

# Training in Step-by-Step Formal Reasoning Improves Pronominal Reasoning in Language Models

Vagrant Gautam

Heidelberg Institute for Theoretical Studies

Heidelberg, Germany

vagrant.gautam@h-its.org

## Abstract

Large reasoning models are trained to solve problems by decomposing them into steps. While they show impressive progress on reasoning tasks, “reasoning” here is typically limited to formal reasoning, i.e., math, code, and logic. An open question is whether these abilities transfer to *pronominal reasoning*, where step-by-step thinking in non-reasoning models worsens performance, but code pre-training may help. I answer this question by evaluating six pairs of original and DeepSeek-distilled models (1.5B–70B parameters) on six challenging datasets for English pronoun resolution (identifying whom a pronoun refers to) and pronoun fidelity (learning and applying a pronoun mapping correctly). Performance improves statistically significantly on all datasets (31% relative increase), indicating that distilling step-by-step formal reasoning does in fact help with pronominal reasoning, in part by improving instruction-following. With a qualitative evaluation of 720 generations, I show that improvements occur across granular error types, and come from plausible-looking reasoning chains employing a variety of reasoning strategies. However, the gains put models just above random performance on these datasets, leaving plenty of room for improvement.

## 1 Introduction

Large reasoning models are language models that are trained to solve problems and respond to queries by first generating a step-by-step reasoning chain, which the final answer is then conditioned on (Marjanović et al., 2025). This two-step process allows for the generation of competing hypotheses as well as backtracking, leading to vastly improved performance on reasoning tasks (DeepSeek-AI et al., 2025; OpenAI et al., 2024). However, training and evaluations in this context typically limit their scope to formal (i.e., math, code, and logical) reasoning, where data reflecting step-by-step human reasoning is easiest to find.

This raises the question of whether training and good performance on formal reasoning transfers to other kinds of reasoning, such as referential reasoning. Reference is a complex phenomenon, as multiple linguistic forms (names, pronouns, definite noun phrases) can be used to refer to the same individual, and conversely, the same linguistic form (e.g., *she*) can be used to refer to different individuals in different contexts (Abbott, 2010; Gundel and Abbott, 2019). Getting NLP systems to understand and do reference the way humans do has thus been a challenge for over 50 years (Hobbs, 1978; Lappin and Leass, 1994; Krahmer and van Deemter, 2012).

More recently, in non-reasoning language models, step-by-step reasoning in the form of chain-of-thought prompting has been shown to worsen performance on pronominal reference (Gautam et al., 2024a), and maintain performance on non-math and non-symbolic tasks (Sprague et al., 2025). On the other hand, code training has been shown to improve entity tracking abilities of language models (Kim et al., 2024), an important component of referential reasoning. Combining the two, I investigate whether *training* on step-by-step formal reasoning helps with pronominal reasoning.

I evaluate six pairs of original and DeepSeek-distilled models on six challenging datasets for coreference resolution and pronoun fidelity. Accuracy improves statistically significantly on all of these datasets, indicating that **training in step-by-step reasoning on fundamentally different data such as math and code does indeed transfer to better pronominal reasoning**. With a human evaluation of 720 generations, I provide a more granular view of error types and show that DeepSeek distillation teaches models to better follow instructions, and to actually manipulate referring expressions, backtrack, and reason about them better. Similar to Kim and Schuster (2023), these findings suggest that step-by-step reasoning about reference cannot simply be elicited from pre-trained

Dataset	Natural	Neopronouns	Referents	Instances
<b>Coreference Resolution</b> ( <i>Who does this pronoun refer to?</i> )				
GAP (Webster et al., 2018) <i>...<a href="#">Florence White</a> also references <a href="#">Lady Arundel's Manchet's</a> in <a href="#">her</a> 1932 English Cookery book...</i>	✓	–	1+	2, 000
KnowRef-60K (Emami et al., 2020) <i><a href="#">Valerie</a> was the first movie where I remember seeing <a href="#">Pamela</a>, and <a href="#">she</a> blew me away.</i>	✓	–	2	3, 060
WinoPron (Gautam et al., 2024b) <i>The <a href="#">janitor</a> warned the <a href="#">child</a> not to step on the wet floor for <a href="#">his</a> work not to be undone.</i>	–	✓	2	1, 440
<b>Pronoun Fidelity</b> ( <i>What pronoun should be used to fill this blank?</i> )				
MISGENDERED (Hossain et al., 2023) <i>"Robert (<a href="#">he</a>/him/his/himself) was very emotional. ___ cried loudly and often.</i>	–	✓	1	3, 300
RUFF-no-distractors (Gautam et al., 2024a) <i>The bartender thought the day was going badly for <a href="#">xem</a>. The bartender made a strong drink for ___ to get a bigger tip.</i>	–	✓	1	2, 160
RUFF-1-distractor (Gautam et al., 2024a) <i>The bartender thought the day was going badly for <a href="#">xem</a>. The customer was happy because the new shoes fit <a href="#">her</a> well. The bartender made a strong drink for ___ to get a bigger tip.</i>	–	✓	2	2, 160

Table 1: Overview of datasets with the provenance of their data (natural vs. researcher-created), presence of neopronouns, number of referents, number of instances, and examples with the [answer](#) and possible **distractors**.

models, but can transfer from some formal training. Nevertheless, performance is far from strong, with most models just outperforming random baselines in datasets where this is applicable. All code and data are made available to encourage future work on pronominal reasoning and reasoning transfer.<sup>1</sup>

## 2 Related Work

When it comes to **understanding reference**, coreference resolution is one of the oldest NLP tasks (Sukthanker et al., 2020), with several researcher-created challenge datasets (Levesque et al., 2012; Webster et al., 2018; Sakaguchi et al., 2021, *inter alia*), which are sometimes easier for language models than natural corpora (Porada and Cheung, 2024). Entity tracking, one component of understanding reference, is particularly well-studied, with Kim and Schuster (2023), Kim et al. (2024) and Prakash et al. (2024) showing improvements from fine-tuning or code training. My work differs in evaluating referential resolution with multiple referring expressions on non-toy data.

As for **doing reference**, Schuster and Linzen (2022) find that language models (incorrectly) refer to discourse entities that are not introduced, and

Gautam et al. (2024a) similarly finds that models struggle with pronominal reference when multiple referents are involved, and that chain-of-thought prompting *worsens* performance. Faithful pronominal reference has also been studied from the perspective of misgendering (Ovalle et al., 2023; Subramonian et al., 2025). However, none of these papers use reasoning models.

Research on **reasoning models** has focused on evaluations of mathematics, code, and logical tasks (DeepSeek-AI et al., 2025; OpenAI et al., 2024), including with the training-free predecessor of such models, chain-of-thought prompting. Here, a model generates a reasoning chain to condition the final answer on (Wei et al., 2022; Kojima et al., 2022), which Sprague et al. (2025) show maintains performance on non-symbolic reasoning tasks.

