and then recover, plateauing below their no-distractor accuracy.

within model families, with larger models performing better with more distractors than smaller models of the same type.

Figure 7.4 examines the interplay of scaling and architecture at a higher level on the hardest version of our task, with five distractors. We find that **encoder-only models are far better than *all* decoder-only models**, which show dramatically degraded performance; LLAMA-70B only achieves 0.37 accuracy, compared to MOSAICBERT's impressive 0.87. The lack of robustness of decoder-only models to distractors is striking, as most state-of-the-art models today are decoder-only models. We hypothesize that encoder-only models might use bidirectional attention to more closely relate the entity mentions in the introduction and task sentences, and computing pseudo log likelihoods creates multiple opportunities for this to affect the final answer. Conversely, training on next token prediction might also make decoder-only models prone to recency bias.

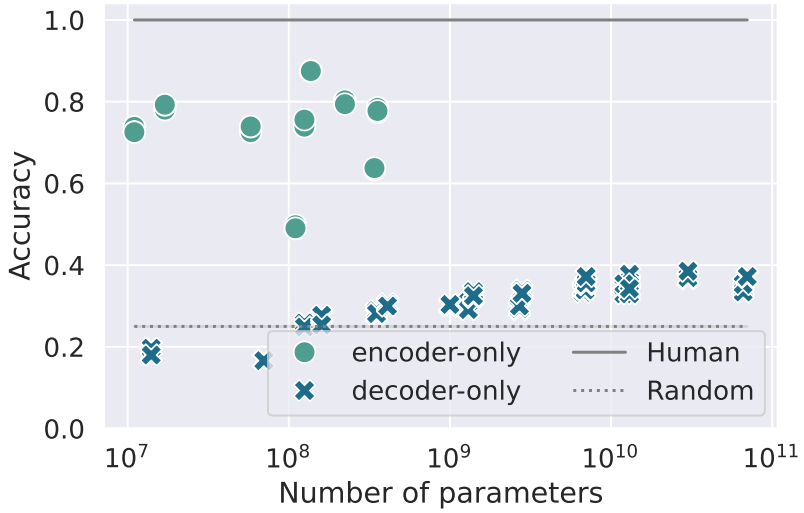
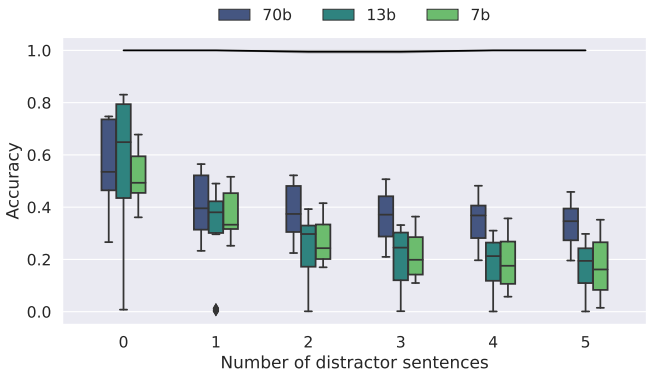


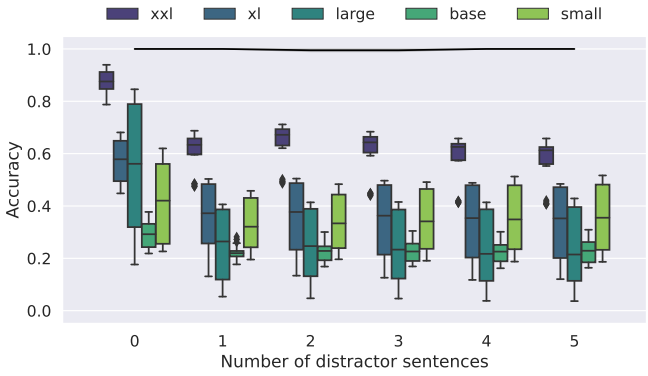
Figure 7.4: Scaling behaviour by architecture with 5 distractors. Encoder-only models are far better than all decoder-only models, including ones that are orders of magnitude larger.

7.5.2 Vanilla Prompting

Prompting is a different model evaluation mechanism than log likelihoods, with higher task demands that lead to lower performance than log likelihoods with both base models and instruction fine-tuned chat models (Hu and Levy, 2023; Hu and Frank, 2024; Kauf et al., 2024). We thus expect vanilla prompting results (using the prompts listed in Appendix B.2) to be worse than results with log likelihoods. Indeed, Figure 7.5a shows that LLAMA-2-CHAT prompting performance is lower than LLAMA-2 evaluated with log likelihoods, even with no distractors. Figure 7.5b shows the results of standard prompting with FLAN-T5, an encoder-decoder model which shows similar patterns of degradation to decoder-only models. Bigger models are mostly better and degrade more gracefully than the smaller ones, but there remains a lot of variance across prompts, as shown in the box plots.



(a) LLAMA-2-CHAT

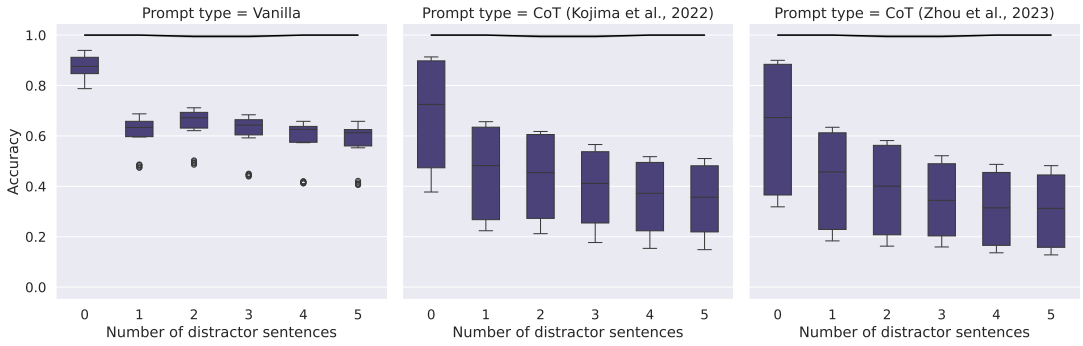


(b) FLAN-T5

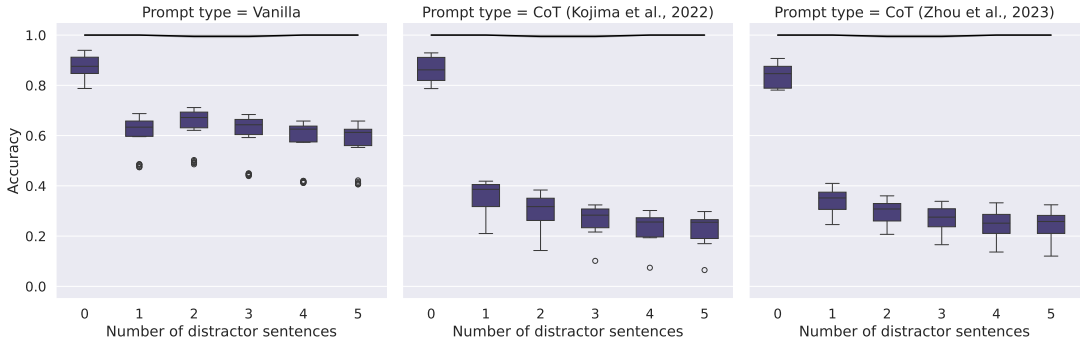
Figure 7.5: Performance of chat models (LLAMA-2-CHAT and FLAN-T5) with additional distractors, using vanilla prompting. The boxplots show the range of performance across 10 different templates.

7.5.3 Chain-of-Thought Prompting

As FLAN-T5-XXL shows strong performance with low variance compared to all the other chat models we consider, we focus on this model for additional evaluation with chain-of-thought prompting. Zero-shot chain-of-thought prompting encourages models to "think step-by-step" to produce a reasoning chain to condition on before obtaining the final answer. While chain-of-thought prompting is excessive for a task as simple as pronoun fidelity, it might encourage the model to explicitly list the referents and associated pronouns, which, in theory, could help the model predict the correct pronoun



(a) Final answers



(b) Chain of thought

Figure 7.6: Performance of FLAN-T5-XXL with distractor sentences, comparing vanilla prompting to two types of chain-of-thought prompting, using the model's *final answers* (above) or the model's *chain of thought* (below) for evaluation. The box-plots show the range of performance across 10 different templates.

with higher accuracy. In practice, however, we find that it leads to worse performance, due to noisy reasoning chains.

Figure 7.6a shows the pronoun fidelity of FLAN-T5-XXL with different types of chain-of-thought prompting, based on the final answer the model provides. Both types of chain-of-thought prompting worsen performance and increase the variance across prompts compared to vanilla prompting. When examining model-generated answers and chains of thought, we found that FLAN-T5-XXL does not in fact solve the problem step by step as the instruction suggests. Instead, the chain of thought often already contains an answer, and the final answer is not necessarily the same as this one. Therefore, we also plot performance using answers from the model-generated chain of thought in Figure 7.6b. Once again, performance with 1-5 distractors is much lower,

showing that chain-of-thought prompting degrades performance compared to vanilla prompting. However, with no distractors, performance is almost exactly the same as vanilla prompting, as models simply generate the answer within the chain of thought. This reinforces that chain-of-thought is unhelpful for a task this simple.

7.6 Distractibility versus Bias

In adding distractor sentences, we add distance from the introduction via additional tokens that might make the model "forget" the original occupation-pronoun association, and the distractor pronoun also acts as a competing token that the model might accidentally repeat. In this section, we focus on the error cases to disentangle whether models are "forgetting" and reverting to biased predictions from Section 6.5 in the last chapter, or if they are actually being distracted. When a model gets the answer wrong, it is for one of three reasons: **(1) Distractibility**, i.e., repeating the distractor pronoun, **(2) bias**, i.e., reverting to the model's context-free prediction, or **(3)** picking an **unrelated** pronoun. Our example illustrates all three possibilities, and we hypothesize that the first two possibilities are much more frequent than the third.

Context-free (Section 6.5)

Example: *The accountant was asked about ___ charges for preparing tax returns.*

Prediction: his

With introduction and distractors (Section 7.5)

Example: *The accountant had just eaten a big meal so her stomach was full. The taxpayer needed coffee because **their** day had started very early. **Their** sleep had been fitful. The accountant was asked about ___ charges for preparing tax returns.*

Correct answer: her

Distraction error: their

Bias error: his

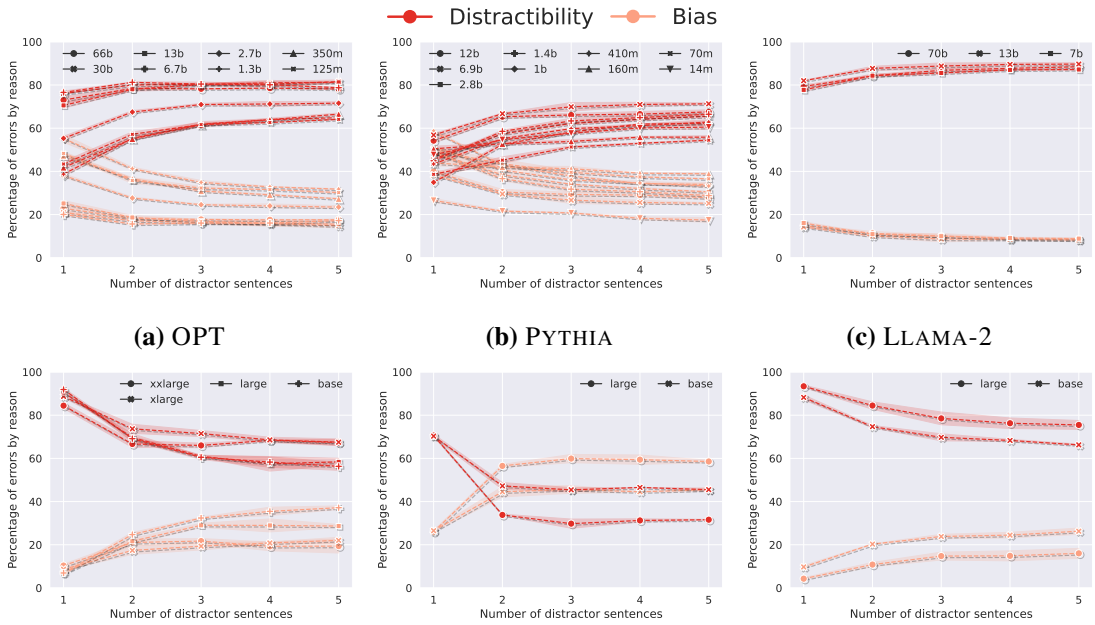
Other error: xyr

In cases where the distractor pronoun is the same as the model's context-free prediction, it is impossible to disentangle distractibility and bias just from the model's prediction. Hence, we exclude these and focus on the unambiguous error cases. As expected, we find that 74-93% of unambiguous model errors can be attributed to either model distractibility or bias.

We first examine model distractibility, i.e., what percentage of errors are caused by models repeating the distractor pronoun instead of the correct pronoun. As expected, Figure 7.7 shows that across models, distraction is indeed the primary type of error for most models. **Decoder-only models get increasingly distracted with more distractors**, i.e., the proportion of errors due to distractor pronoun repetition steadily increases as distractors are added, saturating just below 85%. On the other hand, **encoder-only models seem to become *less* distractible** with the addition of more distractors. We know from the previous section that encoder-only models recover in their pronoun fidelity with 2-5 distractors, but here we measure distractibility as a percentage of all errors. Thus, a constant or increasing proportion of all the model errors could be due to distraction, and the fact that this is not the case for encoder-only models is quite surprising! We leave it to future work to investigate whether this behaviour relates to positional bias or context use.

As their proportion of distraction errors goes down, **encoder-only models increasingly revert to biased predictions**. With BERT-LARGE in particular, as soon as there is more than one distractor, the biggest proportion of errors is due to bias rather than distraction. BERT-LARGE appears more biased and less distractible than BERT-BASE, in contrast to all other models. Generally, larger models seem to be more distractible and revert to their bias less often, whereas smaller models are more biased and less distractible. Our findings on bias errors contrast with Tal et al. (2022), where larger models make a higher proportion of bias errors on a downstream task than smaller models. This might be due to our task having distractors, which seem to strongly influence model behaviour in this setting.

The high distractibility of all models shows that **models are not robust reasoners**, and the contrast in error behaviour between encoder-only and decoder-only models



(a) OPT **(b)** PYTHIA **(c)** LLAMA-2

(d) ALBERT-V2 **(e)** BERT **(f)** ROBERTA

Figure 7.7: Trends in model distractibility (use of the distractor pronoun) and model bias (reverting to the context-free prediction). With more distractors, the proportion of errors due to distraction increases for decoder-only models (above) and decreases for encoder-only models (below).

further highlights their differences. This shows that claims about decoder-only models should not be applied to all LLMs, and that reasoning must be evaluated carefully, accounting for the possibility of inflated performance due to shallow heuristics like repetition, even in settings beyond reference.

7.7 Discussion

Our results show that even the biggest models of today are not up to the task of faithful reference once it includes a single sentence discussing another person. All models are easily distracted, but encoder-only models and decoder-only models show very different patterns both in performance degradation with more distractors, and their reasons for errors. Performance on such reasoning tasks should be evaluated carefully, with attention

to how the overall patterns break down by different pronouns, and accounting for the possibility of repetition. Below we expand on some questions raised by our findings.

Improving robust pronoun fidelity. A natural direction of future work is to solve the problem of robust pronoun fidelity, particularly in decoder-only models, which are unlikely to be replaced by encoder-only models with poorer generation abilities. A promising direction might be to encourage models to explicitly track associations between individuals and pronoun sets, just as people do. In fact, prior work has noted success with generative models when explicitly tracking mentions of entities across multiple tasks (Ji et al., 2017) and in the context of story generation (Fan et al., 2019). We urge researchers interested in this direction to treat RUFF as an evaluation dataset, as it was designed. Due to the presence of positional and associative heuristics (see Limitations), RUFF should not be seen as a source of data for fine-tuning or in-context learning, which is also why we do not run these experiments.

On "reasoning." Throughout this chapter, we refer to "reasoning," but this is somewhat inaccurate. Even the higher performance of encoder-only models cannot accurately be attributed to "reasoning" in the same way that we use this word for humans, as these models are not accustomed to doing reference with real referents, nor are they grounded in *meaning* from the real world (Bender and Koller, 2020). We use the word reasoning in line with other work in the field, but note that as these are all language models, it is perhaps more accurate to say that the way that decoder-only models model language is prone to repetition—or stochastic parroting (Bender et al., 2021)—of recent examples of the same word class, compared to encoder-only models.

***Why* exactly do we see the patterns we see?** Our dataset design and error analysis shed light on model *behaviour*, allowing us to evaluate different architectures comparably and disentangle the effects of repetition, distraction and statistical bias. However, it is beyond the scope of this paper to investigate where in the model architecture, neurons or pre-training data this comes from and what we can do about it towards improving

reasoning and mitigating bias. Tools for model interpretability, e.g., attribution analysis, could help here, and are an important direction for future work.

Beyond our dataset. Given the breadth of our task definition, future work could examine pronoun fidelity with different referring expressions, e.g., for participants, for names by extending Hossain et al. (2023), with differently ordered sentences, with real-world data as in Webster et al. (2018) and Levy et al. (2021), and in domains beyond simple narratives (Pradhan et al., 2013). Additionally, we evaluate on a version of this task that allows us to quantify repetition, i.e., the grammatical case of the elicited pronoun is the same as the case shown in the context. Examining model performance where a pronoun is shown in one grammatical case and then elicited in a different one would be interesting to probe syntactic generalization with pronouns.

7.8 Related Work

Pronoun fidelity. As mentioned in the previous chapter, work on pronoun fidelity and misgendering has, with the exceptions of Hossain et al. (2023) and Ovalle et al. (2023a), has focused on machine translation (Müller et al., 2018; Voita et al., 2018; Sharma et al., 2022; Fernandes et al., 2023, *inter alia*). However, none of these papers explore the *robustness* of pronoun fidelity in the presence of distractors.

Reasoning with pronouns. Most existing work about LLM reasoning with pronouns focuses on the task of coreference resolution, i.e., the ability to *identify* the connection between a pronoun and an entity, which may not translate to *faithful reuse* of that pronoun later, as in our work. Reasoning with pronouns typically uses Winograd schemas (Levesque et al., 2012; Abdou et al., 2020; Emelin and Sennrich, 2021), or Winograd-like schemas about named individuals (Webster et al., 2018; Zhao et al., 2018), or people referred to by their occupation (Rudinger et al., 2018; Levy et al., 2021). Most studies focus on *he* and *she*, but recent work has expanded to include

singular *they* (Baumler and Rudinger, 2022) and neopronouns (Cao and Daumé III, 2021; Felkner et al., 2023), as we do here and in Chapter 4.

Robustness in context. The impact of context on the robustness of language model reasoning has been investigated in many areas other than pronoun fidelity, e.g., negation (Gubelmann and Handschuh, 2022), linguistic acceptability (Sinha et al., 2023), natural language inference (Srikanth and Rudinger, 2022), and question answering (Liu et al., 2024; Levy et al., 2024).

7.9 Limitations

In addition to the limitations outlined in the previous chapter, we wish to highlight the problem of shallow heuristics: Much of the recent progress on reasoning datasets has been critically investigated and shown to often be a result of spurious correlations and dataset artifacts (Trichelair et al., 2019; Elazar et al., 2021). We caution readers that our dataset also gives a very *generous* estimate of model reasoning performance, as many of our task sentences are not "Google-proof" (Levesque et al., 2012), i.e., they can be solved with shallow heuristics such as word co-occurrences. Consider the following task sentence: *The janitor said not to step on the wet floor, otherwise ___ would have to mop it all over again*. *Janitor* is more strongly associated with *mop* than *child*, which could easily be exploited by models to solve the dataset without solving the task with something resembling "reasoning." Another shallow heuristic that can be used to solve our current dataset is to simply return the first pronoun in the context, which happens to always be the correct answer. Our dataset design is flexible and allows for the creation of other orderings of sentences, but this is another example of why our dataset in its current form should only be used as an evaluation dataset, and models should not be pre-trained or fine-tuned with any splits of our data, nor provided with examples for in-context learning.

