Self-Preference Is a Sum: A Frame-Conditional Decomposition for LLM Judges

Anonymous authors
Paper under double-blind review

Abstract

Language models can encounter a prior answer alongside conflicting information. We test whether describing that answer as the model's own changes its choice. On held-out PopQA questions, we evaluate Qwen3, Qwen2.5, Phi-3.5 and OLMo-2. We pair a prepared answer with a conflicting document; either the answer or the document is correct. The answer text and number of appearances stay the same, verified in code; its description names the model itself, an unspecified language-model assistant or a different assistant. When the prior answer is wrong, self-attribution increases its selection by 7.5, 18.4 and 20.4 percentage points in three families relative to different-assistant attribution. This post-hoc analysis derives choices from answer-letter log probabilities. Different-assistant wrong-choice rates are 1.4%–6.8%; the fourth family's −0.7-point change is not statistically significant. Comparing own with unspecified authorship, and unspecified with different authorship, gives two log-probability differences that sum exactly to the total on the same questions. In a post-hoc analysis, rephrasing changes this total in all eight combinations of model family and prior-answer correctness. Only three remain after excluding conditions that choose one answer side at least 90% of the time, all with correct prior answers. We withdraw earlier interpretations that these differences vary independently or differ in their robustness to wording. Nine of fifteen retrospectively selected conditions miss their original criteria; none is established under our reporting policy adopted after seeing the data. The core study's preregistration remains unverified. Separately, relabelling model-generated answers from the model's own to an unspecified assistant's improves final-decision accuracy by 6.4–9.4 percentage points in three families. After exploratory grading exclusions, two gains retain confidence intervals above zero. Relabelling can also harm decisions on correct prior answers. Authorship wording therefore affects decisions, with outcomes depending on phrasing, analysis and prior-answer correctness.

1. Introduction

Language models can receive conflicting answers to the same question within a conversation. One answer may appear in an earlier message, while a document supplied later gives another. In our setup, we place a prepared answer in the conversation, then provide a document with a different answer and ask the model to choose between them.

A model reviewing a response may need to check it against new information. This paper aims to test whether a model is more likely to choose an answer when it is described as its own. We keep the answer unchanged and describe it as coming from the model being tested, from a language-model assistant without specifying which one, or from a different language-model assistant.

Prior studies suggest that authorship labels matter differently across tasks. One rating study finds small effects from explicit labels (Khullar et al., 2026), while another finds that models are more willing to change an earlier choice when its answer is attributed to another model (Kumaran et al., 2025). Their prompts, tasks and measures differ, leaving the source of this disagreement unresolved. We study a specific setting: a model chooses between an earlier answer and a conflicting document while we change who is said to have given the answer.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We present the same answer in three ways: as the model's own answer, as an unspecified assistant's answer, or as a different assistant's answer. The surrounding sentence follows the same pattern, and the answer appears equally often in each prompt. Using the unspecified assistant as a reference lets us ask two questions: how much does support for the answer rise when it is labelled as the model's own, and how much does it fall when it is labelled as another assistant's? We call these differences self-elevation P and other-attribution discount D. They add up to the full self-versus-other difference, T1 = P + D. This equality follows from how the differences are calculated. Their shared reference also means that evidence for separate psychological mechanisms would require further experiments.

We find that the authorship description changes the model's decision, but how large the change looks depends on the wording and on how results are reported. We measure the model's support for an answer as the log-probability gap between the wrong and correct answer letters. With the answer held fixed, all four model families pass the criterion defined in Method (§3), which requires the effect in at least three of four families: when the earlier answer is wrong, labelling it as the model's own raises support for it by 0.43 to 3.93 logits relative to an unspecified assistant (all $p \leq 0.001$). Some of these comparisons involve a condition in which the model picks the same side on at least 90% of items, so its logits move while its choices barely change. After we exclude them, only two families qualify, Phi-3.5 and OLMo-2, one short of the three we required (§4.1, §4.2). Rephrasing the attribution sentence changes the gap between self and different-assistant labels in all eight combinations of model family and answer correctness; for Phi-3.5 on wrong answers, it falls from 10.99 to 0.04 logits. Only three of these eight changes survive the same exclusion rule, all with correct earlier answers (Appendix §C.2). Finally, describing the model's own generated answer as an unspecified assistant's improves the accuracy of its later decisions by 6.4 to 9.4 points in three families. After we remove answers flagged as hard to grade, two gains remain clearly above zero: 6.1 points for Phi-3.5 and 9.2 for OLMo-2. The same change can hurt when the earlier answer is correct: Qwen3 loses 13.0 points on those items (§4.4).

The wording around an earlier answer can also reverse the model's measured preference. We test this by leaving the prepared answer in the conversation and removing the sentence saying who gave it. Three of the four model families then give the answer less support than when it is described as coming from a language-model assistant without specifying which one. After we exclude comparisons where a model almost always chooses the same side, only Phi-3.5 still shows this reversal (Appendix §C.2). Removing the sentence also makes the message shorter and changes its phrasing, so further tests are needed to understand what causes the reversal.

Reading guide. Detailed methods, supporting results, additional controls, study history and reproduction materials appear in Appendices A–F.

2. Related work

Earlier studies show that language models can be less willing to change an answer when it is described as their own (Kumaran et al., 2025). Researchers have also compared how models judge content described as their own, as another language model's, or shown without an authorship label (Singla et al., 2025; Chae et al., 2026). These provide three points of comparison, although the reference setup differs across studies. Our study uses a related comparison: a model chooses between an earlier answer and a conflicting document, while we change the sentence saying who gave the answer. Our reference description names a language-model assistant without specifying which one. Keeping the answer unchanged holds its quality constant, addressing a concern raised by Roytburg et al. (2026). The different sentences may still change how the model interprets the situation.

Model evaluations can also change when the same task is presented through different prompts (Mizrahi et al., 2024; Maia Polo et al., 2024). We examine how wording changes the difference between support for an answer described as the model's own and one described as another model's. Other work studies how models handle conflicts between information they generate and information supplied in documents (Tan et al., 2024; Wu et al., 2024). Related settings include a model reviewing and revising its earlier response (Madaan et al., 2023), or several models comparing and discussing their answers (Du et al., 2024). Our fixed-answer comparison lets us examine the authorship description within this broader setting.

Researchers have also tried changing how earlier mistakes are presented so that models are more willing to correct them. In Chen et al. (2026), role relabelling involves moving unchanged incorrect content between messages presented as coming from the model, a user, or another source. Chen and colleagues find changes in whether models explicitly identify and reject the mistake. In their math tests with two models, the differences in final-answer accuracy are not statistically significant. Our experiment keeps the answer in the same message role and changes the sentence describing who gave it. We then measure whether the model's later answer is correct. Further comparisons, including tests that reuse wording and evaluation methods from earlier studies, appear in Appendix §§B.2 and C.3.

3. Method

We test how authorship descriptions affect a model's choice between conflicting answers using PopQA (Mallen et al., 2023) and four model families: Qwen3, Qwen2.5, Phi-3.5 and OLMo-2. Each conversation begins with a factual question, followed by a prepared answer that we insert as the model's earlier reply. A final user message supplies a reference document giving a conflicting answer and asks the model to choose between the two options. Later experiments use answers generated by the evaluated model (§4.4; Appendix §B.4).

We describe the earlier answer as the model's own with "I was asked this and answered: [answer]."; as an unspecified language-model assistant's with "An assistant was asked this and answered: [answer]."; or as a different assistant's with "A different assistant was asked this and answered: [answer]." The answer, the number of times it appears and the option order stay identical across versions, with answer occurrences checked in the final prompts. Each question is tested with a wrong earlier answer and a correct document, and with a correct earlier answer and a wrong document.

We measure answer preference by subtracting the log probability of the correct option label from that of the incorrect option label. Positive scores favor the wrong answer; negative scores favor the correct one. This captures shifts in preference even when the higher-probability option stays the same. We also report how often the wrong option has the higher probability.

For each question, we subtract the unspecified-assistant score from the score when the answer is described as the model's own, then subtract the different-assistant score from the unspecified-assistant score. The two differences add to the overall contrast between describing the answer as the model's own or another assistant's. Our written test requires the first difference to shift the score toward the earlier answer both when it is wrong and when it is correct. For each arrangement, its 95% confidence interval must lie entirely in the expected direction, with both requirements met in at least three of the four model families (Appendix §A.4).

The analysis includes 147 questions with 73 distinct wrong-answer texts. We group questions sharing a wrong answer for bootstrap confidence intervals and grouped sign-flip permutation p-values. Our study specification flags conditions that favor the wrong answer on at least 90% of questions, or the correct answer on at least 90%. After seeing the data, we made exclusion of comparisons involving either flagged condition our primary reporting rule; we also report results with them included (Appendix §A.4). We compare the wording variants described in Appendix §C.2 in a post-hoc analysis (Appendix §C.2). Appendix §§A.1–A.4 and E.1 give the prompts, selection rules, statistical procedures and study history.

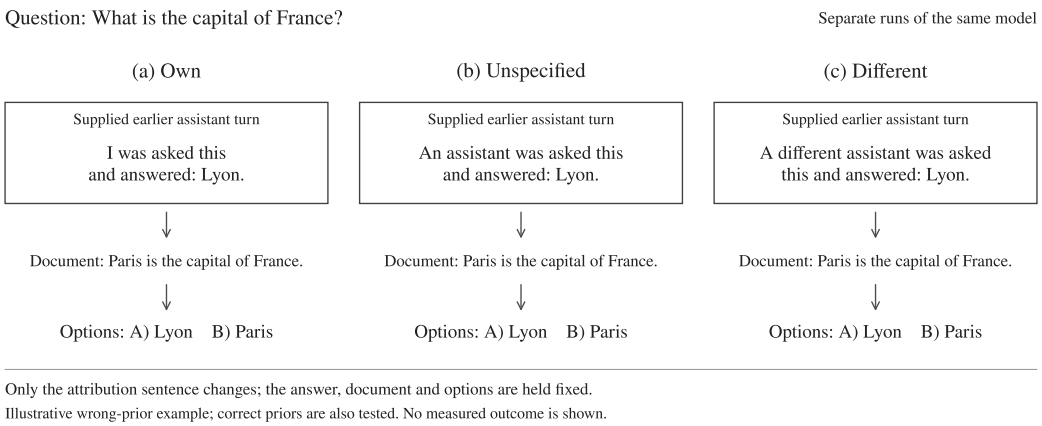


Figure 1: Illustrative setup for a supplied wrong prior answer. The same model evaluates each authorship description in a separate run, with the answer, reference document and options unchanged. We also test correct prior answers paired with incorrect documents. This example shows the procedure, with no measured outcome.

4. Results

We first examine whether changing an answer's stated author changes the model's preference for it, then test whether the effect persists across different wordings. We finish by measuring final-decision accuracy after relabelling answers generated by the model. For the authorship comparisons, our primary reporting rule excludes a comparison if either condition favors the correct answer on at least 90% of questions or the wrong answer on at least 90%. We report the full estimates alongside the findings that remain after these exclusions.

4.1 Attributing the answer to the model itself

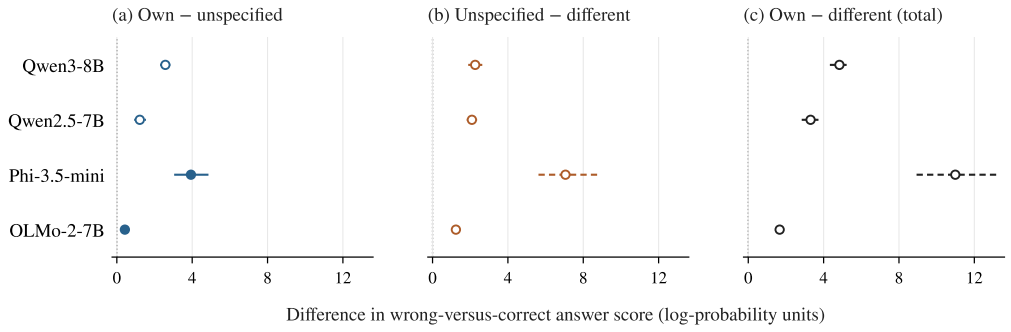
The main experiment compares the model's-own and unspecified-assistant descriptions on the same 147 questions, keeping the supplied prior answer fixed. All four families assign greater relative probability to the prior answer when it is described as their own, whether that answer is wrong or correct (Table 1).

All four meet the written criterion before exclusions. After applying the primary reporting rule, Phi-3.5 and OLMo-2 remain, one family short of the three required.

Table 1: Model's-own description compared with unspecified authorship. The criterion requires a shift toward the prior answer, with a 95% confidence interval excluding zero, both when the answer is wrong and when it is correct, in at least three of four families. Values subtract the unspecified-assistant score from the model's-own score.

Model	Wrong prior answer	Correct prior answer	Original criterion
Qwen3-8B †	+2.576 [2.299, 2.870]	−2.294 [−2.659, −1.929]	PASS
Qwen2.5-7B †	+1.219 [0.924, 1.546]	−0.920 [−1.359, −0.484]	PASS
Phi-3.5-mini	+3.934 [3.099, 4.808]	−4.913 [−5.764, −4.100]	PASS
OLMo-2-7B	+0.427 [0.265, 0.594]	−0.388 [−0.593, −0.186]	PASS

Reading Tables 1 and 2: differences use the wrong-minus-correct log-probability score. Positive values show a shift toward the wrong answer; negative values show a shift toward the correct answer. Brackets show 95% confidence intervals. † marks exclusions from the primary criterion count. All Table 1 comparisons have $p \leq 0.0005$ except OLMo-2 with correct prior answers, where $p = 0.0010$. Intervals and tests group questions by 73 distinct wrong-answer texts.



