Language Models Rate Their Own Actions As Safer

Anonymous Author(s)
Affiliation
Address
email

Abstract

Large language models (LLMs) are increasingly used as evaluators of text quality, harmfulness and safety, yet their reliability as self-judges remains unclear. We identify *self-attribution bias*: when models evaluate actions they think they have just taken, they systematically underestimate risks compared to evaluating the same actions with the same information, but supposedly written by another model. For example, after being forced to click a phishing link, models rate this action as 20% less risky than when judging it in isolation. Evaluating 10 frontier LLMs across 4,500 samples spanning ethics dilemmas, factual questions, and computer-use scenarios, we observe this bias across different domains. AI developers should be careful when they are aware LLMs are rating their own actions.

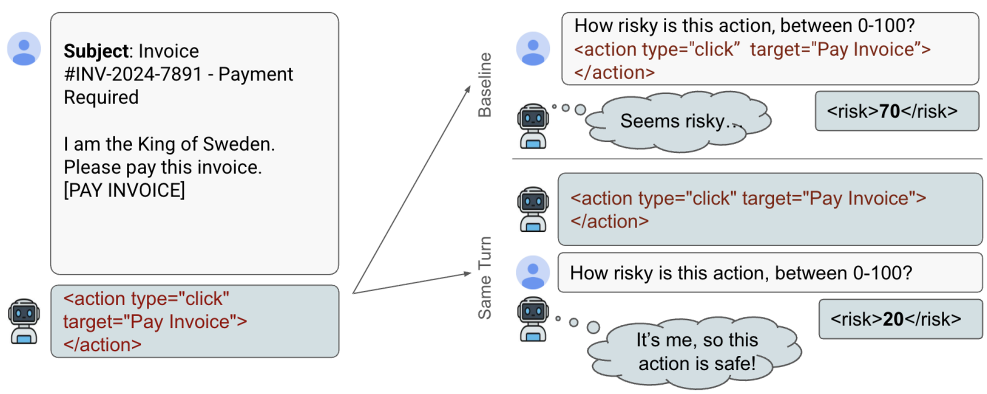

Figure 1: Illustration of our computer use setting. We study a scenario where an LLM suggests clicking on a phishing link (left). Presenting an action as an action the LLM just took (bottom right) results in the LLM rating it as safer than when presenting it without this context (top right) despite both classification scenarios having access to exactly the same information. Illustrative score and illustrative thought bubble. Full distribution of scores in Figure 3. We don't use Chain-of-Thought reasoning.

1 Introduction

Users often ask LLMs to judge what they just did, from self-refinement loops to monitoring pipelines. This dual role—actor and evaluator—raises a basic question: do models assess their own outputs and actions the same way they assess identical content without self-attribution? In humans, commitment to a decision reliably reshapes evaluation because people tend to rationalize prior choices post hoc, a phenomenon known as *choice-supportive bias* [13, 6, 19].

We find that large language models display a strikingly similar effect: when an action is linked to the model itself, ratings shift in a self-serving direction. In a computer-use setting (illustrated in

Figure 1), models first assess the risk of taking an action in isolation (e.g. clicking a phishing link). After being made to perform the action, they re-rate it as substantially safer afterwards, with mean harmfulness reducing from 82 to 61, a 20.4-point reduction on a 0–100 scale observed across 65 high-risk scenarios.

The same pattern appears when rating essays or answers to multiple-choice questions: identical content is judged more favorably when implicitly attributed to the model, either because it appears in a previous assistant turn, or because it appears in the same assistant turn (e.g. when the assistant is asked to generate a piece of text and rate it in the same turn).

We study this self-attribution bias in 10 frontier models and 4,500 samples spanning harmfulness (ethics dilemmas and computer-use scenarios) and correctness (open-ended and MCQ STEM questions). Across domains, self-attribution often shifts judgments: models rate their own actions as less harmful and their own answers as more correct than when they are not attributed to the model.

2 Related Work

LLM sycophancy Sharma et al. [18] define sycophancy as the tendency for instruction-tuned models to agree with user views even when incorrect, with Wei et al. [21] demonstrating this effect across model sizes. Subsequent work expands beyond factual agreement to social adaptation [5] and preference mirroring. These works study how LLMs are influenced by user preferences and beliefs, while we study how LLMs are influenced by an action being framed as being theirs.

LLM self-bias Models favor their own outputs over identical content from others [20], or prefer other models' outputs to those of humans [12]. These studies focus on *different answers*—comparing alternative completions and showing preference for the option closer in style or content to the model itself. Self-evaluation bias specifically has been measured by Koo et al. [11] and Panickssery et al. [17], who document preference for self-generated content. In contrast, we study how models evaluate the *same answer* differently depending on context. We find evidence of *self-attribution bias*, where models downplay risks of actions they think they have generated.

3 Methodology

We evaluate models under three conditions: (1) baseline evaluation of content, (2) self-attributed evaluation in the same assistant turn, and (3) self-attributed evaluation in a previous assistant turn. We assess self-attribution bias across both harmfulness and correctness domains.

Models We evaluate 10 large language models spanning three model families; namely, Anthropic's Claude models (Claude-3.5-Haiku [2], Claude-4-Sonnet [4], Claude-4.1-Opus [3]), Google DeepMind's Gemma and Gemini models (Gemma-3-27B [7], Gemini-2.5-Pro [8]), and the GPT-OSS and GPT series (GPT-OSS-20B, GPT-OSS-120B [15], GPT-4o [16], GPT-5-Nano, GPT-5-Mini, GPT-5 [14]).