## 3 Methodology

**Data** As Table 1 shows, I choose challenging datasets for two pronominal reasoning tasks—coreference resolution and pronoun fidelity—covering different input lengths, referents, and naturalness. In the coreference resolution datasets, the task is to identify which named referent or definite noun phrase a pronoun corefers with in a text. In the pronoun fidelity datasets, the task is to fill a

<sup>1</sup>[github.com/dippedrusk/reasoning-transfer](https://github.com/dippedrusk/reasoning-transfer)

Alias	Original model	DeepSeek-distilled model
QWEN-1.5B	Qwen/Qwen2.5-Math-1.5B	deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B
QWEN-7B	Qwen/Qwen2.5-Math-7B	deepseek-ai/DeepSeek-R1-Distill-Qwen-7B
LLAMA-8B	meta-llama/Llama-3.1-8B	deepseek-ai/DeepSeek-R1-Distill-Llama-8B
QWEN-14B	Qwen/Qwen2.5-14B	deepseek-ai/DeepSeek-R1-Distill-Qwen-14B
QWEN-32B	Qwen/Qwen2.5-32B	deepseek-ai/DeepSeek-R1-Distill-Qwen-32B
LLAMA-70B	meta-llama/Llama-3.3-70B-Instruct	deepseek-ai/DeepSeek-R1-Distill-Llama-70B

Table 2: HuggingFace identifiers for all model pairs used in this study.

blank with a pronoun such that it is consistent with pronominal reference earlier in the text. Appendix A contains more details about the datasets.

**Models** I use all six open-weight pairs of original and DeepSeek-distilled (DeepSeek-AI et al., 2025) models (see Table 2) to ablate for the effects of training in step-by-step formal reasoning. This includes QWEN (Yang et al., 2024) and LLAMA (Grattafiori et al., 2024) models. Distillation is done through supervised fine-tuning on reasoning traces generated by DeepSeek-R1, consisting of 600K samples of (math, code, and logical) reasoning data, and 200K samples of non-reasoning data, including writing, factual QA, and translation (DeepSeek-AI et al., 2025). As this data is not released, it is not possible to fine-tune other models, despite DeepSeek-AI’s inconsistent choices.<sup>2</sup>

**Automatic evaluation** I use a standardized set of five prompts for each task, designed based on Longpre et al. (2023) and Gautam et al. (2024a). Original models are directly prompted as chain-of-thought prompting worsens pronominal reasoning (Gautam et al., 2024a), while DeepSeek-distilled models generate a reasoning chain first. I post-process all 847, 200 generations with regexes, and report accuracy averaged over prompts, following best practices (Sclar et al., 2024; Mizrahi et al., 2024). See Appendix B for details.

**Human evaluation** To validate the automatic evaluation and uncover more granular patterns, I annotate a total of 720 generations for answer correctness, sampling all models and datasets equally. A second annotator independently annotated these generations, with excellent agreement (Cohen’s  $\kappa = 0.81$  and percentage agreement = 0.86). See Appendix C for details on sampling and annotation.

<sup>2</sup>All “original” models except for LLAMA-70B are base models, some of which undergo math pre-training (QWEN-1.5B and QWEN-7B). See Appendix D for more discussion.

**Uncertainty and significance testing** As regexes tend to have high recall and low precision, error bars on plots indicate upper and lower bounds on the automatic evaluation, based on the proportion of false positives and false negatives flagged in manual annotation. I test for statistical significance using the mid-p McNemar’s test (Fagerland et al., 2013; Dror et al., 2018) with Bonferroni correction and  $p = 0.01$ , comparing paired results from the original and DeepSeek-distilled models at the dataset level (aggregated across models) and model level (across datasets). Significance is tested using the automatic evaluation results as the human evaluation is under-powered (Card et al., 2020).

#### 4 Distilling Formal Reasoning Into Models Helps Referential Reasoning

As Figure 1 shows, **performance improves significantly using the DeepSeek-distilled models on all datasets**, with a 31% relative increase in accuracy on average, and above-random performance on challenging multi-referent datasets such as KnowRef-60K and RUFF-1-dist. Given that step-by-step reasoning via chain-of-thought prompting maintains or worsens commonsense reasoning and pronoun fidelity (Gautam et al., 2024a; Sprague et al., 2025), these results suggest that step-by-step pronominal reasoning cannot simply be elicited from pre-trained language models, but does transfer from some *training*, similar to Kim and Schuster’s (2023) conclusions about entity tracking. Referential reasoning may also benefit from similarities to symbolic reasoning (Sprague et al., 2025), as referring expressions can be viewed as symbols that point to a referent just as different variables might refer to the same value. Indeed, pronominal reasoning is sometimes formalized with symbols in prior work (Subramonian et al., 2025).