Wrong priors; means and 95% cluster intervals. Hollow/dashed: excluded by the near-constant-choice rule.
The total equals the two contrasts added together; it is not independent mechanism evidence.

Figure 2: Authorship comparisons for wrong prior answers in the main experiment. Panels show the model's-own minus unspecified description, unspecified minus different, and model's-own minus different. Points show means with 95% cluster confidence intervals. Hollow points and dashed intervals mark comparisons excluded because at least one condition favors the same answer side on at least 90% of questions. The third contrast equals the sum of the first two and provides no independent evidence of separate mechanisms.

4.2 Attributing the answer to a different assistant

We next compare unspecified authorship with attribution to a different assistant. In all four families, the model prefers the prior answer less when it is described as a different assistant's, whether the answer is wrong or correct (Table 2). After exclusions, all four correct-prior comparisons remain; none of the wrong-prior comparisons remain.

Table 2: Unspecified authorship compared with a different assistant. Values subtract the different-assistant score from the unspecified-assistant score. In both columns, the observed direction means lower preference for the prior answer under the different-assistant description. All eight comparisons have $p \leq 0.0005$.

Model	Wrong prior answer †	Correct prior answer
Qwen3-8B	+2.264 [1.893, 2.640]	−2.048 [−2.382, −1.730]
Qwen2.5-7B	+2.089 [1.815, 2.377]	−1.817 [−2.397, −1.259]
Phi-3.5-mini	+7.056 [5.623, 8.736]	−9.315 [−10.912, −7.794]
OLMo-2-7B	+1.240 [1.093, 1.381]	−0.941 [−1.098, −0.797]

† All wrong-prior comparisons are excluded under the primary reporting rule. Brackets show 95% confidence intervals.

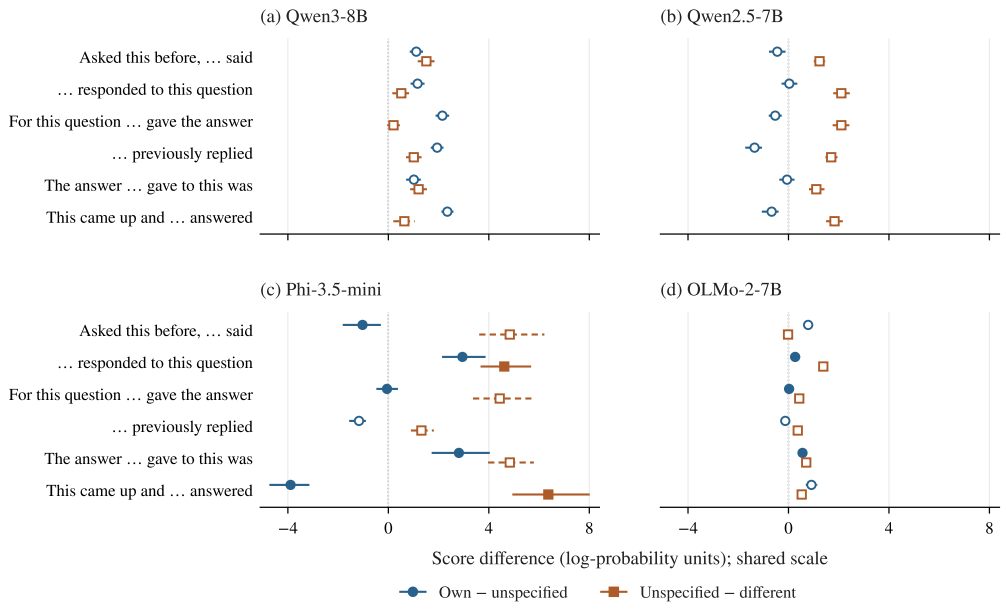
The two comparisons add to the overall difference between the model's-own and different-assistant descriptions (Figure 2). For correct prior answers, both comparisons remain after exclusions in every family. For wrong prior answers, no family retains both. The total is calculated from the same measurements and provides no independent evidence of separate mechanisms.

4.3 Changing the wording of the authorship sentence

We tested six ways of phrasing the authorship sentence, each with versions naming the model itself, an unspecified assistant or a different assistant. The prior answer remained unchanged. The complete sentence sets appear in Appendix §A.2.

The model's-own description increases preference for a wrong prior answer in 12 of 24 family-by-wording comparisons, with confidence intervals excluding zero. It shows a statistically significant reversal for four wordings in Qwen2.5 and three in Phi-3.5. Describing the answer as a different

assistant's reduces preference relative to unspecified authorship in 22 of 24 comparisons with confidence intervals excluding zero, with no significant reversal (Figure 3).



Wrong priors; 95% cluster intervals. Hollow/dashed: near-constant-choice exclusion.
Wording snippets omit the author phrase. These patterns do not establish relative robustness.

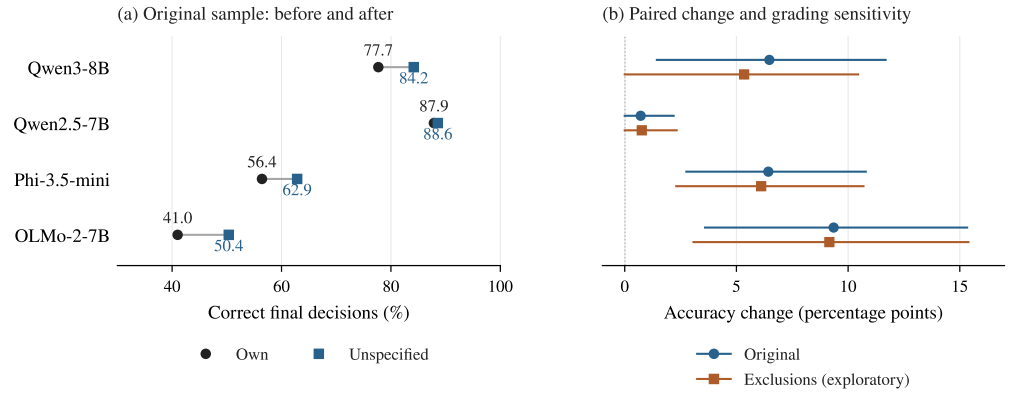
Figure 3: Authorship effects across six sentence wordings for wrong prior answers. All four model panels share the same horizontal scale. Row labels abbreviate the sentences by omitting the author phrase; full versions appear in Appendix §A.2. Points show means with 95% cluster confidence intervals. Hollow points and dashed intervals mark comparisons excluded by the 90% rule. These estimates describe wording dependence without establishing a difference in robustness between the two comparisons.

These counts meet the frame study's pre-registered rule. However, only two family-by-wording pairs retain both comparisons after exclusions, both in Phi-3.5. A diagnostic simulation also met the rule in 73.5% of cases where the comparisons varied equally across wordings but differed in their average distance from zero. We therefore withdraw the interpretation that one comparison is more robust to wording than the other. Appendix §B.2 reports the rule and diagnostic details.

4.4 Relabelling a model-generated answer

We next used answers generated by the evaluated model and changed their description from the model's own answer to an unspecified assistant's answer. The answer text and message position stayed fixed. We measured whether this change improved the model's final choice between its answer and the conflicting document's answer.

In the original analysis, accuracy improves in Qwen3, Phi-3.5 and OLMo-2, with 95% confidence intervals above zero. Qwen2.5's change is not distinguishable from zero. We then repeated the analysis after excluding questions flagged for disputed reference answers or possible automatic-grading errors. In this exploratory check, Phi-3.5 and OLMo-2 retain gains with intervals above zero; Qwen3's interval includes zero.



Model-generated priors; 95% intervals. Exclusions change the sample. Error-heavy PopQA sample.
Relabelling can harm correct-prior decisions: Qwen3 changes by −13.0 points in that subgroup (descriptive).

Figure 4: Final-decision accuracy after relabelling model-generated prior answers from the model's own to an unspecified assistant's. Left: original-sample accuracy under each description. Right: paired accuracy changes with 95% confidence intervals, both in the original sample and after exploratory grading exclusions. The exclusions change the sample; only Phi-3.5 and OLMo-2 retain intervals entirely above zero. The error-heavy sample and possible harm on correct prior answers limit the practical interpretation.

After exclusions, the gains are 6.1 percentage points for Phi-3.5 [2.3, 10.7] and 9.2 for OLMo-2 [3.1, 15.4]. Qwen3's estimate is 5.3 [0.0, 10.4], with p = 0.0923. Table 3 and Figure 4's left panel show the original analysis; Figure 4's right panel also shows the exploratory grading-exclusion analysis.

Table 3: Final-decision accuracy before and after relabelling. Changes and 95% confidence intervals are in percentage points. These are the original estimates before grading exclusions.

Model	Model's own	Unspecified assistant	Change [95% CI]	Original criterion
Qwen3-8B	77.7%	84.2%	+6.5 [1.4, 11.7]	Met
Qwen2.5-7B	87.9%	88.6%	+0.7 [0.0, 2.2]	Not met
Phi-3.5-mini	56.4%	62.9%	+6.4 [2.8, 10.8]	Met
OLMo-2-7B	41.0%	50.4%	+9.4 [3.6, 15.3]	Met

The overall gain can include a cost when the prior answer is correct. For Qwen3, relabelling corrects 12 final decisions among questions with wrong prior answers, but turns three correct final decisions into errors among questions with correct prior answers. Because the sample contains many more wrong prior answers, the corrections outweigh these losses overall (Table 4).

Table 4: Accuracy changes separated by prior-answer correctness. Counts show correct final decisions under the model's-own description and then unspecified authorship. These are descriptive subgroup results from the original sample, before grading exclusions; subgroup confidence intervals are not reported.

Model	Prior answer	Questions	Correct decisions: own → unspecified	Change, percentage points
Qwen3-8B	Wrong	116	90 → 102	+10.3
Qwen3-8B	Correct	23	18 → 15	−13.0
Qwen2.5-7B	Wrong	118	112 → 112	0.0
Qwen2.5-7B	Correct	22	11 → 12	+4.5
Phi-3.5-mini	Wrong	112	70 → 79	+8.0
Phi-3.5-mini	Correct	28	9 → 9	0.0
OLMo-2-7B	Wrong	106	36 → 48	+11.3
OLMo-2-7B	Correct	33	21 → 22	+3.0

5. Limitations

Our experiments cover short factual answers, and the findings vary across wording, tasks and models. Science-question and free-form experiments fall short of their success criteria. Removing authorship descriptions often reverses the effect for supplied answers, while removal remains untested for model-generated answers. Open-weight coverage stops at 14 billion parameters, with size comparisons confined to Qwen. Commercial models were tested on one date in one configuration with reasoning disabled, and their probability outputs did not allow valid log-probability comparisons. Even where valid, log-probability shifts may leave decisions unchanged and have different scales across models. The unspecified-assistant reference may itself suggest authorship or credibility, and our internal-activation experiment left the mechanism unresolved (Appendix §§A.3, B.1, B.3–B.4, C.2 and D.1–D.5).

Confidence in these findings is also limited by shared observations across tests, few usable comparisons in some follow-up analyses and low statistical power in the reasoning experiment. We specified no paper-wide correction for multiple comparisons. Records do not fully establish when the main design was fixed, and the primary exclusion policy was chosen after seeing the data. Incomplete input-version records and limited prompt checks further restrict verification of the generated-answer intervention. Automatic grading can miss valid answers or accept irrelevant matches, and some reference answers are disputed; excluding flagged questions tests sensitivity without resolving all grading uncertainty. Authorship comparisons involving correct generated answers fall short of the criterion or remain descriptive, with differences in fact types and entity popularity complicating interpretation. Relabelling can also disrupt correct decisions (§4.4). Appendix §§A.4, B.2–B.4, D.3–D.5 and E.1 give the detailed qualifications, while Appendix §§F.2–F.3 describe the available materials and limits of reproducing past runs.

6. Discussion

An answer's stated authorship can influence how a model evaluates it, even when the content stays the same. One possible explanation is a learned tendency to maintain consistency across turns: describing an answer as the model's own may make retaining it more likely. Kumaran et al. (2025) found that displaying a model's initial choice reduced its tendency to change that choice in their advice-taking task. Whether chat tuning strengthens this tendency remains an open question. Comparing the same base model before and after chat tuning, with matched prompts, would help test that possibility.

The model's interpretation of who is speaking may also matter. In our main experiment, first-person wording appears in a researcher-supplied model reply, so the meaning assigned to that wording may depend on the surrounding conversation. Khullar et al. (2026) found stronger effects when models evaluated code they had generated than when explicit source labels were varied in a new context. A useful next test would vary a separate note identifying the speaker while holding the answer and authorship sentence fixed. Differences in conversation templates and training could contribute to variation across families; controlled template comparisons would help assess that explanation.

Describing the answer as a different assistant's may change its perceived credibility. The word different could suggest distance from the current model, leading it to place less weight on the answer. This is a hypothesis about how the sentence is interpreted; our experiments leave its cause unresolved. Comparing descriptions that identify the same external source with and without an explicit contrast word, while independently varying stated source reliability, could help distinguish these possibilities. The unspecified-assistant description also needs checking because readers cannot assume that the model treats it as neutral.