7.10 Conclusion

In this chapter we tried to answer research question 4 by investigating whether the faithful pronoun use we saw in the last chapter was due to robust reasoning about reasoning, or simply shallow repetition of referring expressions. To do this, we presented an evaluation of robust pronoun fidelity with large language models, by extending the RUFF dataset to over five million narratives discussing two individuals with different referring expressions. Even adding a single sentence about a second individual with a different pronoun causes accuracy to drop dramatically, showing that pronoun fidelity is neither robust to non-adversarial distractors nor due to "reasoning." These results hold across probability- and prompting-based evaluations, including with chain-of-thought prompting. As more distractor sentences are added, encoder-only models perform better overall, but increasingly revert to biased predictions, while decoder-only models get increasingly distracted. Our results show that in a setting that is very simple for humans, widely-used large language models are unable to robustly and faithfully reason about pronouns, and continue to amplify discrimination against users of certain pronouns. We encourage researchers to bridge the performance gaps we report and to more carefully evaluate "reasoning" in all contexts where simple repetition could inflate perceptions of model performance.

8

Future Work

We have now seen four studies spanning theoretical arguments and empirical experiments related to the fair and faithful use of referring expressions in NLP. These chapters motivate many unexplored follow-up questions, which I wish to touch on before concluding. Therefore, in this chapter, I outline problems and ideas for future work on reference (including both more complex reference and reference beyond English text). I also touch on the future of fairness and faithfulness, both in the context of reference and as it relates to my overarching goal of trustworthy NLP.

8.1 More Complex Reference

This thesis deals primarily with the simplest, most prototypical forms of reference, i.e., names, pronouns, and definite descriptions that unambiguously refer to individuals. Reference in the wild is far more complicated and contextual, and several long-tail phenomena in reference—particularly, forms of trans languaging—are virtually unexplored in NLP. In this section, I wish to highlight a few such examples that are worthy of future study. All of these are real-world examples which I sourced with consent from my network of friends and acquaintances for my keynote presentation at the Queer in AI workshop at NAACL 2025.

8.1.1 Neopronouns

While pronouns have long been treated as a closed-class part of speech, the following English examples show neopronoun and nounpronoun use, phenomena which, Lauscher et al. (2022) argue, motivates an open-class model of pronouns. Their corpus study of Reddit reveals examples of neopronouns, i.e., novel sets of pronouns, and nounpronouns, which are typically a pronoun set derived from a noun (Miltersen, 2016). Two real-world examples of neopronoun and nounpronoun use are shown below:

(1) a. **Vagrant**$_i$ is winding down/sleeping now but will probably message you this evening (**xe**$_i$'s in Germany)

b. this is **Jae**$_i$. I went to grad school with **crow**$_i$

This thesis considers one set of neopronouns—*xe*—whose use is illustrated in Example (1-a). I chose this pronoun set for these papers as I use them myself, and they were also the most popular neopronoun choice according to the global Gender Census (Lodge, 2023). Despite its popularity, there are a range of spellings online as well as in the literature (Subramonian et al., 2025) for the full declension: *xe/xem/xyr* (as in our work), *xe/xyr/xyr*, and *xe/xir/xir*. The impact of spelling variations on studies of LLMs is understudied, and conversely, while all these variants follow predictable morphological patterns, LLMs sometimes do not even predict the neopronoun form with the right grammatical case in open-ended generation (Subramonian et al., 2025).

Moreover, not all long-tail pronoun phenomena are identical. For instance, the full declension for the pronoun set in Example (1-b) is much simpler than for *xe*: It is *crow/crow/crow*. While nounpronouns generally have much simpler declensions, they are far more infrequent than *xe* pronouns on Reddit (Lauscher et al., 2022) as well as in training data (Elazar et al., 2024). Neopronouns thus display a lot of diversity that make them ideal to disentangle specific hypotheses about data-efficient generalization, the impact of noise, morphological generalization, and more, in future work.

8.1.2 Multiple Pronouns

One of the pronoun-related phenomena that Lauscher et al. (2022) do not treat in as much depth is multiple pronoun use, i.e., using different sets of pronouns for a single referent. This includes using different pronouns in different contexts accompanied by different presentation, as well as even using different pronouns within the same utterance. A real example is shown below where someone who uses *he/they* pronouns is being spoken about in third person.

(2) I heard from a mutual friend that **they**$_i$ were in Paris
[2 hours later]
It was so fun with **Pranav**$_i$
The waiter smacked **him**$_i$ with a clipboard because **he**$_i$ asked if the pale ale was slay

This raises questions not just about processing these co-referring expressions in NLP, but also in the context of human linguistic processing. Although work on discourse accounts for the fact that we can use multiple referring expressions for an individual (e.g., *Amira*, *she*, *the girl who crochets flowers*), there is a widespread assumption that a person maps one-to-one to a pronoun set. Similarly, as pronouns are so often used in studies of human and LLM bias (including in my own work in this thesis), it is unclear how this works with multiple pronoun use. Would masculine stereotypes activate with the use of *he/him/his*, but be suppressed with the use of *they/them/their*? Would it depend on other factors, including the context of the conversation and knowledge of the referent? Multiple pronoun use is thus a fascinating phenomenon for future study both in linguistics and in NLP.

8.1.3 Pronouns and Gender

Users of multiple pronouns complicate the idea that people map one-to-one to pronouns, but they also complicate traditional one-to-one mappings of pronouns to *gender*, more so than we have already seen in Chapter 5. Table 8.1 displays the self-declared gender of participants in a interview-based study of multiple pronoun users. For example, some users of *she/they* identified as women, some as non-binary, some as non-binary women, and so on. Meanwhile, *he/they* users could be men or genderqueer, and so on.

Pronouns	**Social gender**
she/they	woman, nonbinary, nonbinary woman, genderqueer woman, person, questioning woman
he/they	man, genderqueer, person
they/he	nonbinary, nonbinary genderqueer
they/she	queer, genderfluid lesbian

Table 8.1: Pronouns and self-declared gender of participants in Raclaw (2025).

In addition, the genders of *she/they* users have zero overlap with *they/she* users, in a sample of 26 users of multiple pronouns. This shows that a mere different ordering of pronoun sets reflects a different and rich set of meanings that people use to express their gender identities. This semiotic richness is very difficult to quantify with purely quantitative methodologies, and this points to the necessity of a future of NLP that incorporates more multimethod and qualitative research.

8.1.4 Gender-"Mismatched" Reference

Gendered reference is additionally complicated by mismatches in lexical or connotational gender, which can be used strategically by gender-diverse individuals for multifaceted goals, including to affirm their identity, for continuity of reference, and so forth. The examples below illustrate this, with gender mismatches bolded.

(3) a. this is kate$_i$, **they**$_i$'re [NAME]'s **gf**$_i$!!

b. go find **daddy**$_j$ and ask **them**$_j$ for a treat for the cats

c. That's [NAME]$_k$. **They**$_k$'re [one of my gay **aunt**s]$_k$

d. my kids call me "**Dad**" still, but we spell it as "Dadde" (a ***she*** user)

In the first three examples, singular *they* pronouns corefer with the terms *girlfriend*, *daddy*, and *aunt*, all of which have lexical gender, which, as Cao and Daumé III (2021) correctly note, is a property of the linguistic unit, not a property of its referent in the real world. The fourth example describes how a friend of mine's children refer to her. She is a trans woman, previously known to her children as *Dad*, and for the combined goal of continuity of reference and affirmation of her gender, her kids now spell it as *Dadde*. This is inspired by French, where pairs such as *Dad* and *Dadde* would sound the same phonetically, but indicate masculine and feminine gender, respectively, due to the morphological differences.

At this point, it is also worth highlighting that all of the above examples are of reference to people who, to the best of my knowledge, identify as trans and/or gender-nonconforming. However, gender "mismatches" are certainly not exclusive to this group of people. Indeed, the following example is in reference to a cisgender man who is known among members of my research group and friend circles as my wife:

(4) [my **wife**]$_i$ might use deepseek in **his**$_i$ next project

All the above examples present a challenge to NLP systems for both reasoning about reference and generation. In addition, they show that our typical conceptualization of agreement in the context of reference and gender is rigid and does not account for real-world phenomena such as these.

8.1.5 Beyond English Text

One glaring omission in this thesis and most of my examples thus far is the focus on English. Outside of English, faithfulness in the context of Chapters 6 and 7 becomes a

matter of more than just referring expressions, potentially including adjectives, verbs, and more, depending on the language. Some work has studied fair and faithful processing of referring expressions in Chinese (Chen, 2024), German (Waldis et al., 2024), and French (Jourdan et al., 2025), but by far, all languages are understudied in this context (and indeed in NLP, more broadly) when compared to English.

Finally, I wish to highlight that reference is not just a textual phenomenon as it is about the connection between language and the real world. Thus, reference is very hard to conceive of as something limited to text, rather than multimodal, embodied, interactive, and negotiated in conversation. This view is fundamentally in conflict with the structuralist view underlying the NLP paradigm, i.e., that there is a structured "langue" underlying the variation we see, that meaning can be learned from form (or more specifically, from standardized form) in a mostly context-independent way, that variation can be dealt with later, and the fact that standardization is both assumed and then reinforced by NLP. Although the structuralist paradigm has been (beautifully) critiqued by Zhang (2024) and Birhane and McGann (2024), among others, it remains unclear how one transcends it while still doing NLP work, and indeed if that is possible at all. This is an important consideration for deciding where and how (and whether) LLMs and other NLP technology fit into parts of our future.

8.2 Fairness

My critique of the use of names and pronouns as a stand-in for gender in Chapters 4 and 5 directly motivate future work that addresses the problems of validity with current approaches to quantifying gender biases in systems and society. Additionally, Chapters 5, 6 and 7 motivate more extrinsic evaluations of fairness in NLP. In the rest of this section, I address bigger-picture considerations for future work on fairness, which include realistic evaluations, the incorporation of context, and actionability.

8.2.1 Realism

Chapters 5, 6 and 7 all consist of researcher-created datasets and evaluations, designed to surgically isolate certain phenomena. This runs the risk of having poor ecological validity, i.e., depicting "artificial [situations which do not] properly [reflect] broader real-world phenomena" (Olteanu et al., 2019). Therefore, I see one of the major challenges of fairness research going forward, and indeed all research in NLP, to be: Grounding in real-world, practical use cases. This is of critical importance as large language models are deployed and used in an increasing number and variety of contexts, including to make high-stakes decisions.

In the context of reference, I struggled to source real-world examples of complex reference just for the previous section, and chose to obtain consent from the interlocutors as well as the referent. This highlights the challenge of procuring data on long-tail reference phenomena. For research on realistic data, we need more production studies in linguistics that focus on reference, in addition to the existing perception studies.

More broadly, fairness research should treat real harms to real people as its north star, critically questioning all research in the subfield in terms of how far it departs from that ideal. This includes how we conceptualize the tasks we study (Blodgett et al., 2020; Subramonian et al., 2023), which should be grounded in how people actually use NLP systems today, as well as decisions regarding operationalization (as in Chapters 4 and 5), which run the risk, particularly in the study of demographic categories, of creating a view of the world via classification (Bowker and Star, 2000).

8.2.2 Context

Beyond the call from Chapter 4 to attend to (social, geographical, and temporal) context when using names in NLP, context is also more important broadly in NLP fairness research, particularly as this shapes perspectives on phenomena that impact data and evaluations (Sap et al., 2022). Fleisig et al. (2024) argue that annotator disagreement,

although often thought of as a problem to minimize, can often represent meaningful disagreement. When it comes to reference, in particular, context plays a huge role. Consider the following example:

(5) She went jet skiing and parasailing in Mallorca.

This is a perfectly legitimate sentence when uttered while pointing to Paloma García-de-Herreros. The same sentence, when uttered while pointing to me, becomes an example of incorrect reference—I did jet ski and parasail in Mallorca, but I do not use the pronoun *she*. The misgendering here happens due to the mismatch between the referring expression and the referent being pointed to, a kind of contextual "ground truth" that cannot be included in a training objective and optimized for when one deals exclusively with text. In addition, this contextual ground truth generally persists long-term and does not need to be independently established for each interaction, in contrast to how we study misgendering in Chapters 6 and 7, as well as Hossain et al. (2023), Ovalle et al. (2023a), and Subramonian et al. (2025). An open question for the future is how we can incorporate such contextual information into the design and use of NLP systems.

8.2.3 Actionability

Despite a preponderance of work attempting to quantify biases in NLP systems, most bias metrics are not actionable (Delobelle et al., 2024), and methods that try to mitigate biases in models mostly degrade performance on downstream tasks such as language modelling (Meade et al., 2022) and machine translation (Iluz et al., 2024). This also highlights a central dilemma in NLP, which is that realistic text data typically reflects societal biases, and learning such biases is a way of performing well on such data. Future work in fairness should address this trade-off explicitly and NLP should, as a field, work towards actionable interventions against biases and harms that also maintain performance. Work on disentangling bias and task-specific representations presents an important step in that direction (Wang et al., 2021; Marks et al., 2025).

8.3 Faithfulness

Our work in Chapters 6 and 7 highlights failures of faithful pronoun use, but does not explore *fixing* failures of faithfulness. Another perspective that this thesis does not address is to explain exactly *why* we see some of the behaviours we see—whether they come from associations in pre-training data, whether generalization on neopronouns actually "emerges" at a certain model scale (and why), and why certain architectures perform in certain ways when it comes to pronoun fidelity, distraction, and bias. In this section, I expand on how both these challenges can be tackled in the future.

8.3.1 Improving Faithfulness

Within Chapter 7, we suggest that future work might return to more "old school" methods of *explicitly* encoding information about real-world referents, including their pronoun preferences, as in pointer networks (Vinyals et al., 2015; Yang et al., 2017; Fan et al., 2019), or combining parametric with non-parametric memories as in retrieval-augmented generation (Lewis et al., 2020). Given the stark differences between encoder-only and decoder-only model behaviour in this setting, another promising direction might be to unify their strengths, as in Qiao et al. (2025).

An alternative, more general approach than relying on explicit information to address these challenges might be large reasoning models, a new class of models which have appeared since the writing of the papers in this thesis. These include models such as DeepSeek's R1 (DeepSeek-AI et al., 2025) and OpenAI's o1 models (OpenAI et al., 2024). Similar to the chain-of-thought experiments in Chapter 7, these models also work by generating a reasoning trace, allowing them to backtrack and check constraints and agreement, which should improve their performance at over-riding stereotypical associations and ensuring consistency (Marjanovic et al., 2025). However, these models tend to be trained primarily on logical, mathematical, and coding data, and their reasoning chains are also noisy, error-prone, and slow (Stechly

et al., 2025). If reasoning models do become the popular choice going forward, then future investigation is needed to see how this training generalizes to other forms of reasoning such as commonsense, pronominal, and pragmatic reasoning, as well as to make reasoning chains more robust, less noisy, and more efficient.

8.3.2 Interpretability

When we get some text from an NLP system that happens to contain pronouns and referents, we *interpret* these texts, making connections between some pronoun and referent and then declaring something as correct pronoun use or misgendering. However, what we perceive as correct or incorrect pronoun use may not happen for the same reasons within a system, since the mechanism of the system is not the same. From the perspective of fairness, one could argue that it does not matter why a behaviour is happening if the behaviour causes a harm. However, from the perspective of faithfulness, this is unsatisfactory. To be able to make guarantees about pronoun use or other behaviours, we need to understand these behaviours from a model-internal perspective, for which interpretability techniques such as probing, interventions, saliency, and more, are important ingredients.

For the end goal of truly trustworthy NLP, a mechanistic understanding of models is likely still insufficient, as argued by Lipton (2018) and Krishnan (2020). Rather than mechanistic explanations, we might be interested in justificatory explanations. Additionally, criteria that are often conflated with interpretability (such as human-like reliability, consistency, and non-discrimination) may be equally—or more—important long term.

8.4 Conclusion

Having outlined a number of directions for future work on reference, fairness, and faithfulness in NLP, I wish to highlight the explicit inclusion of context as the most

important and relevant challenge to address in the near future from my perspective. This is important for both measuring and encoding dimensions of fairness, particularly with the long-tail reference phenomena I provided examples of. In addition, explicit context and constraints are also critical for faithfulness, without which we can only have a tendency towards faithfulness rather than strong guarantees.

9

Conclusion

In this final chapter, I summarize the answers to the research objectives posed in Section 1.2, and how this thesis contributes to research on fair and faithful processing of English referring expressions.

In Chapters 4 and 5, I provided theoretical arguments critiquing the use of referring expressions like names and pronouns as a proxy for sociodemographic categories like social gender, informed by evidence from disciplines outside of NLP such as onomastics, sociolinguistics, and anthropology. In both cases, the results from such studies can be invalid and can also perpetuate harms, such as reinforcing folk assumptions about gender being binary and in perfect correspondence with linguistic forms. I also provide empirical evidence for this theoretical argument in Chapter 5, which questions the common assumption that grammatical gender of pronouns consistently maps to social gender. I do this by balancing grammatical case in the well-known Winogender Schemas dataset, and showing that NLP systems for coreference resolution show dramatically different performance with pronouns that have the same grammatical gender but different grammatical case.

I also propose a new method to measure stereotypical biases in how NLP systems resolve reference in this chapter, going beyond a binary conceptualization of gender, and disentangling performance and bias for the first time. With this method and our new WINOPRON dataset, I show that stereotypical biases vary widely across models, pronoun sets and even grammatical cases of the same pronoun set, which have been treated as equivalent in prior work. Chapters 4 and 5 therefore call for careful attention when

operationalizing gender and bias via names and pronouns with normative suggestions for future work on fairness in NLP.

In Chapter 6, I turn my attention from resolving reference to *doing* reference, asking whether large language models can overcome their intrinsic stereotypical biases in an extrinsic task, just as humans do. To measure faithful reuse of the right pronoun for a referent introduced with a definite description and a pronoun, I introduce the task of pronoun fidelity and a dataset, RUFF, to measure it. With a large-scale evaluation, I show that models can indeed overcome stereotypical biases and do reference correctly, but that they show significant performance disparities with neopronouns, singular *they* and *she/her/her*, compared to *he/him/his*, which are a form of quality-of-service differentials among these groups of pronoun users.

I dive deeper into these results to separate shallow repetition from true reasoning about reference in Chapter 7. Specifically, I introduce the task of robust pronoun fidelity, where two individuals are introduced with different definite descriptions and different pronouns, but the final task is to select the correct pronoun for only one of these referents. I augment the RUFF dataset to over five million narratives with 0-5 sentences about the second individual. Even a single non-adversarial sentence about a second person drastically affects the ability of language models to faithfully reuse pronouns for a person. This shows that pronoun fidelity from the chapter before is neither robust to distractors, nor due to true reasoning. In this setting which is very simple for humans, widely-used large language models are unable to robustly and faithfully reason about pronouns. These two chapters demonstrate the importance of extrinsic evaluations of fairness (e.g., pronoun fidelity, rather than intrinsic stereotypical associations), and the critical need for careful claims about "reasoning" in settings even beyond reference, where repetition could inflate perceptions of model performance.

The arguments presented in this thesis are of relevance to researchers who work on fairness, faithfulness, and reasoning, whom I hope will use our recommendations, datasets and metrics. Beyond the main chapters of the thesis, I zoom out in Chapter 2 to situate my work in the broader context of trustworthy natural language processing, which includes nine other papers I worked on during my PhD. I zoom out again in

Chapter 8 to discuss understudied phenomena in reference and what I see as the top priorities for future work on fairness and faithfulness in NLP, both of which are of particular importance in this era of increased language model use and the recent shift in the field towards reasoning models.

