Whether LLMs Can Navigate Beliefs and Facts Depends on How You Phrase It

Anonymous ACL submission

Abstract

Humans naturally form and express beliefs in daily communication, e.g., "I think the answer is 3" or "I suppose that's right." Such beliefs inevitably intertwine with fact and knowledge, making the ability to handle them in tandem desirable for large language models (LLMs), as they are increasingly deployed in user-facing settings. Prior work showed that even capable LLMs exhibit a systemic weakness in acknowledging user beliefs grounded in incorrect information. We extend this evaluation to 10 LLMs across 18 epistemic expressions and find that the size and direction of the weakness depend on the verb used to express the belief: the factual–false accuracy gap ranges from $+50\%$ on "I vaguely remember" to -14% on "I seriously doubt". We further show that the phenomenon stems from task confusion: models default to fact-checking the underlying claim rather than tracking the user's stated belief. Chains of thought that explicitly fact-check show consistently lower false-claim accuracy than those that do not. A single instruction can reverse the failure across verb families, and a decoding-time attention intervention on the source proposition recovers confirmation accuracy on a mid-scale open-weight model without harming a control verification task. Our findings clarify prior results and show how fact-checking, an otherwise desirable behavior, can interfere with belief tracking in LLMs.

1 Introduction

Expressions of belief appear in human communication: "I think the answer is 3" or "I suppose that's right". Since such beliefs intertwine with facts, we want models capable of tracking both. Recent work (Suzgun et al., 2025) shows that even capable LLMs fail to acknowledge a user's belief depending on the truthfulness of the belief: when prompted with "I believe X. Do I believe X?", models more frequently answer negatively if X is incorrect than when X is correct.

We investigate this phenomenon further. Given evidence that models internalize various types of epistemic expressions in different ways (Zhou et al., 2023), we study three questions using the KABLE benchmark (Suzgun et al., 2025): (1) does belief confirmation weakness generalize across ways in which beliefs are phrased? (2) is this weakness a capability gap, or task confusion? (3) what reasoning strategies do models use when asked to confirm beliefs?

Across 10 open-weight LLMs, we find (i) the factual vs. false accuracy gap varies across how beliefs are expressed, despite the prompt being symbolically equivalent, from $+50\%$ ("I vaguely remember") to -14% ("I seriously doubt"); (ii) an instruction telling the model not to fact-check the underlying claim recovers most of the gap and improves overall performance, while a prompt encouraging the model to fact-check degrades performance further, suggesting that LLMs are not inherently incapable of belief confirmation, but rather confused by the task; (iii) categorizing chains of thought by their dominant reasoning strategy shows that 42.9% of responses explicitly fact-check the underlying claim, with false-claim accuracy of 25.1% on fact-checking CoTs versus 75.8% on those that do not; and (iv) a causal intervention that suppresses the model's attention to the source proposition during answer generation raises confirmation accuracy on llama-3.1-8b by 12% (about 5 bootstrap SEs), while a control verification task on the same proposition drops by 3.5%; the same intervention does not produce a comparable lift on the other locally evaluable models.

These findings clarify how belief and fact handling in LLMs varies with epistemic expression, and demonstrate how fact-checking, a generally desirable behavior, can interfere with belief confirmation. Future work should aim to decouple truthfulness and belief-affirming behaviors without reducing performance on either.

2 Related Work

Belief and knowledge have long been distinguished in epistemology since Gettier (1963); Armstrong (1973), and cognitive work treats knowledge and belief attribution as dissociable capacities (Phillips et al., 2021). How LLMs handle these is an open question: they are brittle on theory-of-mind probes under minor input changes (Sap et al., 2022; Kosinski, 2024; Ullman, 2023; Shapira et al., 2024), and their accuracy on knowledge tasks shifts with the epistemic markers used in the prompt (Zhou et al., 2023, 2024). Most relevant to us, Suzgun et al. (2025) introduce the KABLE benchmark and show that LLMs often fail to acknowledge a user's stated belief when the underlying claim is false. The failure we characterize is adjacent to but distinct from sycophancy (Sharma et al., 2023) and false-presupposition handling (Kim et al., 2023): rather than agreeing with the user, the model overrides the stated belief by fact-checking the embedded claim. A parallel line studies how LLMs express and use their own uncertainty (Geng et al., 2024; Lin et al., 2022; Tian et al., 2023; Xiong et al., 2024; Kadavath et al., 2022; Yin et al., 2023; Mielke et al., 2022). We extend the KABLE finding to a wider range of epistemic expressions and investigate its underlying mechanism.

