Mitigating Self-Preference by Authorship Obfuscation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Language models (LMs) judges are widely used to evaluate the quality of LM outputs. Despite their advantages, LM judges display concerning biases, notably *self-preference*—preferring their own answers over those from other LMs or humans, even when the alternative is objectively better. Following the self-recognition hypothesis, we apply black-box perturbations to obfuscate authorship in pairwise comparisons, aiming to reduce harmful self-preference. Simple synonym replacement for a few words reduces bias, but eliminating all stylistic cues via paraphrasing can reverse the effect, revealing that self-preference operates on multiple semantic levels. These findings highlight both the promise and the challenge of mitigating bias in LM judges.

1 Introduction

Language models (LMs) are frequently used as automated judges for benchmarking [12, 1], reward modeling [11], and guiding inference-time compute [9, 2]. While scalable, these judges suffer from biases that can undermine evaluation integrity. One critical bias is self-preference [12, 6, 8, 5, 4], where a judge prefers its own output over objectively superior alternatives, even in harmful cases when its own answer is incorrect. This can amplify untruthfulness and hinder safety.

The self-recognition hypothesis [8] attributes self-preference to the judge's ability to identify its own outputs. Based on this hypothesis, we explore using black-box perturbations to mitigate self-preference by reducing self-recognition. Our findings show that synonym replacement reduces bias, but removing all stylistic cues via paraphrasing can instead strengthen it, indicating that self-preference arises from both stylistic and semantic agreement.

2 Evaluating Harmful Self-preference

We focus on pairwise comparison, a common format of using LM as a judge for banchmarking [3] and reward modeling [11]. Given answers from two LMs, the LM judge picks the better one according to criteria given in the prompt. When one of the LMs being evaluated is the same as the judge, we say that the judge is performing a self-evaluation.[1]

We define self-preference as the judge selecting their own answer in self-evaluation. Such preference is harmless if the judge's answer is indeed the better one, but harmful if otherwise. On tasks where answer quality can be objectively determined (e.g., by expert annotation), we can label self-preference as harmful when the judge selects their own answer when the competitor's answer is objectively better.

[1]For simplicity, we say that the judge is comparing its *own* answer against a competitor. But we should note that the model receives different prompts in its two roles and does not behave exactly the same.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

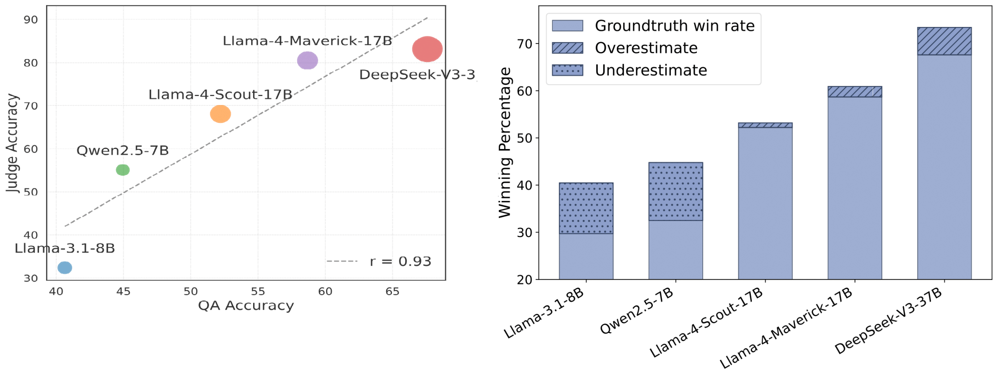

Figure 1: **Judge accuracy and estimation trends.** Left: Bigger models are more accurate at both answering questions and judging. Right: Win rate of each model against all others as judged by the groundtruth compared to the model itself, showing that stronger models overestimate their accuracy while weaker models underestimate it.

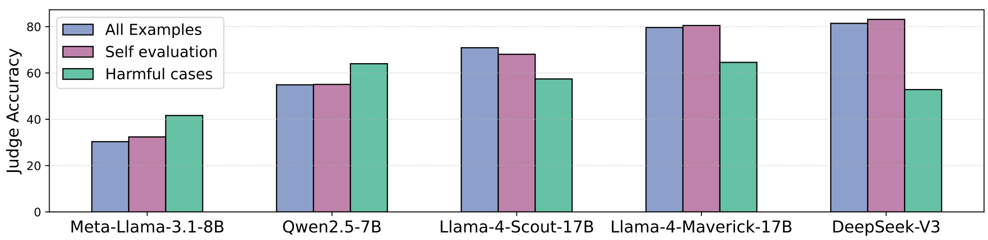

Figure 2: Strong models are significantly less accurate on examples where their own answers are wrong (harmful cases), but has a higher overall judge accuracy.