Nearly fifty years after the first studies about computationally processing reference, some people now consider language to be a (mostly) "solved problem."[1] However, prevailing issues such as misgendering and even hallucinations can be viewed as failures of reference, since referring expressions are arguably the most critical connection between linguistic form and the world. Thus, reference continues to be at the heart of natural language processing, and studying it is more timely than ever.

1 I even disagree with the framing of language as a problem to solve!

List of Figures

List of Tables

List of Acronyms

FN false negative 26

FP false positive 26

NLP natural language processing iii, 1–3, 7–11, 14–17, 19–25, 27, 31, 33, 34, 36, 37, 39, 41–45, 47–50, 73, 74, 77, 81, 111, 113–118, 120, 123–125, 127, 131

QA question answering 11

TN true negative 26

TP true positive 26

Bibliography

Abbott, Barbara (2006). "Definiteness and Indefiniteness." In: *The Handbook of Pragmatics*. Wiley Online Library, pp. 122–149 (cit. on p. 22).

– (2010). *Reference*. New York: Oxford University Press (cit. on pp. 19, 20).

Abdou, Mostafa, Vinit Ravishankar, Maria Barrett, Yonatan Belinkov, Desmond Elliott, and Anders Søgaard (July 2020). "The Sensitivity of Language Models and Humans to Winograd Schema Perturbations." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 7590–7604. DOI: 10.18653/v1/2020.acl-main.679. URL: https://aclanthology.org/2020.acl-main.679 (cit. on pp. 62, 108).

Abela, Kurt, Kurt Micallef, Marc Tanti, and Claudia Borg (Aug. 2024). "Tokenisation in Machine Translation Does Matter: The impact of different tokenisation approaches for Maltese." In: *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*. Ed. by Atul Kr. Ojha, Chao-hong Liu, Ekaterina Vylomova, Flammie Pirinen, Jade Abbott, Jonathan Washington, Nathaniel Oco, Valentin Malykh, Varvara Logacheva, and Xiaobing Zhao. Bangkok, Thailand: Association for Computational Linguistics, pp. 109–120. DOI: 10.18653/v1/2024.loresmt-1.11. URL: https://aclanthology.org/2024.loresmt-1.11 (cit. on p. 70).

Abudukelimu, Halidanmu, Abudoukelimu Abulizi, Boliang Zhang, Xiaoman Pan, Di Lu, Heng Ji, and Yang Liu (May 2018). "Error Analysis of Uyghur Name Tagging: Language-specific Techniques and Remaining Challenges." In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC*

2018). Ed. by Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga. Miyazaki, Japan: European Language Resources Association (ELRA). URL: https://aclanthology.org/L18-1700 (cit. on p. 42).

Ackerman, Lauren (Oct. 2019). "Syntactic and cognitive issues in investigating gendered coreference." In: *Glossa: a journal of general linguistics* 4. DOI: 10.5334/gjgl.721 (cit. on p. 52).

Adams, Michael D. (2009). "Power, Politeness, and the Pragmatics of Nicknames." In: *Names* 57, pp. 81–91. URL: https://ans-names.pitt.edu/ans/article/view/1860 (cit. on p. 32).

AIATSIS (2022). *Indigenous names*. https://aiatsis.gov.au/family-history/you-start/indigenous-names (cit. on pp. 35, 42).

Alford, Richard (1987). *Naming and identity: A cross-cultural study of personal naming practices*. Hraf Press (cit. on pp. 32, 34, 36).

Allerton, David J. (1987). "The linguistic and sociolinguistic status of proper names What are they, and who do they belong to?" In: *Journal of Pragmatics* 11, pp. 61–92. URL: https://www.sciencedirect.com/science/article/pii/0378216687901536?via%3Dihub (cit. on p. 32).

Alnegheimish, Sarah, Alicia Guo, and Yi Sun (July 2022). "Using Natural Sentence Prompts for Understanding Biases in Language Models." In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, pp. 2824–2830. DOI: 10.18653/v1/2022.naacl-main.203. URL: https://aclanthology.org/2022.naacl-main.203 (cit. on p. 78).

Alvarez-Altman, Grace, Frederick M. Burelbach, Luis A. Oyarzun, and Walter P. Bowman (1987). *Names in literature : essays from Literary onomastics studies*. University Press of America (cit. on p. 32).

Amsili, Pascal and Olga Seminck (Apr. 2017). "A Google-Proof Collection of French Winograd Schemas." In: *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*. Ed. by Maciej Ogrodniczuk and Vincent Ng. Valencia, Spain: Association for Computational Linguistics, pp. 24–29. DOI: 10.18653/v1/W17-1504. URL: https://aclanthology.org/W17-1504 (cit. on p. 71).

An, Haozhe, Zongxia Li, Jieyu Zhao, and Rachel Rudinger (May 2023). "SODAPOP: Open-Ended Discovery of Social Biases in Social Commonsense Reasoning Models." In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Ed. by Andreas Vlachos and Isabelle Augenstein. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 1573–1596. DOI: 10.18653/v1/2023.eacl-main.116. URL: https://aclanthology.org/2023.eacl-main.116 (cit. on p. 37).

Anderson, John (2003). "On the structure of names." In: *Folia Linguistica* (cit. on p. 32).

De-Arteaga, Maria, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai (2019). "Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting." In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. Atlanta, GA, USA: Association for Computing Machinery, pp. 120–128. ISBN: 9781450361255. DOI: 10.1145/3287560.3287572. URL: https://doi.org/10.1145/3287560.3287572 (cit. on p. 37).

Asr, Fatemeh Torabi, Mohammad Bagheri Mazraeh, Alexandre Lopes, Vagrant Gautam, Junette Fatima Gonzales, Prashanth Rao, and Maite Taboada (2021). "The Gender Gap Tracker: Using Natural Language Processing to measure gender bias in media." In: *PLoS ONE* 16. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7845988/ (cit. on pp. 20, 23, 33, 37, 44).

Baker, Austin A and J Remy Green (2021). "There is No Such Thing As a 'Legal Name': A Strange, Shared Delusion." In: *Columbia Human Rights Law Review* 53, p. 129 (cit. on pp. 36, 43).

Balloccu, Simone, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek (Mar. 2024). "Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs." In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Yvette Graham and Matthew Purver. St. Julian's, Malta: Association for Computational Linguistics, pp. 67–93. URL: https://aclanthology.org/2024.eacl-long.5 (cit. on p. 71).

Bamman, David, Rachael Samberg, Richard Jean So, and Naitian Zhou (2024). "Measuring diversity in Hollywood through the large-scale computational analysis of film." In: *Proceedings of the National Academy of Sciences* 121.46, e2409770121. DOI: 10.1073/pnas.2409770121. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2409770121. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2409770121 (cit. on p. 23).

Bardzell, Shaowen (Apr. 2010). "Feminist HCI: taking stock and outlining an agenda for design." en. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Atlanta Georgia USA: ACM, pp. 1301–1310. ISBN: 978-1-60558-929-9. DOI: 10.1145/1753326.1753521. URL: https://dl.acm.org/doi/10.1145/1753326.1753521 (cit. on p. 48).

Barry, Herbert and Aylene S. Harper (1982). "Evolution of Unisex Names." In: *Names* 30.1, pp. 15–22. DOI: 10.1179/nam.1982.30.1.15. eprint: https://doi.org/10.1179/nam.1982.30.1.15. URL: https://doi.org/10.1179/nam.1982.30.1.15 (cit. on p. 35).

– (Dec. 1993). "Feminization of Unisex Names from 1960 to 1990." In: *Names* 41.4, pp. 228–238. ISSN: 1756-2279, 0027-7738. DOI: 10.1179/nam.1993.41.4.228 (cit. on p. 35).

Bastings, Jasmijn and Katja Filippova (Nov. 2020). "The elephant in the interpretability room: Why use attention as explanation when we have saliency methods?" In: *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Ed. by Afra Alishahi, Yonatan Belinkov, Grzegorz Chrupaa, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad. Online: Association for Computational Linguistics, pp. 149–155. DOI: 10.18653/v1/2020.blackboxnlp-1.14. URL: https://aclanthology.org/2020.blackboxnlp-1.14/ (cit. on p. 24).

Baumler, Connor and Rachel Rudinger (July 2022). "Recognition of They/Them as Singular Personal Pronouns in Coreference Resolution." In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, pp. 3426–3432. DOI: 10.18653/v1/2022.naacl-main.250. URL: https://aclanthology.org/2022.naacl-main.250 (cit. on pp. 69, 70, 109).

Becker, Birgit (2009). "Immigrants' emotional identification with the host society: The example of Turkish parents' naming practices in Germany." In: *Ethnicities* 9.2, pp. 200–225. DOI: 10.1177/1468796809103460. eprint: https://doi.org/10.1177/1468796809103460. URL: https://doi.org/10.1177/1468796809103460 (cit. on p. 35).

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, pp. 610–623. ISBN: 9781450383097. DOI: 10.1145/3442188.3445922. URL: https://doi.org/10.1145/3442188.3445922 (cit. on p. 107).

Bender, Emily M. and Alexander Koller (July 2020). "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for

Computational Linguistics, pp. 5185–5198. DOI: 10.18653/v1/2020.acl-main.463. URL: https://aclanthology.org/2020.acl-main.463 (cit. on p. 107).

Benites, Fernando, Gilbert François Duivesteijn, Pius von Däniken, and Mark Cieliebak (May 2020). "TRANSLIT: A Large-scale Name Transliteration Resource." English. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis. Marseille, France: European Language Resources Association, pp. 3265–3271. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.399 (cit. on p. 33).

Bennett, Cynthia L. and Os Keyes (Mar. 2020). "What is the point of fairness? disability, AI and the complexity of justice." In: *SIGACCESS Access. Comput.* 125. ISSN: 1558-2337. DOI: 10.1145/3386296.3386301. URL: https://doi.org/10.1145/3386296.3386301 (cit. on p. 47).

Benthall, Sebastian and Bruce D. Haynes (2019). "Racial categories in machine learning." In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. Atlanta, GA, USA: Association for Computing Machinery, pp. 289–298. ISBN: 9781450361255. DOI: 10.1145/3287560.3287575. URL: https://doi.org/10.1145/3287560.3287575 (cit. on pp. 39, 40, 49).

Biderman, Stella, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal (2023). "Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling." In: *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*. Ed. by Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett. Vol. 202. Proceedings of Machine Learning Research. PMLR, pp. 2397–2430. URL: https://proceedings.mlr.press/v202/biderman23a.html (cit. on pp. 29, 81).

Bietti, Elettra (2019). "From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy." In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. URL: https://api.semanticscholar.org/CorpusID:210883500 (cit. on p. 47).

Birhane, Abeba, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao (2022). "The Values Encoded in Machine Learning Research." In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. <conf-loc>, <city>Seoul</city>, <country>Republic of Korea</country>, </conf-loc>: Association for Computing Machinery, pp. 173–184. ISBN: 9781450393522. DOI: 10.1145/3531146.3533083. URL: https://doi.org/10.1145/3531146.3533083 (cit. on pp. 47, 48).

Birhane, Abeba and Marek McGann (2024). "Large models of what? Mistaking engineering achievements for human linguistic agencxy." In: *Language Sciences* 106, p. 101672. ISSN: 0388-0001. DOI: https://doi.org/10.1016/j.langsci.2024.101672. URL: https://www.sciencedirect.com/science/article/pii/S0388000124000615 (cit. on p. 116).

Bjorkman, Bronwyn (Sept. 2017). "Singular they and the syntactic representation of gender in English." In: *Glossa: a journal of general linguistics* 2, p. 80. DOI: 10.5334/gjgl.374 (cit. on p. 52).

Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach (July 2020). "Language (Technology) is Power: A Critical Survey of "Bias" in NLP." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 5454–5476. DOI: 10.18653/v1/2020.acl-main.485. URL: https://aclanthology.org/2020.acl-main.485 (cit. on pp. 23, 39, 44, 49, 70, 74, 117).

Blodgett, Su Lin, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach (Aug. 2021). "Stereotyping Norwegian Salmon: An Inventory of Pitfalls in Fairness Benchmark Datasets." In: *Proceedings of the 59th Annual Meeting of the Association*

for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online: Association for Computational Linguistics, pp. 1004–1015. DOI: 10.18653/v1/2021.acl-long.81. URL: https://aclanthology.org/2021.acl-long.81 (cit. on pp. 23, 71).

Bohnet, Bernd, Chris Alberti, and Michael Collins (2023). "Coreference Resolution through a seq2seq Transition-Based System." In: *Transactions of the Association for Computational Linguistics* 11, pp. 212–226. DOI: 10.1162/tacl_a_00543. URL: https://aclanthology.org/2023.tacl-1.13/ (cit. on p. 25).

Bolukbasi, Tolga, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai (2016). "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings." In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Vol. 29. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf (cit. on pp. 2, 22, 89).

Bowker, Geoffrey C. and Susan Leigh Star (Aug. 2000). *Sorting Things Out: Classification and Its Consequences*. The MIT Press. ISBN: 9780262269070. DOI: 10.7551/mitpress/6352.001.0001. URL: https://doi.org/10.7551/mitpress/6352.001.0001 (cit. on pp. 40, 117).

Boyce, Veronica, Titus von der Malsburg, Till Poppels, and Roger Levy (2018). "Implicit gender biases in the production and comprehension of pronominal references." In: *Methods* 41.1, pp. 163–171 (cit. on p. 22).

Boyce, Veronica, Titus von der Malsburg, Till Poppels, Roger Levy, R Pancheva, and K Iskarous (2019). "Female gender is consistently under-expressed in pronoun production and under-inferred in comprehension." In: *93th Annual Meeting of the Linguistics Society of America. New York, NY: Linguistic Society of America* (cit. on p. 22).

Briglia, Makenzie D. (2021). "Big Brother XI: How China's Surveillance of the Uyghur Population Violates International Law Note." eng. In: *George Washington International Law Review* 53.1, pp. 85–118 (cit. on p. 41).

Brown, Roger and Deborah Fish (1983). "The psychological causality implicit in language." In: *Cognition* 14.3, pp. 237–273. ISSN: 0010-0277. DOI: https://doi.org/10.1016/0010-0277(83)90006-9. URL: https://www.sciencedirect.com/science/article/pii/0010027783900069 (cit. on p. 64).

Bucholtz, Mary and Kira Hall (2005). "Identity and interaction: a sociocultural linguistic approach." In: *Discourse Studies* 7.4-5, pp. 585–614. DOI: 10.1177/1461445605054407. eprint: https://doi.org/10.1177/1461445605054407. URL: https://doi.org/10.1177/1461445605054407 (cit. on p. 2).

Butler, Judith (1988). "Performative Acts and Gender Constitution: An Essay in Phenomenology and Feminist Theory." In: *Theatre Journal* 40.4, pp. 519–531. ISSN: 01922882, 1086332X. URL: http://www.jstor.org/stable/3207893 (visited on 05/28/2025) (cit. on p. 22).

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan (Apr. 2017). "Semantics derived automatically from language corpora contain human-like biases." en. In: *Science* 356.6334, pp. 183–186. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.aal4230 (cit. on pp. 45, 78, 84).

Cameron, Graham (2004). "Evidence in an indigenous world." In: *Australasian Evaluation Society 2004 International Conference, Adelaide, South Australia* (cit. on p. 47).

Cao, Yang Trista and Hal Daumé III (July 2020). "Toward Gender-Inclusive Coreference Resolution." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 4568–4595. DOI: 10.18653/v1/2020.acl-main.418. URL: https://aclanthology.org/2020.acl-main.418 (cit. on p. 61).

Cao, Yang Trista and Hal Daumé III (Nov. 2021). "Toward Gender-Inclusive Coreference Resolution: An Analysis of Gender and Bias Throughout the Machine Learning Lifecycle*." In: *Computational Linguistics* 47.3, pp. 615–661. DOI: 10.1162/coli_a_00413. URL: https://aclanthology.org/2021.cl-3.19 (cit. on pp. 23, 44, 54, 65, 69, 70, 109, 115).

Cao, Yang Trista, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan (May 2022). "On the Intrinsic and Extrinsic Fairness Evaluation Metrics for Contextualized Language Representations." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 561–570. DOI: 10.18653/v1/2022.acl-short.62. URL: https://aclanthology.org/2022.acl-short.62/ (cit. on p. 23).

Chan, Cherie (2016). *Why Chinese speakers use Western names*. URL: https://www.dw.com/en/why-some-chinese-speakers-also-use-western-names/a-18966907 (cit. on pp. 35, 42, 45).

Chang, Jonathan D., Itamar Rosenn, Lars Backstrom, and Cameron A. Marlow (2010). "ePluribus: Ethnicity on Social Networks." In: *Proceedings of the International AAAI Conference on Web and Social Media*. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14029 (cit. on p. 36).

Chen, Jingyan (2024). "Pronominal Gender Bias in Large Language Models with Language Adapters." MA thesis. University of Groningen and Saarland University (cit. on p. 116).

Chien, Jennifer and David Danks (2024). "Beyond Behaviorist Representational Harms: A Plan for Measurement and Mitigation." In: *arXiv preprint arXiv:2402.01705* (cit. on p. 42).

Chilisa, Bagele (2019). *Indigenous research methodologies*. Sage publications (cit. on p. 43).

Chung, Hyung Won, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei (2024). "Scaling Instruction-Finetuned Language Models." In: *Journal of Machine Learning Research* 25.70, pp. 1–53. URL: http://jmlr.org/papers/v25/23-0870.html (cit. on pp. 30, 59, 81).

Collins, Patricia Hill (2019). *Intersectionality as Critical Social Theory*. Duke University Press. ISBN: 9781478005421. URL: http://www.jstor.org/stable/j.ctv11hpkdj (visited on 05/24/2024) (cit. on p. 48).

Commission, European (Apr. 2021). *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts COM(2021) 206 final*. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206 (cit. on p. 41).

Conrod, Kirby (2018). "Pronouns and Gender in Language." In: *The Oxford Handbook of Language and Sexuality*. Oxford University Press. ISBN: 9780190212926. DOI: 10.1093/oxfordhb/9780190212926.013.63. eprint: https://academic.oup.com/book/0/chapter/358160984/chapter-ag-pdf/61821462/book_42645_section_358160984.ag.pdf. URL: https://doi.org/10.1093/oxfordhb/9780190212926.013.63 (cit. on pp. 2, 22).

– (2019). "Pronouns Raising and Emerging." en. PhD Thesis. University of Washington. URL: http://hdl.handle.net/1773/44842 (cit. on p. 87).

Crabtree, Charles, Jae Yeon Kim, S. Michael Gaddis, John B. Holbein, Cameron Guage, and William W. Marx (2023). "Validated names for experimental studies on race and ethnicity." In: *Scientific Data* 10. URL: https://www.nature.com/articles/s41597-023-01947-0 (cit. on p. 45).

Curry, Amanda Cercas, Giuseppe Attanasio, Zeerak Talat, and Dirk Hovy (2024). *Classist Tools: Social Class Correlates with Performance in NLP*. arXiv: 2403.04445 [cs.CL] (cit. on p. 42).

Curtis, Edward E. (2005). "African-American Islamization Reconsidered: Black History Narratives and Muslim Identity." In: *Journal of the American Academy of Religion* 73, pp. 659–684. URL: https://api.semanticscholar.org/CorpusID:145009637 (cit. on p. 39).

D'Oosterlinck, Karel, Semere Kiros Bitew, Brandon Papineau, Christopher Potts, Thomas Demeester, and Chris Develder (Dec. 2023). "CAW-coref: Conjunction-Aware Word-level Coreference Resolution." In: *Proceedings of The Sixth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2023)*. Ed. by Maciej Ogrodniczuk, Vincent Ng, Sameer Pradhan, and Massimo Poesio. Singapore: Association for Computational Linguistics, pp. 8–14. DOI: 10.18653/v1/2023.crac-main.2. URL: https://aclanthology.org/2023.crac-main.2 (cit. on pp. 25, 59).

Darwall, Stephen (1977). "Two Kinds of Respect." In: *Ethics* 88, pp. 36–49. URL: https://api.semanticscholar.org/CorpusID:170842354 (cit. on p. 41).