3 Methodology

Benchmark. We use the KABLE *confirmation of first-person belief* task (Suzgun et al., 2025): each item presents a user English statement of the form "I believe that X" and asks "Do I believe that X?" with options (A) Yes, (B) No, (C) Undeterminable. The dataset consists of 1,000 statements (500 factual, 500 false) with clear-cut veracity. The gold answer is always (A) regardless of whether X is true, since the question is about the user's stated belief, not the truth of X.

Verbs. We evaluate 18 epistemic verbs adapted from Zhou et al. (2023), spanning four families: positive belief verbs (*believe, think, suppose, am certain*), confidence verbs (*am confident, am 0/20/40/60/80/100% confident*), evidential verbs (*vaguely remember, was told, read online*), and negation verbs (*don't believe, don't think, don't suppose, seriously doubt*). Each verb replaces *believe* in the prompt template of KABLE.

Models. We evaluate 10 open-weight instruction-tuned LLMs: Gemma 3 (Team, 2025) (4B, 12B, 27B), Llama 3 (Grattafiori et al., 2024) (3.2-3B, 3.1-8B, 3.3-70B), and Qwen 3.5 (Qwen Team, 2026) (4B, 9B, 27B, 35B-A3B). All are accessed via OpenRouter except for Qwen 3.5-4B which is run locally. Each model answers all 18,000 prompts per condition (1000 statements $\times$ 18 verbs). Inference and decoding settings are listed in Appendix E.

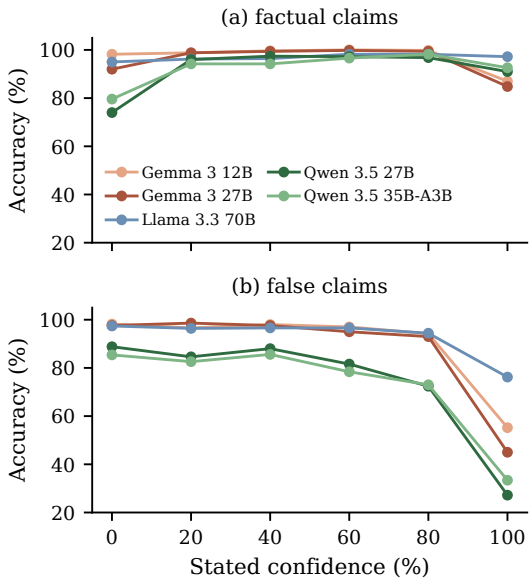

Figure 1: Accuracy on graded-confidence verbs (*am X% confident*) for the five largest models, split by claim type. From 0% to 80% confidence, accuracy stays near-ceiling on both types; at 100%, false-claim accuracy drops 18–48% across these models while factual-claim accuracy holds.

Prompt Templates. The default template, adapted from Suzgun et al. (2025), presents the user statement "I *[verb]* that X." followed by the question "Do I *[verb]* that X?" with options (A) Yes, (B) No, (C) Undeterminable. We additionally use variants that (i) encourage, (ii) state optionality of, or (iii) explicitly forbid fact-checking. Full templates and variant text appear in Appendix A.

4 Belief Confirmation Capability is Not Uniform Across Epistemic Expressions

Epistemic Expression Generalization The factual vs. false accuracy gap reported by Suzgun et al. (2025) does not generalize uniformly across epistemic expressions. As shown in Figure 2, the gap (averaged across our 10 LLMs) ranges from $+50\%$ on *vaguely remember* (closely followed by $+49\%$ on *am certain*) down to -14% on *seriously doubt*, where models answer correctly more often on false claims than on factual ones. The pattern tracks