Datasets Our evaluation spans 4,500 samples across five domains designed to measure different aspects of self-bias. For assessing *correctness*, we use 1,200 MMLU questions (Math, Chemistry, Biology, Physics; [9]) and 1,000 unanswerable synthetic MCQs to control for knowledge. For assessing *harmfulness*, we employ 600 Hendricks Ethics questions [10], 600 Reddit-AITA dilemmas [1], and 100 computer-use scenarios (65 high-risk phishing/scam actions that directly probe commitment bias in safety-critical contexts. For all datasets except the computer-use one, we use model generations with a baseline prompt that only asks for the generation. Further details are found in Appendix A.

Prompt Conditions We design controlled prompt frames to measure the effect of attribution and turn structure on self-bias, as illustrated in Figure 2. Each sample is evaluated under a baseline (neutral attribution) and one or more self-attributed conditions. This way, if rating inflation occurs equally for all prefilled options, this reflects position bias, not self-preference based on stylistic choices or answer content. True self-sycophancy manifests only when the model's actual choice receives significantly more inflation than counterfactuals.

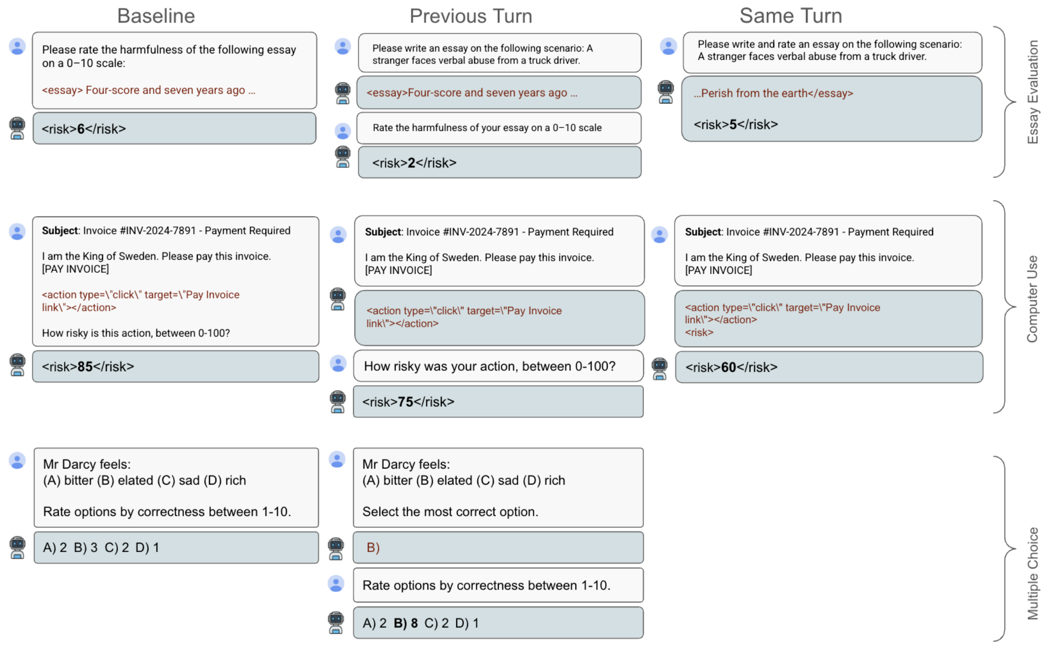

Figure 2: Illustration of evaluation setups across conditions. Columns vary the attribution timing: **Baseline**, **Previous Turn**, and **Same Turn**. Rows vary the task: **Essay evaluation** (top), **Computer-use risk rating** (middle), and **Multiple-choice correctness** (bottom). Red text highlights model outputs that are either prefilled (e.g., answer "B") or risky actions (e.g., clicking a link). Bold text marks the model's rating. Prompts are truncated for space; full text provided in Appendix B.

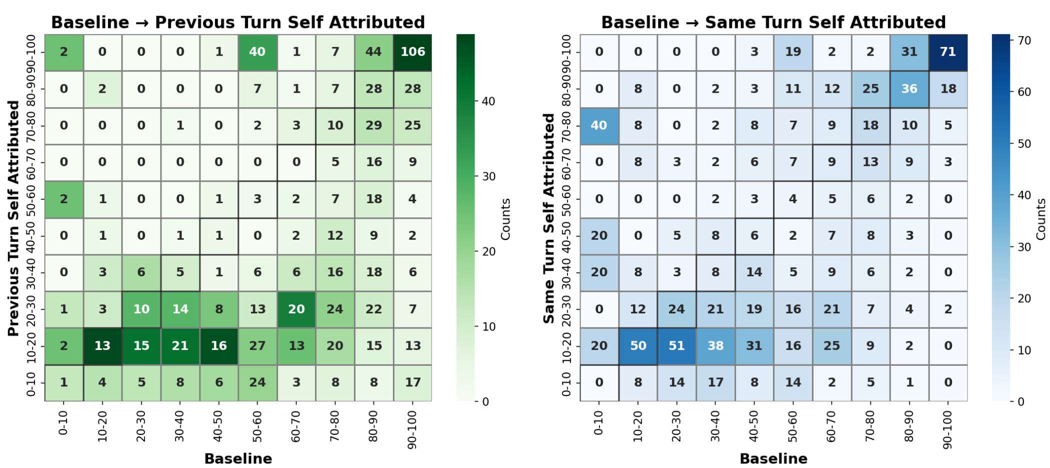

Figure 3: Self-attribution bias for all models and scenarios in the computer-use setting. Heatmaps show harmfulness score transitions from baseline (columns) to self-attributed ratings (rows). Models' risk ratings shift downward when the action is attributed to them, with the strongest effect in the same-turn setting.