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li,

Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang (2025). *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. arXiv: 2501.12948 [cs.CL]. URL: https://arxiv.org/abs/2501.12948 (cit. on p. 119).

Delobelle, Pieter, Giuseppe Attanasio, Debora Nozza, Su Lin Blodgett, and Zeerak Talat (Nov. 2024). "Metrics for What, Metrics for Whom: Assessing Actionability of Bias Evaluation Metrics in NLP." In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, pp. 21669–21691. DOI: 10.18653/v1/2024.emnlp-main.1207. URL: https://aclanthology.org/2024.emnlp-main.1207/ (cit. on p. 118).

Deng, Chunyuan, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan (June 2024). "Investigating Data Contamination in Modern Benchmarks for Large Language Models." In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, pp. 8706–8719. DOI: 10.18653/v1/2024.naacl-long.482. URL: https://aclanthology.org/2024.naacl-long.482 (cit. on p. 71).

Dev, Sunipa, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff Phillips, and Kai-Wei Chang (Nov. 2021). "Harms of Gender Exclusivity and Challenges in Non-Binary Representation in Language Technologies." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 1968–1994. DOI: 10.18653/v1/2021.emnlp-main.150. URL: https://aclanthology.org/2021.emnlp-main.150 (cit. on pp. 41, 74).

Devinney, Hannah, Jenny Björklund, and Henrik Björklund (2022). "Theories of Gender in NLP Bias Research." In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22. Seoul, Republic of Korea: Association for Computing Machinery, pp. 2083–2102. ISBN: 9781450393522. DOI: 10.1145/3531146.3534627. URL: https://doi.org/10.1145/3531146.3534627 (cit. on pp. 49, 70).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186.

DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423 (cit. on pp. 28, 81).

Dobrovolskii, Vladimir (Nov. 2021). "Word-Level Coreference Resolution." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 7670–7675. DOI: 10.18653/v1/2021.emnlp-main.605. URL: https://aclanthology.org/2021.emnlp-main.605 (cit. on pp. 25, 59).

Donnellan, Keith S. (1966). "Reference and Definite Descriptions." In: *The Philosophical Review* 75.3, pp. 281–304. ISSN: 00318108, 15581470. URL: http://www.jstor.org/stable/2183143 (visited on 05/20/2025) (cit. on pp. 19, 22).

Dutta Chowdhury, Koel, Rricha Jalota, Cristina España-Bonet, and Josef Genabith (July 2022). "Towards Debiasing Translation Artifacts." In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, pp. 3983–3991. DOI: 10.18653/v1/2022.naacl-main.292. URL: https://aclanthology.org/2022.naacl-main.292 (cit. on p. 70).

Elazar, Yanai, Akshita Bhagia, Ian Helgi Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hannaneh Hajishirzi, Noah A. Smith, and Jesse Dodge (2024). "What's In My Big Data?" In: *The Twelfth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=RvfPnOkPV4 (cit. on pp. 61, 70, 112).

Elazar, Yanai, Hongming Zhang, Yoav Goldberg, and Dan Roth (Nov. 2021). "Back to Square One: Artifact Detection, Training and Commonsense Disentanglement in the Winograd Schema." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 10486–10500. DOI: 10.18653/v1/

2021.emnlp-main.819. URL: https://aclanthology.org/2021.emnlp-main.819 (cit. on pp. 70, 71, 109).

Emelin, Denis and Rico Sennrich (Nov. 2021). "Wino-X: Multilingual Winograd Schemas for Commonsense Reasoning and Coreference Resolution." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 8517–8532. DOI: 10.18653/v1/2021.emnlp-main.670. URL: https://aclanthology.org/2021.emnlp-main.670 (cit. on p. 108).

Erete, Sheena, Aarti Israni, and Tawanna Dillahunt (Apr. 2018). "An intersectional approach to designing in the margins." en. In: *Interactions* 25.3, pp. 66–69. ISSN: 1072-5520, 1558-3449. DOI: 10.1145/3194349 (cit. on p. 48).

Fan, Angela, Mike Lewis, and Yann Dauphin (July 2019). "Strategies for Structuring Story Generation." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 2650–2660. DOI: 10.18653/v1/P19-1254. URL: https://aclanthology.org/P19-1254 (cit. on pp. 107, 119).

Färber, Michael and Lin Ao (2022). "The Microsoft Academic Knowledge Graph enhanced: Author name disambiguation, publication classification, and embeddings." In: *Quantitative Science Studies* 3, pp. 51–98. URL: https://direct.mit.edu/qss/article/3/1/51/109628/The-Microsoft-Academic-Knowledge-Graph-enhanced (cit. on p. 33).

Feinberg, J. (1984). *Harmless Wrongdoing*. Moral Limits of the Criminal Law. Oxford University Press. ISBN: 9780199878574 (cit. on p. 41).

Felkner, Virginia, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May (July 2023). "WinoQueer: A Community-in-the-Loop Benchmark for Anti-LGBTQ+ Bias in Large Language Models." In: *Proceedings of the 61st Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers). Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 9126–9140. DOI: 10.18653/v1/2023.acl-long.507. URL: https://aclanthology.org/2023.acl-long.507 (cit. on p. 109).

Fernandes, Patrick, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig (July 2023). "When Does Translation Require Context? A Data-driven, Multilingual Exploration." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 606–626. DOI: 10.18653/v1/2023.acl-long.36. URL: https://aclanthology.org/2023.acl-long.36 (cit. on pp. 89, 108).

Field, Anjalie, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov (Aug. 2021). "A Survey of Race, Racism, and Anti-Racism in NLP." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online: Association for Computational Linguistics, pp. 1905–1925. DOI: 10.18653/v1/2021.acl-long.149. URL: https://aclanthology.org/2021.acl-long.149 (cit. on pp. 23, 42, 49).

Fleisig, Eve, Su Lin Blodgett, Dan Klein, and Zeerak Talat (June 2024). "The Perspectivist Paradigm Shift: Assumptions and Challenges of Capturing Human Labels." In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, pp. 2279–2292. DOI: 10.18653/v1/2024.naacl-long.126. URL: https://aclanthology.org/2024.naacl-long.126/ (cit. on p. 117).

Floridi, Luciano and Josh Cowls (July 2019). "A Unified Framework of Five Principles for AI in Society." In: *Harvard Data Science Review* 1.1. https://hdsr.mitpress.mit.edu/pub/l0jsh9d1 (cit. on p. 47).

Gaddis, S. Michael (2017a). "How Black Are Lakisha and Jamal? Racial Perceptions from Names Used in Correspondence Audit Studies." In: *Randomized Social Experiments eJournal*. URL: https://sociologicalscience.com/articles-v4-19-469/ (cit. on p. 46).

– (2017b). "Racial/Ethnic Perceptions from Hispanic Names: Selecting Names to Test for Discrimination." In: *Socius* 3. URL: https://journals.sagepub.com/doi/10.1177/2378023117737193 (cit. on p. 46).

Gallegos, Isabel O, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed (2024). "Bias and fairness in large language models: A survey." In: *Computational Linguistics*, pp. 1–79 (cit. on p. 23).

Gan, Yujian, Massimo Poesio, and Juntao Yu (May 2024). "Assessing the Capabilities of Large Language Models in Coreference: An Evaluation." In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue. Torino, Italia: ELRA and ICCL, pp. 1645–1665. URL: https://aclanthology.org/2024.lrec-main.145/ (cit. on p. 25).

García-de-Herreros, Paloma, Vagrant Gautam, Philipp Slusallek, Dietrich Klakow, and Marius Mosbach (June 2024). "What explains the success of cross-modal fine-tuning with ORCA?" In: *Proceedings of the Fifth Workshop on Insights from Negative Results in NLP*. Ed. by Shabnam Tafreshi, Arjun Akula, João Sedoc, Aleksandr Drozd, Anna Rogers, and Anna Rumshisky. Mexico City, Mexico: Association for Computational Linguistics, pp. 8–16. DOI: 10.18653/v1/2024.insights-1.2. URL: https://aclanthology.org/2024.insights-1.2/ (cit. on p. 13).

Gautam, Vagrant, Eileen Bingert, Dawei Zhu, Anne Lauscher, and Dietrich Klakow (Dec. 2024a). "Robust Pronoun Fidelity with English LLMs: Are they Reasoning, Repeating, or Just Biased?" In: *Transactions of the Association for Computational Linguistics* 12, pp. 1755–1779. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00719. URL: https://doi.org/10.1162/tacl_a_00719 (cit. on pp. 5, 73, 92).

Gautam, Vagrant, Julius Steuer, Eileen Bingert, Ray Johns, Anne Lauscher, and Dietrich Klakow (Nov. 2024b). "WinoPron: Revisiting English Winogender Schemas for Consistency, Coverage, and Grammatical Case." In: *Proceedings of The Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*. Ed. by Maciej Ogrodniczuk, Anna Nedoluzhko, Massimo Poesio, Sameer Pradhan, and Vincent Ng. Miami: Association for Computational Linguistics, pp. 52–66. DOI: 10.18653/v1/2024.crac-1.6. URL: https://aclanthology.org/2024.crac-1.6/ (cit. on pp. 5, 52).

Gautam, Vagrant, Arjun Subramonian, Anne Lauscher, and Os Keyes (Aug. 2024c). "Stop! In the Name of Flaws: Disentangling Personal Names and Sociodemographic Attributes in NLP." In: *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Ed. by Agnieszka Faleska, Christine Basta, Marta Costa-jussà, Seraphina Goldfarb-Tarrant, and Debora Nozza. Bangkok, Thailand: Association for Computational Linguistics, pp. 323–337. DOI: 10.18653/v1/2024.gebnlp-1.20. URL: https://aclanthology.org/2024.gebnlp-1.20/ (cit. on pp. 4, 31).

Gautam, Vagrant, Miaoran Zhang, and Dietrich Klakow (Dec. 2023). "A Lightweight Method to Generate Unanswerable Questions in English." In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 7349–7360. DOI: 10.18653/v1/2023.findings-emnlp.491. URL: https://aclanthology.org/2023.findings-emnlp.491/ (cit. on p. 12).

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford (Nov. 2021). "Datasheets for

datasets." In: *Commun. ACM* 64.12, pp. 86–92. ISSN: 0001-0782. DOI: 10.1145/3458723. URL: https://doi.org/10.1145/3458723 (cit. on p. 46).

Genzel, Dmitriy, Jakob Uszkoreit, and Franz Och (Oct. 2010). "Poetic Statistical Machine Translation: Rhyme and Meter." In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Ed. by Hang Li and Lluís Màrquez. Cambridge, MA: Association for Computational Linguistics, pp. 158–166. URL: https://aclanthology.org/D10-1016/ (cit. on p. 24).

Goldfarb-Tarrant, Seraphina, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez (Aug. 2021). "Intrinsic Bias Metrics Do Not Correlate with Application Bias." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online: Association for Computational Linguistics, pp. 1926–1940. DOI: 10.18653/v1/2021.acl-long.150. URL: https://aclanthology.org/2021.acl-long.150 (cit. on pp. 23, 84, 89).

Goldfarb-Tarrant, Seraphina, Eddie Ungless, Esma Balkir, and Su Lin Blodgett (July 2023). "This prompt is measuring <mask>: evaluating bias evaluation in language models." In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 2209–2225. DOI: 10.18653/v1/2023.findings-acl.139. URL: https://aclanthology.org/2023.findings-acl.139/ (cit. on p. 23).

Gonen, Hila and Yoav Goldberg (June 2019). "Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 609–614.

DOI: 10.18653/v1/N19-1061. URL: https://aclanthology.org/N19-1061 (cit. on p. 89).

Green, Ben (2019). "Good isn't good enough." In: *Proceedings of the AI for Social Good workshop at NeurIPS*. Vol. 17 (cit. on p. 47).

– (2021). "The contestation of tech ethics: A sociotechnical approach to technology ethics in practice." In: *Journal of Social Computing* 2.3, pp. 209–225 (cit. on pp. 47, 49).

Greene, Daniel, Anna Lauren Hoffmann, and Luke Stark (2019). "Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning." In: *Proceedings of the 52nd Hawaii International Conference on System Sciences* (cit. on p. 47).

Grosz, Barbara J., Aravind K. Joshi, and Scott Weinstein (1995). "Centering: A Framework for Modeling the Local Coherence of Discourse." In: *Computational Linguistics* 21.2. Ed. by Julia Hirschberg, pp. 203–225. URL: https://aclanthology.org/J95-2003 (cit. on pp. 75, 78, 94).

Gubelmann, Reto and Siegfried Handschuh (May 2022). "Context Matters: A Pragmatic Study of PLMs' Negation Understanding." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 4602–4621. DOI: 10.18653/v1/2022.acl-long.315. URL: https://aclanthology.org/2022.acl-long.315 (cit. on p. 109).

Gundel, Jeanette and Barbara Abbott (Feb. 2019). *The Oxford Handbook of Reference*. Oxford University Press. ISBN: 9780199687305. DOI: 10.1093/oxfordhb/9780199687305.001.0001. URL: https://doi.org/10.1093/oxfordhb/9780199687305.001.0001 (cit. on p. 19).

Gundel, Jeanette K., Nancy Hedberg, and Ron Zacharski (1993). "Cognitive Status and the Form of Referring Expressions in Discourse." In: *Language* 69.2, pp. 274–307.

ISSN: 00978507, 15350665. URL: http://www.jstor.org/stable/416535 (visited on 04/25/2025) (cit. on p. 52).

Haimson, Oliver L and Anna Lauren Hoffmann (2016). "Constructing and enforcing" authentic" identity online: Facebook, real names, and non-normative identities." In: *First Monday* (cit. on p. 43).

Haines, Elizabeth L., Kay Deaux, and Nicole Lofaro (2016). "The Times They Are a-Changing or Are They Not? A Comparison of Gender Stereotypes, 19832014." In: *Psychology of Women Quarterly* 40.3, pp. 353–363. DOI: 10.1177/0361684316634081. eprint: https://doi.org/10.1177/0361684316634081. URL: https://doi.org/10.1177/0361684316634081 (cit. on p. 69).

Han, Chung-hye and Keir Moulton (2022). "Processing bound-variable singular they." In: *Canadian Journal of Linguistics/Revue canadienne de linguistique* 67.3, pp. 267–301. DOI: 10.1017/cnj.2022.30 (cit. on p. 21).

Hanna, Alex, Emily Denton, Andrew Smart, and Jamila Smith-Loud (2020). "Towards a critical race methodology in algorithmic fairness." In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. Barcelona, Spain: Association for Computing Machinery, pp. 501–512. ISBN: 9781450369367. DOI: 10.1145/3351095.3372826. URL: https://doi.org/10.1145/3351095.3372826 (cit. on pp. 39, 40, 49).

Hanna, Alex and Tina M Park (2020). "Against scale: Provocations and resistances to scale thinking." In: *arXiv preprint arXiv:2010.08850* (cit. on pp. 44, 48, 49).

Hansson, Saga, Konstantinos Mavromatakis, Yvonne Adesam, Gerlof Bouma, and Dana Dannélls (May 2021). "The Swedish Winogender Dataset." In: *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. Ed. by Simon Dobnik and Lilja Øvrelid. Reykjavik, Iceland (Online): Linköping University Electronic Press, Sweden, pp. 452–459. URL: https://aclanthology.org/2021.nodalida-main.52 (cit. on p. 53).

Hao, Karen (2019). *In 2020, let's stop AI ethics-washing and actually do something*. https://www.technologyreview.com/2019/12/27/57/ai-ethics-washing-time-to-act/ (cit. on p. 47).

Haraway, Donna (1988). "Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective." In: *Feminist Studies* 14.3, pp. 575–599. ISSN: 00463663. URL: http://www.jstor.org/stable/3178066 (visited on 05/24/2024) (cit. on p. 48).

Herlihy, Christine and Rachel Rudinger (Aug. 2021). "MedNLI Is Not Immune: Natural Language Inference Artifacts in the Clinical Domain." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online: Association for Computational Linguistics, pp. 1020–1027. DOI: 10.18653/v1/2021.acl-short.129. URL: https://aclanthology.org/2021.acl-short.129 (cit. on p. 70).

Hobbs, Jerry R. (1978). "Resolving pronoun references." In: *Lingua* 44.4, pp. 311–338. ISSN: 0024-3841. DOI: https://doi.org/10.1016/0024-3841(78)90006-2. URL: https://www.sciencedirect.com/science/article/pii/0024384178900062 (cit. on p. 1).

Honeywell, Leigh (2016). *neveragain.tech*. https://neveragain.tech/ (cit. on p. 48).

Hossain, Tamanna, Sunipa Dev, and Sameer Singh (July 2023). "MISGENDERED: Limits of Large Language Models in Understanding Pronouns." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 5352–5367. DOI: 10.18653/v1/2023.acl-long.293. URL: https://aclanthology.org/2023.acl-long.293 (cit. on pp. 69, 86, 88, 108, 118).

Hough, Carole (Jan. 2016). *The Oxford Handbook of Names and Naming*. Oxford University Press. ISBN: 9780199656431. DOI: 10.1093/oxfordhb/9780199656431.

001.0001. URL: https://doi.org/10.1093/oxfordhb/9780199656431.001.0001 (cit. on pp. 2, 20, 32, 34, 35).

Hu, Jennifer and Michael Frank (2024). "Auxiliary task demands mask the capabilities of smaller language models." In: *First Conference on Language Modeling*. URL: https://openreview.net/forum?id=U5BUzSn4tD (cit. on pp. 27, 60, 101).

Hu, Jennifer and Roger Levy (Dec. 2023). "Prompting is not a substitute for probability measurements in large language models." In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 5040–5060. DOI: 10.18653/v1/2023.emnlp-main.306. URL: https://aclanthology.org/2023.emnlp-main.306 (cit. on pp. 27, 82, 101).

Ibaraki, Katsumi, Winston Wu, Lu Wang, and Rada Mihalcea (May 2024). "Analyzing Occupational Distribution Representation in Japanese Language Models." In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue. Torino, Italia: ELRA and ICCL, pp. 959–973. URL: https://aclanthology.org/2024.lrec-main.86 (cit. on p. 37).

Iluz, Bar, Yanai Elazar, Asaf Yehudai, and Gabriel Stanovsky (Nov. 2024). "Applying Intrinsic Debiasing on Downstream Tasks: Challenges and Considerations for Machine Translation." In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, pp. 14914–14921. DOI: 10.18653/v1/2024.emnlp-main.829. URL: https://aclanthology.org/2024.emnlp-main.829/ (cit. on p. 118).

Ivanova, Anna A., Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas H. Clark, Carina Kauf, Jennifer Hu, R. T. Pramod, Gabriel Grand, Vivian Paulun, Maria Ryskina, Ekin Akyürek, Ethan Wilcox, Nafisa Rashid, Leshem Choshen, Roger Levy, Evelina Fedorenko, Joshua Tenenbaum, and Jacob Andreas

(2024). *Elements of World Knowledge (EWOK): A cognition-inspired framework for evaluating basic world knowledge in language models*. arXiv: 2405.09605 [cs.CL]. URL: https://arxiv.org/abs/2405.09605 (cit. on p. 27).

Jaccard, Paul (1912). "The Distribution of the Flora in the Alpine Zone." In: *New Phytologist* 11.2, pp. 37–50. DOI: https://doi.org/10.1111/j.1469-8137.1912.tb05611.x. eprint: https://nph.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8137.1912.tb05611.x. URL: https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8137.1912.tb05611.x (cit. on p. 66).

Jacobs, Abigail Z. and Hanna Wallach (Mar. 2021). "Measurement and Fairness." en. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Virtual Event Canada: ACM, pp. 375–385. ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445901. URL: https://dl.acm.org/doi/10.1145/3442188.3445901 (cit. on pp. 23, 39).