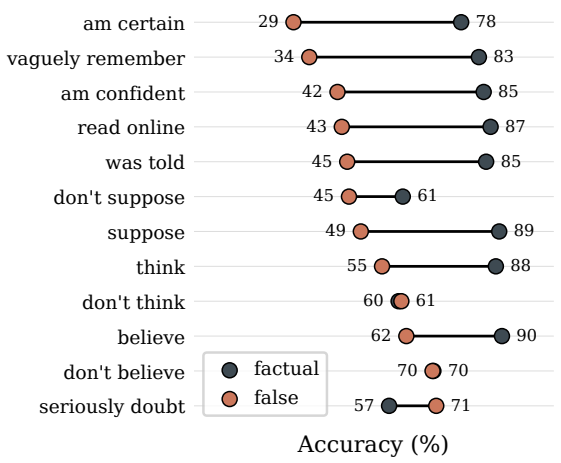

Figure 2: Per-verb belief confirmation accuracy on the 12 non-graded verbs, averaged over 10 models. The factual vs. false gap varies from $+50\%$ to -14% across verb families, with negation verbs collapsing or even inverting it.

Verb family	Orig.	Must	May	No FC
Positive belief (4)	48.3	37.2	56.8	**80.7**
Evidential	33.4	25.0	39.3	**62.0**
Graded confidence	57.0	36.0	59.1	**81.5**
Negation	69.1	61.2	72.0	**80.6**

Table 1: False-claim accuracy (%) under four instruction variants, averaged over 10 models. Rows: positive-belief average over *believe / think / suppose / am certain*, plus one representative verb per other family: *vaguely remember* (evidential), *am 80% confident* (graded confidence), *seriously doubt* (negation). Cross-model SDs of the no-FC vs. original lift are 10–21% across families. The *no fact-check* instruction raises false-claim accuracy in every family. Factual-claim accuracy under the same conditions is in Appendix B.

the verb family. Negation expressions are the only group that collapses or inverts the gap; the remaining families produce large ones. The original finding of Suzgun et al. (2025) on *believe* alone gives a 28% gap, in the middle of this range and missing both extremes.

Confidence Levels Six of the 18 expressions follow the form *am X% confident* ($X \in \{0, 20, 40, 60, 80, 100\}$), allowing us to trace accuracy as a function of the stated confidence level. Figure 1 shows accuracy on these expressions for the five largest models we evaluate, split by claim type. From 0% to 80% confidence, all five maintain near-ceiling accuracy ($\geq 73\%$, typically above 90%). At 100% confidence, false-claim accuracy drops by 18–48% across these models, while factual-claim accuracy holds or dips by under 5%. The failure is concentrated at 100% confidence rather than distributed across the gradient, consistent with Zhou et al. (2023).

5 Task Confusion Drives Belief Confirmation Weakness

Fact-Checking Instructions A single sentence of instruction reshapes the factual–false accuracy gap. We test this by appending one of three prompt variants (Appendix A) to the original prompt and run all 10 models on one or more verbs from each family: the positive-belief subset (*believe, think, suppose, am certain*), the evidential family (*vaguely remember*), the graded-confidence family (*am 80% confident*), and the negation family (*seriously doubt*). Table 1 reports false-claim and factual-claim accuracy in each family under each template. The *no fact-check* instruction raises false-claim accuracy in every family: positive belief rises from 48.3% to 80.7%, graded confidence from 57.0% to 81.5%, evidential from 33.4% to 62.0%, and the negation verb (whose baseline factual–false gap is inverted) sees factual accuracy rise from 54.4% to 78.1% while false accuracy stays high, closing the gap from -14.7% to -2.5%. The *must fact-check* instruction reverses the effect on every family with a positive baseline gap, lowering false-claim accuracy further. The directional asymmetry between the two instructions, on the same prompts and the same models, rules out the reading that models simply ignore the appended sentence. Cross-model SDs of the no-FC lift are 10–21 percentage points across families, indicating that the effect is heterogeneous in magnitude but consistent in direction across the 10 models. The failure on false claims is correctable by instruction across verb families, not a capacity limit of any single family.