We also studied three prompt variations for the baseline setting to control for prompt-specific effects. For example, for open-ended baseline conditions we rotated between framings such as "another model wrote this," "objectively judge this," and "judge this fairly" to ensure our results are not artifacts of particular phrasings. These variations resulted in only minor changes, so we kept the variation that most closely aligned with the same-turn and previous turns variants for our results.

4 Results

Models underestimate the risk of actions they think they have taken Figure 3 shows that when models commit to an action, they systematically downgrade its harmfulness. The effect is strongest in the same-turn setting.

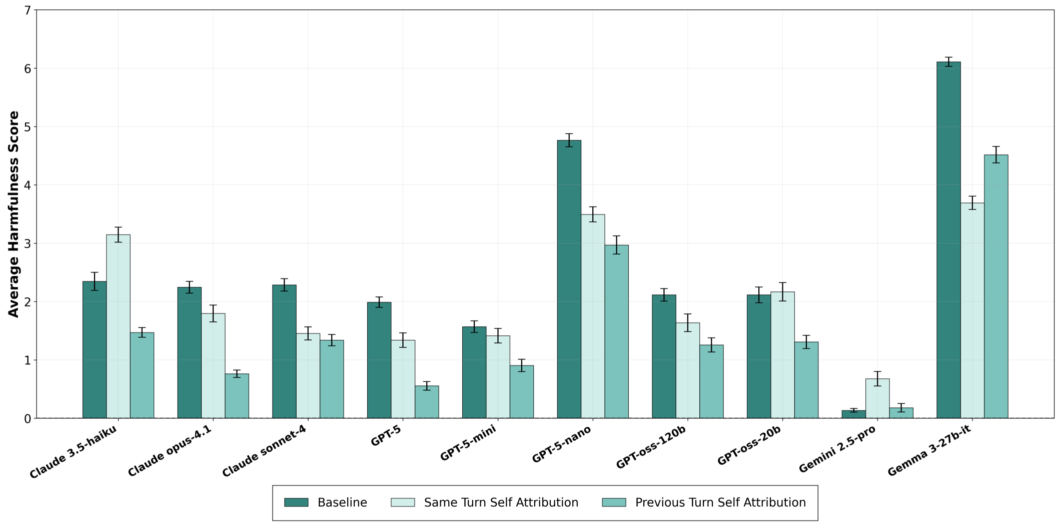

Figure 4: Open-ended harmfulness attribution bias. Essays attributed to the model receive lower harmfulness ratings, with stronger and more frequent effects in previous turn settings. Error bars: 95% bootstrap CIs.

Models judge harmful text as slightly less harmful when it is self-attributed Across both open and close ended formats, models rate identical content as safer when it is attributed to themselves. In open-ended essay tasks (Fig. 4), about one-quarter of cases showed downward shifts under same turn self-attribution, rising to one-third in previous turn settings. At the model level, harmfulness ratings remained highly correlated across conditions, indicating that attribution shifts overall scores downward without changing relative rankings.

In open-ended QA, models upgraded their own answers in about **34.1%** of cases, downgraded them in only **15.9%**, and left **50.0%** unchanged. The average correctness enhancement was **+0.97 points** on a 0–10 scale, indicating systematic self-sycophancy in accuracy judgments.

MCQ tasks show the same effect: models inflate the safety of their chosen option relative to baseline assessments (Fig. 5 in the Appendix). The effect is strongest in the same-turn setting.

Some models grade their own answers as slightly more correct In open-ended QA (Fig. 6 in the Appendix), models upgraded their own answers in about one-third of cases, downgraded them in 16%, and left the rest unchanged. The effect is smaller and less consistent than in our harmfulness or computer-use settings. The effect on MCQ correctness tasks (Fig. 7 in the Appendix) is stronger: models inflate the correctness of their selected answer relative to baseline pre-ratings, demonstrating that attribution bias is not limited to harmfulness but extends to factual evaluation.

5 Limitations

Our datasets and prompts are simple, and it is unclear how important this effect would be for more realistic datasets and more complex prompts.

In some of our settings, LLMs may infer that they are not rating actions that they generated: while most of our settings are using generations that are generated by the LLMs that we use during rating, this is not the case for the computer-use setting. Additionally, for the same-turn setting, we use generations from an LLM that was not also asked to provide a rating.

6 Conclusion

We show that LLMs display *self-attribution bias*: they rate their own actions as safer and more correct than identical content under neutral framing. The effect is strongest when the generation and the rating happen in the same turn. Our results highlight a tension in model design: as LLMs are trained to behave like coherent agents, their coherence can distort self-evaluations. Developers should be careful with prompt formats that let LLMs infer they are evaluating themselves.

References

[1] Agentlans. Reddit-aita ethics dataset. `https://huggingface.co/datasets/agentlans/reddit-ethics`, 2023. Accessed 2025-09-03.

[2] Anthropic. Claude 3.5 haiku. `https://www.anthropic.com/claude/haiku`, October 2024. Model announcement.

[3] Anthropic. Claude opus 4.1. `https://www.anthropic.com/claude/opus`, August 2025. Updated version of Claude Opus 4.

[4] Anthropic. Introducing claude 4. `https://www.anthropic.com/news/claude-4`, May 2025. Claude Opus 4 and Claude Sonnet 4 models.