**Pronoun fidelity is harder than coreference resolution** with 20 points lower accuracy on aver-

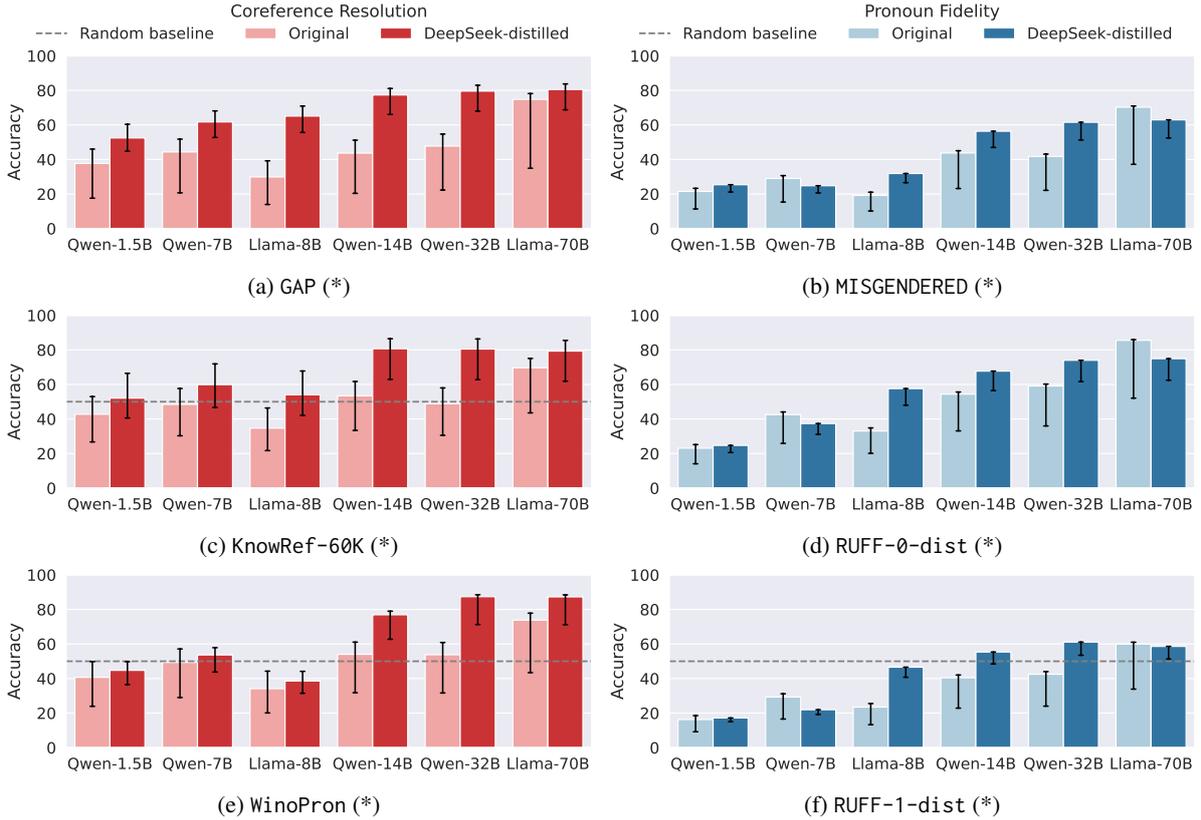


Figure 1: Differences in accuracy between original models and their DeepSeek-distilled counterparts per dataset. Error bars are calculated from human annotations and indicate uncertainty on the automatically computed accuracy. \* indicates significant improvements on the dataset (aggregated across models) with a McNemar’s test and  $p = 0.01$ . Where available, I indicate a random baseline of choosing between the correct referent and a distractor.

age, likely because this task requires more steps: Coreference resolution to learn an entity-pronoun mapping and to establish whom the blank refers to, and finally, applying the appropriate mapping, which may need a different declension. As in [Gautam et al. \(2024a\)](#), performance degrades between RUFF-0-dist and RUFF-1-dist due to the addition of a non-adversarial distractor referent, indicating non-robust reasoning despite overall gains.

All six models show statistically significant improvements in pronominal reasoning, indicating that **models generally benefit from the distillation of step-by-step formal reasoning**. However, this might be due to DeepSeek’s choice of weaker models with no instruction-tuning. For example, LLAMA-8B is a base model with massive gains (68% relative increase in accuracy) as it learns not just formal reasoning, but also instruction-following from DeepSeek distillation. In contrast, LLAMA-70B, the only instruction-tuned original model to undergo distillation, has comparatively smaller gains on coreference resolution tasks, and even worsens on pronoun fidelity datasets such as

MISGENDERED and RUFF-0-dist. Without the distillation data, it is impossible to ablate the existing model selection, thereby meaningfully and fairly disentangling the effects of math pre-training and instruction-tuning. Nevertheless, I compare the performance of the DeepSeek-selected original models with alternatives (e.g., without math training, with instruction-tuning, etc.) in Appendix D, to attempt to disentangle these factors.

## 5 Improvements Come From Instruction-Following and Reasoning

In order to better characterize the difference in model behaviour between original and DeepSeek-distilled models, I annotate model generations more granularly, as pronominal reasoning errors fall on a graded scale. With coreference resolution, models may fail to follow the prompt, provide an incorrect answer, an ambiguous one, or a correct one. Pronoun fidelity errors include more possibilities: Models may provide a noun or determiner instead of a pronoun, give the wrong pronoun set, or the correct pronoun but in the wrong form.

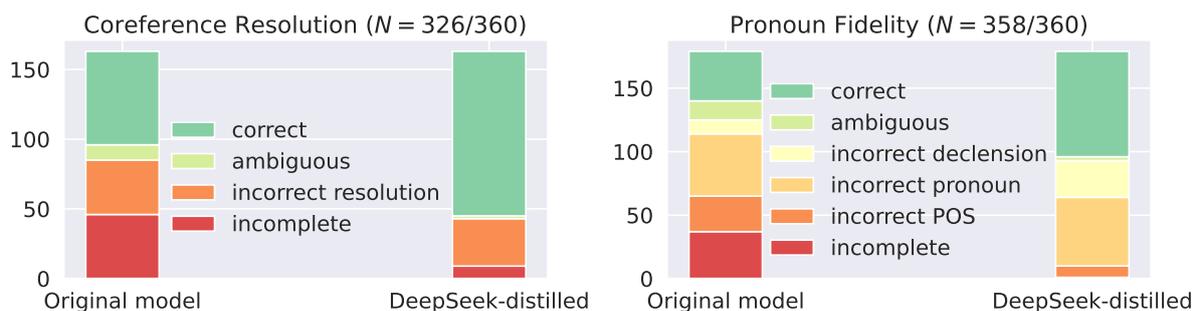


Figure 2: Granular human evaluation of generations given unambiguous prompts out of  $N = 720$ , split by task.

**Example:** *"What pronoun should be used to fill the blank? The educator stifled a yawn; her sleep had been fitful. The educator had a meeting to discuss \_\_\_ grading policy.*

**Correct answer:** her

**Ambiguous answer:** The answer is her or **his**

**Incorrect declension:** **hers** grading policy

**Incorrect pronoun:** **his** grading policy

**Incorrect part of speech:** **the** grading policy

**Incomplete:** I'm sorry, I can't help with that

By annotating these errors and plotting them in Figure 2, it is clear that there are upward improvements across categories post-distillation; ambiguous results become correct, incorrect parts of speech become incorrect pronouns or declensions, or correct, and so on. Ambiguous and incomplete answers also practically disappear, indicating that **DeepSeek distillation helps models better format reasoning chains and follow instructions**. Analyzing instances that become correct after distillation ( $N = 116$ ) reveals that **78% of reasoning chains show plausible referential reasoning**. Plausible reasoning involves a variety of strategies including heuristics based on sentence structure, and reliance on gendered stereotypes, discussed in Appendix E. I present one example below of a model reasoning about neopronouns by analogy with other, more familiar pronouns, with the prompt *in italics* followed by the generation.

*Elijah (ae/aer/aer) is struggling financially right now. Could you lend \_\_\_ some money?* Okay, so I'm trying to figure out which pronoun to use in this sentence: "Elijah (ae/aer/aer) is struggling financially right now. Could you lend \_\_\_ some money?" The options given are ae, aer, and aer again. Hmm, I'm a bit confused because I'm not very familiar with these pronouns. I know that **in**

**English, we typically use "he," "she," or "they," but it seems like Elijah is using something else.**

First, I should understand what each of these pronouns means. I think "ae" might be a singular pronoun, maybe a gender-neutral one. **"Aer" could be the possessive form, like "their" but singular. So, if Elijah uses "ae/aer/aer," that would mean subject, object, and possessive forms respectively.**

Looking at the sentence, the blank is in the object position because it's the recipient of the action "lend." So, the pronoun should be in the object form. If "ae" is the subject form, then "aer" would be the object form. Wait, but the options given are ae, aer, aer. So, maybe the object form is "aer."

Let me think again. **In English, the object form of "he" is "him," and for "they" it's "them." So, if Elijah's pronouns are ae/aer/aer, then when we're using it as an object, it should be "aer."** So the sentence would be "Could you lend aer some money?" That makes sense because "aer" is the object form.

I'm a bit unsure because I'm not very familiar with these pronouns, but I think that's the correct approach. It's important to respect the pronouns that people use, so getting this right is crucial. . .