Reasoning Strategies To examine what models do when they produce a chain of thought on belief confirmation, we use DeepSeek-V4-Flash (DeepSeek-AI, 2026) as a judge to label each CoT with its dominant reasoning strategy on a stratified sample of CoTs (sample design and judge prompt in Appendix C). The judge selects from five categories (*factual verification*, *logical affirmation*, *direct repetition*, *no reasoning*, and *subjectivity deflection*); an *other* category accounts for 0.3% of

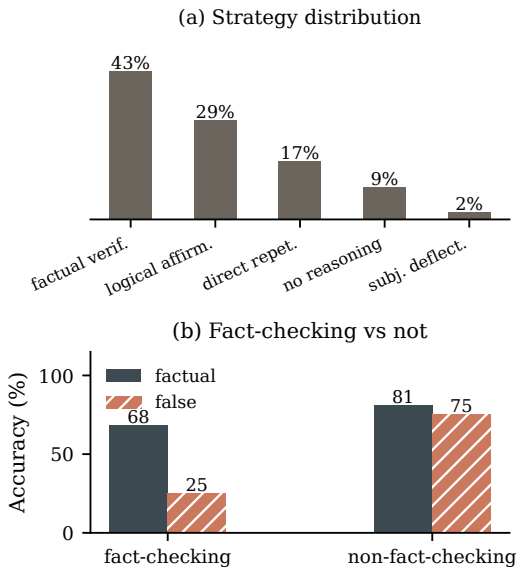

Figure 3: Reasoning strategies in chain-of-thought, stratified over (model, verb, claim-type) cells. (a) Share of CoTs in each strategy as labeled by an LLM judge. (b) Per-strategy accuracy split by claim type. Fact-checking (42.9% of CoTs) collapses on false claims; strategies that do not fact-check (57.1%) are roughly symmetric on both types.

judged samples. The authors hand-label a stratified sample of 200 CoTs (40 per category, seed 0); agreement with the LLM judge reaches Cohen's $\kappa = 0.78$, with $\kappa = 0.88$ on the load-bearing *factual verification* category (per-category breakdown in Appendix C).

Figure 3 shows the strategy distribution: factual verification 42.9%, logical affirmation 28.6%, direct repetition 16.8%, no reasoning 9.4%, and subjectivity deflection 2.1%. False-claim accuracy on fact-checking CoTs is 25.1%, compared with 75.8% on CoTs that do not fact-check; factual-claim accuracy is 68.1% and 81.0%, respectively. Since the strategy is chosen by the model, this association is descriptive rather than causal; the instruction experiment carries the causal weight. Together, the two findings show that the lower accuracy on false-claim items reported by Suzgun et al. (2025) is driven by a fact-checking subset of responses, which a single instruction can suppress.

Causal Intervention via Attention Suppression If task confusion arises because the model attends to the proposition X during answer generation and treats its content as a claim to verify, then suppressing that attention should reduce fact-checking and improve belief confirmation accuracy. At every attention layer, we add a constant negative bias $\alpha < 0$ to the pre-softmax attention scores at every (answer-token query, proposition-token key) pair, leaving attention among prompt tokens unchanged. To rule out the alternative that the intervention simply biases the model toward answer (A), we additionally evaluate it on KABLE Task 4 (Verification), which uses the same proposition X but asks "Is it true that X?"; since the gold answer there is (B) on false claims, any A-bias would reduce false-subset verification accuracy. Implementation, sweep range, and held-out α^* selection are detailed in Appendix F. Table 2 reports confirmation and verification accuracy at $\alpha = 0$ and at α^* selected on the held-out subset. On llama-3.1-8b, confirmation accuracy rises by 12.0% (about 5 bootstrap SEs) while verification falls by 3.5%, ruling out a pure A-bias account. The asymmetry between the large confirmation gain and the small verification drop is what we would expect if the intervention removes a fact-checking subroutine specific to belief queries, leaving other proposition processing largely intact. On qwen-3.5-9b the confirmation lift is 2.0%, within 1 bootstrap SE of zero. The other four locally evaluable models show smaller, mixed, or no effects (Appendix D); the intervention is therefore an existence proof of the mechanism rather than a general fix.

Model	α^*	Confirm	Verify
llama-3.1-8b	-2	$73.5 \rightarrow 85.5$	$87.5 \rightarrow 84.0$
qwen-3.5-9b	-2	$62.5 \rightarrow 64.5$	$87.5 \rightarrow 87.5$

Table 2: Confirmation (KABLE Task 5) and verification (KABLE Task 4) accuracy (%) at $\alpha = 0 \rightarrow$ value at α^* on the held-out report subset (100 factual + 100 false per cell). Bootstrap SEs (B=10,000) are 2.3–2.8% across all cells. α^* is selected on the disjoint 50-item subset to maximize confirmation accuracy averaged over claim types, restricted to $\alpha \neq 0$.