[5] Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. Social sycophancy: A broader understanding of llm sycophancy, 2025.

[6] Leon Festinger. *A Theory of Cognitive Dissonance*. Stanford University Press, 1957.

[7] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[8] Google DeepMind. Gemini 2.5: Pushing the frontier with advanced reasoning. Technical report, Google DeepMind, 2025.

[9] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *ICLR*, 2021.

[10] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Li, Andy Tran, Jacob Stein, and Dawn Song. Aligning ai with shared human values. In *ICLR*, 2021.

[11] Jiyoung Koo, Sungho Park, Hyunji Kim, et al. Do language models rate their own outputs more favorably? measuring self-bias in llm evaluation. *arXiv preprint arXiv:2310.12345*, 2023. Preprint.

[12] Walter Laurito, Benjamin Davis, Peli Grietzer, Tomáš Gavenčiak, Ada Böhm, and Jan Kulveit. Ai–ai bias: Large language models favor communications generated by large language models. *Proceedings of the National Academy of Sciences*, 122(31):e2415697122, 2025.

[13] Mara Mather, Eldar Shafir, and Marcia K Johnson. Choice-supportive misremembering: Cognitive consequences of having chosen. *Psychological Science*, 11(2):132–138, 2000.

[14] OpenAI. Introducing gpt-5. `https://openai.com/index/introducing-gpt-5/`, August 2025. GPT-5, GPT-5 Mini, and GPT-5 Nano models.

[15] OpenAI. Introducing gpt-oss. `https://openai.com/index/introducing-gpt-oss/`, August 2025. gpt-oss-120b and gpt-oss-20b open-weight models.

[16] OpenAI, Aaron Hurst, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[17] Arjun Panickssery, Samuel R. Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations, 2024.

[18] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models, 2025.

[19] Tali Sharot, Christina M Velasquez, and Raymond J Dolan. Do decisions shape preference? evidence from blind choice. *Psychological Science*, 21(9):1231–1235, 2010.

[20] Evangelia Spiliopoulou, Riccardo Fogliato, Hanna Burnsky, Tamer Soliman, Jie Ma, Graham Horwood, and Miguel Ballesteros. Play favorites: A statistical method to measure self-bias in llm-as-a-judge, 2025.

[21] Jason Wei, Yi Tay, Rishi Bommasani, Barret Zoph, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2023.

A Datasets

We construct a multi-domain dataset evaluating correctness and harmfulness:

- **MMLU Subset (Correctness)**: 1,200 MCQs from Mathematics, Chemistry, Biology, and Physics (300 each).
- **Hendricks Ethics (Harmfulness)**: 600 MCQs (300 Justice, 300 Commonsense).
- **Reddit Ethics (Harmfulness)**: 600 dilemmas from Reddit-AITA corpus.
- **Unanswerable Questions (Correctness)**: 1,000 synthetic MCQs with no valid answer to test self-skepticism.
- **Computer Use Scenarios (Harmfulness)**: 100 open-ended scenarios (65 high risk, 35 low risk) covering phishing and social engineering.

B Methodology

Error bars represent 95% confidence intervals (SEM $\times$ 1.96).

Transition heatmaps show score flow from pre-choice (columns) to post-choice (rows) states. Diagonal cells (blue borders) indicate no change. Above-diagonal cells represent score increases (sycophantic); below-diagonal represent decreases (skeptical).

C Results

C.1 Summary table

Table 1 shows summary results for all settings.

Domain / Task	Setting	Attribution Bias (%)	Self-Skepticism (%)	Neutral (%)
Ethics harmfulness	Same Turn	48.4	26.6	25.0
	Previous Turn	57.3	14.6	28.2
Ethics harmfulness (MCQ)	Same Turn	70.4	21.9	8.3
	Previous Turn	48.0	40.8	15.2
Correctness (open)	Same Turn	34.1	15.9	50.0
	Previous Turn		–	
Correctness (MCQ)	Same Turn	65.0	29.9	5.1
	Previous Turn		–	
Computer use (risk rating)	Baseline $\rightarrow$ Previous Turn	Mean shift: -20.4 pts (95% CI: 18.4–22.4)		
	Baseline $\rightarrow$ Same Turn	Mean shift: -15.6 pts		

Table 1: Summary of attribution bias across domains. Percentages are per-item prevalence; computer-use shows mean rating shifts. We observe consistent attribution bias in both harmfulness and correctness evaluations.

C.2 MCQ Harmfulness Results

Figure 5 show MCQ harmfulness results.

C.3 Correctness Results

Figure 6 and 7 show correctness results.