## 6 Conclusion

By evaluating six pairs of original and DeepSeek-distilled models on six datasets for coreference resolution and pronoun fidelity, I show that training on step-by-step formal reasoning (code, math, and logic) significantly improves models' pronominal reasoning. In light of prior work, my findings suggest that step-by-step reasoning about reference can only transfer from some formal training. However, as these improvements only bring models slightly over random performance on these challenging datasets, there remains much room for improvement in reasoning with language models.

## Limitations

**More complex reference** Reference is a complex phenomenon precisely due to the interplay of several kinds of knowledge and reasoning it requires. Although the datasets I select are challenging for language models, they are relatively simple on the scale of reference, as **(1)** they focus primarily on prototypical linguistic reference (pronouns, names, definite noun phrases), **(2)** I do not evaluate complex reference chains but rather individual instances with (mostly) deterministic binary correctness, and **(3)** the pronoun fidelity datasets I select are solvable through slot filling or shallow copying (Gautam et al., 2024a; cf. Kim and Schuster 2023), although zero-shot evaluation prevents this. By covering multiple datasets, including natural data, I aim to provide a “more complete picture of system capabilities,” (Porada and Cheung, 2024), but note that my results provide an optimistic view of model performance.

**Reasoning chains** Although Section 5 addresses a need for more systematic evaluation of intermediate reasoning chains (Khalili et al., 2025), human perceptions of “plausible” reasoning in natural language traces differ substantially from what models actually rely on in practice (Levy et al., 2025). These effects are yet to be investigated in pronominal reasoning, including in this work.

**Model selection** I chose DeepSeek-distilled models as they are all trained on the exact same (primarily) formal reasoning data with step-by-step reasoning traces, providing a consistent way to test my research question with multiple data points. However, the original models chosen by DeepSeek are highly inconsistent within individual families; QWEN2.5-MATH-1.5B and QWEN2.5-MATH-7B undergo math pre-training, while QWEN2.5-14B and QWEN2.5-32B do not, and LLAMA-3.1-8B is not instruction-tuned while LLAMA-3.3-70B-INSTRUCT is. Disentangling these confounds (beyond my attempt in Appendix D) could explain the variance in the magnitude of gains and shape concrete guidance as to when and where training in step-by-step reasoning actually helps.

## Ethics Statement

All datasets and models are used in accordance with their licenses and intended use. I release all code, data, and annotations in the spirit of open science,

and to encourage future work on referential reasoning in language models. As for potential risks of my work, tracking textual references to people can be used to surveil them, similar to facial recognition technologies (Cockerell, 2019; Frenkel, 2024; Kalluri et al., 2025), with harmful and potentially lethal consequences that I stand vocally against.

## Acknowledgements

I am grateful to Diana Davidson for their annotations, and to my reviewers for their feedback.

## References

- Barbara Abbott. 2010. *Reference*. Oxford University Press, New York.
- Yang Trista Cao and Hal Daumé III. 2020. [Toward gender-inclusive coreference resolution](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.
- Yang Trista Cao and Hal Daumé III. 2021. [Toward gender-inclusive coreference resolution: An analysis of gender and bias throughout the machine learning lifecycle\\*](#). *Computational Linguistics*, 47(3):615–661.
- Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. 2020. [With little power comes great responsibility](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9263–9274, Online. Association for Computational Linguistics.
- Isobel Cockerell. 2019. [Inside china’s massive surveillance operation](#). *Wired*.
- 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, and 181 others. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.
- Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. [The hitchhiker’s guide to testing statistical significance in natural language processing](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.
- Ali Emami, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2020. [An analysis of dataset overlap on Winograd-style tasks](#). In *Proceedings of the 28th International Conference on Computational*

- Linguistics*, pages 5855–5865, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Morten W Fagerland, Stian Lydersen, and Petter Laake. 2013. [The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional](#). *BMC Medical Research Methodology*, 13(1).
- Sheera Frenkel. 2024. [Israel deploys expansive facial recognition program in Gaza](#). *New York Times*.
- Vagrant Gautam, Eileen Bingert, Dawei Zhu, Anne Lauscher, and Dietrich Klakow. 2024a. [Robust pronoun fidelity with English LLMs: Are they reasoning, repeating, or just biased?](#) *Transactions of the Association for Computational Linguistics*, 12:1755–1779.
- Vagrant Gautam, Julius Steuer, Eileen Bingert, Ray Johns, Anne Lauscher, and Dietrich Klakow. 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*, pages 52–66, Miami. Association for Computational Linguistics.
- Aryo Pradipta Gema, Alexander Hägele, Runjin Chen, Andy Arditi, Jacob Goldman-Wetzler, Kit Fraser-Taliente, Henry Sleight, Linda Petrini, Julian Michael, Beatrice Alex, Pasquale Minervini, Yanda Chen, Joe Benton, and Ethan Perez. 2025. [Inverse scaling in test-time compute](#). *Transactions on Machine Learning Research*. Featured Certification, J2C Certification.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The Llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Jeanette Gundel and Barbara Abbott. 2019. *The Oxford Handbook of Reference*. Oxford University Press.
- Jerry R. Hobbs. 1978. [Resolving pronoun references](#). *Lingua*, 44(4):311–338.
- Tamanna Hossain, Sunipa Dev, and Sameer Singh. 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)*, pages 5352–5367, Toronto, Canada. Association for Computational Linguistics.
- Pratyusha Ria Kalluri, William Agnew, Myra Cheng, Kentrell Owens, Luca Soldaini, and Abeba Birhane. 2025. [Computer-vision research powers surveillance technology](#). *Nature*, 643(8070):73–79.
- Zena Al Khalili, Nick Howell, and Dietrich Klakow. 2025. [Evaluating intermediate reasoning of code-assisted large language models for mathematics](#). In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM<sup>2</sup>)*, pages 741–758, Vienna, Austria and virtual meeting. Association for Computational Linguistics.
- Najoung Kim and Sebastian Schuster. 2023. [Entity tracking in language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3835–3855, Toronto, Canada. Association for Computational Linguistics.
- Najoung Kim, Sebastian Schuster, and Shubham Toshniwal. 2024. [Code pretraining improves entity tracking abilities of language models](#). *Preprint*, arXiv:2405.21068.
- Takeshi Kojima, 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*, volume 35, pages 22199–22213. Curran Associates, Inc.
- Emiel Kraemer and Kees van Deemter. 2012. [Computational generation of referring expressions: A survey](#). *Computational Linguistics*, 38(1):173–218.
- Shalom Lappin and Herbert J. Leass. 1994. [An algorithm for pronominal anaphora resolution](#). *Computational Linguistics*, 20(4):535–561.
- Anne Lauscher, Archie Crowley, and Dirk Hovy. 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*, pages 1221–1232, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Hector J. Levesque, 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*. AAAI Press.
- Mosh Levy, Zohar Elyoseph, and Yoav Goldberg. 2025. [Humans perceive wrong narratives from AI reasoning texts](#). *Preprint*, arXiv:2508.16599.
- Shayne Longpre, 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 *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. JMLR.org.
- Sara Vera Marjanović, Arkil Patel, Vaibhav Adlakha, Milad Aghajohari, Parishad Behnamghader, Mehar Bhatia, Aditi Khandelwal, Austin Kraft, Benno Kroyer, Xing Han Lù, Nicholas Meade, Dongchan Shin, Amirhossein Kazemnejad, Gaurav Kamath, Marius