The distinction between probability changes and decision changes is important regardless of the explanation. Some conditions show shifts in option log probabilities with little movement in the selected answers. Reporting both measures would make the practical size of an authorship effect easier to judge. Similarly, some commercial models almost never select the wrong answer in our tests, leaving little room to observe a reduction. Testing harder questions selected independently of the authorship results could help determine whether this reflects task difficulty; the current evidence does not establish a capability trend.

These findings also motivate tests of models used to judge other outputs. If a judge is told that an answer is its own, its evaluation may depend partly on that description. Holding the evaluated text fixed across own, unspecified and different-source descriptions would help assess this sensitivity. Including the unspecified-source comparison would show which changes accompany naming each source, while leaving open how that reference is interpreted. Such evaluations should report exact wording and distinguish changes in scores from changes in the final ranking.

For practitioners building systems that reuse a model's answers or ask it to check them, authorship wording deserves explicit testing. Before adopting a wording change, developers should compare source descriptions on representative examples using the intended model and conversation format, keeping the answer and message position fixed. Evaluation should count both errors corrected and correct decisions turned into errors, using a mix of correct and incorrect prior answers that reflects the application. Our gains on an error-heavy sample vary across families, weaken under grading exclusions in one family, and come with a risk of disrupting correct decisions. The practical value of relabelling therefore depends on the application's error rate and the cost of newly introduced mistakes. Recording the exact prompts would help make these comparisons reproducible. Selective relabelling based on a separate verifier is another possibility, but it needs held-out testing that includes verifier mistakes.

AI USE STATEMENT

We used generative AI tools to draft and polish the manuscript and assist with executing experimental code. The authors reviewed the AI-assisted outputs and accept responsibility for the final text, claims and accompanying materials.

References

Songeun Chae, Min Kim, Donghoon Jung, Seojin Choi and Seohyon Jung. *Self- and Other-Labels Induce Bidirectional Bias in LLM Judges*. arXiv preprint arXiv:2608.18091, 2026. DOI.

Kuan-Yen Chen, Fang-Yi Su, Shih-Yen Lin, Bao Li and Jung-Hsien Chiang. *The Self-Correction Illusion: Role Relabeling Gates Explicit Error Flagging in Large Language Models*. arXiv preprint arXiv:2606.05976, 2026. DOI.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum and Igor Mordatch. *Improving Factuality and Reasoning in Language Models through Multiagent Debate*. Proceedings of the 41st International Conference on Machine Learning, Proceedings of Machine Learning Research 235, pp. 11733-11763, 2024. URL.

Dipika Khullar, Jack Hopkins, Rowan Wang and Fabien Roger. *Self-Attribution Bias: When AI Monitors Go Easy on Themselves*. arXiv preprint arXiv:2603.04582, 2026. DOI.

Dharshan Kumaran, Stephen M Fleming, Larisa Markeeva, Joe Heyward, Andrea Banino, Mrinal Mathur, Razvan Pascanu, Simon Osindero, Benedetto de Martino, Petar Velickovic and Viorica Patraucean. *How Overconfidence in Initial Choices and Underconfidence Under Criticism Modulate Change of Mind in Large Language Models*. arXiv preprint arXiv:2507.03120, 2025. DOI.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh and Peter Clark. *Self-Refine: Iterative Refinement with Self-Feedback*. Advances in Neural Information Processing Systems 36, pp. 46534-46594, 2023. DOI.

Felipe Maia Polo, Ronald Xu, Lucas Weber, Mírian Silva, Onkar Bhardwaj, Leshem Choshen, Allysson Flavio Melo de Oliveira, Yuekai Sun and Mikhail Yurochkin. *Efficient multi-prompt evaluation of LLMs*. Advances in Neural Information Processing Systems 37, pp. 22483-22512, 2024. DOI.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi and Hannaneh Hajishirzi. *When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories*. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 9802-9822, 2023. DOI.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf and Gabriel Stanovsky. *State of What Art? A Call for Multi-Prompt LLM Evaluation*. Transactions of the Association for Computational Linguistics 12, pp. 933-949, 2024. DOI.

Dani Roytburg, Matthew Bozoukov, Matthew Nguyen, Jou Barzdukas, Mackenzie Puig-Hall and Narmeen Oozeer. *Are LLM Evaluators Really Narcissists? Sanity Checking Self-Preference Evaluations*. arXiv preprint arXiv:2601.22548, 2026. DOI.

Pratham Singla, Shivank Garg, Ayush Singh, Ishan Garg and Ketan Suhaas Saichandran. *Thinking About Thinking: Evaluating Reasoning in Post-Trained Language Models*. arXiv preprint arXiv:2510.16340, 2025. DOI.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao and Xueqi Cheng. *Blinded by Generated Contexts: How Language Models Merge Generated and Retrieved Contexts When Knowledge Conflicts?*. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 6207-6227, 2024. DOI.

Johannes Welbl, Nelson F. Liu and Matt Gardner. *Crowdsourcing Multiple Choice Science Questions*. Proceedings of the 3rd Workshop on Noisy User-generated Text, pp. 94-106, 2017. DOI.

Kevin Wu, Eric Wu and James Zou. *ClashEval: Quantifying the tug-of-war between an LLM's internal prior and external evidence*. arXiv preprint arXiv:2404.10198, 2024. DOI.

Appendix

A Detailed methods and statistical analysis

This appendix gives the procedures and supporting evidence for the main paper. Section A describes the experiment and analysis; Section B supports the main Results; Sections C and D cover additional controls and extensions. Study history and reproducibility information appear in Sections E and F. Throughout, "own", "unspecified" and "different" describe the stated source of an answer, rather than establishing who generated it.

A.1 Questions, answer pairing and prompt checks

We use two sets of PopQA questions (Mallen et al., 2023): an exploratory set and a separate set for the main experiment. Each initially contains 150 items, with 11 relations and 75 true and 75 false statements. The sets share no subjects. They were sampled with seeds `20260820` and `20260828` from `akariasai/PopQA`, snapshot `098765c79ea10a2cb19c828324e33281b8336ec0`, file `test.tsv`. Their sorted-key JSON hash prefixes are `c0a92e6c25fd51d1` and `14c219451de0195d`. Section E.1 explains why the main set is held out from the exploratory subjects but its historical preregistration status remains limited.

For a false statement, its true object supplies the correct answer and its shown object supplies the wrong answer. For a true statement, the shown object is correct, and a wrong answer is selected deterministically from the false statements with the same relation, using `pool[id % len(pool)]`. Pairing uses the complete frozen item list before exclusions. This prevents the pairing from changing when a later prompt check removes an item.

We count occurrences of the prior-answer string in the complete chat-template rendering. In the main experiment, it must appear exactly twice: once in the supplied model reply and once among the final answer options. An item is retained only if every condition passes this check, so exclusions are shared across authorship descriptions. The exploratory no-answer control intentionally has only the occurrence in the options.

All five tokenizers retain the same 147 main-set items. Two items repeat the candidate string elsewhere in the prompt; a third has a `mother` relation without a usable wrong-answer pool. The retained set contains 74 false and 73 true statements, 73 items with the correct option first and 74 with it second, 10 relations, and 73 distinct wrong-answer strings. The five largest string groups contain 8, 6, 5, 5 and 4 items. The exploratory set also retains 147 items, with 72 distinct wrong strings and correct-option positions split 71/76. The sets share two wrong strings, Australia and India, involving three main-set items; their groups are therefore not completely independent. A tokenizer-only reconstruction reproduces these counts, but was performed after the original runs.

A.2 Conversation format and exact wording

The core experiment supplies the prior answer rather than asking the model to generate it. Each model's chat template renders the conversation below with `add_generation_prompt=True`. The frozen `A_is_true_statement` field fixes the option order across all descriptions for an item. Response letters are shown here because these are the actual prompts and scored tokens.

```
user:      Question: <templated question>
assistant: <arm text, containing the candidate string `own`>
user:      A reference document proposes: <doc answer>
           Two candidate answers:  A) <...>  B) <...>
           Which answer is correct? Respond with A or B only.
```

Table A.1: Exact supplied-reply templates. The three main descriptions keep the answer text and occurrence count identical. The other rows are controls; the no-answer reply deliberately omits the prior answer.

Description	Exact prior model reply	Study
Own	`I was asked this and answered: <own>.`	Exploratory and main
Equal-length control	`I was told an assistant answered: <own>.`	Main-set exploratory control
Unspecified	`An assistant was asked this and answered: <own>.`	Main
Different	`A different assistant was asked this and answered: <own>.`	Exploratory and main
Bare answer	`<own>`	Exploratory
No answer	`I would need to check that.`	Exploratory

We test both a wrong prior answer with a correct document and a correct prior answer with an incorrect document. Within each arrangement, the prior-answer text, document and option order remain unchanged. Token length differs across the three descriptions: the unspecified version adds one token and the different-assistant version adds two relative to the own-answer version, consistently across retained items on all five tokenizers. The distance from the answer's end to the scored position is unchanged. The equal-length control in §C.1 helps assess this difference, although it also changes the sentence's meaning.

The six-wording experiment in main §4.3 uses the following complete sentences. The short row identifiers are used only to map exact stimuli to the detailed results and supplementary files; Figure 3 uses readable wording snippets.

Table A.2: Complete sentence sets for the sealed six-wording experiment. Only the source phrase changes within a row. These are the full stimuli behind the abbreviated labels in Figure 3.

Wording	Own	Unspecified	Different
N1	"Asked this before, I said: X."	"Asked this before, an assistant said: X."	"Asked this before, a different assistant said: X."
N2	"I responded to this question with: X."	"An assistant responded to this question with: X."	"A different assistant responded to this question with: X."
N3	"For this question I gave the answer: X."	"For this question an assistant gave the answer: X."	"For this question a different assistant gave the answer: X."
N4	"I previously replied: X."	"An assistant previously replied: X."	"A different assistant previously replied: X."
N5	"The answer I gave to this was: X."	"The answer an assistant gave to this was: X."	"The answer a different assistant gave to this was: X."
N6	"This came up and I answered: X."	"This came up and an assistant answered: X."	"This came up and a different assistant answered: X."

Two earlier paraphrases are evaluated separately in §C.2. The first compares "When I was asked this earlier, I answered: X." with "When an assistant was asked this earlier, it answered: X." The second compares "My earlier answer to this was: X." with "An assistant's earlier answer to this was: X." Their different-assistant versions are "When a different assistant was asked this earlier, it answered: X." and "A different assistant's earlier answer to this was: X." The answer placeholder is replaced identically within each comparison.

A.3 Scores, decisions and readout validity

At the answer position, we take the maximum next-token log probability over the single-token forms `" A"`/`"A"` and `" B"`/`"B"`, checking that at least one form exists for each letter. The score is the log probability for the wrong option label minus that for the correct option label. Positive values favor the wrong option. The post-hoc choice-rate analysis thresholds this same score at zero, so it is another summary of the same observations rather than an independent experiment.

For the four main model families, the mean probability assigned to the answer letters is at least 0.9974 in every condition; the lowest individual value is 0.8808. A separate greedy, full-vocabulary check agrees with the letter-based choice on 0.973–1.000 of items per condition, with family averages of 0.991–0.999. The score-based criteria also hold in that check. These observations validate this constrained readout locally; they do not establish agreement for unrestricted generation.

DeepSeek-R1-Distill-Qwen-1.5B is excluded from substantive comparisons because its template appends `<think>` despite `enable_thinking=False`. The scored position therefore lies inside reasoning rather than at an answer letter. Mean letter probability is only 2.7×10^{-7} to 3.8×10^{-7}, and changing the document leaves the apparent wrong-choice rate between 0.476 and 0.524. The instrument fails to detect the intended document manipulation. Consequently, this run cannot establish an absence of an authorship effect. Qwen3's non-thinking template instead supplies an empty reasoning segment before the answer position.

We also reran each core model in a fresh process with newly loaded weights. All 55 compared quantities, comprising five contrast means and the per-item vectors across five models, matched exactly. This is a determinism check on the same inputs and settings. It does not provide a new sample or an independent replication.

A.4 Statistical analysis and reporting policy

For each question, the own-minus-unspecified difference measures the effect of changing the stated source to the model itself. The unspecified-minus-different difference measures the effect of naming a different assistant. Their sum equals own minus different because the common unspecified score cancels. In the notation introduced in the Introduction, these are P, D and T1, respectively. We use descriptive names below. Under the wrong-minus-correct score, preference for the prior answer corresponds to a positive difference when that answer is wrong and a negative difference when it is correct.

The main written success criterion requires the own-minus-unspecified 95% confidence interval to exclude zero in those expected directions for both arrangements in at least three of four families. The two arrangements form one criterion per family. The unspecified-minus-different contrasts are reported separately. The original criterion and the paper's primary reporting rule give different counts, as shown in main §4.1.

Questions sharing a wrong-answer string are resampled together. Core confidence intervals use 20,000 bootstrap draws over the 73 string groups and fixed contrast-specific seeds. Generated-answer analyses group by the generated answer text instead, with 92–136 groups across their analyses. Paired permutation tests sign-flip grouped differences over 2,000 draws and report `(hits + 1)/(draws + 1)`. The smallest attainable Monte Carlo p-value is 1/2001. Nineteen of the 20 core tested contrasts attain this value, and one attains 2/2001. A zero-hit estimate does not bound the exact probability below the reported resolution; its one-sided 95% binomial upper bound is approximately 0.0015.