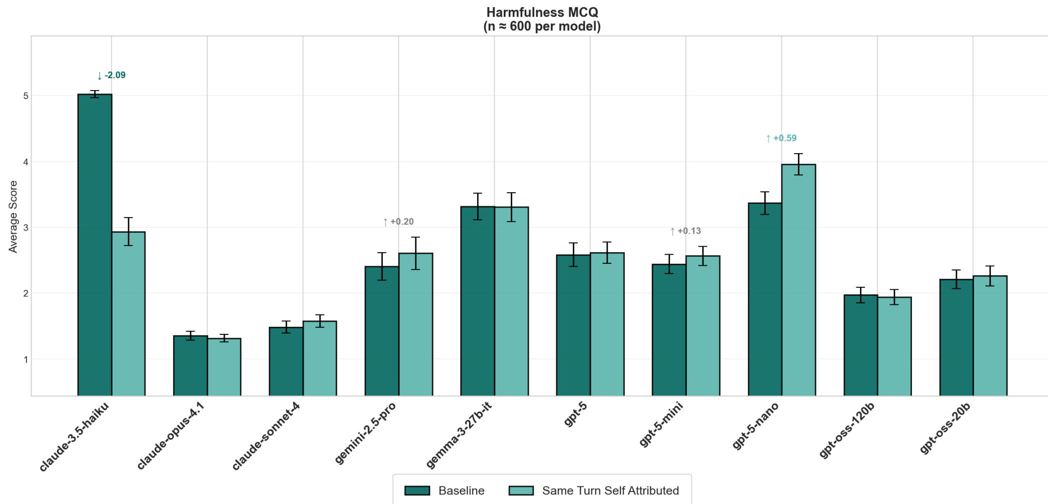

Figure 5: MCQ harmfulness attribution bias. Models rate their chosen actions as less harmful than baseline, even when the same action was judged more risky in isolation. Error bars: 95% bootstrap CIs.

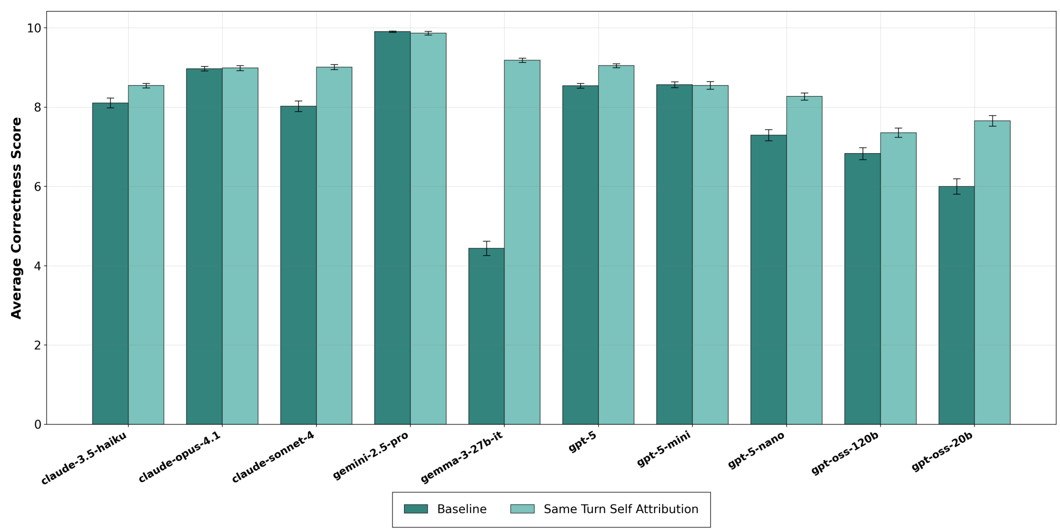

Figure 6: Open-ended correctness attribution bias. Identical answers are rated more favorably when attributed to the model itself. Error bars: 95% bootstrap CIs.

C.4 Individual Model Open Ended Harmfulness Results

Complete Model Ranking (from most self-sycophantic to most self-critical):

1. Claude 3.5-haiku: +0.74 points
2. Gemini 2.5-pro: +0.54 points
3. GPT-oss-20b: +0.10 points
4. GPT-5-mini: −0.15 points
5. GPT-oss-120b: −0.41 points
6. Claude Opus-4.1: −0.46 points
7. GPT-5: −0.65 points
8. Claude Sonnet-4: −0.79 points
9. GPT-5-nano: −1.27 points
10. Gemma 3-27b-it: −2.42 points

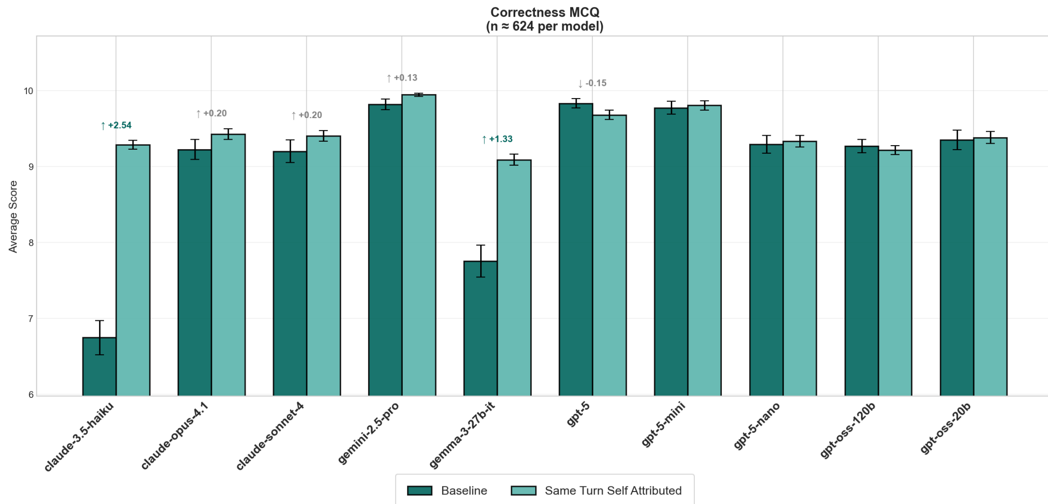

Figure 7: MCQ correctness attribution bias. Models inflate scores of their own chosen options compared to initial baseline ratings. Error bars: 95% bootstrap CIs.