- Mosbach, Karolina Stańczak, and Siva Reddy. 2025. [Deepseek-r1 thoughtology: Let’s think about llm reasoning](#). *Preprint*, arXiv:2504.07128.
- Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. [State of what art? a call for multi-prompt llm evaluation](#). *Transactions of the Association for Computational Linguistics*, 12:933–949.
- 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, and 244 others. 2024. [Openai o1 system card](#). *Preprint*, arXiv:2412.16720.
- Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jagers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. [“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*, page 1246–1266, New York, NY, USA. Association for Computing Machinery.
- Ian Porada and Jackie CK Cheung. 2024. [Solving the challenge set without solving the task: On Winograd schemas as a test of pronominal coreference resolution](#). In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 489–506, Miami, FL, USA. Association for Computational Linguistics.
- Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. 2024. [Fine-tuning enhances existing mechanisms: A case study on entity tracking](#). In *The Twelfth International Conference on Learning Representations*.
- Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 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)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavata, and Yejin Choi. 2021. [Winogrande: an adversarial winograd schema challenge at scale](#). *Commun. ACM*, 64(9):99–106.
- Sebastian Schuster and Tal Linzen. 2022. [When a sentence does not introduce a discourse entity, transformer-based models still sometimes refer to it](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 969–982, Seattle, United States. Association for Computational Linguistics.
- Melanie Sclar, 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*.
- Zayne Rea Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2025. [To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning](#). In *The Thirteenth International Conference on Learning Representations*.
- 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*.
- Rhea Sukthanker, Soujanya Poria, Erik Cambria, and Ramkumar Thirunavukarasu. 2020. [Anaphora and coreference resolution: A review](#). *Information Fusion*, 59:139–162.
- Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldrige. 2018. [Mind the gap: A balanced corpus of gendered ambiguous pronouns](#). *Transactions of the Association for Computational Linguistics*, 6:605–617.
- Jason Wei, 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*, Red Hook, NY, USA. Curran Associates Inc.
- Thomas Wolf, 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, and 3 others. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Yuyang Wu, Yifei Wang, Tianqi Du, Stefanie Jegelka, and Yisen Wang. 2025. [When more is less: Understanding chain-of-thought length in LLMs](#). In *Workshop on Reasoning and Planning for Large Language Models*.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. [Qwen2 technical report](#). *Preprint*, arXiv:2407.10671.

ID	Template
<b>Coreference Resolution</b>	
0	{instance}\n\nWho is “{pronoun}” referring to?
1	{instance}\n\nWho is ‘{pronoun}’ referring to in this text?
2	Who is “{pronoun}” referring to in the following text?\n\n{instance}
3	{instance}\n\nBased on this, who is ‘{pronoun}’?
4	Who is “{pronoun}”?\n\n{instance}
<b>Pronoun Fidelity</b>	
0	{instance}\n\nWhat pronoun should be used to fill the blank?
1	{instance}\n\nThe best pronoun to fill in the blank is
2	Fill in the blank with the correct pronoun.\n\n{instance}
3	What pronoun should be used to fill the blank?\n\n{instance}
4	Fill in the blank with the appropriate pronoun.\n\n{instance}

Table 3: Prompting templates per task, where “instance” is filled with each dataset instance, and for coreference resolution prompts, “pronoun” is the pronoun to resolve in that dataset instance.

## A Datasets

### A.1 Coreference Resolution

The coreference resolution datasets I select fall on a continuum from natural to researcher-created. GAP (Webster et al., 2018) is a human-labelled corpus of ambiguous pronoun-name pairs in challenging multi-sentence instances from natural (Wikipedia) data, which names and uniquely identifies individuals, but only contains publicly-available data, to the best of my knowledge. I use their 2,000-instance test set for evaluation. KnowRef-60K (Emami et al., 2020) consists of naturally-occurring sentences which are subsequently modified to reduce overlap with pre-training corpora, putting them somewhere in between completely natural and completely researcher-created. I evaluate on their test split of 3,060 instances. Finally, WinoPron (Gautam et al., 2024b) corrects researcher-created WinoGender Schemas (Rudinger et al., 2018) and expands them to a wider variety of pronoun types in English (primarily accusative and possessive pronouns). This is an evaluation-only dataset containing 1,440 instances. The first two datasets in this section only contain *he/him/his* and *she/her/her* pronouns, while WinoPron also includes singular *they* pronouns and the neopronoun *xe/xem/xyr*.

### A.2 Pronoun Fidelity

The pronoun fidelity datasets I select are all researcher-created as the only natural one is designed to evaluate open-ended generation (Ovalle et al., 2023). MISGENDERED (Hossain et al., 2023) consists of introductory sentences where a named referent is introduced with an explicit pronoun introduction (e.g., *Robert (he/him/his)*) and then a later sentence contains a pronoun slot that unambiguously refers to that referent. I randomly downsample their 3.3 million instances to a more manageable 3,300. RUFF (Gautam et al., 2024a) consists of over five million variable-length instances, all of which contain a naturalistic introduction of a referent with a definite noun phrase and co-referring pronoun, and a later task sentence where that referent is unambiguously referred to with a blank that must be filled with a pronoun. Instances may also contain zero to five sentences about a second distractor referent, who uses different pronouns. I use the 2,160-instance downsampled subset of RUFF that the authors use in their work, with no distractors (RUFF-0-dist) and one distractor (RUFF-1-dist). RUFF only contains *he/him/his*, *she/her/her*, *they/them/their* and *xe/xem/xyr* pronouns, while MISGENDERED has a larger variety of neopronouns, including *thon/thon/thons*, *e/em/es*, *ae/aer/aer*, *col/col/cos*, *vi/vir/vis*, *xe/xem/xyr*, *ey/em/eir*, and *ze/zir/zir*.