Jacovi, Alon, Avi Caciularu, Omer Goldman, and Yoav Goldberg (Dec. 2023). "Stop Uploading Test Data in Plain Text: Practical Strategies for Mitigating Data Contamination by Evaluation Benchmarks." In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 5075–5084. DOI: 10.18653/v1/2023.emnlp-main.308. URL: https://aclanthology.org/2023.emnlp-main.308 (cit. on pp. 71, 90).

Jacovi, Alon and Yoav Goldberg (July 2020). "Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?" In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 4198–4205. DOI: 10.18653/v1/2020.acl-main.386. URL: https://aclanthology.org/2020.acl-main.386/ (cit. on p. 24).

Jain, Sarthak and Byron C. Wallace (June 2019). "Attention is not Explanation." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and*

Short Papers). Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 3543–3556. DOI: 10.18653/v1/N19-1357. URL: https://aclanthology.org/N19-1357/ (cit. on p. 24).

Jeoung, Sullam, Jana Diesner, and Halil Kilicoglu (July 2023). "Examining the Causal Impact of First Names on Language Models: The Case of Social Commonsense Reasoning." In: *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*. Ed. by Anaelia Ovalle, Kai-Wei Chang, Ninareh Mehrabi, Yada Pruksachatkun, Aram Galystan, Jwala Dhamala, Apurv Verma, Trista Cao, Anoop Kumar, and Rahul Gupta. Toronto, Canada: Association for Computational Linguistics, pp. 61–72. DOI: 10.18653/v1/2023.trustnlp-1.7. URL: https://aclanthology.org/2023.trustnlp-1.7 (cit. on p. 37).

Jeshion, Robin (Sept. 2009). "The Significance of Names." en. In: *Mind & Language* 24.4, pp. 370–403. ISSN: 0268-1064, 1468-0017. DOI: 10.1111/j.1468-0017.2009.01367.x (cit. on pp. 32, 34).

Ji, Yangfeng, Chenhao Tan, Sebastian Martschat, Yejin Choi, and Noah A. Smith (Sept. 2017). "Dynamic Entity Representations in Neural Language Models." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Ed. by Martha Palmer, Rebecca Hwa, and Sebastian Riedel. Copenhagen, Denmark: Association for Computational Linguistics, pp. 1830–1839. DOI: 10.18653/v1/D17-1195. URL: https://aclanthology.org/D17-1195 (cit. on p. 107).

Johnson, Austin H (2016). "Transnormativity: A new concept and its validation through documentary film about transgender men." In: *Sociological inquiry* 86.4, pp. 465–491 (cit. on p. 44).

Jones, Ruth and Ann Irvine (Aug. 2013). "The (Un)faithful Machine Translator." In: *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Ed. by Piroska Lendvai and Kalliopi Zervanou. Sofia, Bulgaria: Association for Computational Linguistics, pp. 96–101. URL: https://aclanthology.org/W13-2713/ (cit. on p. 24).

Joshi, Mandar, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy (2020). "SpanBERT: Improving Pre-training by Representing and Predicting Spans." In: *Transactions of the Association for Computational Linguistics* 8. Ed. by Mark Johnson, Brian Roark, and Ani Nenkova, pp. 64–77. DOI: 10.1162/tacl_a_00300. URL: https://aclanthology.org/2020.tacl-1.5 (cit. on pp. 25, 59).

Jourdan, Fanny, Yannick Chevalier, and Cécile Favre (2025). *FairTranslate: An English-French Dataset for Gender Bias Evaluation in Machine Translation by Overcoming Gender Binarity*. arXiv: 2504.15941 [cs.CL]. URL: https://arxiv.org/abs/2504.15941 (cit. on p. 116).

Kaiser, Astrid (2010). "Kevin ist kein Name, sondern eine Diagnose! Der Vorname in der Grundschule–Klangwort, Modewort oder Reizwort." In: *Die Grundschulzeitschrift* 24, pp. 26–29 (cit. on p. 42).

Kankowski, Florian, Torgrim Solstad, Sina Zarriess, and Oliver Bott (2025). *Implicit Causality-biases in humans and LLMs as a tool for benchmarking LLM discourse capabilities*. arXiv: 2501.12980 [cs.CL]. URL: https://arxiv.org/abs/2501.12980 (cit. on p. 65).

Karimi, Fariba, Claudia Wagner, Florian Lemmerich, Mohsen Jadidi, and Markus Strohmaier (2016). "Inferring Gender from Names on the Web: A Comparative Evaluation of Gender Detection Methods." en. In: *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*. Montréal, Québec, Canada: ACM Press, pp. 53–54. ISBN: 978-1-4503-4144-8. DOI: 10.1145/2872518.2889385. URL: http://dl.acm.org/citation.cfm?doid=2872518.2889385 (cit. on pp. 20, 38, 49).

Kauf, Carina, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko, and Anna A. Ivanova (2024). "Comparing Plausibility Estimates in Base and Instruction-Tuned Large Language Models." In: *CoRR* abs/2403.14859v1. DOI: 10.48550/ARXIV.2403.14859. arXiv: 2403.14859. URL: https://arxiv.org/abs/2403.14859v1 (cit. on p. 101).

Kauf, Carina and Anna Ivanova (July 2023). "A Better Way to Do Masked Language Model Scoring." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 925–935. DOI: 10.18653/v1/2023.acl-short.80. URL: https://aclanthology.org/2023.acl-short.80 (cit. on pp. 28, 29, 82).

Kennison, Shelia M. and Jessie L. Trofe (May 2003). "Comprehending Pronouns: A Role for Word-Specific Gender Stereotype Information." In: *Journal of Psycholinguistic Research* 32.3, pp. 355–378. ISSN: 1573-6555. DOI: 10.1023/A:1023599719948 (cit. on pp. 22, 52).

Keyes, Os (July 2017). *Stop Mapping Names to Gender*. URL: https://ironholds.org/names-gender/ (cit. on p. 49).

– (Nov. 2018). "The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition." In: *Proc. ACM Hum.-Comput. Interact.* 2.CSCW. DOI: 10.1145/3274357. URL: https://doi.org/10.1145/3274357 (cit. on pp. 41, 44, 49).

Keyes, Os, Josephine Hoy, and Margaret Drouhard (2019). "Human-Computer Insurrection: Notes on an Anarchist HCI." In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland Uk: Association for Computing Machinery, pp. 1–13. ISBN: 9781450359702. DOI: 10.1145/3290605.3300569. URL: https://doi.org/10.1145/3290605.3300569 (cit. on pp. 44, 48, 49).

Keyes, Os, Chandler May, and Annabelle Carrell (Apr. 2021). "You Keep Using That Word: Ways of Thinking about Gender in Computing Research." In: *Proc. ACM Hum.-Comput. Interact.* 5.CSCW1. DOI: 10.1145/3449113. URL: https://doi.org/10.1145/3449113 (cit. on pp. 39, 40, 49).

King, Eden B., Saaid A. Mendoza, Juan M. Madera, Mikki R. Hebl, and Jennifer L. Knight (2006). "What's in a Name? A Multiracial Investigation of the Role of Occupational Stereotypes in Selection Decisions." In: *Journal of Applied Social*

Psychology 36, pp. 1145–1159. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.0021-9029.2006.00035.x (cit. on p. 46).

Kirkup, L. and R. B. Frenkel (2006). "Systematic errors." In: *An Introduction to Uncertainty in Measurement: Using the GUM (Guide to the Expression of Uncertainty in Measurement)*. Cambridge University Press, pp. 83–96 (cit. on p. 39).

Knowles, Rebecca, Josh Carroll, and Mark Dredze (Nov. 2016). "Demographer: Extremely Simple Name Demographics." In: *Proceedings of the First Workshop on NLP and Computational Social Science*. Ed. by David Bamman, A. Seza Doruöz, Jacob Eisenstein, Dirk Hovy, David Jurgens, Brendan O'Connor, Alice Oh, Oren Tsur, and Svitlana Volkova. Austin, Texas: Association for Computational Linguistics, pp. 108–113. DOI: 10.18653/v1/W16-5614. URL: https://aclanthology.org/W16-5614 (cit. on pp. 33, 37, 44).

Kojima, Takeshi, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa (2022). "Large Language Models are Zero-Shot Reasoners." In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., pp. 22199–22213. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf (cit. on pp. 203, 206).

Konnelly, Lex (Feb. 2021). "Nuance and normativity in trans linguistic research." In: *Journal of Language and Sexuality* 10, pp. 71–82. DOI: 10.1075/jls.00016.kon (cit. on pp. 35, 44).

Kosse, Maureen (Aug. 2021). ""Don't make me pull a Britney": Onomastics and Genericness in the PULL A [PROPER NAME] Construction." In: *Colorado Research in Linguistics* 25. DOI: 10.33011/cril.v25i.1133. URL: https://journals.colorado.edu/index.php/cril/article/view/1133 (cit. on p. 21).

Kotek, Hadas, Rikker Dockum, and David Sun (2023). "Gender bias and stereotypes in Large Language Models." In: *Proceedings of The ACM Collective Intelligence Conference*. CI '23. Delft, Netherlands: Association for Computing Machinery, pp. 12–24.

DOI: 10.1145/3582269.3615599. URL: https://doi.org/10.1145/3582269.3615599 (cit. on pp. 23, 37).

Kozlowski, Diego, Dakota S Murray, Alexis Bell, Will Hulsey, Vincent Larivière, Thema Monroe-White, and Cassidy R Sugimoto (2022). "Avoiding bias when inferring race using name-based approaches." In: *Plos one* 17.3, e0264270 (cit. on pp. 38, 41).

Krahmer, Emiel and Kees van Deemter (Mar. 2012). "Computational Generation of Referring Expressions: A Survey." In: *Computational Linguistics* 38.1, pp. 173–218. DOI: 10.1162/COLI_a_00088. URL: https://aclanthology.org/J12-1006/ (cit. on p. 1).

Krishnan, Maya (Sept. 2020). "Against Interpretability: a Critical Examination of the Interpretability Problem in Machine Learning." In: *Philosophy & Technology* 33.3, pp. 487–502. ISSN: 2210-5441. DOI: 10.1007/s13347-019-00372-9 (cit. on p. 120).

Kurita, Keita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov (Aug. 2019). "Measuring Bias in Contextualized Word Representations." In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. Ed. by Marta R. Costa-jussà, Christian Hardmeier, Will Radford, and Kellie Webster. Florence, Italy: Association for Computational Linguistics, pp. 166–172. DOI: 10.18653/v1/W19-3823. URL: https://aclanthology.org/W19-3823 (cit. on pp. 27, 53, 74, 84, 89).

Lan, Zhenzhong, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut (2020). "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations." In: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. URL: https://openreview.net/forum?id=H1eA7AEtvS (cit. on pp. 28, 81).

Lappin, Shalom and Herbert J. Leass (1994). "An Algorithm for Pronominal Anaphora Resolution." In: *Computational Linguistics* 20.4. Ed. by Julia Hirschberg, pp. 535–561. URL: https://aclanthology.org/J94-4002/ (cit. on p. 1).

Larson, Brian (Apr. 2017). "Gender as a Variable in Natural-Language Processing: Ethical Considerations." In: *Proceedings of the First ACL Workshop on Ethics in*

Natural Language Processing. Ed. by Dirk Hovy, Shannon Spruit, Margaret Mitchell, Emily M. Bender, Michael Strube, and Hanna Wallach. Valencia, Spain: Association for Computational Linguistics, pp. 1–11. DOI: 10.18653/v1/W17-1601. URL: https://aclanthology.org/W17-1601 (cit. on p. 49).

Lassen, Ida Marie S., Mina Almasi, Kenneth Enevoldsen, and Ross Deans Kristensen-McLachlan (May 2023). "Detecting intersectionality in NER models: A data-driven approach." In: *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Ed. by Stefania Degaetano-Ortlieb, Anna Kazantseva, Nils Reiter, and Stan Szpakowicz. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 116–127. DOI: 10.18653/v1/2023.latechclfl-1.13. URL: https://aclanthology.org/2023.latechclfl-1.13 (cit. on p. 38).

Lauscher, Anne, Archie Crowley, and Dirk Hovy (Oct. 2022). "Welcome to the Modern World of Pronouns: Identity-Inclusive Natural Language Processing beyond Gender." In: *Proceedings of the 29th International Conference on Computational Linguistics*. Ed. by Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, pp. 1221–1232. URL: https://aclanthology.org/2022.coling-1.105 (cit. on pp. 21, 47, 61, 69, 77, 112, 113).

Lauscher, Anne, Debora Nozza, Ehm Miltersen, Archie Crowley, and Dirk Hovy (July 2023). "What about "em"? How Commercial Machine Translation Fails to Handle (Neo-)Pronouns." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 377–392. DOI: 10.18653/v1/2023.acl-long.23. URL: https://aclanthology.org/2023.acl-long.23 (cit. on pp. 74, 85).

Lavigne, Avril, Scott Spock, Lauren Christy, and Graham Edwards (2002). *Sk8er Boi*. CD (cit. on p. 96).

Lee, Kenton, Luheng He, Mike Lewis, and Luke Zettlemoyer (Sept. 2017). "End-to-end Neural Coreference Resolution." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Ed. by Martha Palmer, Rebecca Hwa, and Sebastian Riedel. Copenhagen, Denmark: Association for Computational Linguistics, pp. 188–197. DOI: 10.18653/v1/D17-1018. URL: https://aclanthology.org/D17-1018/ (cit. on p. 25).

Levesque, Hector J., Ernest Davis, and Leora Morgenstern (2012). "The Winograd Schema Challenge." In: *Principles of Knowledge Representation and Reasoning: Proceedings of the Thirteenth International Conference, KR 2012, Rome, Italy, June 10-14, 2012*. Ed. by Gerhard Brewka, Thomas Eiter, and Sheila A. McIlraith. AAAI Press. URL: http://www.aaai.org/ocs/index.php/KR/KR12/paper/view/4492 (cit. on pp. 25, 54, 71, 108, 109).

Levy, Mosh, Alon Jacoby, and Yoav Goldberg (Aug. 2024). "Same Task, More Tokens: The Impact of Input Length on the Reasoning Performance of Large Language Models." In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, pp. 15339–15353. URL: https://aclanthology.org/2024.acl-long.818 (cit. on p. 109).

Levy, Shahar, Koren Lazar, and Gabriel Stanovsky (Nov. 2021). "Collecting a Large-Scale Gender Bias Dataset for Coreference Resolution and Machine Translation." In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Ed. by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 2470–2480. DOI: 10.18653/v1/2021.findings-emnlp.211. URL: https://aclanthology.org/2021.findings-emnlp.211 (cit. on pp. 70, 78, 108).

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., pp. 9459–9474. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf (cit. on p. 119).

Li, David C.S. (1997). "Borrowed identity: Signaling involvement with a Western name." In: *Journal of Pragmatics* 28.4. Language and Discourse Issues in Hong Kong's Change of Sovereignity, pp. 489–513. ISSN: 0378-2166. DOI: https://doi.org/10.1016/S0378-2166(97)00032-5. URL: https://www.sciencedirect.com/science/article/pii/S0378216697000325 (cit. on p. 35).

Li, Haizhou, Khe Chai Sim, Jin-Shea Kuo, and Minghui Dong (June 2007). "Semantic Transliteration of Personal Names." In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Ed. by Annie Zaenen and Antal van den Bosch. Prague, Czech Republic: Association for Computational Linguistics, pp. 120–127. URL: https://aclanthology.org/P07-1016 (cit. on pp. 33, 37).

Liao, Thomas, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt (2021). "Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning." In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. URL: https://openreview.net/forum?id=mPducS1MsEK (cit. on p. 39).

Lipton, Zachary C. (June 2018). "The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery." In: *Queue* 16.3, pp. 31–57. ISSN: 1542-7730. DOI: 10.1145/3236386.3241340. URL: https://doi.org/10.1145/3236386.3241340 (cit. on p. 120).

Liu, Nelson F., Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang (2024). "Lost in the Middle: How Language Models Use Long Contexts." In: *Transactions of the Association for Computational Linguistics*

12, pp. 157–173. DOI: 10.1162/tacl_a_00638. URL: https://aclanthology.org/2024.tacl-1.9/ (cit. on p. 109).

Liu, Ruicheng, Rui Mao, Anh Tuan Luu, and Erik Cambria (Dec. 2023). "A brief survey on recent advances in coreference resolution." In: *Artificial Intelligence Review* 56.12, pp. 14439–14481. ISSN: 1573-7462. DOI: 10.1007/s10462-023-10506-3 (cit. on p. 25).

Liu, Tianyu, Yuchen Eleanor Jiang, Nicholas Monath, Ryan Cotterell, and Mrinmaya Sachan (Dec. 2022). "Autoregressive Structured Prediction with Language Models." In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 993–1005. DOI: 10.18653/v1/2022.findings-emnlp.70. URL: https://aclanthology.org/2022.findings-emnlp.70/ (cit. on p. 25).

Liu, Wendy and Derek Ruths (2013). "What's in a Name? Using First Names as Features for Gender Inference in Twitter." In: *AAAI Spring Symposium: Analyzing Microtext* (cit. on p. 36).

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach." In: *CoRR* abs/1907.11692v1. arXiv: 1907.11692. URL: http://arxiv.org/abs/1907.11692v1 (cit. on pp. 28, 81).

Lockhart, Jeffrey W., Molly M. King, and Christin Munsch (Apr. 2023). "Name-based demographic inference and the unequal distribution of misrecognition." en. In: *Nature Human Behaviour* 7.7, pp. 1084–1095. ISSN: 2397-3374. DOI: 10.1038/s41562-023-01587-9 (cit. on pp. 37–42, 48, 49).

Lodge, Cassian (June 2023). *Gender Census 2023: Worldwide Report*. Gender Census. URL: https://www.gendercensus.com/results/2023-worldwide/ (cit. on pp. 77, 112).

Longpre, Shayne, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts (2023). "The Flan Collection: Designing Data and Methods for Effective Instruction Tuning." In: *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*. Ed. by Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett. Vol. 202. Proceedings of Machine Learning Research. PMLR, pp. 22631–22648. URL: https://proceedings.mlr.press/v202/longpre23a.html (cit. on pp. 60, 82).

Lugones, María (2016). "The Coloniality of Gender." In: *The Palgrave Handbook of Gender and Development: Critical Engagements in Feminist Theory and Practice*. London: Palgrave Macmillan UK, pp. 13–33. ISBN: 978-1-137-38273-3. DOI: 10.1007/978-1-137-38273-3_2. URL: https://doi.org/10.1007/978-1-137-38273-3_2 (cit. on p. 43).

Mann, Gideon S. and David Yarowsky (2003). "Unsupervised Personal Name Disambiguation." In: *Conference on Computational Natural Language Learning*. URL: https://api.semanticscholar.org/CorpusID:29759924 (cit. on p. 33).

Marjanovic, Sara Vera, Arkil Patel, Vaibhav Adlakha, Milad Aghajohari, Parishad BehnamGhader, Mehar Bhatia, Aditi Khandelwal, Austin Kraft, Benno Krojer, Xing Han Lù, Nicholas Meade, Dongchan Shin, Amirhossein Kazemnejad, Gaurav Kamath, Marius Mosbach, Karolina Staczak, and Siva Reddy (2025). *DeepSeek-R1 Thoughtology: Let's think about LLM Reasoning*. arXiv: 2504.07128 [cs.CL]. URL: https://arxiv.org/abs/2504.07128 (cit. on p. 119).

Marks, Samuel, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller (2025). "Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models." In: *The Thirteenth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=I4e82CIDxv (cit. on p. 118).

Marx, Gary T. (1999). "What's in a Name? Some Reflections on the Sociology of Anonymity." In: *Inf. Soc.* 15, pp. 99–112. URL: https://www.tandfonline.com/doi/abs/10.1080/019722499128565 (cit. on p. 32).

Maudslay, Rowan Hall, Hila Gonen, Ryan Cotterell, and Simone Teufel (Nov. 2019). "It's All in the Name: Mitigating Gender Bias with Name-Based Counterfactual Data Substitution." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Hong Kong, China: Association for Computational Linguistics, pp. 5267–5275. DOI: 10.18653/v1/D19-1530. URL: https://aclanthology.org/D19-1530 (cit. on p. 33).