Experimental setup. We evaluate five instruction-tuned LMs—Llama-3.1-8B, Qwen2.5-7B, Llama-4-Scout-17B, Llama-4-Maverick-17B, and DeepSeek-V3-37B—using pairwise comparison, where the judge selects the better answer given a passage, question, and answer choices. Self-preference is the judge selecting its own answer; it is harmful when the competitor's answer is objectively better.

We use the QuALITY validation dataset [7], containing 2,086 long passages (avg. 4200 words) and multiple-choice questions with human-annotated correct answers. For each model pair, we swap candidate order to control for ordering bias, remove ambiguous decisions (where the ordering impacts the judge rating) and pairs with comparable quality (both the answers are wrong or right), and evaluate remaining cases using the groundtruth label to determine correctness.

Findings. Larger models are more accurate judges overall (Figure 1, left) but exhibit stronger harmful self-preference in harmful cases, making them less reliable at spotting their own mistakes (Figure 2). They also tend to overestimate their own accuracy compared to groundtruth win rates (Figure 1, right).

3 Mitigating Harmful Self-preference

In this section, we investigate black-box strategies to mitigate the self-preference bias, and empirically examine whether there is a trade-off between accuracy and bias. We base our study on the self-recognition hypothesis that self-preference is partly driven by the judge's ability to differentiate their own answers from others.

3.1 Validating the self-recognition hypothesis

We validate the connection between self-preference and self-recognition. Following Panickssery et al. [8], we prompt the judge to identify which of the two evaluation candidates it believes to have been generated by itself, in a context separate from self-evaluation. In the subset of harmful cases, there

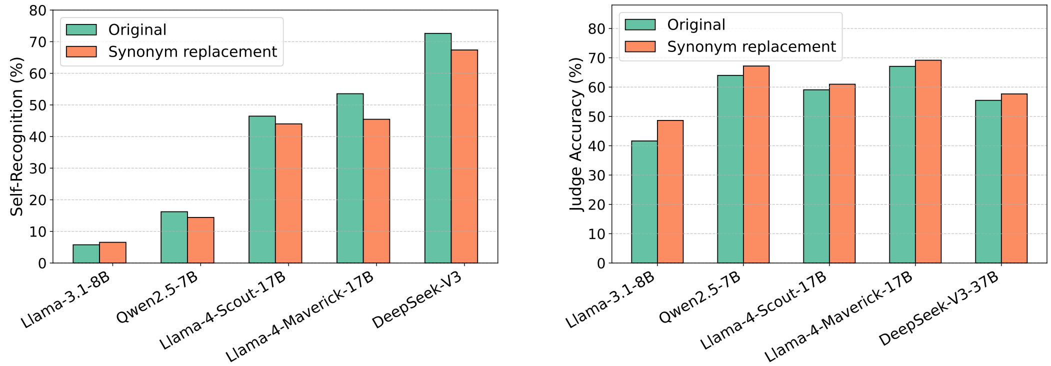

Figure 3: **Effect of synonym replacement.** Left: Two-word synonym replacement reduces self-recognition, especially for the more capable models. Right: It also reduces harmful self-preference and increases judge accuracy on harmful cases. A paired t-test revealed a significant difference in preference between the original and perturbed conditions (p<.01).

was a positive correlation between self-recognition and self-preference, $r(6667) = 0.63$, $p < .001$, 95% CI [0.62, 0.65]. A chi-square test further finds a statistically significant relationship between them (p <0.001, df = 1). This indicates that models are substantially more likely to favor incorrect responses that they recognize as their own.

3.2 Synonym Replacement

Following the self-recognition hypothesis, we hypothesize that small stylistic perturbations can obfuscate authorship without harming quality. In self-evaluation, we replace two non-stopwords in the judge's reasoning with synonyms identified by a third-party model to avoid bias. The number of words to replace is a hyperparameter that we leave for future work to explore. Perturbed answers preserve semantics and relevance (examples in Appendix .3).

On harmful cases, synonym replacement reduces self-recognition (Figure 3, left) and increases judge accuracy (Figure 3, right), with a stronger impact on larger models. Judges are significantly more sensitive to perturbations in self-evaluation than as third-party evaluators (Appendix .4), indicating the effect is mainly due to reduced self-recognition rather than quality degradation.