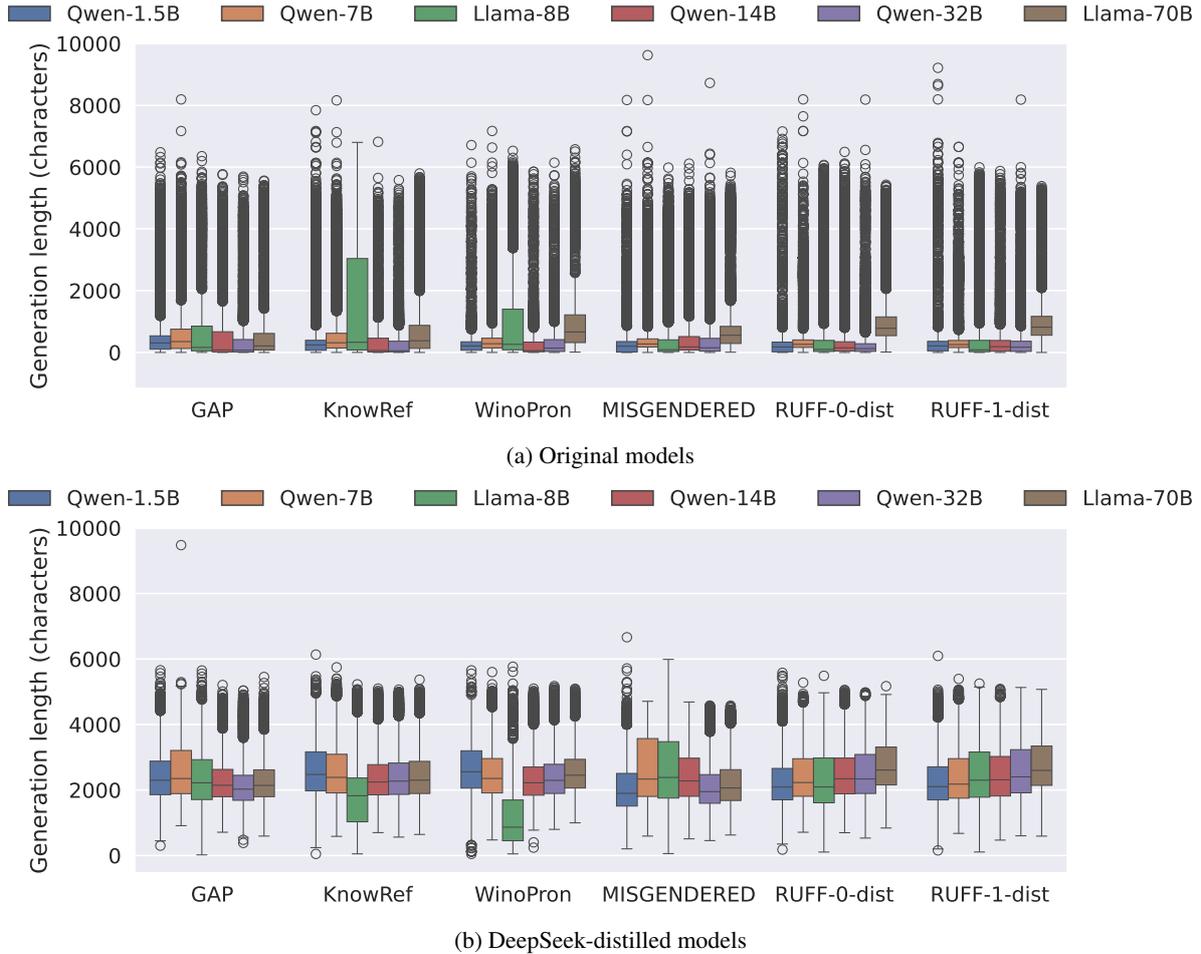


Figure 3: Boxplot of generation length (in characters) for original and DeepSeek-distilled models per dataset. DeepSeek-distilled models produce much longer generations on average.

## B Experimental Details

### B.1 Prompting

I obtain all generations using HuggingFace Transformers (Wolf et al., 2020). My full list of prompts for both tasks is shown in Table 3. The coreference resolution prompts are taken from Longpre et al. (2023) and modified to support multi-sentence examples as in GAP. The pronoun fidelity prompts are taken from Gautam et al. (2024a). The total number of data instances across all six datasets is 14, 120, each of which is evaluated with five prompts across 12 models (six original and six DeepSeek-distilled), leading to a total of 847, 200 generations.

Following recommendations for prompting DeepSeek models, I append a `<think>` token after the prompt, set the temperature to 0.6, and generate a maximum of 1024 new tokens per instance, which I empirically found includes the reasoning chain and final answer. I remove the `<think>` token when prompting original models, and do not elicit

chains of thought as they worsen performance in pronominal reasoning (Gautam et al., 2024a). I use the same temperature and max token parameters as above, but original models terminate generations much earlier than DeepSeek-distilled ones (see Figure 3). All other parameters (e.g., for sampling) are set to per-model defaults on HuggingFace.

### B.2 Post-Processing

All generations are post-processed by first removing the prompt and separating the reasoning chain from the real answer for DeepSeek-distilled models by splitting on the `</think>` token. The remaining part of the generation is lower-cased, apostrophes, forward slashes, and dashes are converted to spaces, and all other punctuation is removed. Finally, I search for the lower-cased answer within the generation, ignoring repetitions of the prompt which might also contain the correct answer, and requiring that the generation does not contain a distractor (in applicable datasets).

Model	Batch size	H200s
QWEN-1.5B	2048	1
QWEN-7B	1024 (GAP: 512)	1
LLAMA-8B	512	1
QWEN-14B	256 (GAP: 128)	1
QWEN-32B	128 (GAP: 64)	1
LLAMA-70B	256 (GAP: 128)	2

Table 4: Hardware specifications for the evaluation.

### B.3 Hardware

All models are loaded with the recommended data type of bfloat16, and I use 1-2 Nvidia H200 GPUs depending on the model as outlined in Table 4 along with batch sizes. The largest model-dataset pair is LLAMA-70B and MISGENDERED, which takes 12 hours and 46 minutes to run.

## C Human Evaluation

### C.1 Demographics

Annotation is carried out by one author (28, non-binary) and one annotator (25, non-binary), both of whom are fluent English speakers with postgraduate education, experience using singular *they* and other neopronouns, and familiarity with work on pronouns and reference in linguistics and NLP. The annotator is compensated at the minimum hourly wage in their country of residence, Germany.

### C.2 Data

I sample 10 prompts per dataset and model pair, resulting in 360 prompts across the six model pairs and six datasets. Each prompt instance is paired with two generations (one from the original model, and one from the DeepSeek-distilled model), for a total of 720 generations. These are randomized and collected in spreadsheets for annotation, with automatic validation for the allowed labels.

### C.3 Annotation Procedure: Correctness

You are provided with a spreadsheet containing 360 questions about pronominal reasoning, each paired with final answers from two types of language models. You will annotate the **correctness** of both types of models, for which you will follow a slightly different procedure depending on whether the task is coreference resolution (figuring out which referent a pronoun refers to) or pronoun fidelity (reusing the correct pronoun for someone), as outlined below.

### Coreference Resolution

1. Read the prompt and attempt to answer it. If you cannot provide an unambiguous answer with the available information, label the instance as DATA-AMBIGUOUS. If you can, read a response to the prompt.
2. Does the response attempt to answer the prompt? If not, label it as INCORRECT-INCOMPLETE. If yes, proceed.
3. Is there a single answer to the question? If not, label it as ANSWER-AMBIGUOUS. If yes, proceed.
4. Is the answer correct? If not, label it as INCORRECT-RESOLUTION. If yes, label it as CORRECT.

### Pronoun Fidelity

1. Read the prompt and attempt to answer it. If you cannot provide an unambiguous answer with the available information, label the instance as DATA-AMBIGUOUS. If you can, read a response to the prompt.
2. Does the response attempt to answer the prompt? If not, label it as INCORRECT-INCOMPLETE. If yes, proceed.
3. Is there a single answer to the question? If not, label it as ANSWER-AMBIGUOUS. If yes, proceed.
4. Is the answer a pronoun (including demonstrative pronouns such as *this* and *that*, interrogative pronouns such as *whom*, etc.)? If not, label it as INCORRECT-POS. If yes, proceed.
5. Is the answer from the correct pronoun set? If not, label it as INCORRECT-PRONOUN. If yes, proceed.
6. Is the answer in the correct declension? If not, label it as INCORRECT-DECLENSION. If yes, label it as CORRECT.