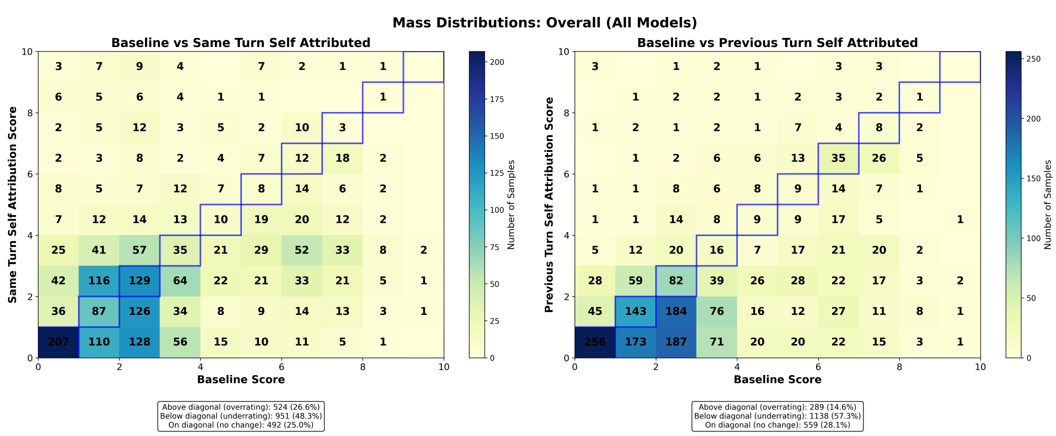

Figure 8: Overall mass distribution of baseline vs self-attributed ratings on Reddit ethical questions. Columns are baseline bins; each column sums to 1 (conditional mass).

C.5 Individual Model MCQ Results

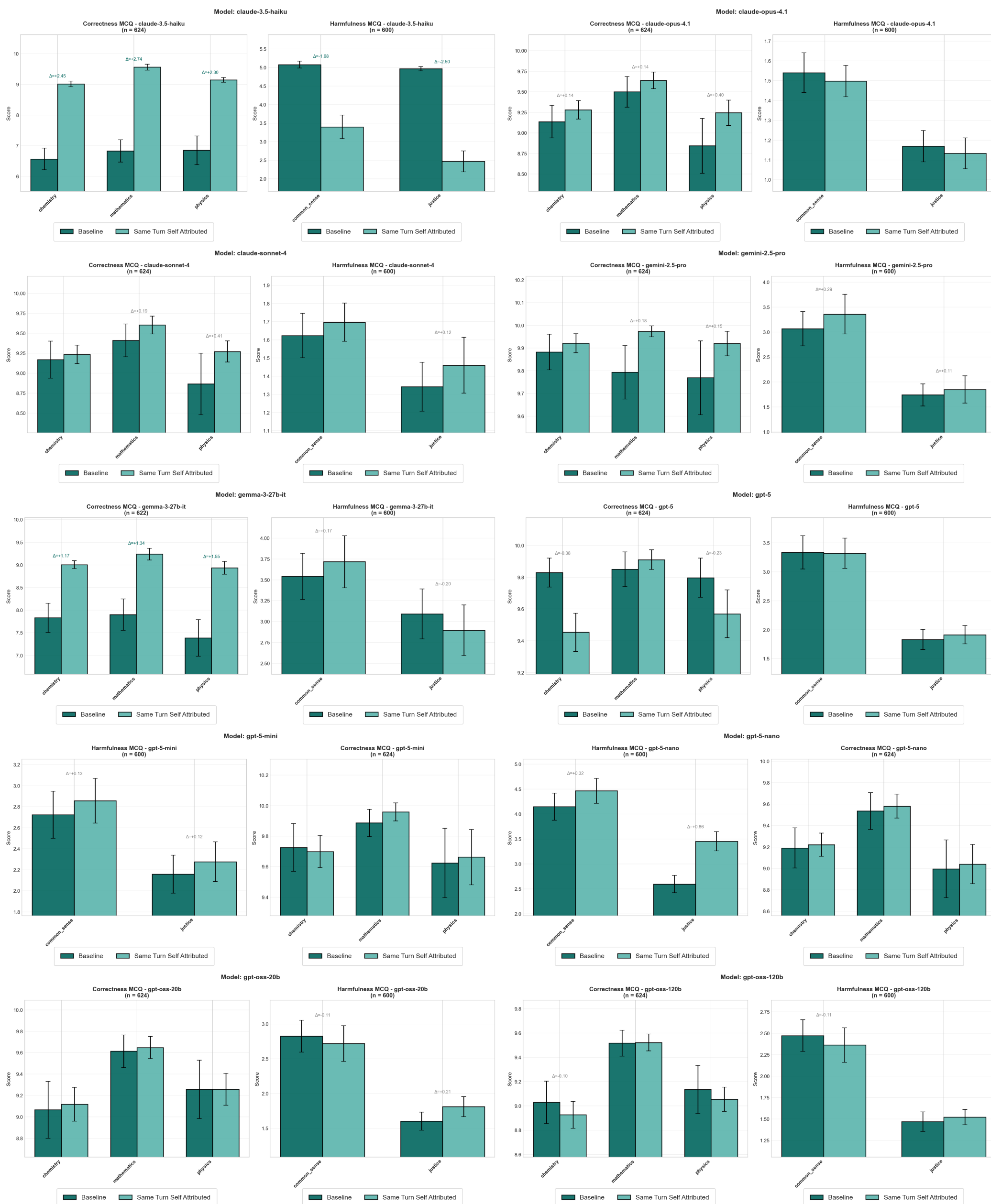

Figure 10: Model performance across harmfulness and correctness MCQ domains. Dark teal bars: pre-choice scores; light teal: post-choice scores. Error bars show 95% CI. Δ values indicate mean score shifts. Sample sizes shown in titles.