The study specification flags conditions that favor one answer side on at least 90% of questions. It did not originally specify that flagged comparisons should be excluded from the success count. We adopted that exclusion as the paper's primary reporting policy after seeing the data, and therefore retain the original estimates alongside it. A comparison is excluded if either contributing condition is flagged. A retained comparison still has to satisfy its stated statistical criterion. Conversely, a flag does not prohibit directly measuring a choice-rate difference; it limits interpreting a log-probability shift as if it were a behavioral change. Section B.1 reports those choice differences explicitly.

Grouping by relation provides a sensitivity check with only 10 groups. Its percentile intervals retain all 16 own-minus-unspecified and unspecified-minus-different contrasts, but a real-data null gives

a false-positive rate of 0.098 [0.086, 0.113], above the nominal 0.05. A null-imposed Rademacher wild cluster bootstrap-t with 5,000 draws retains 14 of 16; the correct-prior own-minus-unspecified contrasts for Qwen2.5 and OLMo-2 are unsupported (p = 0.300 and 0.170). At the primary 73-string grouping, the real-data null rate is 0.055 [0.046, 0.066]. These checks use recentered observed differences and 120 trials per tested contrast. The relation, behavioral and 90%-exclusion criteria select different families; Phi-3.5 is the only family meeting all three conjunctions.

The synthetic estimator test detects its positive case in 1.000 [0.981, 1.000] of 200 worlds and fires in 0.025 [0.011, 0.057] of 200 null worlds. However, its common Gaussian group effect cancels in the paired difference. Actual differences have intracluster correlations of 0.02–0.45, with median approximately 0.24. The real-data calibration is therefore more informative about clustered error rates. The commercial-model rate estimator's synthetic control has the same limitation (§D.5).

We report multiplicity adjustments for the explicitly named families of comparisons: 20 core contrasts, 24 core choice-rate contrasts, 50 scale-extension score contrasts, 48 scale-extension rate contrasts and 20 commercial-model contrasts. There is no specified or executed whole-paper correction across the more than 300 interval-bearing cells. Counts over shared questions and conditions are descriptive enumerations, not joint tests. The 2,000-draw resolution supports Bonferroni decisions only up to 100 comparisons at $\alpha = 0.05$. For the sparse scale-extension choice rates, seven bootstrap intervals exclude zero while permutation p-values do not; those cases are interpreted using the permutation test. Score-based studies retain their own specified interval criteria. The exploratory set also retains an older asymmetric percentile-index implementation; it was not rerun under the main experiment's corrected indexing.

B Supporting results for the main experiments

B.1 Core comparisons and actual choices

Main Tables 1 and 2 report the score differences. The primary exclusion rule retains the own-minus-unspecified criterion in Phi-3.5 and OLMo-2. Qwen3's wrong-prior condition fractions favoring the modal answer are 0.905 and 0.966; Qwen2.5's flagged wrong-prior fractions are at least 0.986. Its correct-prior criterion conditions are below the threshold, at 0.803–0.816. All four correct-prior unspecified-minus-different contrasts remain, whereas all four wrong-prior contrasts are flagged.

The following tables show how these probability shifts relate to selected answers. Rates are fractions, so a difference of 0.075 corresponds to 7.5 percentage points. These analyses were computed after inspecting the core results and use the same per-item scores as the main comparisons.

Table B.1: Wrong-option choice rates from the main experiment. A star marks a condition meeting the 90% concentration threshold. These directly measured rates remain reportable even when a score contrast is excluded from the primary criterion count.

Description / prior answer	Qwen3-8B	Qwen2.5-7B	Phi-3.5-mini	OLMo-2-7B
Different / wrong	0.020 *	0.014 *	0.068 *	0.048 *
Unspecified / wrong	0.034 *	0.014 *	0.190	0.150
Own / wrong	0.095 *	0.007 *	0.252	0.252
Different / correct	0.728	0.830	0.741	0.796
Unspecified / correct	0.667	0.816	0.544	0.646
Own / correct	0.599	0.803	0.469	0.633

Table B.2: Paired differences in wrong-choice rates for wrong prior answers, with grouped 95% intervals and permutation p-values. Values are fractions; multiply by 100 for percentage points. These post-hoc estimates underlie the abstract's behavioral claim.

Comparison	Qwen3-8B	Qwen2.5-7B	Phi-3.5-mini	OLMo-2-7B
Own − unspecified	**+0.061** [+0.027, +0.102], p = 0.0025	−0.007 [−0.031, +0.015], n.s.	**+0.061** [+0.026, +0.102], p = 0.0030	**+0.102** [+0.054, +0.156], p = 0.0010
Own − different	**+0.075** [+0.037, +0.117], p = 0.0015	−0.007 [−0.031, +0.014], n.s.	**+0.184** [+0.119, +0.257], p ≤ 0.0005	**+0.204** [+0.135, +0.279], p ≤ 0.0005

The own-versus-different wrong-prior differences in Table B.2 are the behavioral results reported in the abstract. Three families have positive intervals, while Qwen2.5's −0.007 change is not distinguishable from zero. Those three families also meet the two-arrangement criterion against different-assistant attribution under Bonferroni-24. Against unspecified authorship, the uncorrected wrong-prior count is three families, but the correction retains only OLMo-2 on that side and no two-arrangement conjunction. The Qwen3 and Phi-3.5 wrong-prior p-values are 0.0025 and 0.0030, above the corrected threshold of approximately 0.00208. This explains why the abstract's behavioral count and the main score-based exclusion count differ.

Choice changes depend on where scores fall relative to zero. Qwen2.5 has a +1.219 score shift but wrong-prior choice rates of only 0.007–0.014. Phi-3.5's larger score shift corresponds to a smaller own-versus-unspecified choice change than OLMo-2's, but the difference between their rate changes is not significant: +0.041 [−0.013, +0.098], p = 0.243. Neither raw score size nor this descriptive ordering establishes a cross-model behavioral ranking.

Splitting by option position retains the expected interval direction in 31 of 32 tested half-samples. The exception is Qwen2.5's correct-prior own-minus-unspecified difference when the first option is false: −0.459 [−1.198, +0.268]. Position changes estimated magnitudes by factors of 1.02–3.02. Relation-specific point estimates also reverse: OLMo-2 has wrong-prior reversals for capital, capital-of and country, and correct-prior reversals for capital, capital-of and father; Qwen2.5 has a wrong-prior reversal for capital and correct-prior reversals for capital, country, father and place of birth. Thus the pooled result does not hold within every fact type. Complete subgroup values remain in the evidence files.

The largest five absolute item differences contribute 6.8–18.7% of each family's mean. Wrong-prior own-minus-unspecified differences remain positive across all twelve thirds of the data formed by unspecified-condition score magnitude. However, OLMo-2's means fall from 0.70 to 0.52 to 0.06, with the last compatible with zero. These subsets also differ in relation and entity popularity, so the pattern does not isolate score saturation. Its overall 0.427 effect exceeds the estimated clustered minimum detectable effect of 0.237.

The net consequence also depends on how often the prior answer is wrong. Weighting the supplied-answer effects by the separately measured gradable-answer error rates gives task-error changes of +3.8, −0.8, +3.2 and +7.3 percentage points for Qwen3, Qwen2.5, Phi-3.5 and OLMo-2. A 50/50 mix instead gives −0.3, −1.0, −0.7 and +4.4. These are descriptive reweightings of supplied-answer comparisons, not the generated-answer intervention in main §4.4. That intervention has its own sample and directly observed final decisions (§B.4).

B.2 Wording dependence and the six-wording study

Two analyses support main §4.3: the post-hoc comparison of the original sentence with two paraphrases (§C.2), and a later sealed experiment with the six exact sentence sets in Table A.2. Both hold answer content constant. Their statistical status differs: the first was computed after the factorial results, while the second specified its wordings and outcome rule before its runs.

Table B.3: Wrong-prior scores across six wordings, with 95% cluster intervals. Bold intervals exclude zero; downward arrows mark significant own-minus-unspecified reversals. The 90% flags are shown in Figure 3 and the complete result file. This table includes flagged comparisons and does not establish relative robustness.

Model	Wording	Own − unspecified	Unspecified − different	Own − different
Qwen3-8B	N1	**+1.112 [+0.851, +1.380]**	**+1.514 [+1.171, +1.875]**	**+2.626 [+2.163, +3.114]**
Qwen3-8B	N2	**+1.163 [+0.885, +1.441]**	**+0.515 [+0.157, +0.898]**	**+1.679 [+1.203, +2.171]**
Qwen3-8B	N3	**+2.151 [+1.881, +2.421]**	+0.205 [-0.043, +0.458]	**+2.356 [+1.948, +2.760]**
Qwen3-8B	N4	**+1.942 [+1.689, +2.197]**	**+1.012 [+0.705, +1.322]**	**+2.954 [+2.523, +3.407]**
Qwen3-8B	N5	**+1.016 [+0.707, +1.292]**	**+1.205 [+0.869, +1.551]**	**+2.221 [+1.766, +2.661]**
Qwen3-8B	N6	**+2.349 [+2.114, +2.592]**	**+0.637 [+0.198, +1.059]**	**+2.986 [+2.490, +3.455]**
Qwen2.5-7B	N1	**-0.446 [-0.776, -0.117]** ↓	**+1.237 [+1.009, +1.464]**	**+0.792 [+0.422, +1.159]**
Qwen2.5-7B	N2	+0.032 [-0.275, +0.351]	**+2.104 [+1.781, +2.441]**	**+2.136 [+1.700, +2.585]**
Qwen2.5-7B	N3	**-0.527 [-0.786, -0.269]** ↓	**+2.104 [+1.757, +2.446]**	**+1.577 [+1.172, +1.968]**
Qwen2.5-7B	N4	**-1.350 [-1.722, -1.017]** ↓	**+1.701 [+1.469, +1.951]**	**+0.350 [+0.028, +0.678]**
Qwen2.5-7B	N5	-0.058 [-0.365, +0.239]	**+1.108 [+0.816, +1.434]**	**+1.050 [+0.624, +1.476]**
Qwen2.5-7B	N6	**-0.670 [-1.062, -0.315]** ↓	**+1.833 [+1.495, +2.190]**	**+1.163 [+0.718, +1.560]**
Phi-3.5-mini	N1	**-1.027 [-1.775, -0.340]** ↓	**+4.833 [+3.616, +6.212]**	**+3.806 [+2.643, +5.132]**
Phi-3.5-mini	N2	**+2.949 [+2.180, +3.829]**	**+4.616 [+3.712, +5.643]**	**+7.565 [+6.117, +9.202]**
Phi-3.5-mini	N3	-0.054 [-0.436, +0.340]	**+4.434 [+3.370, +5.696]**	**+4.379 [+3.075, +5.855]**
Phi-3.5-mini	N4	**-1.168 [-1.561, -0.734]** ↓	**+1.316 [+0.898, +1.813]**	+0.148 [-0.446, +0.871]
Phi-3.5-mini	N5	**+2.810 [+1.765, +3.998]**	**+4.835 [+3.971, +5.792]**	**+7.645 [+5.854, +9.628]**
Phi-3.5-mini	N6	**-3.893 [-4.695, -3.185]** ↓	**+6.369 [+4.978, +7.975]**	**+2.476 [+1.506, +3.649]**
OLMo-2-7B	N1	**+0.783 [+0.658, +0.911]**	-0.017 [-0.100, +0.063]	**+0.766 [+0.633, +0.890]**
OLMo-2-7B	N2	**+0.265 [+0.112, +0.419]**	**+1.386 [+1.239, +1.529]**	**+1.651 [+1.469, +1.834]**
OLMo-2-7B	N3	+0.024 [-0.115, +0.157]	**+0.429 [+0.308, +0.553]**	**+0.453 [+0.226, +0.682]**
OLMo-2-7B	N4	-0.128 [-0.272, +0.020]	**+0.369 [+0.251, +0.478]**	**+0.241 [+0.048, +0.429]**
OLMo-2-7B	N5	**+0.558 [+0.430, +0.672]**	**+0.707 [+0.576, +0.840]**	**+1.265 [+1.059, +1.458]**
OLMo-2-7B	N6	**+0.914 [+0.690, +1.147]**	**+0.526 [+0.407, +0.646]**	**+1.440 [+1.137, +1.751]**

Table B.3 gives the wrong-prior means and intervals behind Figure 3. Own minus unspecified has 12 positive intervals and seven significantly negative intervals, with four reversals in Qwen2.5 and three in Phi-3.5. Unspecified minus different has 22 positive intervals, one additional positive point estimate with an interval spanning zero, and one negative point estimate with an interval spanning zero. It has no statistically significant negative contrast on this side. After the 90% exclusions, only two family-by-wording pairs retain both comparisons, both in Phi-3.5. They are the second and sixth

wordings; the latter contains a reversed own-minus-unspecified difference. Retention is therefore not equivalent to confirming both expected directions.

The original rule required the unspecified-minus-different contrast to meet its directional requirement in at least 18 of 24 wrong-prior comparisons without a significant reversal, while the own-minus-unspecified contrast showed reversals in at least two families. The observed cells meet that rule. Earlier paraphrase results had already reversed the own-minus-unspecified difference in the same two families, but the preregistration did not disclose that antecedent. With only two jointly retained pairs, rescaling the count threshold leaves an uninformative result under the primary reporting rule.

A later diagnostic explains why the original label cannot establish greater wording robustness for one contrast. When two simulated comparisons have equal spread and per-cell uncertainty but differ in their mean distance from zero, the rule is met in 73.5% of 4,000 draws. It is also met in 99.6% of worlds with genuine spread asymmetry at matched location and 7.1% with no difference. The rule responds to both properties. A coefficient-of-variation argument offered earlier does not resolve this: it is algebraically tied to the same mean-to-spread ratio. We therefore retain the observed cells while withdrawing the robustness-asymmetry interpretation, consistently with main §4.3.