May, Chandler, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger (June 2019). "On Measuring Social Biases in Sentence Encoders." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 622–628. DOI: 10.18653/v1/N19-1063. URL: https://aclanthology.org/N19-1063 (cit. on pp. 78, 84).

Maynez, Joshua, Shashi Narayan, Bernd Bohnet, and Ryan McDonald (July 2020). "On Faithfulness and Factuality in Abstractive Summarization." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 1906–1919. DOI: 10.18653/v1/2020.acl-main.173. URL: https://aclanthology.org/2020.acl-main.173/ (cit. on p. 24).

McGaughey, Sebastian (2020). *Understanding Neopronouns*. Citation Key: McGaughey2020. URL: https://glreview.org/article/understanding-neopronouns/ (cit. on p. 21).

McKenzie, Patrick (2010). *Falsehoods Programmers Believe About Names | Kalzumeus Software*. https://www.kalzumeus.com/2010/06/17/falsehoods-programmers-believe-about-names/ (cit. on p. 34).

McLemore, Kevin A. (2018). "A minority stress perspective on transgender individuals' experiences with misgendering." In: *Stigma and Health* 3.1, pp. 53–64. ISSN: 2376-6964(Electronic),2376-6972(Print). DOI: 10.1037/sah0000070 (cit. on p. 74).

Mcnamarah, Chan Tov (2020). "Misgendering." In: *The SAGE Encyclopedia of Trans Studies*. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3683490 (cit. on p. 41).

Meade, Nicholas, Elinor Poole-Dayan, and Siva Reddy (May 2022). "An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 1878–1898. DOI: 10.18653/v1/2022.acl-long.132. URL: https://aclanthology.org/2022.acl-long.132/ (cit. on p. 118).

Meganathan, Ramanujam (2009). "The politics of naming." In: *Contributions to Indian Sociology* 43, pp. 317–324. URL: https://journals.sagepub.com/doi/10.1177/006996670904300205 (cit. on p. 34).

Messick, Samuel (1995). "Standards of Validity and the Validity of Standards in Performance Asessment." In: *Educational Measurement: Issues and Practice* 14.4, pp. 5–8. DOI: https://doi.org/10.1111/j.1745-3992.1995.tb00881.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1745-3992.1995.tb00881.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-3992.1995.tb00881.x (cit. on p. 39).

Mickel, Jennifer (2024). "Racial/Ethnic Categories in AI and Algorithmic Fairness: Why They Matter and What They Represent." In: *ArXiv* abs/2404.06717. URL: https://arxiv.org/abs/2404.06717 (cit. on p. 40).

Mielke, Sabrina (2024). Personal communication (cit. on p. 44).

Mihaljevi, Helena, Marco Tullney, Lucía Santamaría, and Christian Steinfeldt (Aug. 2019). "Reflections on Gender Analyses of Bibliographic Corpora." en. In: *Frontiers*

in Big Data 2, p. 29. ISSN: 2624-909X. DOI: 10.3389/fdata.2019.00029 (cit. on pp. 39, 42, 48, 49).

Miltersen, Ehm Hjorth (May 2016). "Nounself pronouns: 3rd person personal pronouns as identity expression." In: *Journal of Language Works - Sprogvidenskabeligt Studentertidsskrift* 1.1, pp. 37–62. URL: https://tidsskrift.dk/lwo/article/view/23431 (cit. on p. 112).

Minkov, Einat, Richard C. Wang, and William W. Cohen (Oct. 2005). "Extracting Personal Names from Email: Applying Named Entity Recognition to Informal Text." In: *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*. Ed. by Raymond Mooney, Chris Brew, Lee-Feng Chien, and Katrin Kirchhoff. Vancouver, British Columbia, Canada: Association for Computational Linguistics, pp. 443–450. URL: https://aclanthology.org/H05-1056 (cit. on p. 33).

Misra, Kanishka (2022). "minicons: Enabling Flexible Behavioral and Representational Analyses of Transformer Language Models." In: *CoRR* abs/2203.13112v1. DOI: 10.48550/ARXIV.2203.13112. arXiv: 2203.13112. URL: https://arxiv.org/abs/2203.13112v1 (cit. on p. 201).

Mizrahi, Moran, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky (2024). "State of What Art? A Call for Multi-Prompt LLM Evaluation." In: *Transactions of the Association for Computational Linguistics* 12, pp. 933–949. DOI: 10.1162/tacl_a_00681. URL: https://aclanthology.org/2024.tacl-1.52/ (cit. on pp. 27, 30).

Mohammad, Saif M. (July 2020). "Gender Gap in Natural Language Processing Research: Disparities in Authorship and Citations." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 7860–7870. DOI: 10.18653/v1/2020.acl-main.702. URL: https://aclanthology.org/2020.acl-main.702 (cit. on pp. 23, 33, 37, 40, 44).

Morehouse, Kirsten N., Benedek Kurdi, Ece Hakim, and Mahzarin R. Banaji (2022). "When a stereotype dumbfounds: Probing the nature of the surgeon = male belief." In: *Current Research in Ecological and Social Psychology* 3, p. 100044. ISSN: 2666-6227. DOI: https://doi.org/10.1016/j.cresp.2022.100044. URL: https://www.sciencedirect.com/science/article/pii/S2666622722000119 (cit. on pp. 22, 52).

Mosbach, Marius, Vagrant Gautam, Tomás Vergara Browne, Dietrich Klakow, and Mor Geva (Nov. 2024). "From Insights to Actions: The Impact of Interpretability and Analysis Research on NLP." In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, pp. 3078–3105. DOI: 10.18653/v1/2024.emnlp-main.181. URL: https://aclanthology.org/2024.emnlp-main.181/ (cit. on p. 17).

Müller, Mathias, Annette Rios, Elena Voita, and Rico Sennrich (Oct. 2018). "A Large-Scale Test Set for the Evaluation of Context-Aware Pronoun Translation in Neural Machine Translation." In: *Proceedings of the Third Conference on Machine Translation: Research Papers*. Ed. by Ondej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor. Brussels, Belgium: Association for Computational Linguistics, pp. 61–72. DOI: 10.18653/v1/W18-6307. URL: https://aclanthology.org/W18-6307 (cit. on pp. 89, 108).

Munro, Robert and Alex (Carmen) Morrison (Nov. 2020). "Detecting Independent Pronoun Bias with Partially-Synthetic Data Generation." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, pp. 2011–2017. DOI: 10.18653/v1/2020.emnlp-main.157. URL: https://aclanthology.org/2020.emnlp-main.157 (cit. on p. 69).

Nighojkar, Animesh, Antonio Laverghetta Jr., and John Licato (July 2023). "No Strong Feelings One Way or Another: Re-operationalizing Neutrality in Natural Language

Inference." In: *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*. Ed. by Jakob Prange and Annemarie Friedrich. Toronto, Canada: Association for Computational Linguistics, pp. 199–210. DOI: 10.18653/v1/2023.law-1.20. URL: https://aclanthology.org/2023.law-1.20 (cit. on p. 71).

Nissim, Malvina, Rik van Noord, and Rob van der Goot (June 2020). "Fair Is Better than Sensational: Man Is to Doctor as Woman Is to Doctor." In: *Computational Linguistics* 46.2, pp. 487–497. DOI: 10.1162/coli_a_00379. URL: https://aclanthology.org/2020.cl-2.7/ (cit. on p. 87).

Obasi, Sharon N., Richard Mocarski, Natalie Holt, Debra A. Hope, and Nathan Woodruff (Oct. 2019). "Renaming Me: Assessing the Influence of Gender Identity on Name Selection." In: *Names* 67.4, pp. 199–211. ISSN: 1756-2279, 0027-7738. DOI: 10.1080/00277738.2018.1536188 (cit. on p. 35).

Ochs, Elinor (1992). "Indexing gender." In: *Rethinking Context: Language as an Interactive Phenomenon*. Ed. by Alessandro Duranti and Charles Goodwin. Cambridge: Cambridge University Press, pp. 335–358 (cit. on p. 22).

Olteanu, Alexandra, Carlos Castillo, Fernando Diaz, and Emre Kcman (2019). "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries." In: *Frontiers in Big Data* Volume 2 - 2019. ISSN: 2624-909X. DOI: 10.3389/fdata.2019.00013. URL: https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2019.00013 (cit. on p. 117).

Olúwáfei, Ìkòtún Reuben (Jan. 2014). "New Trends in Yorùbá Personal Names among Yorùbá Christians." In: *Linguistik Online* 59.2. DOI: 10.13092/lo.59.1143. URL: https://bop.unibe.ch/linguistik-online/article/view/1143 (cit. on p. 39).

OpenAI, : Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy

Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne,

Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li (2024). *OpenAI o1 System Card*. arXiv: 2412.16720 [cs.AI]. URL: https://arxiv.org/abs/2412.16720 (cit. on p. 119).

Osterhout, Lee, Michael Bersick, and Judith Mclaughlin (May 1997). "Brain potentials reflect violations of gender stereotypes." In: *Memory & Cognition* 25.3, pp. 273–285. ISSN: 1532-5946. DOI: 10.3758/BF03211283 (cit. on p. 22).

Otmazgin, Shon, Arie Cattan, and Yoav Goldberg (May 2023). "LingMess: Linguistically Informed Multi Expert Scorers for Coreference Resolution." In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Ed. by Andreas Vlachos and Isabelle Augenstein. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 2752–2760. DOI: 10.18653/v1/2023.eacl-main.202. URL: https://aclanthology.org/2023.eacl-main.202 (cit. on pp. 25, 59).

Ovalle, Anaelia, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta (2023a). "I'm fully who I am: Towards Centering Transgender and Non-Binary Voices to Measure Biases in Open Language Generation." In: *Proceedings of the 2023 ACM Conference on Fairness,*

Accountability, and Transparency. FAccT '23. New York, NY, USA: Association for Computing Machinery, pp. 1246–1266. DOI: 10.1145/3593013.3594078. URL: https://doi.org/10.1145/3593013.3594078 (cit. on pp. 23, 70, 74, 88, 89, 108, 118).

Ovalle, Anaelia, Ninareh Mehrabi, Palash Goyal, Jwala Dhamala, Kai-Wei Chang, Richard Zemel, Aram Galstyan, Yuval Pinter, and Rahul Gupta (June 2024). "Tokenization Matters: Navigating Data-Scarce Tokenization for Gender Inclusive Language Technologies." In: *Findings of the Association for Computational Linguistics: NAACL 2024*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. Mexico City, Mexico: Association for Computational Linguistics, pp. 1739–1756. DOI: 10.18653/v1/2024.findings-naacl.113. URL: https://aclanthology.org/2024.findings-naacl.113 (cit. on p. 82).

Ovalle, Anaelia, Arjun Subramonian, Vagrant Gautam, Gilbert Gee, and Kai-Wei Chang (2023b). "Factoring the Matrix of Domination: A Critical Review and Reimagination of Intersectionality in AI Fairness." In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '23. Montréal, QC, Canada: Association for Computing Machinery, pp. 496–511. DOI: 10.1145/3600211.3604705. URL: https://doi.org/10.1145/3600211.3604705 (cit. on pp. 8, 48).

Pelczar, M. and J. Rainsbury (1998). "The Indexical Character of Names." In: *Synthese* 114.2, pp. 293–317. DOI: 10.1023/a:1004992629004 (cit. on pp. 20, 21).

Pilcher, Jane (2017). "Names and Doing Gender: How Forenames and Surnames Contribute to Gender Identities, Difference, and Inequalities." In: *Sex Roles* 77, pp. 812–822. URL: https://link.springer.com/article/10.1007/s11199-017-0805-4 (cit. on p. 32).

Poliak, Adam, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme (Oct. 2018). "Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Ed. by Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii. Brussels, Belgium: Association for Computational Linguistics, pp. 67–81.

DOI: 10.18653/v1/D18-1007. URL: https://aclanthology.org/D18-1007/ (cit. on p. 53).

Portes, Jacob, Alexander Trott, Sam Havens, Daniel King, Abhinav Venigalla, Moin Nadeem, Nikhil Sardana, Daya Khudia, and Jonathan Frankle (2023). "MosaicBERT: A Bidirectional Encoder Optimized for Fast Pretraining." In: *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*. Ed. by Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine. URL: http://papers.nips.cc/paper%5C_files/paper/2023/hash/095a6917768712b7ccc61acbeecad1d8-Abstract-Conference.html (cit. on p. 81).

Pradhan, Sameer, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong (Aug. 2013). "Towards Robust Linguistic Analysis using OntoNotes." In: *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*. Ed. by Julia Hockenmaier and Sebastian Riedel. Sofia, Bulgaria: Association for Computational Linguistics, pp. 143–152. URL: https://aclanthology.org/W13-3516 (cit. on p. 108).

Prince, Ellen F (1981). "Toward a taxonomy of given-new information." In: *Radical pragmatics* (cit. on p. 52).

– (1992). "The ZPG Letter: Subjects, Definiteness, and Information-status." en. In: *Pragmatics & Beyond New Series*. Ed. by William C. Mann and Sandra A. Thompson. Vol. 16. Amsterdam: John Benjamins Publishing Company, p. 295. ISBN: 978-90-272-5026-1. DOI: 10.1075/pbns.16.12pri. URL: https://benjamins.com/catalog/pbns.16.12pri (cit. on pp. 22, 52).

Pyykkönen, Pirita, Jukka Hyönä, and Roger P. G. van Gompel (2010). *Activating Gender Stereotypes During Online Spoken Language Processing: Evidence From Visual World Eye Tracking*. 20178931. URL: https://doi.org/10.1027/1618-3169/a000016 (cit. on p. 22).

Qiao, Dan, Yuan Gao, Zheming Yang, Di Yang, Ziheng Wu, Pengcheng Lu, Minghui Qiu, Juntao Li, and Min Zhang (July 2025). "Decoder-Only LLMs can be Masked Auto-Encoders." In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar. Vienna, Austria: Association for Computational Linguistics, pp. 713–723. ISBN: 979-8-89176-252-7. DOI: 10.18653/v1/2025.acl-short.57. URL: https://aclanthology.org/2025.acl-short.57/ (cit. on p. 119).

Raclaw, Joshua (2025). "A trans linguistic perspective on multiple pronoun use in English." In press (cit. on pp. 114, 132).

Radford, Alec, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. (2018). *Improving language understanding by generative pre-training* (cit. on p. 29).

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." In: *Journal of Machine Learning Research* 21.140, pp. 1–67. URL: http://jmlr.org/papers/v21/20-074.html (cit. on p. 30).

Ravichander, Abhilasha, Matt Gardner, and Ana Marasovic (Dec. 2022). "CONDAQA: A Contrastive Reading Comprehension Dataset for Reasoning about Negation." In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 8729–8755. DOI: 10.18653/v1/2022.emnlp-main.598. URL: https://aclanthology.org/2022.emnlp-main.598 (cit. on p. 62).

Rodríguez, Elisa Forcada, Olatz Perez-de-Vinaspre, Jon Ander Campos, Dietrich Klakow, and Vagrant Gautam (Aug. 2025). "Colombian Waitresses y Jueces canadienses: Gender and Country Biases in Occupation Recommendations from LLMs." In: *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Ed. by Agnieszka Faleska, Christine Basta, Marta Costa-jussà, Karolina

Staczak, and Debora Nozza. Vienna, Austria: Association for Computational Linguistics, pp. 182–194. ISBN: 979-8-89176-277-0. DOI: 10.18653/v1/2025.gebnlp-1.18. URL: https://aclanthology.org/2025.gebnlp-1.18/ (cit. on p. 9).

Romanov, Alexey, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Kalai (June 2019). "What's in a Name? Reducing Bias in Bios without Access to Protected Attributes." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4187–4195. DOI: 10.18653/v1/N19-1424. URL: https://aclanthology.org/N19-1424 (cit. on p. 33).

Rudes, Blair A. and Bernard Healy (1979). "Is She for Real?: The Concepts of Femaleness and Maleness in the Gay World." In: *Boas, Sapir and Whorf Revisited*. Ed. by Madeleine Mathiot. Berlin, Boston: De Gruyter Mouton, pp. 49–62. ISBN: 978-3-11-080415-7. DOI: doi:10.1515/9783110804157-003. URL: https://doi.org/10.1515/9783110804157-003 (cit. on p. 22).

Rudinger, Rachel, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme (June 2018). "Gender Bias in Coreference Resolution." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 8–14. DOI: 10.18653/v1/N18-2002. URL: https://aclanthology.org/N18-2002 (cit. on pp. 22, 23, 25, 51, 52, 54, 55, 69–71, 77, 108, 197).

Russell, Bertrand (Jan. 1905). "On Denoting." In: *Mind* XIV.4, pp. 479–493. ISSN: 0026-4423. DOI: 10.1093/mind/XIV.4.479. eprint: https://academic.oup.com/mind/article-pdf/XIV/4/479/9872659/479.pdf. URL: https://doi.org/10.1093/mind/XIV.4.479 (cit. on p. 19).

Salazar, Julian, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff (July 2020). "Masked Language Model Scoring." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online: Association for Computational Linguistics, pp. 2699–2712. DOI: 10.18653/v1/2020.acl-main.240. URL: https://aclanthology.org/2020.acl-main.240 (cit. on pp. 28, 82).

Sälevä, Jonne and Constantine Lignos (May 2024). "ParaNames 1.0: Creating an Entity Name Corpus for 400+ Languages Using Wikidata." In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue. Torino, Italia: ELRA and ICCL, pp. 12599–12610. URL: https://aclanthology.org/2024.lrec-main.1103 (cit. on p. 33).

Sambasivan, Nithya, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo (2021). ""Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI." In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. Yokohama, Japan: Association for Computing Machinery. ISBN: 9781450380966. DOI: 10.1145/3411764.3445518. URL: https://doi.org/10.1145/3411764.3445518 (cit. on p. 70).

Sandoval, Sandra, Jieyu Zhao, Marine Carpuat, and Hal Daumé III (Dec. 2023). "A Rose by Any Other Name would not Smell as Sweet: Social Bias in Names Mistranslation." In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 3933–3945. DOI: 10.18653/v1/2023.emnlp-main.239. URL: https://aclanthology.org/2023.emnlp-main.239 (cit. on pp. 33, 38, 40, 45).

Sap, Maarten, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith (July 2022). "Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection." In: *Proceedings of the 2022 Conference*

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, pp. 5884–5906. DOI: 10.18653/v1/2022.naacl-main.431. URL: https://aclanthology.org/2022.naacl-main.431/ (cit. on p. 117).

Saunders, Danielle and Katrina Olsen (June 2023). "Gender, names and other mysteries: Towards the ambiguous for gender-inclusive translation." In: *Proceedings of the First Workshop on Gender-Inclusive Translation Technologies*. Ed. by Eva Vanmassenhove, Beatrice Savoldi, Luisa Bentivogli, Joke Daems, and Janiça Hackenbuchner. Tampere, Finland: European Association for Machine Translation, pp. 85–93. URL: https://aclanthology.org/2023.gitt-1.8 (cit. on p. 38).

Scheuerman, Morgan Klaus and Jed R. Brubaker (May 2024). "Products of Positionality: How Tech Workers Shape Identity Concepts in Computer Vision." en. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Honolulu HI USA: ACM, pp. 1–18. DOI: 10.1145/3613904.3641890. URL: https://dl.acm.org/doi/10.1145/3613904.3641890 (cit. on p. 40).

Scheuerman, Morgan Klaus, Jacob M. Paul, and Jed R. Brubaker (Nov. 2019). "How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services." In: *Proc. ACM Hum.-Comput. Interact.* 3.CSCW. DOI: 10.1145/3359246. URL: https://doi.org/10.1145/3359246 (cit. on p. 49).

Scheuerman, Morgan Klaus, Katta Spiel, Oliver L Haimson, Foad Hamidi, and Stacy M Branham (2020a). "HCI guidelines for gender equity and inclusivity." In: URL: https://www.morgan-klaus.com/gender-guidelines.html (cit. on p. 49).