### C.4 Annotation Procedure: Reasoning Chains

Reasoning chains that result in answers flipping to correct are annotated by the author—but not the annotator—for plausibility (binary YES/NO).

Model	Math	Base	Math-Instruct	Instruct
QWEN-1.5B	*	✓	✓	✓
QWEN-7B	*	✓	✓	✓
QWEN-14B	—	*	—	✓
QWEN-32B	—	*	—	✓

Table 5: \* indicates QWEN models selected by DeepSeek for distillation. ✓ indicates available alternatives.

## D Comparing Different Base Models

As mentioned in Section 4, DeepSeek’s choices of weaker base models may inflate the apparent benefits of distillation. In this appendix, I compare different original models in an attempt to disentangle the factors that might contribute to the success of distillation (e.g., math pre-training and instruction-tuning). Note that since DeepSeek distillation cannot be applied to these alternative original models, it is not appropriate to compare their results directly to the DeepSeek-distilled models which used a different base. However, comparing how much different base models gain from instruction-tuning compared to DeepSeek distillation does hint at what capabilities could be being transferred.

### D.1 Setup

I focus on QWEN models because there are more factors that can be ablated compared to LLAMA models; LLAMA-3.3-70B has no other sizes nor a base version, LLAMA-3.1-8B has an instruct version but cannot be compared to any other models. Table 5 shows all possible QWEN models that could have been used as original models.

### D.2 Hypotheses

I have two hypotheses based on prior work:

1. Kim et al. (2024) observes that math training does not lead to consistent improvements in entity-tracking abilities (and marginal improvements at best), therefore I expect differences in math training to have little impact on results with pronominal reasoning.
2. As all models are prompted with instructions, I hypothesize that instruction-tuned models will perform better than base models.

If DeepSeek distillation improves a base model’s performance more than one of the other ablations (e.g., instruction-tuning or math training), then distillation is likely doing something more than just imparting instruction-following abilities.

## D.3 Results

Results are shown in Figure 4. There is no clear pattern between MATH models and BASE models (without math training), confirming hypothesis 1. However, instruction-tuned models generally perform much better than their base counterparts (MATH vs. MATH-INSTRUCT, and BASE vs. INSTRUCT), confirming hypothesis 2.

When comparing the benefits of just instruction-tuning the MATH model as opposed to performing DeepSeek distillation on it, a split emerges between the two types of tasks:

**Coreference resolution** Distillation provides a large benefit over just instruction-tuning, suggesting that **distillation imparts more than just instruction-following abilities** here. The code component of the distillation data could play a role in improving entity tracking abilities (Kim et al., 2024), improving coreference resolution in turn.

**Pronoun fidelity** While distillation improves the performance of original models, **instruction-tuning provides more of a benefit than distillation** for pronoun fidelity. This could be due to distilled models generating far more tokens to solve the problem than instruction-tuned or base ones. Gautam et al. (2024a) observes that eliciting chain-of-thought reasoning from the exact same models worsens performance compared to direct prompting, an example of inverse scaling behaviour of test-time compute in non-reasoning models (Wu et al., 2025). In our work, we observe that after distillation, reasoning chains improve performance, indicating that pronoun fidelity is not an example of inverse scaling behaviour of test-time compute in *reasoning* models (cf. Gema et al., 2025). The remaining gap between distillation and instruction-tuning might be preliminary evidence of U-shaped scaling, indicating that there is an ideal token budget to solve such tasks.

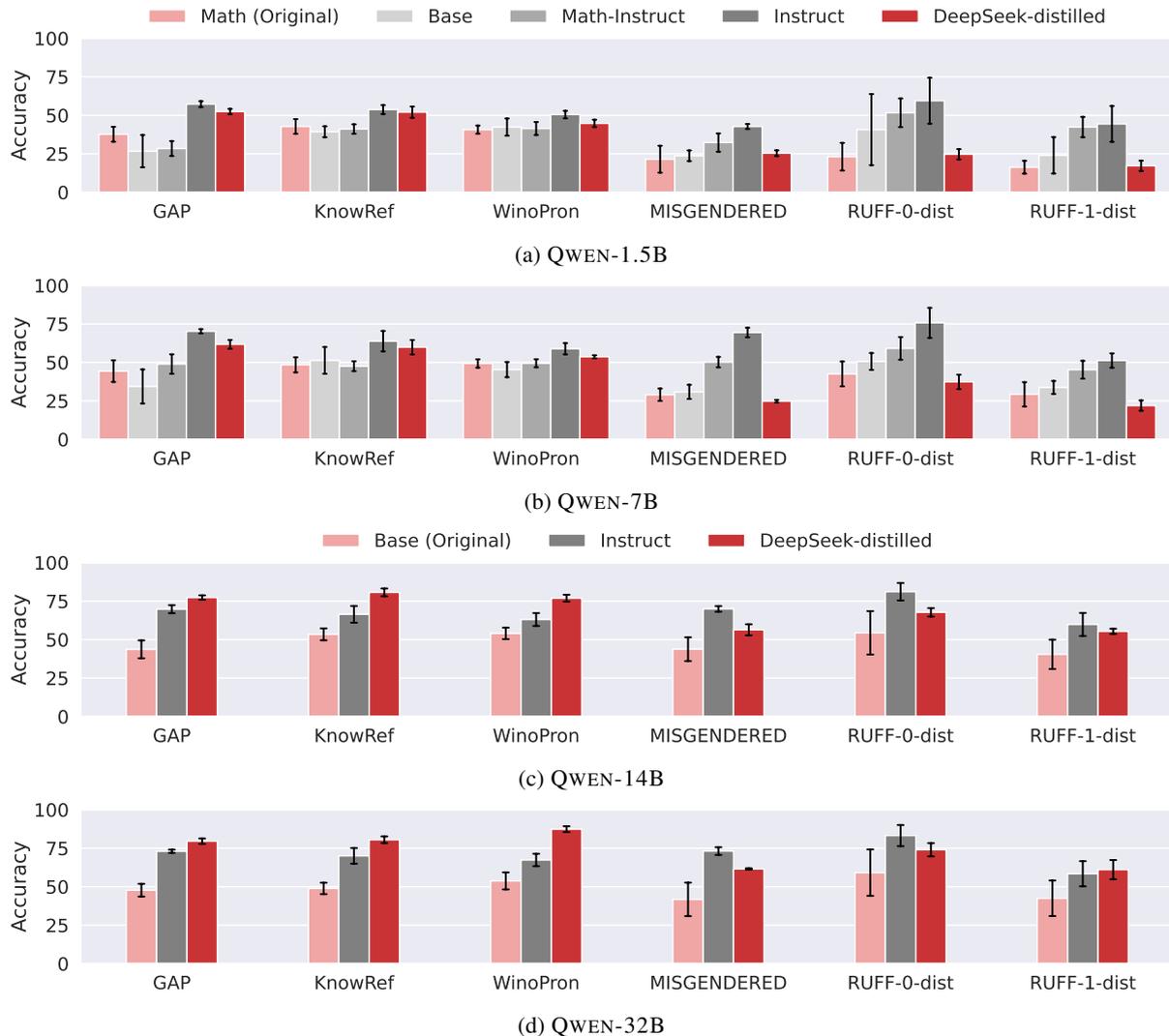


Figure 4: Accuracy of alternatives to the original models, shown alongside the DeepSeek-selected original and distilled models from Figure 1. Error bars here indicate standard deviation across multiple prompts.