Other diagnostics lead to the same limit. Predictors based on the other wordings place all 24 unspecified-minus-different cells outside the region where a confident reversal is reachable, but place eight own-minus-unspecified cells inside it. Four of seven confident reversals lack a counterpart within one confidence-interval half-width. Replacing pooled means while preserving deviations reverses the expected ordering of sign-flip counts. With only seven wordings including the original, the exact frame-level permutation floor is 0.0156, which cannot pass Bonferroni-4. These analyses do not identify intrinsic differences in robustness or independent psychological mechanisms.

Correct-prior results were reported without deciding the study's primary outcome. The unspecified-minus-different contrast has 19 intervals in the expected direction and one significant reversal: OLMo-2 under the first wording, +0.131 [+0.035, +0.226], p = 0.0115. This reversal is unflagged. Own minus unspecified reverses significantly in eight correct-prior comparisons, six in Qwen2.5 and two in Phi-3.5; the total reverses in two. Complete results for both prior-answer arrangements, including flags, are supplied in `results/wording_comparisons.json`. Across the 144 contrast cells, 60 are flagged. The pre-seal null-test history and its unresolved reconstruction issue are documented in §E.1.

B.3 What the shared reference identifies

The unspecified-assistant description names a source and can itself imply credibility or ownership. It has not been shown to be a neutral psychological baseline. On the wrong-prior side, its position between the different-assistant and own-answer scores is 0.468, 0.632, 0.642 and 0.744 across the four families. The equal-length "I was told…" control falls outside that interval in two families (§C.1). These positions make it inappropriate to interpret the three conditions as fixed points along a single psychological axis.

The two paired differences share the unspecified score with opposite signs. Their bootstrap correlations are 0.173, 0.320, 0.552 and 0.261 over 5,000 grouped draws. Summing them cancels that shared score exactly, but neither this identity nor their unequal response to wording identifies independent causal components. A psychological allocation of the total would require knowing how much attribution is already carried by the unspecified condition; this design does not measure that quantity. Earlier numerical "shares" are therefore withdrawn. The supplementary diagnostic record retains the old illustrative calculations for transparency, rather than presenting them as measured mechanisms.

The exploratory and main sets yield similar wrong-prior own-minus-different point estimates: respectively 4.886/4.839 for Qwen3, 3.888/3.308 for Qwen2.5, 13.184/10.990 for Phi-3.5 and 1.689/1.666 for OLMo-2. Their intervals overlap, but this does not establish equivalence or cross-distribution generality. Both sets use PopQA, share relations and have limited wrong-string overlap (§A.1).

B.4 Generated answers, grading and relabelling

Each model first greedily answers the 150 main-set questions. Normalized equality or containment against the reference answer grades these generated answers, after validity checks. An authorship sentence then presents the unchanged generated answer in the prior model-reply position. We distinguish this score experiment from the final-decision intervention in main §4.4, which generates and parses the model's subsequent option choice.

The stages of sample selection are important. The gradable pools contain 144, 145, 147 and 149 answers; their wrong-answer counts are 118, 121, 115 and 112. These give the previously reported 75.2–83.4% error range. Subsequent shared prompt-occurrence checks produce the paired samples in main Table 4: 139, 140, 140 and 139, with 116, 118, 112 and 106 wrong prior answers. The corresponding paired-sample error range is 76.3–84.3%. These percentages describe different selection stages.

Table B.4: Own-minus-unspecified score differences for model-generated prior answers. Positive scores favor the generated answer on both sides. Correct-answer subsets below 30 items are descriptive; the remaining correct-answer criterion is unmet.

Model	Wrong generated answer	Items / groups	Correct generated answer
Qwen3-8B	**+2.476** [+2.174, +2.782], $p \leq 0.0005$	116 / 96	+2.467 [+1.381, +3.512], $p = 0.0030$ (desc., n = 23)
Qwen2.5-7B	**+0.983** [+0.729, +1.242], $p \leq 0.0005$	118 / 110	−0.375 [−1.435, +0.643], n.s. (desc., n = 22)
Phi-3.5-mini	**+4.531** [+3.672, +5.480], $p \leq 0.0005$	112 / 92	+3.321 [+1.825, +5.038], $p = 0.0010$ (desc., n = 28)
OLMo-2-7B	**+0.575** [+0.433, +0.719], $p \leq 0.0005$	106 / 104	+0.133 [−0.148, +0.430], n.s. (criterion, n = 33: **fails**)

Unlike the wrong-minus-correct score elsewhere, Table B.4 is oriented toward the generated answer: positive values mean greater preference for that answer whether it is right or wrong. The own-minus-unspecified contrast for wrong generated answers has a positive interval in all four families before exclusions and three afterward. Qwen2.5's contributing conditions are flagged; Qwen3's flagged different-assistant condition lies outside this comparison. Mean answer-letter probability is at least 0.9996. The correct-generated-answer subsets in Qwen3, Qwen2.5 and Phi-3.5 contain fewer than 30 items and are descriptive under the study's rule. OLMo-2's larger correct subset fails its criterion. Thus no family establishes the original two-sided conjunction, despite the wrong-answer results.

The grading check flags disputed reference answers and possible automatic-grading errors, including containment matches caused by echoing a subject. Excluding all flagged questions leaves the wrong-generated-answer score contrasts positive, with means 2.404, 1.058, 4.478 and 0.581. Two estimates rise and two fall, so this does not establish a uniform attenuation bias. The correct-answer score subsets were not recomputed in this sensitivity analysis. Informal hand-audit error percentages lack a saved emitting analysis and are not presented as validated error-rate estimates.

For the generated final decisions, parsing succeeds in every analyzed condition. The original paired gains appear in main Table 3 and Figure 4. Removing the grading-flagged questions leaves 131 pairs per model; the exploratory estimates are shown below. This tests sensitivity to excluding questionable items rather than supplying independently corrected labels.

Table B.5: Exploratory grading-exclusion sensitivity for final-decision accuracy. Changes and 95% intervals are in percentage points. Original estimates appear in main Table 3 and Figure 4.

Model	Original paired items	After exclusions	Accuracy change after exclusions
Qwen3-8B	139	131	+5.3 [0.0, 10.4]
Qwen2.5-7B	140	131	+0.8 [0.0, 2.3]
Phi-3.5-mini	140	131	+6.1 [2.3, 10.7]
OLMo-2-7B	139	131	+9.2 [3.1, 15.4]

Qwen3's exclusion interval touches zero and its exact grouped sign-flip p-value is 0.0923. Phi-3.5 and OLMo-2 remain interval-positive, while Qwen2.5 remains inconclusive. The original and exclusion results answer related questions on different samples, so both remain visible. The saved final-decision artifacts omit item-input and generation-input hashes, and the rendered-prompt manipulation check covers only the first valid item. Stored decisions support the arithmetic, but later harness repairs cannot establish the historical execution path.

Content selection further limits comparison between correct and incorrect generated answers. Correct-answer items are 6–9 times more popular and concentrated in geography; supplied-answer tests on those same items reproduce much of the pattern. OLMo-2 also emits an "Answer:" prefix in 74 of 149 gradable strings. The prefix is retained identically across descriptions, but its interaction with authorship wording remains untested. Removing authorship framing from genuinely generated answers has not been evaluated; §C.2 concerns supplied answers only.

Main Table 4 separates corrected decisions from harms on correct priors. For Qwen3, the observed +10.3-point gain on wrong priors and −13.0-point change on correct priors imply a descriptive break-even error prevalence of approximately 55.8%. A 50/50 mixture would instead give a 1.3-point accuracy loss. This calculation uses the generated-answer intervention, whereas §B.1's reweighting uses supplied-answer scores. Neither demonstrates a deployment policy that can identify when relabelling will help.

C Additional wording controls and comparisons with prior studies

C.1 Pronouns, source words and their combinations

The exploratory equal-length control compares the own-answer sentence with "I was told an assistant answered: X." Both use first-person wording and have the same token length. This control was added after the core results. The eight score differences retain their expected signs before exclusions, with four remaining under the primary policy.

Table C.1: Own-description minus equal-length first-person control, with 95% cluster intervals. All eight original contrasts have the expected direction; four remain under the primary exclusions. The control is exploratory.

Model	Wrong prior answer	Correct prior answer
Qwen3-8B	**+2.263** [+1.828, +2.690], $p \leq 0.0005$	**−2.765** [−3.176, −2.344], $p \leq 0.0005$
Qwen2.5-7B	**+1.108** [+0.766, +1.477], $p \leq 0.0005$	**−0.628** [−1.128, −0.092], $p = 0.0235$
Phi-3.5-mini	**+12.119** [+9.713, +14.743], $p \leq 0.0005$	**−15.289** [−17.836, −12.831], $p \leq 0.0005$
OLMo-2-7B	**+1.680** [+1.412, +1.951], $p \leq 0.0005$	**−1.002** [−1.344, −0.658], $p \leq 0.0005$

The control changes the meaning of the sentence as well as its attribution. Relative to a scale running from the different-assistant score to the own-answer score, it lies at 0.532, 0.665, −0.103 and −0.008. Its location beyond the different-assistant condition in Phi-3.5 and OLMo-2 shows that

first-person form alone does not define an ordered attribution axis. Relation-grouped inference also makes Qwen2.5's correct-prior interval cross zero. User-voiced first-person wording and identifying the evaluated model by name remain untested.

A separate control replaces "A different assistant…" with "Another assistant…" or "A second assistant…". Six of eight two-arrangement conjunctions meet the original rule, compared with two after exclusions; 10 of 16 individual comparisons remain eligible. All 14 significant original comparisons have the expected direction. However, three families' "second assistant" ratios fall below half the original magnitude, and the correct-prior Qwen3 and Qwen2.5 estimates point in the opposite direction without excluding zero. The results therefore depend on the particular source description even when the original directional count is met.

Using the bare name "Assistant B was asked this and answered: X." gives the expected two-sided result in three families before exclusions and only Phi-3.5 afterward. OLMo-2 reverses on both sides: −0.168 [−0.292, −0.045] for wrong priors and +0.182 [+0.074, +0.291] for correct priors. Qwen2.5 has an exactly zero wrong-choice rate under the bare-name condition despite a nonzero score contrast. These observations rule out describing the effect as invariant to source wording.

We also add the word "assistant" to the own-answer sentence in two forms: "As the assistant, I was asked this and answered: X." and "I, the assistant, was asked this and answered: X." Both are compared with the same unspecified-assistant description.

Table C.2: Own-minus-unspecified contrasts when both descriptions contain the word assistant. Intervals are 95% cluster intervals. The first form meets the two-sided criterion in three families and the second in four before exclusions; each retains two afterward.

Model	First form: wrong	First form: correct	Second form: wrong	Second form: correct
Qwen3-8B	**+3.324** [+2.863, +3.788]	**−2.066** [−2.615, −1.518]	**+2.459** [+2.104, +2.850]	**−2.247** [−2.621, −1.884]
Qwen2.5-7B	**+0.858** [+0.626, +1.094]	−0.190 [−0.759, +0.420], n.s.	**+0.980** [+0.716, +1.273]	**−0.571** [−1.131, −0.025], p = 0.0330
Phi-3.5-mini	**+2.629** [+1.885, +3.401]	**−1.990** [−2.814, −1.243]	**+2.585** [+1.783, +3.435]	**−3.182** [−4.095, −2.334]
OLMo-2-7B	**+1.405** [+1.148, +1.661]	**−0.962** [−1.252, −0.682]	**+0.499** [+0.379, +0.623]	**−0.202** [−0.384, −0.016], p = 0.0360

The first form meets the original two-sided criterion in three families and the second in four, but each retains only two after exclusions. Across the two forms, 12 of 16 individual comparisons and four of eight conjunctions remain eligible. Ratios to the original own-minus-unspecified contrast range from 0.40 to 3.29. Thus including "assistant" can preserve the directional contrast while changing its size. Token length still differs by one or two tokens, and the word's effect is not separable from the surrounding formulation.

Crossing "another assistant" with the two earlier paraphrases gives four of eight expected two-sided conjunctions and three significant correct-prior reversals: Qwen3 under the second paraphrase (+1.022), Qwen2.5 under the first (+0.277) and OLMo-2 under the second (+0.138). No wrong-prior interval reverses significantly, although the Qwen3 and OLMo-2 second-paraphrase point estimates are negative. The original six-of-eight, no-reversal criterion fails, and all eight conjunctions are excluded under the primary rule. Testing source words and sentence structures separately therefore does not establish invariance when they are combined.

C.2 Earlier paraphrases and removing the description

The two exact paraphrase pairs in §A.2 retain the same questions, answer contents and scoring procedure. Their own-minus-unspecified results are shown below. "Uncertain" denotes an interval spanning zero; a reversal denotes an interval in the direction opposite to the prior-answer criterion.

Table C.3: Own-minus-unspecified means for two paraphrases, shown as wrong-prior / correct-prior values. Direction and uncertainty annotations follow the original interval rule; the Phi-3.5 exception is explained below.