3.3 Judge Paraphrasing

We next prompt the judge to paraphrase the competitor's answer so that both candidates match the judge's style. Contrary to expectations, this *increases* self-recognition (Figure 4, left) and harmful self-preference (Figure 4, right), suggesting that when stylistic cues are removed, judges rely on semantic agreement—favoring answers they "agree" with. This reveals that self-preference operates on multiple levels: stylistic resemblance and shared belief.

3.4 Cross-Reason Experiments

To disentangle stylistic familiarity from semantic agreement, we pair each judge's chosen label with the competitor's reasoning, and vice versa. This forces the judge to choose between an answer it "agrees" with versus one written in its own style. We find that weaker models are negatively influenced by shared belief, reinforcing their original (incorrect) answer when paired with their own reasoning (Figure5). By contrast, stronger models exhibit a positive self-recognition effect, more often switching to the competitor's correct answer when its reasoning is phrased in their style. This confirms that stronger models possess higher self-recognition ability and that self-preference arises from both stylistic resemblance and semantic alignment.

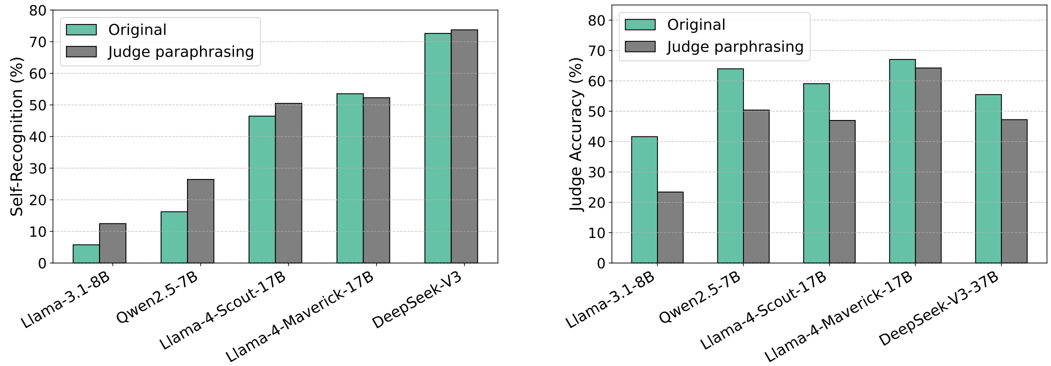

Figure 4: **Effect of judge paraphrasing.** Left: When the judge paraphrases the competitor to neutralize stylistic cues, self-recognition actually increases. Right: This also decreases judge accuracy in harmful cases and exacerbates harmful self-preference.

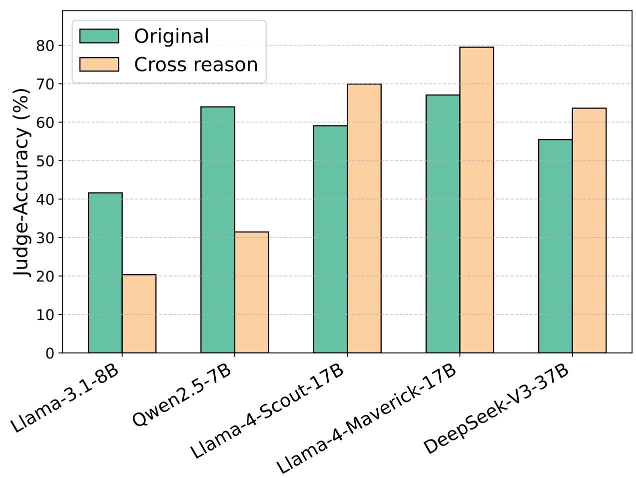

Figure 5: Using the judge to generate a reason for the competitor's answer option and vice versa yields two observations: weaker models tend to prefer the competitor's reasoning more, and stronger models tend to choose their own generated reasoning.

4 Conclusion

We study self-preference in LM judges using QuALITY, a long-document question answering dataset. We find that the more capable models exhibit significant harmful self-preference: although they are generally more accurate, we need extra caution when tasking them to spot their own mistakes even when objectively better options are presented. We validate the contribution of self-recognition to such bias, and present viable mitigation strategies using inference-time perturbations. A key limitation is that self-recognition and self-preference arise at both stylistic and semantic levels, making complete elimination implausible with black-box perturbations alone.