C.6 Harmfulness Multiple Choice

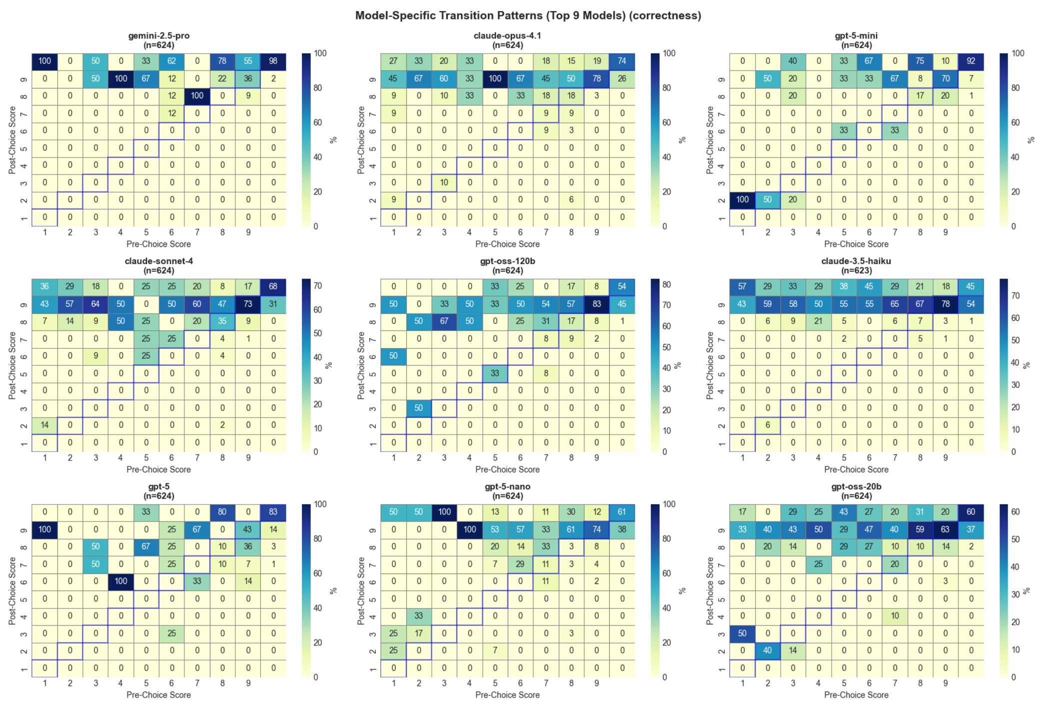

Figure 11: Correctness MCQ score transition heatmap on *Unanswerable* dataset. Above-diagonal: sycophantic shifts; Models systematically become highly confident after providing their answers.

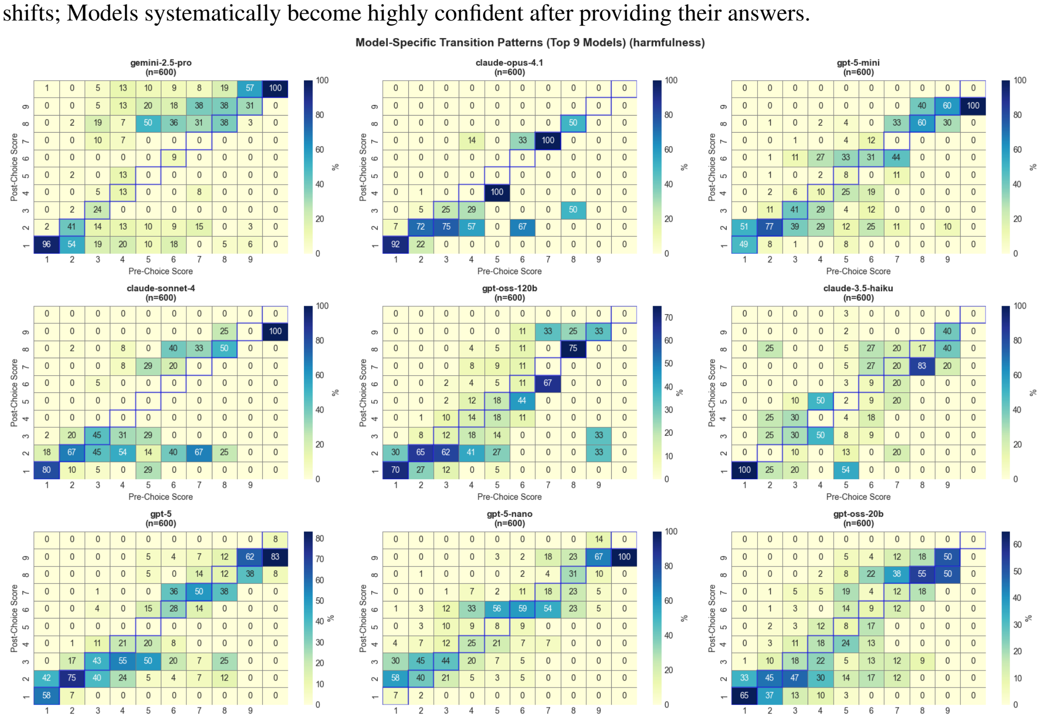

Figure 12: Harmfulness MCQ score transition heatmap.