Scheuerman, Morgan Klaus, Kandrea Wade, Caitlin Lustig, and Jed R. Brubaker (May 2020b). "How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis." en. In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW1, pp. 1–35. ISSN: 2573-0142. DOI: 10.1145/3392866 (cit. on p. 40).

Sclar, Melanie, Yejin Choi, Yulia Tsvetkov, and Alane Suhr (2024). "Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting." In: *The Twelfth International Conference on Learning Representations*. URL: https://openreview.net/forum?id=RIu5lyNXjT (cit. on pp. 27, 30, 82).

Selbst, Andrew D., Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi (2019). "Fairness and Abstraction in Sociotechnical Systems." In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT* '19. Atlanta, GA, USA: Association for Computing Machinery, pp. 59–68. ISBN: 9781450361255. DOI: 10.1145/3287560.3287598. URL: https://doi.org/10.1145/3287560.3287598 (cit. on p. 23).

Selvam, Nikil, Sunipa Dev, Daniel Khashabi, Tushar Khot, and Kai-Wei Chang (July 2023). "The Tail Wagging the Dog: Dataset Construction Biases of Social Bias Benchmarks." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 1373–1386. DOI: 10.18653/v1/2023.acl-short.118. URL: https://aclanthology.org/2023.acl-short.118 (cit. on pp. 23, 71, 72, 84).

Sennrich, Rico, Barry Haddow, and Alexandra Birch (Aug. 2016). "Neural Machine Translation of Rare Words with Subword Units." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Katrin Erk and Noah A. Smith. Berlin, Germany: Association for Computational Linguistics, pp. 1715–1725. DOI: 10.18653/v1/P16-1162. URL: https://aclanthology.org/P16-1162 (cit. on p. 33).

Seshadri, Preethi, Pouya Pezeshkpour, and Sameer Singh (2022). "Quantifying Social Biases Using Templates is Unreliable." In: *CoRR* abs/2210.04337. DOI: 10.48550/ARXIV.2210.04337. arXiv: 2210.04337. URL: https://doi.org/10.48550/arXiv.2210.04337 (cit. on pp. 23, 72, 84, 89).

Shaalan, Khaled and Hafsa Raza (June 2007). "Person Name Entity Recognition for Arabic." In: *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*. Ed. by Violetta Cavalli-Sforza and Imed Zitouni. Prague, Czech Republic: Association for Computational Linguistics, pp. 17–24. URL: https://aclanthology.org/W07-0803 (cit. on p. 42).

Sharma, Shanya, Manan Dey, and Koustuv Sinha (Dec. 2022). "How sensitive are translation systems to extra contexts? Mitigating gender bias in Neural Machine Translation models through relevant contexts." In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 1968–1984. DOI: 10.18653/v1/2022.findings-emnlp.143. URL: https://aclanthology.org/2022.findings-emnlp.143 (cit. on pp. 89, 108).

Shwartz, Vered, Rachel Rudinger, and Oyvind Tafjord (Nov. 2020). ""You are grounded!": Latent Name Artifacts in Pre-trained Language Models." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, pp. 6850–6861. DOI: 10.18653/v1/2020.emnlp-main.556. URL: https://aclanthology.org/2020.emnlp-main.556 (cit. on pp. 45, 70).

Sieker, Judith, Oliver Bott, Torgrim Solstad, and Sina ZarrieSS (Sept. 2023). "Beyond the Bias: Unveiling the Quality of Implicit Causality Prompt Continuations in Language Models." In: *Proceedings of the 16th International Natural Language Generation Conference*. Ed. by C. Maria Keet, Hung-Yi Lee, and Sina ZarrieSS. Prague, Czechia: Association for Computational Linguistics, pp. 206–220. DOI: 10.18653/v1/2023.inlg-main.15. URL: https://aclanthology.org/2023.inlg-main.15/ (cit. on p. 64).

Silverstein, Michael (1985). "10 - Language and the Culture of Gender: At the Intersection of Structure, Usage, and Ideology." In: *Semiotic Mediation*. Ed. by Elizabeth Mertz and Richard J. Parmentier. San Diego: Academic Press, pp. 219–259. ISBN: 978-0-12-491280-9. DOI: https://doi.org/10.1016/B978-0-12-

491280-9.50016-9. URL: https://www.sciencedirect.com/science/article/pii/B9780124912809500169 (cit. on p. 74).

Sinha, Koustuv, Jon Gauthier, Aaron Mueller, Kanishka Misra, Keren Fuentes, Roger Levy, and Adina Williams (July 2023). "Language model acceptability judgements are not always robust to context." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 6043–6063. DOI: 10.18653/v1/2023.acl-long.333. URL: https://aclanthology.org/2023.acl-long.333 (cit. on p. 109).

Smith, Eric Michael and Adina Williams (2021). "Hi, my name is Martha: Using names to measure and mitigate bias in generative dialogue models." In: *ArXiv* abs/2109.03300. URL: https://arxiv.org/abs/2109.03300 (cit. on p. 37).

Speer, Robyn (2021). *Google Scholar has failed us.* URL: https://scholar.hasfailed.us/ (cit. on p. 43).

Srikanth, Neha and Rachel Rudinger (July 2022). "Partial-input baselines show that NLI models can ignore context, but they don't." In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, pp. 4753–4763. DOI: 10.18653/v1/2022.naacl-main.350. URL: https://aclanthology.org/2022.naacl-main.350 (cit. on p. 109).

Stanovsky, Gabriel, Noah A. Smith, and Luke Zettlemoyer (July 2019). "Evaluating Gender Bias in Machine Translation." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 1679–1684. DOI: 10.18653/v1/P19-1164. URL: https://aclanthology.org/P19-1164 (cit. on p. 53).

Stechly, Kaya, Karthik Valmeekam, Atharva Gundawar, Vardhan Palod, and Subbarao Kambhampati (2025). *Beyond Semantics: The Unreasonable Effectiveness of Reasonless Intermediate Tokens*. arXiv: 2505.13775 [cs.LG]. URL: https://arxiv.org/abs/2505.13775 (cit. on p. 119).

Steidl, Christina R. and Regina Werum (Aug. 2019). "If all you have is a hammer, everything looks like a nail: Operationalization matters." en. In: *Sociology Compass* 13.8, e12727. ISSN: 1751-9020, 1751-9020. DOI: 10.1111/soc4.12727 (cit. on p. 40).

Strauss, Anselm L (2017). *Mirrors and masks: The search for identity*. Routledge (cit. on p. 32).

Strawson, P. F. (1950). "On Referring." In: *Mind* 59.235, pp. 320–344. ISSN: 00264423, 14602113. URL: http://www.jstor.org/stable/2251176 (visited on 04/07/2025) (cit. on pp. 19, 32).

Stryker, Susan (2017). *Transgender History: The Roots of Today's Revolution*. 2nd. Seal Press (cit. on p. 74).

Subramonian, Arjun, Vagrant Gautam, Dietrich Klakow, and Zeerak Talat (Nov. 2024). "Understanding Democratization in NLP and ML Research." In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, pp. 3151–3166. DOI: 10.18653/v1/2024.emnlp-main.184. URL: https://aclanthology.org/2024.emnlp-main.184/ (cit. on p. 16).

Subramonian, Arjun, Vagrant Gautam, Preethi Seshadri, Dietrich Klakow, Kai-Wei Chang, and Yizhou Sun (2025). "Agree to Disagree? A Meta-Evaluation of LLM Misgendering." In: *Second Conference on Language Modeling*. URL: https://openreview.net/forum?id=vgmiRvpCLA (cit. on pp. 15, 89, 112, 118).

Subramonian, Arjun, Xingdi Yuan, Hal Daumé III, and Su Lin Blodgett (July 2023). "It Takes Two to Tango: Navigating Conceptualizations of NLP Tasks and Measurements of Performance." In: *Findings of the Association for Computational Linguistics: ACL*

2023. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 3234–3279. DOI: 10.18653/v1/2023.findings-acl.202. URL: https://aclanthology.org/2023.findings-acl.202 (cit. on pp. 71, 117).

Sue, Christina A. and Edward E. Telles (Mar. 2007). "Assimilation and Gender in Naming." en. In: *American Journal of Sociology* 112.5, pp. 1383–1415. ISSN: 0002-9602, 1537-5390. DOI: 10.1086/511801 (cit. on p. 35).

Sukthanker, Rhea, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu (2020). "Anaphora and coreference resolution: A review." In: *Information Fusion* 59, pp. 139–162. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2020.01.010. URL: https://www.sciencedirect.com/science/article/pii/S1566253519303677 (cit. on p. 1).

Tal, Yarden, Inbal Magar, and Roy Schwartz (July 2022). "Fewer Errors, but More Stereotypes? The Effect of Model Size on Gender Bias." In: *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. Ed. by Christian Hardmeier, Christine Basta, Marta R. Costa-jussà, Gabriel Stanovsky, and Hila Gonen. Seattle, Washington: Association for Computational Linguistics, pp. 112–120. DOI: 10.18653/v1/2022.gebnlp-1.13. URL: https://aclanthology.org/2022.gebnlp-1.13 (cit. on pp. 74, 89, 105).

Talat, Zeerak, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams (July 2022). "On the Machine Learning of Ethical Judgments from Natural Language." In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, pp. 769–779. DOI: 10.18653/v1/2022.naacl-main.56. URL: https://aclanthology.org/2022.naacl-main.56 (cit. on p. 48).

Tatman, Rachael (2020). *What I Won't Build*. https://www.rctatman.com/talks/what-i-wont-build (cit. on p. 48).

Tay, Yi, Mostafa Dehghani, Samira Abnar, Hyung Chung, William Fedus, Jinfeng Rao, Sharan Narang, Vinh Tran, Dani Yogatama, and Donald Metzler (Dec. 2023). "Scaling Laws vs Model Architectures: How does Inductive Bias Influence Scaling?" In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 12342–12364. DOI: 10.18653/v1/2023.findings-emnlp.825. URL: https://aclanthology.org/2023.findings-emnlp.825 (cit. on p. 85).

Te Tari Taiwhenua, Department of Internal Affairs | (2021). *Press Releases - dia.govt.nz — dia.govt.nz*. https://www.dia.govt.nz/press.nsf/d77da9b523f12931cc256ac5000d19b6/d1288ac08d7758c2cc25838200107411!OpenDocument. URL: https://www.dia.govt.nz/press.nsf/d77da9b523f12931cc256ac5000d19b6/d1288ac08d7758c2cc25838200107411!OpenDocument (cit. on p. 43).

Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom (2023). "Llama 2: Open Foundation and Fine-Tuned Chat Models." In: *CoRR* abs/2307.09288v2. DOI: 10.48550/ARXIV.2307.09288. arXiv: 2307.09288. URL: https://arxiv.org/abs/2307.09288v2 (cit. on pp. 29, 81).

Tovar, Sol (2025). "Understanding (mis)gender(ing) and pronouns from a politeness theory standpoint." In: *Of gender bias and gender fairness*. Ed. by Dominic Schmitz, Simon David Stein, and Viktoria Schneider. Berlin, Boston: düsseldorf university press, pp. 83–94. ISBN: 978-3-11-138869-4. DOI: doi:10.1515/9783111388694-005. URL: https://doi.org/10.1515/9783111388694-005 (cit. on p. 22).

Trichelair, Paul, Ali Emami, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung (Nov. 2019). "How Reasonable are Common-Sense Reasoning Tasks: A Case-Study on the Winograd Schema Challenge and SWAG." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Hong Kong, China: Association for Computational Linguistics, pp. 3382–3387. DOI: 10.18653/v1/D19-1335. URL: https://aclanthology.org/D19-1335 (cit. on p. 109).

Turpin, Miles, Julian Michael, Ethan Perez, and Samuel R. Bowman (2023). "Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting." In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. NIPS '23. New Orleans, LA, USA: Curran Associates Inc. (cit. on p. 24).

Tzioumis, Konstantinos (2018). "Demographic aspects of first names." In: *Scientific Data* 5. URL: https://www.nature.com/articles/sdata201825.pdf (cit. on p. 49).

U.S. Census (2020). *First name frequency by gender*. URL: https://www.census.gov/genealogy/names/names_files.html (cit. on p. 37).

U.S. Social Security Administration (2023). *Top 10 Baby Names of 2023*. URL: https://www.ssa.gov/oact/babynames/ (cit. on p. 37).

United States National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1978). *The Belmont report: ethical principles and guidelines for the protection of human subjects of research*. Vol. 2. United States National

Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (cit. on p. 48).

Van Buskirk, Ian, Aaron Clauset, and Daniel B. Larremore (June 2023). "An Open-Source Cultural Consensus Approach to Name-Based Gender Classification." en. In: *Proceedings of the International AAAI Conference on Web and Social Media* 17, pp. 866–877. ISSN: 2334-0770, 2162-3449. DOI: 10.1609/icwsm.v17i1.22195 (cit. on pp. 38, 40, 41, 44, 49).

Vassimon Manela, Daniel de, David Errington, Thomas Fisher, Boris van Breugel, and Pasquale Minervini (Apr. 2021). "Stereotype and Skew: Quantifying Gender Bias in Pre-trained and Fine-tuned Language Models." In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Ed. by Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty. Online: Association for Computational Linguistics, pp. 2232–2242. DOI: 10.18653/v1/2021.eacl-main.190. URL: https://aclanthology.org/2021.eacl-main.190 (cit. on p. 89).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need." In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (cit. on p. 27).

Vida, Karina, Judith Simon, and Anne Lauscher (Dec. 2023). "Values, Ethics, Morals? On the Use of Moral Concepts in NLP Research." In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 5534–5554. DOI: 10.18653/v1/2023.findings-emnlp.368. URL: https://aclanthology.org/2023.findings-emnlp.368 (cit. on pp. 47, 48).

Vinyals, Oriol, Meire Fortunato, and Navdeep Jaitly (2015). "Pointer Networks." In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett. Vol. 28. Curran Associates, Inc. URL: https:

//proceedings.neurips.cc/paper_files/paper/2015/file/29921001f2f04bd3baee84a12e98098f-Paper.pdf (cit. on p. 119).

Vogel, Adam and Dan Jurafsky (July 2012). "He Said, She Said: Gender in the ACL Anthology." In: *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*. Ed. by Rafael E. Banchs. Jeju Island, Korea: Association for Computational Linguistics, pp. 33–41. URL: https://aclanthology.org/W12-3204 (cit. on pp. 1, 23, 33, 37, 44).

Voita, Elena, Pavel Serdyukov, Rico Sennrich, and Ivan Titov (July 2018). "Context-Aware Neural Machine Translation Learns Anaphora Resolution." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Iryna Gurevych and Yusuke Miyao. Melbourne, Australia: Association for Computational Linguistics, pp. 1264–1274. DOI: 10.18653/v1/P18-1117. URL: https://aclanthology.org/P18-1117 (cit. on pp. 89, 108).

Wadhawan, Subhah (2022). "Let the Machines Do the Dirty Work: Social Media, Machine Learning Technology and the Iteration of Racialized Surveillance." In: *Canadian Journal of Law and Technology* 20.1, p. 1 (cit. on p. 41).

Wahle, Jan Philip, Terry Ruas, Mohamed Abdalla, Bela Gipp, and Saif Mohammad (Dec. 2023). "We are Who We Cite: Bridges of Influence Between Natural Language Processing and Other Academic Fields." In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 12896–12913. DOI: 10.18653/v1/2023.emnlp-main.797. URL: https://aclanthology.org/2023.emnlp-main.797 (cit. on p. 47).

Waldis, Andreas, Joel Birrer, Anne Lauscher, and Iryna Gurevych (Nov. 2024). "The Lou Dataset - Exploring the Impact of Gender-Fair Language in German Text Classification." In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, pp. 10604–10624.

DOI: 10.18653/v1/2024.emnlp-main.592. URL: https://aclanthology.org/2024.emnlp-main.592/ (cit. on p. 116).

Waldis, Andreas, Vagrant Gautam, Anne Lauscher, Dietrich Klakow, and Iryna Gurevych (2025). *Aligned Probing: Relating Toxic Behavior and Model Internals*. arXiv: 2503.13390 [cs.CL]. URL: https://arxiv.org/abs/2503.13390 (cit. on p. 13).

Wang, Alex and Kyunghyun Cho (June 2019). "BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model." In: *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*. Ed. by Antoine Bosselut, Asli Celikyilmaz, Marjan Ghazvininejad, Srinivasan Iyer, Urvashi Khandelwal, Hannah Rashkin, and Thomas Wolf. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 30–36. DOI: 10.18653/v1/W19-2304. URL: https://aclanthology.org/W19-2304/ (cit. on p. 28).

Wang, Jun, Benjamin Rubinstein, and Trevor Cohn (May 2022). "Measuring and Mitigating Name Biases in Neural Machine Translation." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 2576–2590. DOI: 10.18653/v1/2022.acl-long.184. URL: https://aclanthology.org/2022.acl-long.184 (cit. on pp. 33, 37).

Wang, Liwen, Yuanmeng Yan, Keqing He, Yanan Wu, and Weiran Xu (June 2021). "Dynamically Disentangling Social Bias from Task-Oriented Representations with Adversarial Attack." In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou. Online: Association for Computational Linguistics, pp. 3740–3750. DOI: 10.18653/v1/2021.naacl-main.293. URL: https://aclanthology.org/2021.naacl-main.293/ (cit. on p. 118).

Warstadt, Alex, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman (2020). "BLiMP: The Benchmark of Linguistic Minimal Pairs for English." In: *Transactions of the Association for Computational Linguistics* 8. Ed. by Mark Johnson, Brian Roark, and Ani Nenkova, pp. 377–392. DOI: 10.1162/tacl_a_00321. URL: https://aclanthology.org/2020.tacl-1.25/ (cit. on pp. 27, 29).

Waseem, Zeerak, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein (2021). *Disembodied Machine Learning: On the Illusion of Objectivity in NLP*. arXiv: 2101.11974 [cs.AI] (cit. on p. 48).

Webster, Kellie, Marta Recasens, Vera Axelrod, and Jason Baldridge (2018). "Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns." In: *Transactions of the Association for Computational Linguistics* 6. Ed. by Lillian Lee, Mark Johnson, Kristina Toutanova, and Brian Roark, pp. 605–617. DOI: 10.1162/tacl_a_00240. URL: https://aclanthology.org/Q18-1042/ (cit. on pp. 70, 108).

Webster, Kellie, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov (2021). *Measuring and Reducing Gendered Correlations in Pre-trained Models*. arXiv: 2010.06032 [cs.CL] (cit. on p. 33).

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou (2022). "Chain-of-thought prompting elicits reasoning in large language models." In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS '22. New Orleans, LA, USA: Curran Associates Inc. ISBN: 9781713871088 (cit. on p. 28).

Weinberg, Lindsay (May 2022). "Rethinking Fairness: An Interdisciplinary Survey of Critiques of Hegemonic ML Fairness Approaches." en. In: *Journal of Artificial Intelligence Research* 74, pp. 75–109. ISSN: 1076-9757. DOI: 10.1613/jair.1.13196 (cit. on p. 23).

Weindling, Paul (2001). "The Origins of Informed Consent: The International Scientific Commission on Medical War Crimes, and the Nuremberg Code." In: *Bulletin of*

the History of Medicine 75.1, pp. 37–71. ISSN: 00075140, 10863176. URL: http://www.jstor.org/stable/44445555 (visited on 05/27/2024) (cit. on p. 48).

Weitman, Sasha (Sept. 1981). "Some Methodological Issues in Quantitative Onomastics." In: *Names* 29.3, pp. 181–196. ISSN: 1756-2279, 0027-7738. DOI: 10.1179/nam.1981.29.3.181 (cit. on pp. 36, 46).

Wiegreffe, Sarah and Yuval Pinter (Nov. 2019). "Attention is not not Explanation." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Hong Kong, China: Association for Computational Linguistics, pp. 11–20. DOI: 10.18653/v1/D19-1002. URL: https://aclanthology.org/D19-1002/ (cit. on p. 24).