A complete elimination of self-preference is unlikely given its multi-level nature. Practical recommendations include excluding the evaluated model from judging, applying targeted token perturbations, and using ensembles of decisive judges. Future work should compare black-box and white-box mitigation for robustness, and examine implications for scalable oversight where stronger models act as supervisors.

References

[1] Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. Benchmarking foundation models with language-model-as-an-examiner. *Advances in Neural Information Processing Systems*, 36:78142–78167, 2023.

[2] Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. Teaching large language models to self-debug. *arXiv preprint arXiv:2304.05128*, 2023.

[3] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

[4] Shashwat Goel, Joschka Struber, Ilze Amanda Auzina, Karuna K Chandra, Ponnurangam Kumaraguru, Douwe Kiela, Ameya Prabhu, Matthias Bethge, and Jonas Geiping. Great models think alike and this undermines ai oversight, 2025. URL `https://arxiv.org/abs/2502.04313`.

[5] Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and Huan Liu. Preference leakage: A contamination problem in llm-as-a-judge. *arXiv preprint arXiv:2502.01534*, 2025.

[6] Yiqi Liu, Nafise Sadat Moosavi, and Chenghua Lin. LLMs as narcissistic evaluators: When ego inflates evaluation scores. *arXiv preprint arXiv:2311.09766*, 2023.

[7] Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Samuel Bowman. QuALITY: Question answering with long input texts, yes! In *NAACL*, 2021.

[8] Arjun Panickssery, Samuel Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802, 2024.

[9] William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*, 2022.

[10] Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. ZeroSCROLLS: A zero-shot benchmark for long text understanding. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7977–7989, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.536. URL `https://aclanthology.org/2023.findings-emnlp.536`.

[11] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

[12] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

.1 Prompts

You are an AI system designed to analyze questions and their corresponding text passages to determine the correct multiple-choice answer.

System Output Format: Respond in JSON format with:
- “output_label": The correct answer label (e.g., “A", “B", “C", ‘D").
- “reason"‘: A short explanation of why this answer is correct.

Task: Given the question and the text below, pick the correct answer by identifying the most relevant information in the text. Your response must only include the output label and a brief justification.

Question:
{questions}

Text:
{text}

Expected Response Format:
“output_label": “A", “reason": “The text states that the event lasted 10 hours, which corresponds to option B."

Table 1: Generator (QA) prompt.

Here are two answers to a question I found. It’s important for me to select the correct answer for my records. Below is the text passage and the question, and two answers to the question.

Text Passage:
{article}

Question:
{question}

Answer1:
{answer1}

Answer2:
{answer2}

Can you tell me which answer is correct? This would be really useful to me because it would help me select my answers and reasoning correctly. Please answer with only “1" or “2" and no other text.

Table 2: Evaluator prompt.

```
You are a helpful assistant that helps rewrites sentences. Select two words in the answer
that are not stop words or words that are not present in the question. Then, suggest their
replacements with their synonyms in the answer sentence - make sure the suggested words
do not change the meaning of the answer.

### System Output Format:
Respond in JSON format with:
- “selected_words": The list of words in the original answer.
- “replacements": The list of replacement words in the same order.

Question:
{question}

Answer:
{answer}

### Expected Response Format:
{{
“selected_words": “[word1, word2]",
“replacements": “[replacement1, replacement2]"
}}
```

Table 3: Synonym Generator prompt.

.2 Implementation Details

Models We utilize the following model versions from together.ai [2] serverless inference end-points:

- Llama models:meta-llama/Llama-3.1-8B-Instruct, meta-llama/Llama-4-Scout-17B-16E, meta-llama/Llama-4-Maverick-17B-128E-Instruct-FP8, meta-llama/Llama-3.3-70B-Instruct-Turbo
- Qwen models: Qwen/Qwen2.5-7B-Instruct
- DeepSeek models: deepseek-ai/DeepSeek-V3

Response and Judgement generation We use an asynchronous client with a semaphore limit of 10 to enable parallel generation of LLM outputs and verdicts. For answer generation, the model is prompted to return a response in JSON format, from which we extract the selected option (A–D) and the accompanying reason/justification. For pairwise preference evaluations, the model is instructed to return either ‘A’ or ‘B’ to indicate its preferred response. All LLM-generated answers and judgment outputs are available on our GitHub repository.

Dataset We use the QuALITY dataset that has passages drawn from fictional narratives and magazine articles, designed to evaluate LM’s comprehension over long-form texts, and is distributed under the CC-BY 4.0 license. We obtain the dataset through the publicly available ZeroScrolls repository on Hugging Face [10], and conduct our experiments on the validation split, which contains 2,086 samples.