Model	First wording: wrong / correct	Second wording: wrong / correct
Qwen3-8B	**+0.814** / **−1.553** (both $p \leq 0.0005$, expected direction)	**+1.252** / **−0.825** (both expected direction)
Qwen2.5-7B	**+0.821** expected direction / −0.300 uncertain	**−0.514** / **+1.183**; **reversed both sides**
Phi-3.5-mini	+0.951 / **−4.406** (first interval excludes zero; see note)	**−0.913** reversed / +0.007 uncertain
OLMo-2-7B	−0.009 uncertain / −0.122 uncertain	**+0.593** / **−0.531** (both expected direction)

The original criterion requires at least six of eight family-by-wording conjunctions with no reversal. Four meet it before exclusions and none afterward. Qwen2.5 and Phi-3.5 reverse under the second wording. Phi-3.5's wrong-prior first-wording interval, 0.951 [0.047, 1.966], excludes zero but its permutation p-value is 0.059. The study's specified interval rule counts that cell; uniformly deferring to permutation inference would instead reduce the conjunction count from four to three. This exception is kept explicit because the sparse-rate policy in §A.4 belongs to a different analysis.

The corresponding own-minus-different totals show how much a one-wording evaluation can change. Table C.4 preserves the post-hoc estimates used in the Introduction and abstract.

Table C.4: Own-minus-different score differences under the original sentence and two paraphrases. Intervals are grouped by the 73 wrong-answer strings. These are post-hoc estimates; all wrong-prior rows and the Phi-3.5 correct-prior paraphrases are flagged.

Model	Prior answer	Original	First paraphrase	Second paraphrase
Qwen3-8B	Wrong	**+4.839** [+4.335, +5.354]	**+2.095** [+1.750, +2.435]	**+1.679** [+1.304, +2.056]
Qwen3-8B	Correct	**−4.342** [−4.827, −3.861]	**−2.395** [−2.755, −2.040]	**−0.592** [−0.939, −0.232]
Qwen2.5-7B	Wrong	**+3.308** [+2.842, +3.802]	**+1.752** [+1.428, +2.081]	**+0.996** [+0.701, +1.269]
Qwen2.5-7B	Correct	**−2.737** [−3.579, −1.901]	−0.615 [−1.249, +0.024]	+0.389 [−0.178, +0.960]
Phi-3.5-mini	Wrong	**+10.990** [+8.960, +13.225]	**+3.236** [+1.804, +4.875]	+0.041 [−0.313, +0.456]
Phi-3.5-mini	Correct	**−14.228** [−16.447, −12.097]	**−8.204** [−10.296, −6.194]	**−1.600** [−2.272, −1.005]
OLMo-2-7B	Wrong	**+1.666** [+1.420, +1.914]	**+0.724** [+0.469, +1.018]	**+0.890** [+0.711, +1.066]
OLMo-2-7B	Correct	**−1.330** [−1.568, −1.098]	**−0.653** [−0.794, −0.499]	**−0.987** [−1.151, −0.821]

Comparing the original and second wording yields interval-excluding-zero differences in all eight family-by-correctness pairs. Only three comparisons remain after the primary exclusions, all with correct priors. Fourteen of the 24 totals are flagged: every wrong-prior total and Phi-3.5's correct-prior totals under both paraphrases. In particular, Phi-3.5's wrong-prior total changes from 10.990 to 0.041, but this is a score change in flagged conditions, not an identified behavioral fold change.

Unequal changes in the two paired contrasts also do not establish independent mechanisms. An additive interaction excludes zero in 12 of 16 comparisons, and a retained-fraction comparison does so in 12 of 16, with its directions split equally. Their joint count is nine, exactly the count expected from those marginal proportions; primary exclusions reduce these counts to four, four and three. Wording-only differences reuse the same condition means. The measurements show dependence on the sentence frame while leaving the psychological allocation unresolved (§B.3).

Finally, we remove the description and leave only the supplied prior answer in the model-reply turn. Relative to unspecified-assistant wording, the wrong-prior score changes are −1.542, −4.872 and −1.619 in Qwen2.5, Phi-3.5 and OLMo-2, with intervals excluding zero and mirrored correct-prior directions. Only Phi-3.5 retains the reversal under the primary rule. Qwen3 instead keeps the original direction, +4.568/−4.145. These are the reversals described in the Introduction. They differ from the original success criterion, which is met by one family before exclusions and none afterward. Removing a whole sentence changes length and phrasing as well as stated attribution; this experiment does not test an unframed answer actually generated by the model.

C.3 Comparisons with earlier attribution studies

The two closest studies differ in task, prompts and measurement. Kumaran et al. (2025) evaluates whether a visible prior choice changes under advice; its different-model description also says the source's accuracy is unknown, and its baseline hides the answer. Khullar et al. (2026) finds little separation among explicit source labels on identical insecure patches, with larger differences for implicit attribution. Its scalar rating task differs from our factual forced choice. We use two transplant experiments to investigate these differences while retaining the contrary finding.

The first puts their source descriptions into our knowledge-conflict task. The prior turn contains the bare answer and the attribution precedes the final user turn. The advice-study pair is used verbatim: "You can see above your original answer to the question." versus "You can see above the original answer to the question which was provided by another LLM (LLM 1) of similar size (in billions of parameters) to you. The accuracy of LLM 1 is unknown." The rating-study prompts are adapted from imperatives to "The answer above is your answer." and "The answer above was generated by another model." This unequal adaptation is a confound.

Table C.5: Prior-study source descriptions transplanted into the knowledge-conflict task. The advice pair is verbatim; the rating pair is adapted from imperatives to declaratives. Values are own minus other with 95% cluster intervals. This adaptation asymmetry limits the comparison.

Model	Advice pair: wrong	Advice pair: correct	Rating pair: wrong	Rating pair: correct
Qwen3-8B	**+3.143** [+2.585, +3.736]	**−5.513** [−6.376, −4.684]	**+7.074** [+6.000, +8.230]	**−9.264** [−10.151, −8.455]
Qwen2.5-7B	−0.519 [−1.230, +0.226]	−0.602 [−1.884, +0.618]	**+0.844** [+0.414, +1.304]	**−1.393** [−1.916, −0.861]
Phi-3.5-mini	+0.641 [−1.832, +3.044]	**−3.699** [−7.008, −0.628]	**+10.973** [+8.582, +13.724]	**−20.017** [−23.807, −16.542]
OLMo-2-7B	**−1.083** [−1.350, −0.808] (reversed)	**+0.762** [+0.355, +1.209] (reversed)	**+1.637** [+1.403, +1.881]	**−2.372** [−2.866, −1.926]

The predicted ordering fails in every family. The adapted rating-study pair produces a larger wrong-prior own-minus-other contrast in all four; the verbatim advice pair reverses significantly on both sides in OLMo-2. The original positive criterion also requires a positive advice contrast and an absolute rating-to-advice ratio below 0.5 in at least three families; it is unmet. Its study-specific exclusion threshold was 98%, reached by no family. Applying the paper's 90% threshold instead excludes all four ratio comparisons. The advice study's different-model sentence is also 24 words longer. These results leave the disagreement unresolved rather than locating it solely in attribution wording.

The reverse transplant keeps our original authorship sentences but changes the final request. One version supplies advice from a model "known to have 70% accuracy" and reads option probabilities after "My final answer is:". The other uses the rating study's correctness rubric, taking an expected rating over digits 1–9 because "10" requires two tokens in three tokenizers. Rating differences are oriented toward the wrong answer. Under unspecified authorship, every model rates correct answers above wrong answers, with intervals excluding zero, providing a correctness check on the rating measure.

Table C.6: Our source descriptions under the advice request. Values are differences in wrong-minus-correct option scores, with 95% cluster intervals. Downward arrows mark a contrast opposite to the expected prior-answer direction.

Model	Prior answer	Own − unspecified	Unspecified − different	Own − different
Qwen3-8B	Wrong	**+3.401 [+2.938, +3.873]**	**+0.563 [+0.297, +0.828]**	**+3.964 [+3.411, +4.540]**
Qwen3-8B	Correct	**-3.105 [-3.548, -2.693]**	**+0.464 [+0.187, +0.717]** ↓	**-2.641 [-3.194, -2.123]**
Qwen2.5-7B	Wrong	**+1.953 [+1.549, +2.373]**	**+3.009 [+2.566, +3.484]**	**+4.963 [+4.306, +5.631]**
Qwen2.5-7B	Correct	**-0.632 [-0.952, -0.332]**	**-2.139 [-2.609, -1.677]**	**-2.771 [-3.404, -2.179]**
Phi-3.5-mini	Wrong	**+1.476 [+1.220, +1.749]**	**+2.012 [+1.736, +2.294]**	**+3.488 [+3.063, +3.950]**
Phi-3.5-mini	Correct	**-0.923 [-1.154, -0.711]**	**-1.490 [-1.685, -1.298]**	**-2.413 [-2.734, -2.127]**
OLMo-2-7B	Wrong	**+1.148 [+0.984, +1.320]**	**+1.344 [+1.210, +1.482]**	**+2.492 [+2.315, +2.674]**
OLMo-2-7B	Correct	**-0.805 [-0.934, -0.677]**	**-1.087 [-1.214, -0.966]**	**-1.891 [-2.095, -1.699]**

Table C.7: Our source descriptions under the correctness-rating request. Differences are expected rating points oriented toward the wrong answer, with 95% cluster intervals. This bounded scale differs from the advice task's log-probability scale.

Model	Prior answer	Own − unspecified	Unspecified − different	Own − different
Qwen3-8B	Wrong	+0.025 [-0.045, +0.099]	**+0.242 [+0.164, +0.324]**	**+0.268 [+0.170, +0.377]**
Qwen3-8B	Correct	+0.015 [-0.064, +0.091]	**-0.209 [-0.290, -0.134]**	**-0.194 [-0.303, -0.093]**
Qwen2.5-7B	Wrong	**+0.363 [+0.257, +0.480]**	**+0.330 [+0.181, +0.476]**	**+0.693 [+0.495, +0.899]**
Qwen2.5-7B	Correct	**-0.206 [-0.327, -0.073]**	**-0.339 [-0.536, -0.157]**	**-0.545 [-0.696, -0.383]**
Phi-3.5-mini	Wrong	**+0.270 [+0.108, +0.439]**	**+0.373 [+0.065, +0.676]**	**+0.643 [+0.249, +1.046]**
Phi-3.5-mini	Correct	+0.003 [-0.058, +0.072]	**-0.179 [-0.317, -0.064]**	**-0.176 [-0.353, -0.015]**
OLMo-2-7B	Wrong	-0.015 [-0.187, +0.149]	**+0.701 [+0.563, +0.849]**	**+0.686 [+0.536, +0.836]**
OLMo-2-7B	Correct	**+0.258 [+0.152, +0.370]** ↓	**-0.775 [-0.883, -0.674]**	**-0.517 [-0.649, -0.380]**

Own-minus-different totals have intervals in the expected direction in all eight arrangements under each readout, although Phi-3.5's correct-prior rating contrast has permutation $p = 0.0575$. The advice task meets the study's movement rule in all four families. For ratings, two families clear the rescaled 0.45-point equivalence boundary, one lies inside and one straddles it; only one clears a literal 0.5-point boundary. Bounded ratings and log-probability differences are distinct scales. The combined outcome remains unresolved, and that same unresolved label occurs in six of twelve null trials. It cannot establish that localization is impossible.

The reverse experiment flags concentration on a particular output letter or digit, whereas the core policy flags concentration on correct or wrong answers. These differ on six of 24 advice conditions: a letter-based rule flags OLMo-2 conditions that are not correctness-concentrated and misses highly correctness-concentrated Qwen2.5 conditions. Rating has no comparable correct/wrong decision axis. These flags cannot be silently substituted. The two transplants vary task and item interpre-

tation along with the readout or attribution position, so a one-factor explanation of the literature disagreement remains open.

Earlier studies use different references: an absent authorship label (Singla et al., 2025), a hidden answer (Kumaran et al., 2025), or unlabeled scores from a separate experiment (Chae et al., 2026). Our unspecified-assistant reference keeps content visible within the same item. Main §2 describes these distinctions without claiming a first discovery. Likewise, message-role relabelling in Chen et al. (2026) changes the chat-template position of erroneous content, whereas our generated-answer intervention changes its source sentence at a fixed position. Their measured correction behavior should not be equated with our final-decision accuracy gains.

D Additional models, tasks and reasoning tests

D.1 Model size and numerical precision

The size extension adds eight Qwen checkpoints to the two known 7–8B Qwen anchors while retaining the same main-set questions and score criterion. It covers 0.5B–14B parameters within Qwen2.5 and Qwen3, rather than eight independent model families. The two 14B models first run in predeclared NF4 quantization and then in a separately specified bfloat16 comparison. The tables use bfloat16 where available. Their 14B paraphrase and choice-rate results remain NF4-only.

Table D.1: Own-minus-unspecified score differences across Qwen sizes, with 95% cluster intervals. Met refers to the original two-sided interval rule; a dagger marks a size with a flagged criterion condition. The surrounding text gives exclusion and multiplicity results.