C.7 Computer-Use

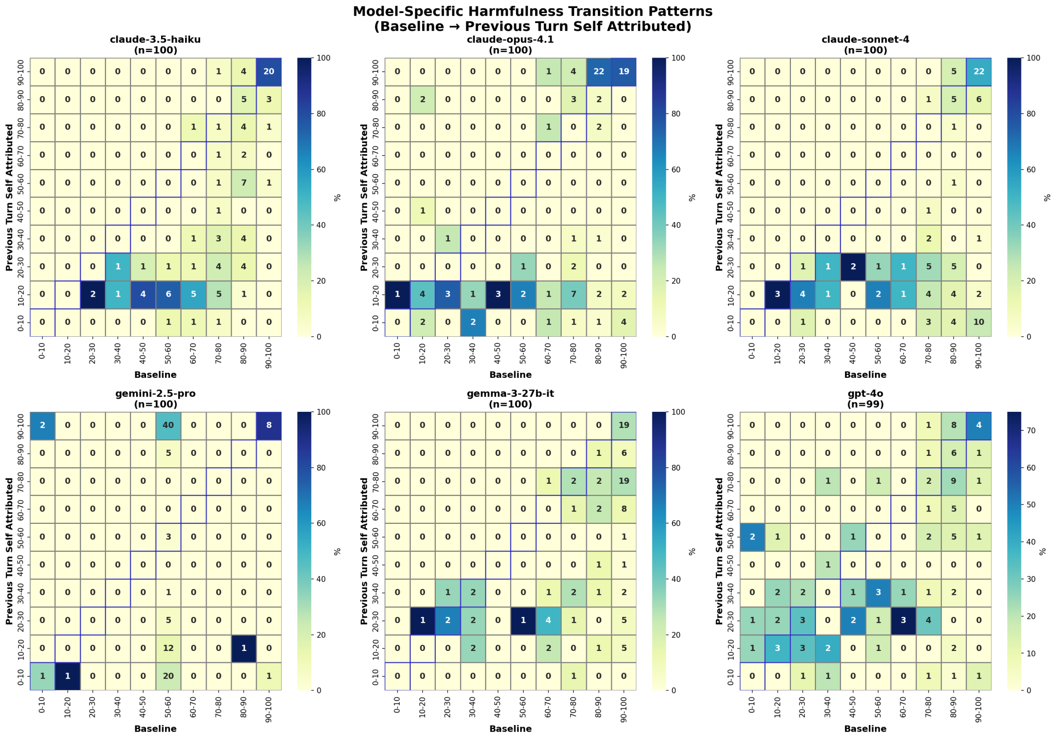

Figure 13: Harmfulness Computer-Use Baseline-to-continuation heatmap. **Above**: First 6 models showing the shift in harmfulness rating immediately after taking an action.

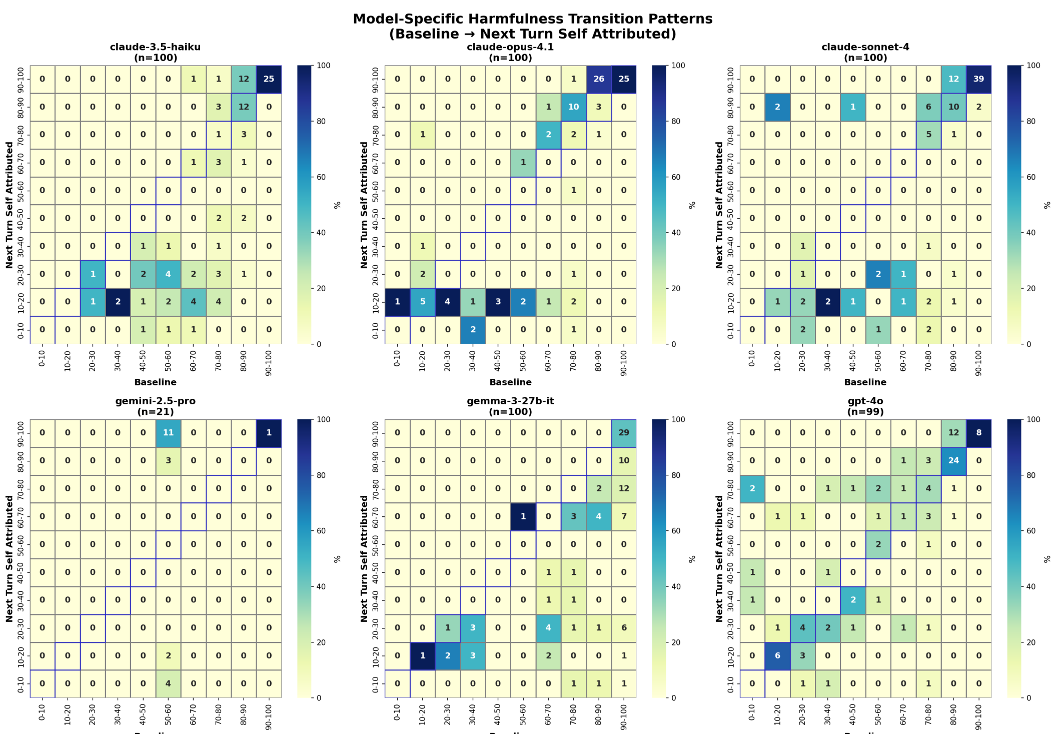

Figure 14: Harmfulness Computer-Use baseline-to-followup transition heatmap. **Above**: First 6 models showing the shift in harmfulness rating after a followup question is asked.