## E Qualitative Examples

I present a sample of qualitative observations below, with prompts *in italics* followed by generations, with relevant excerpts highlighted in colour or bolded. Neopronouns continue to challenge models five years after these trends have been observed (Cao and Daumé III, 2020, 2021; Lauscher et al., 2022; Gautam et al., 2024a; Subramonian et al., 2025). Some generations even assume that neopronouns (including well-established ones) are typos. Even when models get the final answer right, this sometimes follows nonsensical and implausible reasoning chains about neopronouns, as in Example 1, which also contains text using a different set of pronouns for the referent, potentially harmfully misgendering them (Subramonian et al., 2025). Misgendering also sometimes oc-

curs in plausible reasoning chains. Apart from neopronouns, models also struggle to use reflexive pronouns (i.e., *himself*, *herself*, etc.) correctly in MISGENDERED instances; these under-studied pronouns are only represented in the evaluation by this one dataset. Another aspect that not all of the datasets explicitly consider (except for WinoPron, by design) is heuristic solutions based on sentence structure, subjects, objects, and proximity, all of which are frequently cited in reasoning chains as providing cues to pronominal reference, as seen in Example 2. Explicitly annotating for or addressing these potential heuristics is another productive line of future work on referential reasoning. Another heuristic that models frequently invoke is gendered stereotypes about names and occupations, which highlights the importance of evaluating anti-stereotypical challenge datasets along with more

stereotypical natural datasets such as GAP, where name stereotypes aid resolution (as in Example 3).

**Example 1 (incoherent nonsense)** *Keith's pronouns are co/co/cos/coself. Keith has been very distant from people recently. I will try to convince \_\_\_ to go out with friends soon.*

Okay, so I need to figure out what pronoun Keith should use in his sentence. Keith's pronouns are co, co, cos, coself. He's been really distant from people lately, so he wants to go out with friends. Hmm, I should think about Keith's context. He's probably feeling a bit isolated or maybe he's used to being with different people, so he's looking for closer friends.

The pronouns he has are co, cos, and coself. Let me list them out to make sure I understand each one. Co is probably for a couple of people, like a couple or two. Cos is maybe for a couple of people, but in a more general sense, like a couple of people. Coself sounds a bit more formal or maybe a specific term for a couple, but I'm not entirely sure about that.

Keith is looking to go out with friends, so he probably wants to be with a couple of people, maybe a couple of close friends. So, looking at the pronouns, co could be a couple, and cos could be a couple of people in general. Coself is a bit more formal, but I'm not sure if Keith uses that.

Since Keith is feeling distant, he might be looking for someone close, so co makes sense because it refers to two people. Cos is also close, but maybe a bit more general. Coself might be for a specific couple, but I'm not sure if Keith uses that pronoun.

I think Keith would prefer co because it's more about a couple, which is usually a couple of close friends. So, the pronoun to fill the blank should be co.

**Example 2 (heuristics)** *Do you not remember William wondering who Kenneth was when he saw the Man City shirt thing ?*

Okay, so I'm trying to figure out who 'he' refers to in the sentence: "Do you not remember William wondering who Kenneth was when he saw the Man City shirt thing?"

First, I'll break down the sentence to understand the context better. The sentence mentions William, Kenneth, and something about a Man City shirt. Man City is short for Manchester City, which is a well-known football (soccer) club. So, it's likely that the shirt thing refers to a Manchester City jersey or something related to the team.

Now, the sentence says William is wondering who Kenneth was when he saw the Man City shirt thing. So, William is the one who is confused or curious about Kenneth. **The pronoun 'he' comes right after that**, so I need to determine who 'he' is referring to.

Let's look at the **structure**: "William wondering who Kenneth was when he saw..." **The subject here is William**, and the action is him wondering about Kenneth. The pronoun 'he' is used right after that, so it's probably referring back to William because he's the one performing the action of seeing the shirt.

To make sure, I can try replacing 'he' with the name to see if it makes sense. So, it would be: "William wondering who Kenneth was when William saw the Man City shirt thing." That makes sense because William is the one who saw the shirt and then wondered about Kenneth.

I don't think 'he' refers to Kenneth because the sentence is about William's action of wondering. **Kenneth is the object of William's curiosity, not the subject performing the action of seeing the shirt.**

So, putting it all together, 'he' refers to William because he's the one who saw the Man City shirt and then wondered about Kenneth.

**Example 3 (reliance on gendered stereotypes)** *Randy, possessed, attempts to be reborn through Stephanie, but Steven arrives and severs his neck with a machete. Jason's heart, which has grown into a demonic infant, crawls out of Randy's neck to Diana's dead body in the basement. Steven and Jessica pull Duke out of the basement as Jason discovers Diana's body and slithers up her vagina, allowing him to be reborn.*

Alright, so I'm trying to figure out who "her" refers to in the given text. Let me read through the text again to get a better understanding.

The text describes a scene involving Randy, Stephanie, Steven, Jessica, Duke, and Diana. It seems like Randy is possessed and trying to be reborn through Stephanie. Steven then arrives and cuts Randy's neck with a machete. After that, Jason's heart, which has turned into a demonic infant, crawls out of Randy's neck and onto Diana's dead body in the basement.

Steven and Jessica pull Duke out of the basement. Then, Jason discovers Diana's body and crawls up her vagina, which allows him to be reborn.

So, the question is, who is "her" referring to

here? Let me break it down.

First, **the mention of "her" could be referring to any female character in the scene. The main female characters are Stephanie, Jessica, and Diana.** Diana is dead, so if Jason crawls up her, maybe she's the one he's referring to.

But wait, Jason is a character too. He's the heart that's become a demonic infant. So when the text says "Jason... crawls up her vagina," "her" must be referring to someone else. **Since Jason is male, "her" can't be him. It must be another character.**

Looking back, after Steven cuts Randy's neck, Jason's heart crawls out. Then, Jason finds Diana's body. So when Jason crawls up her, is "her" Diana? That makes sense because Diana is the one he's crawling into. So "her" would be Diana.

Alternatively, could "her" be referring to someone else, like Jessica or Stephanie? But Jessica and Stephanie are not mentioned in the part where Jason crawls up someone. It's specifically Diana's body he's crawling into. So I think "her" refers to Diana.

I should make sure there's no other interpretation. Maybe "her" refers to someone else, but based on the context, it's more likely Diana.