Model	Wrong prior answer	Correct prior answer	Original result
Qwen3-0.6B	−0.154 [−0.349, +0.044], p = 0.1449	+0.219 [+0.058, +0.381], p = 0.0125	Correct-prior reversal
Qwen3-1.7B †	**+2.133** [+1.551, +2.776], p ≤ 0.0005	**−1.759** [−2.187, −1.345], p ≤ 0.0005	Met
Qwen3-4B †	−0.034 [−0.389, +0.304], p = 0.8401	+0.402 [−0.100, +0.965], p = 0.1354	Unmet
Qwen3-8B †	**+2.576**	**−2.294**	Met
Qwen3-14B (bf16) †	**+0.734** [+0.453, +1.024], p ≤ 0.0005	**−1.901** [−2.334, −1.471], p ≤ 0.0005	Met; NF4 also met
Qwen2.5-0.5B	+0.003 [−0.072, +0.082], p = 0.9415	+0.013 [−0.064, +0.088], p = 0.7621	**not distinguishable from zero**
Qwen2.5-1.5B †	**+0.203** [+0.053, +0.352], p = 0.0130	**−0.144** [−0.266, −0.013], p = 0.0355	Met
Qwen2.5-3B	**+3.209** [+2.405, +4.082], p ≤ 0.0005	**−4.649** [−5.546, −3.723], p ≤ 0.0005	Met
Qwen2.5-7B †	**+1.219**	**−0.920**	Met
Qwen2.5-14B (bf16) †	**+3.535** [+2.804, +4.370], p ≤ 0.0005	**−3.347** [−4.280, −2.450], p ≤ 0.0005	Met; NF4 also met

Table D.2: Unspecified-minus-different score differences across added Qwen sizes. Positive wrong-prior and negative correct-prior differences have the expected direction. These descriptive contrasts include flagged conditions and uncorrected results; the adjacent text gives the correction and precision qualifications.

Model	Wrong prior answer	Correct prior answer	Original result
Qwen3-0.6B	**+0.902** [+0.720, +1.082], p ≤ 0.0005	**−0.719** [−0.894, −0.544], p ≤ 0.0005	Expected direction
Qwen3-1.7B †	**+0.799** [+0.500, +1.123], p ≤ 0.0005	**−0.988** [−1.327, −0.647], p ≤ 0.0005	Expected direction
Qwen3-4B †	**+1.523** [+1.011, +2.143], p ≤ 0.0005	**−2.037** [−2.817, −1.370], p ≤ 0.0005	Expected direction

Table D.2 (continued)

Model	Wrong prior answer	Correct prior answer	Original result
Qwen3-14B (bf16) †	**−1.639** [−1.974, −1.328], $p \leq 0.0005$	**+1.753** [+1.332, +2.196], $p \leq 0.0005$	Reversed on both sides (bf16)
Qwen2.5-0.5B	**+0.158** [+0.096, +0.218], $p \leq 0.0005$	**−0.134** [−0.183, −0.085], $p \leq 0.0005$	Expected direction
Qwen2.5-1.5B †	+0.276 [+0.050, +0.514], $p = 0.0185$	−0.247 [−0.512, +0.012], $p = 0.0790$	Wrong-prior side only
Qwen2.5-3B	+1.213 [+0.193, +2.219], $p = 0.0200$	+0.296 [−0.911, +1.497], $p = 0.6452$	Wrong-prior side only
Qwen2.5-14B (bf16) †	−0.602 [−1.061, −0.072], $p = 0.0185$	−0.721 [−1.545, +0.180], $p = 0.1249$	Wrong-prior reversal before correction

The own-minus-unspecified criterion is met at five of the six new sizes between 1.5B and 14B before correction, alongside the two known anchors. Bonferroni-50 removes Qwen2.5-1.5B, leaving four of those six. After the primary exclusions, only Qwen2.5-3B meets the criterion. The descriptive unspecified-minus-different comparison also weakens under correction: Qwen2.5's 1.5B and 3B wrong-prior results, as well as the adverse 14B result, fail the same corrected threshold. Applying correction only to contrary findings would give a misleading picture.

The smallest checkpoints illustrate wording dependence. Qwen2.5-0.5B has no interval-positive own-minus-unspecified contrast in the original wording, but its two paraphrases give +0.105 [0.031, 0.177] and +0.135 [0.087, 0.184] for wrong priors, with mirrored correct-prior results. Qwen3-0.6B has four reversed own-minus-unspecified cells across the tested wordings. Although the different-source contrasts keep the expected direction across the six small-model wording combinations, the proposed dissociation between the two effects fails its original robustness criterion. These observations do not identify a model size at which one psychological mechanism appears.

Qwen3-14B reverses the unspecified-minus-different contrast in the original wording at bfloat16 on both sides: −1.639 and +1.753. Quantization magnifies rather than creates that reversal: bfloat16-minus-NF4 differences are +0.344 [0.176, 0.526] and −0.878 [−1.236, −0.535]. Under the first NF4 paraphrase the wrong-prior contrast becomes inconclusive, while the correct-prior reversal remains +0.481 [0.202, 0.773]. Under the second, the wrong-prior contrast returns to the expected direction and the correct-prior contrast is inconclusive. The full two-sided paraphrase criterion therefore fails. Qwen2.5-14B is mixed, with a weak wrong-prior reversal before correction and an inconclusive correct-prior result.

Mean answer-letter probability is at least 0.9705 across these conditions, although the lowest individual bfloat16 Qwen2.5-14B value is 0.262. The prespecified mean threshold is 0.5; it has only been observed failing at the much more extreme DeepSeek readout, leaving intermediate failures less well tested. The same occurrence checks retain 147 questions.

The choice-rate analysis is separate from the score analysis. Seven intervals exclude zero despite permutation p-values of 0.054–0.135 and are treated as inconclusive. Five of eight added checkpoints have no significant own-minus-unspecified rate change on either side before correction; only Qwen2.5-3B has both. Under Bonferroni-48, no own-minus-unspecified rate effect remains. Retained effects are Qwen3-0.6B's own-minus-different wrong-prior contrast and both different-source contrasts, Qwen2.5-0.5B's own-minus-different wrong-prior contrast, Qwen2.5-3B's own-minus-different wrong-prior contrast, and the reversed correct-prior different-source contrast in Qwen3-14B. Complete intervals and decisions remain in the evidence files. Neither family shows a consistent monotonic size trend.

D.2 Other factual questions and free-form answers

On 150 SciQ science questions (Welbl et al., 2017) with authored distractors and 144 wrong-answer groups, Qwen3 and Phi-3.5 meet the original two-sided own-minus-unspecified criterion. Qwen2.5 and OLMo-2 have inconclusive correct-prior comparisons, so the required three-family count is missed. All four wrong-prior score contrasts are positive before exclusions, but all are flagged by the primary rule. Qwen2.5 selects the wrong answer on none of its wrong-prior conditions despite a +0.808 score difference. The unspecified-minus-different contrast has seven of eight expected-

direction intervals before exclusions, with only three correct-prior comparisons retained afterward. This limits the generality of the PopQA result without changing what was measured in that task.

The generated-answer relabelling intervention also transfers incompletely to free-form final answers. With greedy decoding and containment grading, Phi-3.5 gains +0.109 [0.041, 0.177] in accuracy; Qwen3's +0.028 and OLMo-2's −0.014 changes are inconclusive, while Qwen2.5's measured change is zero. The original requirement of at least two positive families is missed. Ungradable outputs are at most 8.3% per condition. The saved outputs lack the modal fields needed for the paper's primary exclusion recount, so that original failed criterion is retained without claiming a new exclusion-adjusted test. Absolute accuracy also depends strongly on response format: OLMo-2 reaches 0.785 here compared with 0.410 under the own-labelled forced-choice condition.

D.3 Deliberate reasoning and position controls

The sealed reasoning experiment compares thinking on and off for Qwen3-8B and Qwen3-14B using the same questions and authorship descriptions. The primary outcome is the paired change in parsed final-letter decisions for correct priors, tested with grouped sign flips. A document-agreement control checks responsiveness. A content-free prefix with thinking off checks whether simply placing text before the final answer changes the readout. A separate step-by-step prompt is secondary. Generation uses batch 8 with SDPA attention; score reads use eager attention at batch 1. Thinking-off cloud comparisons reproduce their anchors exactly for both models.

Table D.3: Paired changes in wrong-choice rate contrasts. Values are fractions with 95% intervals, not log-probability differences. Correct-prior thinking comparisons are primary; wrong-prior and prefix results describe secondary behavior and floor effects.

Model / manipulation	Prior answer	Own − unspecified change	Unspecified − different change
Qwen3-8B / thinking	Wrong	-0.022 [-0.090, +0.043]	-0.043 [-0.104, +0.008]
Qwen3-8B / thinking	Correct	+0.080 [+0.000, +0.158]	+0.043 [-0.051, +0.139]
Qwen3-8B / prefix	Wrong	**-0.068 [-0.109, -0.032]** (p = 0.0025)	-0.014 [-0.042, +0.013]
Qwen3-14B / thinking	Wrong	+0.014 [-0.042, +0.070]	+0.007 [-0.045, +0.062]
Qwen3-14B / thinking	Correct	+0.057 [-0.021, +0.144]	-0.071 [-0.148, +0.000]
Qwen3-14B / prefix	Wrong	+0.000 [+0.000, +0.000] (floor)	+0.000 [-0.027, +0.027]

Neither model establishes a primary change at the specified permutation threshold. For Qwen3-8B, the own-minus-unspecified change is +8.0 [0.0, 15.8] percentage points (p = 0.101, 137 pairs), and the unspecified-minus-different change is +4.3 [−5.1, 13.9] (p = 0.464, 138 pairs). Qwen3-14B gives +5.7 [−2.1, 14.4] (p = 0.227) and −7.1 [−14.8, 0.0] (p = 0.097), each over 141 pairs. Table D.3 reports the stored fractions underlying those percentage-point summaries. The document-agreement controls respond in both models.

The parse threshold was lowered from 0.95 to 0.90 after a pilot in which two of twenty traces reached the 1,024-token cap. The amendment preceded the main launch, but the preregistration retains both numbers. Observed parsing is 0.94–0.97 for 8B and 0.97–1.00 for 14B. Some 8B conditions would have failed the older threshold. Results therefore apply to the paired parsed completions under the amended rule.

The design can resolve changes of approximately 10.6–13.6 percentage points, rather than the earlier amendment's invalid 8–9-point estimate. Approximate power for 5–6-point changes is only 0.25. The 8B own-minus-unspecified estimate is smaller than its 11.3-point minimum detectable effect and its interval allows a sizable change. Likewise, the 14B experiment cannot reliably detect abolition of its approximately five-point reversed different-source effect. The original no-movement labels express unresolved changes, not equivalence or evidence that reasoning leaves the effects intact.

The position control is informative on wrong priors: in Qwen3-8B, the content-free prefix alone reduces the own-minus-unspecified choice-rate contrast by 6.8 [3.2, 10.9] percentage points (p

= 0.0025). Its score contrast remains positive. Several reasoning conditions have near-zero wrong-choice rates, and the 14B prefix-control own-minus-unspecified change is exactly zero on flagged conditions. These floor observations are reported without interpreting them as absence of an effect. Complete secondary score, step-by-step and prefix-control estimates remain in `results/reasoning_details.json`.

D.4 Activation intervention and its failed specificity control

An exploratory intervention on Qwen3-8B tests a direction formed from paired hidden activations at layer 18. The hook check has zero mismatch, and a per-item replacement control shifts the measured score by +0.695 [0.582, 0.806]. Adding the mean direction at strength 2 shifts held-out unspecified-condition scores by +3.346 [2.856, 3.891]; subtracting it from the own-answer condition removes 2.390 of a 2.458 contrast.

However, a norm-matched random direction also shifts the score by +0.837, with an interval above zero. This is 25.0% of the targeted shift and exceeds the specified limit of 20%. A positive mechanistic interpretation therefore fails the specificity requirement. The result also fails the negative criterion, which requires targeted steering to be no larger than random steering: here it is four times larger. The control works, so the machinery-failure criterion does not apply either. The saved program's fallback negative label conflicts with these written conditions. We classify the experiment as inconclusive, preserving that mismatch rather than treating it as evidence for or against a mechanism. No new exclusion-adjusted analysis was performed.

D.5 Commercial-model behavior and floor effects

Four served models were evaluated on 2026-09-04 with the same 147 questions and three descriptions, using four default-sampler draws per item and condition. Reasoning was disabled; Claude used low effort, and temperature was not settable. The models are single served configurations on that date, not fixed future checkpoint identities. Per-item wrong-choice averages use the same 73 wrong-string groups for intervals and permutation inference.

Table D.4: Wrong-choice fractions in the commercial-model experiment. These are supplied-answer, original-wording results; they do not evaluate the generated-answer intervention.

Model	Wrong prior: different / unspecified / own	Correct prior: different / unspecified / own
Claude Opus 5	0.022 / 0.036 / 0.051	0.197 / 0.190 / 0.124
Claude Sonnet 5	0.000 / 0.000 / 0.000	0.714 / 0.796 / 0.765
GPT-5.6 Sol	0.000 / 0.000 / 0.000	0.643 / 0.670 / 0.585
GPT-5.5	0.022 / 0.014 / 0.032	0.359 / 0.391 / 0.253

Table D.5: Behavioral rate differences with 95% cluster intervals and uncorrected p-values. Fractions multiply by 100 to give percentage points. Exact-zero wrong-prior conditions are floors; the text reports Bonferroni-20 results separately.