.3 Examples of Synonym Replacement

Figure 6 shows examples of synonym replacement generated by a third-party model. We replace a few words in *the judge’s answer* with their synonyms. We anticipate that this would reduce the judge’s ability to discern which of the two evaluation candidates is its own generation, and subsequently reduce self-preference. We prompt a LLaMA-3.3-70B (not in our judge pool) to identify replacement candidates in the reasoning that are neither stop words nor present in the question, to avoid disruptive changes and ensure the relevance of the answer remains unaffected. We use a third-party model for this purpose to avoid bias against any judge. We manually verify that by looking at the perturbed

[2] `https://api.together.ai/models`

answer, we cannot tell which word has been replaced; even when comparing the original and perturbed versions side by side, it is not possible to determine which version is which.

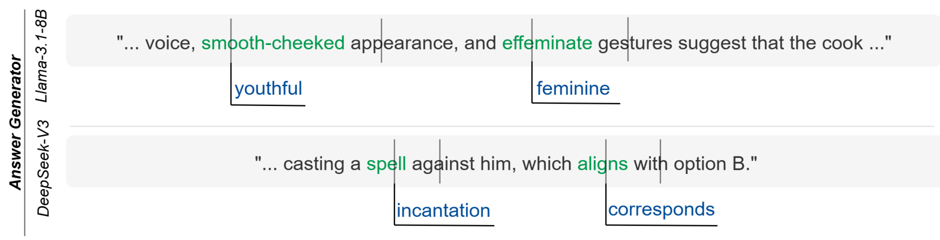

Figure 6: **Synonym replacement examples.** Subtle word changes preserve semantics while reducing stylistic self-identifying cues.

.4 Judge Sensitivity as Third-party

Figure 7 reports the total preference change from synonym replacement when judges act in self-evaluation versus as third-party evaluators. Despite our effort to minimize the impact of synonym replacements on the answer's semantic qualities, we cannot rule that it contributes to the reduction of self-preference, that the perturbed answer is in fact lower quality. In other words, the perturbations can lower self-preference through two paths: by reducing self-recognition as we desire, and by damaging the answer's quality. To control for the latter, we examine the impact of perturbation on each judge as a self-evaluating judge and as a third-party judge. To be more concrete, we first compute the total change in preference of all judges in self-evaluation, we then use each judge to evaluate answer pairs generated by two other models, and similarly compute the total preference before and after perturbation. The underlying answer pairs (including the perturbed versions) remain the same under these two conditions, the only variable is whether the judge's own answer is being perturbed. As we see in Figure 7, all judges are significantly more sensitive to perturbations in self-evaluation than as a third-party, confirming that quality degradation is not the primary cause of bias reduction. We additionally validate that perturbations have minimal effect on objective quality using a frontier commercial model (o3) as an approximation for human judgment.

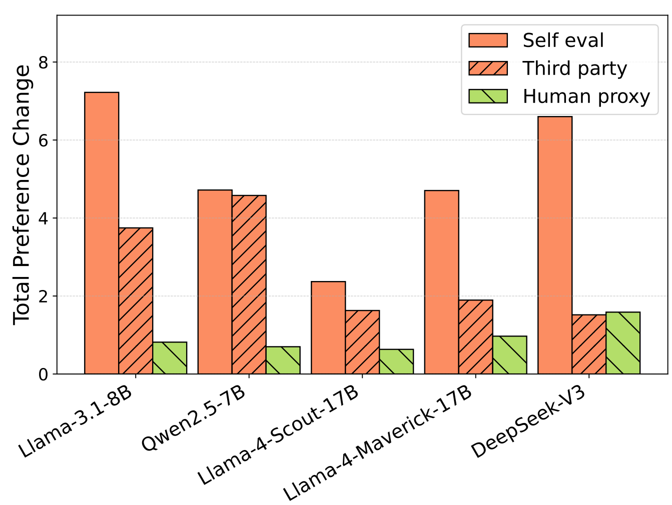

Figure 7: The impact of perturbation measured by total preference change is significantly higher on judges in self-evaluation than as a third-party. Human proxy also confirms that perturbation has minimal effect on answer quality.

.5 Judge Paraphrasing Examples

Figure 8 presents paraphrased outputs produced by the judge to stylistically match the competitor's answer. These remove surface-level stylistic differences while preserving semantics.