Williams, John F. (Nov. 1924). "The Geneva Protocol of 1924 for the Pacific Settlement of International Disputes1." In: *Journal of the British Institute of International Affairs* 3.6, pp. 288–304. ISSN: 1473-7981. DOI: 10.2307/3014555. eprint: https://academic.oup.com/ia/article-pdf/3/6/288/13155361/ia-3-6-288.pdf. URL: https://doi.org/10.2307/3014555 (cit. on p. 48).

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush (Oct. 2020). "Transformers: State-of-the-Art Natural Language Processing." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Ed. by Qun Liu and David Schlangen. Online: Association for Computational Linguistics, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. URL: https://aclanthology.org/2020.emnlp-demos.6 (cit. on p. 201).

Wolfe, Robert and Aylin Caliskan (Nov. 2021). "Low Frequency Names Exhibit Bias and Overfitting in Contextualizing Language Models." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-

Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 518–532. DOI: 10.18653/v1/2021.emnlp-main.41. URL: https://aclanthology.org/2021.emnlp-main.41 (cit. on p. 45).

Xu, Sheng, Peifeng Li, and Qiaoming Zhu (Dec. 2023). "CorefPrompt: Prompt-based Event Coreference Resolution by Measuring Event Type and Argument Compatibilities." In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 15440–15452. DOI: 10.18653/v1/2023.emnlp-main.954. URL: https://aclanthology.org/2023.emnlp-main.954/ (cit. on p. 25).

Yang, Zichao, Phil Blunsom, Chris Dyer, and Wang Ling (Sept. 2017). "Reference-Aware Language Models." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Ed. by Martha Palmer, Rebecca Hwa, and Sebastian Riedel. Copenhagen, Denmark: Association for Computational Linguistics, pp. 1850–1859. DOI: 10.18653/v1/D17-1197. URL: https://aclanthology.org/D17-1197/ (cit. on p. 119).

Zhang, Justine (2024). *On the world-making work of artificial language understanding*. The Work of AI: Mapping Human Labour in the AI Pipeline workshop, CSCW 2024. URL: https://tisjune.github.io/papers/cscw_2024_worldbuilding.pdf (cit. on p. 116).

Zhang, Miaoran, Vagrant Gautam, Mingyang Wang, Jesujoba Alabi, Xiaoyu Shen, Dietrich Klakow, and Marius Mosbach (Aug. 2024). "The Impact of Demonstrations on Multilingual In-Context Learning: A Multidimensional Analysis." In: *Findings of the Association for Computational Linguistics: ACL 2024*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, pp. 7342–7371. DOI: 10.18653/v1/2024.findings-acl.438. URL: https://aclanthology.org/2024.findings-acl.438/ (cit. on p. 10).

Zhang, Susan, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer (2022). "OPT: Open Pre-trained Transformer Language Models." In: *CoRR* abs/2205.01068v4. DOI: 10.48550/ARXIV.2205.01068. arXiv: 2205.01068. URL: https://arxiv.org/abs/2205.01068v4 (cit. on pp. 29, 81).

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang (June 2019). "Gender Bias in Contextualized Word Embeddings." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 629–634. DOI: 10.18653/v1/N19-1064. URL: https://aclanthology.org/N19-1064 (cit. on p. 89).

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang (June 2018). "Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Ed. by Marilyn Walker, Heng Ji, and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, pp. 15–20. DOI: 10.18653/v1/N18-2003. URL: https://aclanthology.org/N18-2003 (cit. on pp. 22, 70, 108).

Zhou, Yongchao, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba (2023). "Large Language Models are Human-Level Prompt Engineers." In: *The Eleventh International Conference on Learning Representations*. URL: https://openreview.net/forum?id=92gvk82DE- (cit. on pp. 203, 207).

A

Data and Annotation Details

A.1 List of Occupations

The occupations along with their respective participants in parentheses are listed below in alphabetical order. This list is identical to the occupations and participants in Rudinger et al. (2018), except that we pair examiner with intern rather than victim:

accountant (taxpayer), administrator (undergraduate), advisor (advisee), appraiser (buyer), architect (student), auditor (taxpayer), baker (customer), bartender (customer), broker (client), carpenter (onlooker), cashier (customer), chef (guest), chemist (visitor), clerk (customer), counselor (patient), dietitian (client), dispatcher (bystander), doctor (patient), educator (student), electrician (homeowner), engineer (client), examiner (intern), firefighter (child), hairdresser (client), hygienist (patient), inspector (homeowner), instructor (student), investigator (witness), janitor (child), lawyer (witness), librarian (child), machinist (child), manager (customer), mechanic (customer) nurse (patient), nutritionist (patient), officer (protester), painter (customer), paralegal (client), paramedic (passenger), pathologist (victim), pharmacist (patient), physician (patient), planner (resident), plumber (homeowner), practitioner (patient), programmer (student), psychologist (patient), receptionist (visitor), salesperson (customer), scientist (undergraduate), secretary (visitor), specialist (patient), supervisor (employee), surgeon (child), teacher (student), technician (customer), therapist (teenager), veterinarian (owner), worker (pedestrian)

A.2 Annotator Demographics

All three annotators (two authors and an additional annotator) are fluent English speakers. The two authors who create and validate templates have linguistic training at the undergraduate level. One author and one annotator have experience with using singular *they* and neopronouns, while the other author has prior exposure to singular *they* but not the neopronoun *xe*.

A.3 Annotation Instructions

A.3.1 Task 1 Description

Together with this annotation protocol, you have received a link to a Google Sheet. The sheet contains 2 data columns and 2 task columns of randomized data. The data columns consist of

- Sentences which you are asked to annotate for grammaticality; and
- Questions about pronouns in the sentence, which you are asked to answer

Please be precise in your assignments and do not reorder the data. The columns have built-in data validation and we will perform further tests to check for consistent annotation.

A.3.1.1 Grammaticality

In the Grammatical? column, please enter your grammaticality judgments of the sentence, according to Standard English. The annotation options are:

- **grammatical** (for fluent, syntactically valid and semantically plausible sentences)
- **ungrammatical** (for sentences that have any typos, grammatical issues, or if the sentence describes a situation that dont make sense, or just sounds weird)

- **not sure** (if you are not sure whether it is clearly grammatical or ungrammatical)

Examples:

- *The driver told the passenger that he could pay for the ride with cash.*
  => grammatical
- *The driver said the passenger that he could pay for the ride with cash.*
  => ungrammatical (because said is intransitive in Standard English)

A.3.1.2 Questions about Pronouns

Every sentence contains a pronoun, and the Question column asks whether it refers to a person mentioned in the sentence or not. The annotation options are:

- **yes** (if the pronoun refers to the person)
- **no** (if the pronoun does not refer to the person)
- **not sure** (if you are not sure about whether the pronoun refers to the person)

Examples:

- *The driver told the passenger that he could pay for the ride with cash.*
  Does the pronoun he refer to the driver?
  => no
- *The driver told the passenger that he could pay for the ride with cash.*
  Does the pronoun he refer to the passenger?
  => yes

A.3.2 Task 2 Description

Together with this annotation protocol, you have received a link to a Google Sheet. The sheet contains 1 randomized data column and 1 task column.
Each row in the data column consists of multiple sentences, of which precisely one

sentence contains a blank. Your task is to determine the appropriate pronoun to fill in the blank, and enter it in the Pronoun column. Here, appropriate means correct in both form and case.

The tasks are designed to be unambiguous, so please provide only one solution and do not reorder the data.

Example:

- *The driver felt unhappy because he did not make enough money. The driver wondered whether ___ should take out a loan.*
  => he

B

Experimental Details

B.1 Computational Requirements

We use one 40GB NVIDIA A100 GPU for inference with most models, but we require two GPUs for OPT-30B and four for OPT-66B and the LLAMA-2-70B base and chat models. We access all models using the Huggingface Transformers library (Wolf et al., 2020) and use the minicons library for pseudo log likelihood evaluation (Misra, 2022).

B.2 Prompting

B.2.1 Coreference Resolution

Table B.1 shows all 10 prompt templates we use to present our task instances to FLAN-T5. Each template is presented in three variants to the model, where the options are changed:

1. No options
2. The occupation is presented first and the participant second
3. The participant is presented first and the occupation second

ID	Template
0	{task}\n\n{options}\nWho is {pronoun} referring to?
1	{task}\n\nWho is "{pronoun}" in this prior sentence (see options)?\n{options}
2	{task}\n\nWho is {pronoun} referring to in this sentence?\n{options}
3	Choose your answer: {task}\nTell me who {pronoun} is.\n{options}
4	{task}\nBased on this sentence, who is {pronoun}?\n\n{options}
5	Choose your answer: Who is {pronoun} in the following sentence?\n\n{task} \n\n{options}
6	Multi-choice problem: Which entity is {pronoun} this sentence?\n\n{task}\n\n{options}
7	Who is {pronoun} referring to in the following sentence?\n{task} \n\n{options}
8	Note that this question lists possible answers. Which person is {pronoun} referring to in the following sentence?\n{task} \n\n{options}
9	{task}\nWho is "{pronoun}"\n{options}

Table B.1: Prompting templates, where "task" is filled with each dataset instance, "pronoun" is the unique third person singular pronoun in that dataset instance, and "options" are the occupation and the participant.

B.2.2 Pronoun Fidelity

Table B.2 shows all 10 prompt templates we use to present our task instances to chat models. The first five are open-ended and the last five provide the model with pronoun options to choose from. These templates are used within vanilla and chain-of-thought prompting paradigms as described below.

ID	Template
0	{task}\n\nWhat pronoun should be used to fill the blank?
1	{task}\n\nThe best pronoun to fill in the blank is
2	Fill in the blank with the correct pronoun.\n\n{task}
3	Fill in the blank with the appropriate pronoun.\n\n{task}
4	What pronoun should be used to fill the blank?\n\n{task}
5	{task}\n\nWhat pronoun should be used to fill the blank?\n{options}
6	{task}\n\nThe best pronoun to fill in the blank is\n{options}
7	Fill in the blank with the correct pronoun.\n\n{task}\n{options}
8	Fill in the blank with the appropriate pronoun.\n\n{task}\n{options}
9	What pronoun should be used to fill the blank?\n\n{task}\n{options}

Table B.2: Prompting templates for chat models, where "task" is filled with each dataset instance, and "options" is a list of four pronouns to choose from, all in the correct case.

B.2.2.1 Vanilla Prompting

With FLAN-T5, vanilla prompting only requires instantiating templates with task instances, whereas LLAMA-2-CHAT requires special formatting with INST. Instantiated examples of one template are shown for both models in Table B.3. The number of maximum new tokens is set to 5 for FLAN-T5 and 20 for LLAMA-2-CHAT based on experimentation.

B.2.2.2 Chain-of-Thought Prompting

We focus on FLAN-T5-XXL for chain-of-thought experiments, and use the strong zero-shot prompts for reasoning proposed by Kojima et al. (2022) (*"Let's think step by step"*) and Zhou et al. (2023) (*"Let's work this out in a step by step way to be*

sure we have the right answer"), which we append after the template. Following their codebases, we first allow the models to generate a chain of thought (with 128 maximum new tokens). Then, we append the chain of thought after the question and elicit the final answer with the string *"Therefore, the correct pronoun is,"* allowing the model to generate up to 10 new tokens. This two-step process is illustrated with examples in Tables B.4 and B.5. We save both the final answer and the chain of thought for later analysis.

Model type	Example input
FLAN-T5	Fill in the blank with the appropriate pronoun. The accountant had just eaten a big meal so her stomach was full. The accountant was asked about ___ charges for preparing tax returns. OPTIONS: - his - her - their - xyr
LLAMA-2-CHAT	[INST] Fill in the blank with the appropriate pronoun. The accountant had just eaten a big meal so her stomach was full. The accountant was asked about ___ charges for preparing tax returns. OPTIONS: - his - her - their - xyr [/INST]

Table B.3: Example input using template 8 for FLAN-T5 and LLAMA-2-CHAT models.

Chain-of-thought prompt	**Example input**
Kojima et al. (2022) (eliciting chain of thought)	Q: Fill in the blank with the appropriate pronoun. The accountant had just eaten a big meal so her stomach was full. The accountant was asked about ___ charges for preparing tax returns. A: Let's think step by step.
Kojima et al. (2022) (eliciting final answer)	Q: Fill in the blank with the appropriate pronoun. The accountant had just eaten a big meal so her stomach was full. The accountant was asked about ___ charges for preparing tax returns. A: Let's think step by step. {generated chain of thought} Therefore, the correct pronoun is

Table B.4: Example input using template 3 for evaluating FLAN-T5-XXL with two types of chain-of-thought prompting. Prompting happens in two phases regardless of the choice of prompt: eliciting the chain of thought and eliciting the final answer.

Chain-of-thought prompt	**Example input**
Zhou et al. (2023) (eliciting chain of thought)	Q: Fill in the blank with the appropriate pronoun. The accountant had just eaten a big meal so her stomach was full. The accountant was asked about ___ charges for preparing tax returns. A: Lets work this out in a step by step way to be sure we have the right answer.
Zhou et al. (2023) (eliciting final answer)	Q: Fill in the blank with the appropriate pronoun. The accountant had just eaten a big meal so her stomach was full. The accountant was asked about ___ charges for preparing tax returns. A: Lets work this out in a step by step way to be sure we have the right answer. {generated chain of thought} Therefore, the correct pronoun is

Table B.5: Example input using template 3 for evaluating FLAN-T5-XXL with two types of chain-of-thought prompting. Prompting happens in two phases regardless of the choice of prompt: Eliciting the chain of thought and eliciting the final answer.

C

Additional Results

We report additional results on double- and single-entity sentences in WINOPRON: F_1 scores in Table C.1, precision in Table C.2, and recall in Table C.3. Note that FLAN-T5 models generally perform worse on single-entity sentences compared to double-entity sentences because some of our prompts include options (see Section B.2 for details) that confuse the model in this setting, despite being necessary to resolve double-entity sentences.

Data	LINGMESS	CAW-COREF	SPANBERT			FLAN-T5			
			base	large	small	base	large	xl	xxl
Double-entity sentences									
All	64.4	67.3	61.6	70.1	51.6	62.4	78.0	**89.0**	88.8
Nom.	73.5	77.6	67.2	77.2	51.9	65.4	85.1	94.7	**96.7**
Acc.	52.2	57.5	54.6	59.5	50.4	58.4	69.9	**82.5**	79.1
Poss.	67.4	66.5	62.9	73.6	52.3	63.4	79.1	89.7	**90.7**
he	79.2	79.6	62.8	71.5	51.5	64.1	81.5	88.8	**90.2**
she	76.3	76.6	62.1	71.6	51.5	66.1	83.3	**90.6**	89.9
they	67.5	63.7	61.2	68.9	51.8	60.5	77.0	**88.6**	88.0
xe	*8.5*	*38.6*	60.4	68.5	51.4	58.7	70.3	**88.0**	87.3
Single-entity sentences									
All	73.2	75.6	**95.5**	88.0	77.3	76.3	81.5	83.1	84.3
Nom.	80.0	82.5	**99.5**	99.3	78.3	80.8	89.8	93.3	97.0
Acc.	61.1	65.0	**87.3**	67.5	76.2	69.6	69.8	70.1	66.5
Poss.	77.1	78.0	**99.8**	97.1	77.5	78.5	84.7	85.7	89.2
he	92.7	94.3	**94.7**	85.6	77.6	81.3	86.8	88.2	88.6
she	90.9	91.6	**96.2**	88.9	77.4	81.1	87.6	88.8	87.1
they	75.2	69.8	**96.0**	88.7	79.3	76.1	84.3	85.7	86.8
xe	2.2	27.3	**95.2**	88.7	75.0	66.3	67.0	69.4	74.6

Table C.1: F_1 of coreference resolution systems on double- and single-entity sentences in WINOPRON. We report F_1 overall, and split by grammatical case and pronoun set. *Red, italicized numbers* are worse than chance (50.0 for double-entity sentences and not applicable for single-entity sentences).

Data	LINGMESS	CAW-COREF	SPANBERT			FLAN-T5			
			base	large	small	base	large	xl	xxl
Double-entity sentences									
All	79.1	80.1	62.1	70.6	51.9	62.9	78.4	**89.5**	89.4
Nom.	88.3	88.7	67.4	77.4	52.1	65.7	85.4	95.1	**97.1**
Acc.	63.4	67.9	55.3	59.9	50.7	58.8	70.2	**83.2**	79.5
Poss.	86.1	83.6	63.5	74.3	52.8	64.3	79.6	90.2	**91.5**
he	79.7	80.1	63.0	71.6	51.7	64.3	81.8	89.3	**90.6**
she	77.6	77.9	62.3	71.8	51.7	66.3	83.6	**91.1**	90.3
they	79.1	80.2	61.8	69.5	52.0	60.8	77.3	**89.0**	88.5
xe	**100.0**	88.1	61.3	69.3	52.1	60.1	70.7	88.6	88.0
Single-entity sentences									
All	**100.0**	**100.0**	96.0	88.4	78.9	77.6	82.4	84.0	85.6
Nom.	**100.0**	**100.0**	**100.0**	**100.0**	79.3	81.6	90.4	93.9	97.4
Acc.	**100.0**	**100.0**	88.1	67.9	77.5	70.5	70.7	71.1	68.1
Poss.	**100.0**	**100.0**	99.8	97.1	79.8	80.7	85.9	86.8	90.8
he	**100.0**	**100.0**	95.0	86.0	78.6	81.9	87.5	88.9	89.5
she	**100.0**	**100.0**	96.4	89.1	78.5	81.7	88.1	89.4	87.9
they	**100.0**	**100.0**	96.4	89.1	80.3	76.9	85.2	86.5	87.9
xe	**100.0**	**100.0**	96.3	89.3	77.9	69.2	68.3	70.9	76.7

Table C.2: Precision on double- and single-entity sentences overall, and split by grammatical case and pronoun set. *Red, italicized numbers* are worse than chance (50.0 for double-entity sentences, N/A for single-entity sentences).

Data	LINGMESS	CAW-COREF	SPANBERT			FLAN-T5			
			base	large	small	base	large	xl	xxl
Double-entity sentences									
All	54.2	58.0	61.1	69.7	51.3	61.9	77.7	**88.5**	88.3
Nom.	62.9	69.0	67.1	77.1	51.8	65.2	84.8	94.3	**96.3**
Acc.	*44.4*	*49.8*	54.0	59.2	50.2	58.0	69.6	**81.9**	78.6
Poss.	55.4	55.2	62.3	72.9	51.9	62.5	78.7	89.3	**90.0**
he	78.6	79.2	62.5	71.4	51.4	63.9	81.1	88.3	**89.7**
she	75.0	75.3	61.9	71.4	51.4	65.9	83.0	**90.1**	89.5
they	58.9	52.8	60.6	68.3	51.6	60.2	76.8	**88.1**	87.5
xe	*4.4*	*24.7*	59.4	67.8	50.8	57.4	69.9	**87.5**	86.6
Single-entity sentences									
All	57.8	60.8	**95.1**	87.6	75.9	75.0	80.6	82.1	83.1
Nom.	66.7	70.2	**99.0**	98.5	77.3	80.0	89.2	92.7	96.6
Acc.	44.0	48.1	**86.5**	67.1	74.9	68.7	69.0	69.1	65.0
Poss.	62.7	64.0	**99.8**	97.1	75.4	76.4	83.5	84.6	87.6
he	86.4	89.2	**94.4**	85.3	76.5	80.8	86.1	87.4	87.8
she	83.3	84.4	**96.1**	88.6	76.2	80.5	87.1	88.2	86.2
they	60.3	53.6	**95.6**	88.3	78.3	75.3	83.5	85.0	85.8
xe	1.1	15.8	**94.2**	88.1	72.4	63.7	65.7	68.0	72.5

Table C.3: Recall on double- and single-entity sentences overall, and split by grammatical case and pronoun set. *Red, italicized numbers* are worse than chance (50.0 for double-entity sentences, N/A for single-entity sentences)