Model	Own − unspecified: wrong	Own − unspecified: correct	Unspecified − different: correct
Claude Opus 5	+0.015 [−0.010, +0.044], p = 0.3538	−0.066 [−0.112, −0.022], p = 0.0035	−0.005 [−0.054, +0.045], p = 0.9010
Claude Sonnet 5	+0.000 [+0.000, +0.000], p = 1.0000 (floor)	−0.031 [−0.074, +0.008], p = 0.1689	+0.080 [+0.041, +0.125], p ≤ 0.0005 (reversed)
GPT-5.6 Sol	+0.000 [+0.000, +0.000], p = 1.0000 (floor)	−0.085 [−0.123, −0.049], p ≤ 0.0005	+0.027 [+0.000, +0.057], p = 0.0790
GPT-5.5	+0.019 [+0.005, +0.035], p = 0.0165	−0.138 [−0.184, −0.094], p ≤ 0.0005	+0.032 [+0.005, +0.059], p = 0.0290 (reversed)

Only behavioral estimates are interpretable here. Claude exposes no logits, and the OpenAI responses often expose a single alternative in the returned top-five probabilities. The old reconstruction

assigns exactly zero to every censored draw, including all 2,399 such GPT-5.5 draws and 2,997 Sol draws. Per-condition censoring ranges from 0.43 to 0.997, so mixtures of these artificial zeros with observed alternatives cannot estimate score effects. The graded positive controls also fail; extending a censoring exception to those controls was an unstated interpretation of the original specification.

The behavioral document-agreement control raises wrong-choice rates by 0.394, 0.825, 0.833 and 0.875 in model order. It also adds a third occurrence of the candidate, whereas the main conditions contain two. It therefore validates responsiveness to the combined document-and-mention manipulation, without isolating a causal ratio between document evidence and authorship. Parsing failures are at most 1.7% per condition and arise from truncation. Opus has 1, 0 and 10 wrong-prior failures under own, unspecified and different descriptions; two items lose all four different-condition draws. Its paired total therefore uses 145 items. The specified removal rule does not establish that truncations are independent of attribution.

Sonnet and Sol produce no wrong choices in any wrong-prior condition, out of 588 draws per condition. These floors leave reductions unmeasurable. Opus and GPT-5.5 have small wrong-prior rates; GPT-5.5's own-minus-unspecified increase of 0.019 [0.005, 0.035] meets only the uncorrected rule. Wrong-prior unspecified-minus-different contrasts are inconclusive in both measurable models: Opus +0.014 [−0.012, 0.044], $p = 0.51$, and GPT-5.5 −0.009 [−0.022, 0.002], $p = 0.28$. The local-model wrong-prior pattern is therefore not replicated across these commercial configurations.

For correct priors, own attribution reduces wrong-document following in three families before multiplicity correction. Under Bonferroni-20, the retained contrasts are Sol's and GPT-5.5's own-minus-unspecified differences and Sonnet's reversed unspecified-minus-different difference. Opus's $p = 0.0035$, GPT-5.5's wrong-prior $p = 0.0165$ and its reversed correct-prior $p = 0.0290$ do not survive. The paper's interpretation follows the stated intervals, correction and floor limits rather than the program's outcome labels.

These runs use only the original sentence frame and supplied priors. They neither test the generated-answer intervention on commercial models nor establish a capability trend. Near-floor estimator tests have power of only 0.43 for a two-point rate change, making the observed 1.9-point GPT-5.5 estimate vulnerable to selection inflation. Synthetic controls omit within-group correlation; a null using the actual group structure gives false-positive rates of 0.05–0.07. Both limits matter when interpreting small effects.

E Study history and evidence status

E.1 Design development and recorded amendments

The exploratory PopQA study initially froze two models and a bare-answer versus different-assistant comparison. That comparison combined attribution, wording and length. The matched own-answer sentence and two further models were added after results were inspected. Its matched comparisons describe design development, and the exploratory percentile-index defect remains unresolved by a rerun. A deterministic repeat of unchanged inputs was initially treated as replication and later withdrawn from that role.

The main experiment introduces subject-disjoint questions, the unspecified-assistant condition and corrected interval indexing. Its saved item set predates the written specification. Later harness edits prevent independent reconstruction of the original ordering from the current checkout, and the core lacks the predating prior-work record and criterion-specific validation files required by the project's sealing procedure. Consequently, the core remains unsealed. Later sealed experiments cannot retrospectively fix that history. The equal-length first-person control and the factorial wording analyses are explicitly post-hoc.

The later six-wording and reverse-transplant designs have committed specifications and seals preceding their harness code, together with prior-work records and outcome-rule tests. The reasoning design also has sealed files. Other additions rely on dated specifications and contemporaneous lab notes rather than an independently reconstructable pre-code chain. The commercial-model validity rule and the reasoning parse threshold were amended after pilots; their actual consequences appear in §D.3 and §D.5.

Before sealing the six-wording study, its initial null configuration gave a false-positive rate of 0.125 and failed a 0.10 threshold. Expanding the configuration from 48 to 192 cells and changing seed placement gave 0.036. The current 48-cell code instead gives 0.021, so the historical 0.125 cannot be reconstructed from its present writer. The larger configuration also reuses eight source vectors: its seven hits lie in four blocks, with paired contrasts sharing sign-flip draws. The effective independent information is smaller than the cell count. Moreover, that null did not fully exercise the reversal clause that decided the observed outcome. The later location diagnostic in §B.2 is therefore essential to its interpretation.

The reverse-transplant advice equivalence band was widened from 0.50 to 0.75 logits before sealing because the narrower band could not reliably classify realistic no-effect noise as equivalent. These changes preceded the relevant experimental results, but they are still design choices informed by validation exercises. The numerical companion preserves reported tables; the historical run records and amendments belong to the separately retained research-code package (§F.2).

E.2 Selected conditions reported in the abstract

The abstract's fifteen conditions were selected retrospectively from the study record. They share observations and experiments; the roster is neither exhaustive nor fifteen independent tests. Nine conditions are missed under their original analyses. Under the paper's later primary reporting policy, thirteen can be evaluated and are unmet; the already-unmet free-form and activation conditions are carried forward without a new exclusion-adjusted calculation. Thus none of the selected conditions is established under that policy, without implying that every row received a fresh statistical test.

Table E.1: The fifteen retrospectively selected conditions cited in the abstract. Counts within rows refer to the specified unit, not independent experiments. Thirteen primary-policy rows are evaluated and unmet; two original failures are carried forward without a new calculation.

Selected condition	Original requirement	Original analysis	Primary reporting policy
Exploratory supplied answers (§E.1)	Both directions in either of two specified models	1 of 2: met	0 of 2: missed
Main own-attribution test (main §4.1)	Both directions in at least three families	4 of 4: met	2 of 4: missed
Generated-answer scores (§B.4)	Both directions in at least three families	0 of 4: missed (wrong-prior only: 4 of 4)	0 of 4: missed
Two paraphrases (§C.2)	At least six of eight conjunctions, no reversal	4 of 8, 2 reversals: missed	0 of 8: missed
Science questions (§D.2)	Both directions in at least three families	2 of 4: missed	0 of 4: missed
Sub-1B dissociation (§D.1)	Survives both paraphrases	0 of 4 rung-frame cells: missed	0 of 4: missed
14B reversal (§D.1)	Survives both paraphrases	0 of 2: missed	0 of 2: missed
Bare prior reply (§C.2)	Expected directions in at least three families	1 of 4: missed	0 of 4: missed
Alternative source words (§C.1)	At least six of eight conjunctions, no reversal	6 of 8: met	2 of 8: missed
Free-form intervention (§D.2)	Positive interval in at least two families, none negative	1 of 4: missed	Not recomputed: modal data absent
Source word / wording cross (§C.1)	At least six of eight conjunctions, no reversal	4 of 8, 3 reversals: missed	0 of 8: missed
Activation specificity (§D.4)	Random shift below 20% of targeted shift	25.0%: missed	Not recomputed: originally unmet
Assistant word: first form (§C.1)	Both directions in at least three families	3 of 4: met	2 of 4: missed
Assistant word: second form (§C.1)	Both directions in at least three families	4 of 4: met	2 of 4: missed
Six-wording rule (§B.2)	Directional-count and reversal rule in §B.2	22 of 24 expected-direction intervals: met	Only two joint pairs: unmet

For the directional comparisons in Table E.1, "both" means the expected interval direction with wrong and correct priors in the same model. The roster omits the generated-final-decision intervention's per-model success conditions, individual size and replication conditions, commercial-model and reverse-transplant outcomes, and the first transplant's additional positive criterion. That transplant criterion is itself unmet (§C.3). The roster was not preregistered.

Removing the three arguable rows, the generated-content conjunction and both size-paraphrase conditions, leaves twelve selected conditions, six missed originally and none established under the primary policy. Counting all four exploratory models instead of the two initially specified would establish one selected condition through Phi-3.5. That alternative leaves twelve evaluated unmet conditions and two carried failures, fourteen in total. These alternatives describe selection sensitivity; they do not replace the roster disclosed in the abstract or summarize the entire project as a single success rate.

F Reproduction materials

F.1 Models, settings and environment

Table F.1: Public model identifiers and recorded checkpoint revision prefixes. DeepSeek is listed for reproducibility, outside the four substantive model families.

Model ID	Revision prefix	Role
Qwen/Qwen3-8B	b968826d	Main
Qwen/Qwen2.5-7B-Instruct	a09a3545	Main
microsoft/Phi-3.5-mini-instruct	2fe19245	Main
allenai/OLMo-2-1124-7B-Instruct	470b1fba	Main
deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B	ad9f0ae0	Invalid readout; §A.3

The added checkpoints are Qwen3 at 0.6B, 1.7B, 4B and 14B and Qwen2.5-Instruct at 0.5B, 1.5B, 3B and 14B. Revision prefixes in ascending size order are `c1899de2`, `70d244cc`, `1cfa9a72`, `40c06982` for Qwen3 and `7ae55760`, `989aa798`, `aa8e7253`, `cf98f3b3` for Qwen2.5. Core settings are `torch.bfloat16`, `device_map="cuda"` and `attn_implementation="eager"`; `enable_thinking=False` is effective for Qwen3. NF4 runs use bfloat16 computation. The 14B bfloat16 comparisons cover the original score experiment, while its later paraphrase and rate results remain quantized.

Environment records report PyTorch 2.11.0+cu128, Transformers 5.15.0 and Accelerate 1.14.0. Local runs use an RTX 5080 Laptop GPU; cloud bfloat16 and reasoning runs use an L40S. Because cross-card bfloat16 results need not be bit-identical, the reasoning experiment repeats thinking-off comparisons on the cloud card. Admission requires agreement with the anchor or, failing that, agreement between repeated cloud runs; otherwise the comparison is invalid. Both tested models match the anchor exactly.

Core bootstrap seeds are 601–607, with permutation seeds offset by seven. Exploratory bootstrap seeds are 501–512. Synthetic clustered tests use base seed 3 and trial seeds `900 + i` in the harness or `90000 + i` in the standalone test. Core choice-rate seeds begin at 7100 and size-extension rate seeds at 8100. Dataset snapshots, sampling seeds and item hashes are in §A.1. Saved compute totals cover particular run groups rather than the whole campaign: 2,525.5 seconds for the listed local core/control/size harnesses, approximately fifteen minutes for the bfloat16 comparison, 14,688 commercial API calls and 6.165 hours of reasoning-run wall time.

F.2 Evidence files and reproduction scope

The accompanying evidence archive uses descriptive relative filenames and a checksum manifest. It includes complete saved table records, both sides of the six-wording study, secondary reasoning measurements and the grading-sensitivity estimates. Numerical values are copied from the saved analyses rather than recomputed to fit the rewritten text. The manifest identifies each file's role and distinguishes historical outcome labels from the current qualified conclusions. The full earlier tables remain accessible even when only the primary comparisons appear in this PDF.

`results/wording_comparisons.json` contains the complete wording cells and flags. `results/reasoning_details.json` retains the decision, score and position-control results removed from the long PDF tables. `results/grading_sensitivity.json` records the exploratory exclusions alongside original estimates. Core subgroup values and size-extension choice rates appear in `results/core_sensitivity.json` and `results/size_choice_rates.json`. `results/archived_tables.json` retains every cell of the previous tables, including withdrawn descriptive diagnostics, with their status marked. `manifest.json` gives file roles and SHA-256 checksums. The evidence archive is a numerical companion, not a claim that every historical run can be reproduced from it alone.

Executing the original experiments additionally requires the model checkpoints, datasets, full historical harness dependencies and, for the API study, access to the served configurations. The existing research-code package and original artifacts are preserved separately. Renaming or condensing the reader-facing appendix does not establish an executable reconstruction of missing historical inputs. No new model observations were generated for this editorial revision.

F.3 Provenance, validation and data use

Several records have incomplete provenance. The generated-final-decision artifacts omit hashes for their item and generated-answer inputs. Some reconstruction and sensitivity records omit execution markers, and the initial frame-null record lacks the historical wrapper needed to recover the original reported null rate. Numerical agreement with stored vectors does not restore those missing links. Historical unrecorded rebinding counts, pre-fix interval estimates and informal grading rates are therefore excluded from the scientific evidence presented here.

Validation of estimators also differs from validation of every outcome label. Tests of both branches of the shipped labeling rule exist for the assistant-word controls, six-wording experiment, reverse transplant and reasoning study. The seven other label rules assessed in the earlier inventory rely on estimator tests or lack complete branch tests. The paper's post-data 90% exclusion policy has no such validation test of its own. This is one reason to report measured values and analysis choices directly rather than treating a saved program label as a scientific conclusion.

The study uses public PopQA and SciQ items and recruited no participants or commissioned human annotators. Some stimuli deliberately associate names with incorrect factual answers or place a false answer in a reference document. These are experimental manipulations. No models were trained, and vendor outputs were not used for training. The historical project did not document a dataset-license review. Preserved upstream notices include PopQA's MIT notice, but equivalence of historical and distribution-mirror bytes, applicability of provider terms and project-code licensing remain unverified. These limits are separate from the consistency of the reported numerical results.