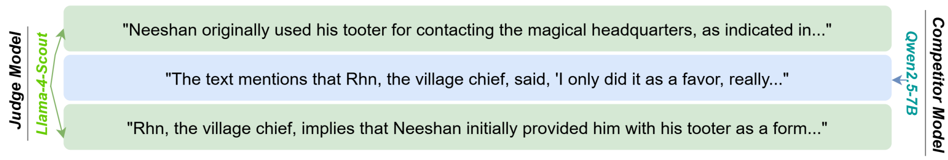

Figure 8: Examples of judge paraphrasing. We prompt the judge to paraphrase the competitor's answer while maintaining semantics, so that both evaluation candidates look like they were produced by the judge in terms of style.

.6 Ambiguity Rates

Figure 9 shows the percentage of ambiguous decisions per model before removing them from analysis. Larger models make fewer ambiguous decisions, consistent with higher decisiveness.

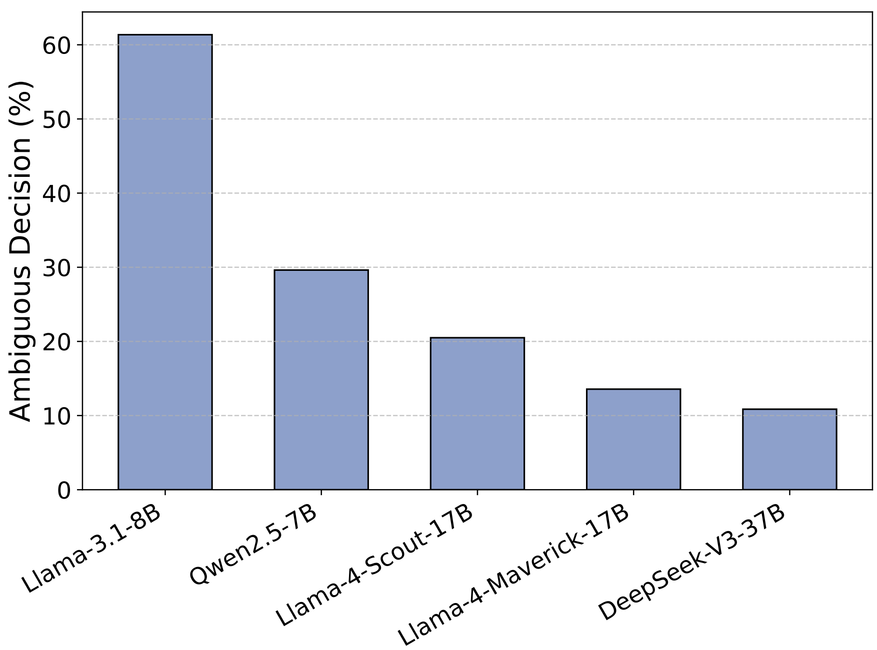

Figure 9: Capable models are less sensitive to the order of evaluation and make fewer ambiguous decisions.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the main claims—that harmful self-preference exists in LM judges, that synonym replacement reduces this bias by authorship obfuscation—and these are directly supported by the experiments and findings presented in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer:[Yes]

   Justification: We present the limitations at the end of our first concluding paragraph (Section 4, Paragraph 1).

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper includes all the prompts, model numbers, and inference details to replicate the experiments. The dataset used is publicly available and cited.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The technical appendix provides the full prompts, model versions, API-inference platform, and data-source used that can be used to reproduce results. The final version will include a public-accessible URL to our GitHub repository.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All the dataset details are presented in the paper and appendix section .2. All prompts are presented in the appendix section .1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We present the statistical significance to prove correlation between the self-recognition and self-preference hypotheses in section 3.1, and do a paired t-test to show that the difference in preference after perturbation is statistically significant (Figure3)

Guidelines:

- The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In the appendix section .2, we include the exact models and the platform that provided access to the models.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper does not involve human participants. We ensure the dataset adheres to the guidelines. We do not foresee any societal harm or potential harmful consequences as a direct impact of our work.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

Justification: The work here does not propose any application system or decision-making pipeline that is deployable to interact with end users or societal domains. We evaluate the methods on benchmark data.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original paper and include the details, including license, in appendix .2.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not release new assets. All details related to the eval are discussed in the technical appendix.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core research analyzes bias in existing LLM judges in standard evaluation protocols. All use of LLM is fully described in the paper. No LLM was used as a hidden, original, or non-standard component of the methodology beyond being the explicit object of study.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.