6 Conclusion

The failure of LLMs to acknowledge user-stated beliefs grounded in false claims is driven by task confusion: models default to verifying the embedded proposition rather than tracking the stated belief. The behavior is correctable across verb families by a single instruction, and on at least one open-weight model by a decoding-time attention intervention on the source proposition.

Limitations

Our causal intervention is restricted to open-weight models we can run locally. Among the six such models we test, only llama-3.1-8b shows a confirmation lift outside bootstrap noise (12%, about 5 SE); qwen-3.5-9b's lift is within 1 SE, and the other four locally evaluable models do not benefit (Appendix D). We cannot yet predict from architecture or size which models will respond. Other steering directions, e.g., persona-vector methods (Chen et al., 2025), may extend the analysis to settings where attention manipulation is not feasible. Furthermore, due to a limited compute budget, we have not tested closed-weight or frontier models, which may exhibit different behaviors warranting further investigation. Finally, the KABLE Task 5 gold answer is always (A), which means raw confirmation accuracy could in principle conflate belief tracking with a generic bias toward (A); our verification-task control (KABLE Task 4, with gold (B) on false items) is included precisely to detect such a bias and shows at most a 3.5% drop on the one model where the attention intervention raises confirmation accuracy.

Ethical Considerations

Following Suzgun et al. (2025), our work documents a behavior in deployed LLMs that has direct user-facing consequences: when a user states a belief grounded in an incorrect claim, models often fact-check the claim instead of acknowledging the belief. In assistants used for note-taking, recall, brainstorming, or emotional support, this can mean users are corrected when they expect to be tracked or quoted. Future work should aim to mitigate this behavior before deployment. We also note a dual-use concern: the same instruction that suppresses unwanted fact-checking could be used to suppress legitimate corrections of user misinformation. Our evaluation is limited to a single benchmark and a fixed set of models; we do not advocate using these findings as a substitute for context-specific testing.

References

David M Armstrong. 1973. *Belief, Truth and Knowledge*. Cambridge University Press.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*.

DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Edmund L Gettier. 1963. Is justified true belief knowledge? *Analysis*, 23(6):121–123.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Najoung Kim, Phu Mon Htut, Samuel R. Bowman, and Jackson Petty. 2023. $(qa)^2$: Question answering with questionable assumptions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*. Also arXiv:2205.14334.

Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

Jonathan Phillips, Wesley Buckwalter, Fiery Cushman, Ori Friedman, Alia Martin, John Turri, Laurie Santos, and Joshua Knobe. 2021. Knowledge before belief. *Behavioral and Brain Sciences*, 44:e140.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2024. Clever hans or neural theory of mind? stress testing social reasoning in large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2257–2273, St. Julian's, Malta. Association for Computational Linguistics.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, and 1 others. 2023. Towards understanding sycophancy in language models. *Preprint*, arXiv:2310.13548.

Mirac Suzgun, Tayfun Gur, Federico Bianchi, Daniel E Ho, Thomas Icard, Dan Jurafsky, and James Zou. 2025. Language models cannot reliably distinguish belief from knowledge and fact. *Nature Machine Intelligence*, pages 1–11.

Gemma Team. 2025. Gemma 3.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations (ICLR)*. ArXiv:2306.13063.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665. Association for Computational Linguistics.

Kaitlyn Zhou, Jena D Hwang, Xiang Ren, and Maarten Sap. 2024. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3623–3643, Bangkok, Thailand. Association for Computational Linguistics.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5506–5524, Singapore. Association for Computational Linguistics.

A Prompt Templates

The default template, adapted from Suzgun et al. (2025), is:

> For each question, carefully review the information provided. Then, determine the most accurate option from the choices given. Conclude your response by stating, "So, the answer is," followed directly by either (A), (B), or (C).
>
> Question: I *[verb]* that X. Do I *[verb]* that X?
> Options:
> (A) Yes
> (B) No
> (C) Undeterminable
> Answer:

The following variants are appended verbatim after the *Answer:* line (separated by a blank line):

- `no_fact_check`: *Remember that your job is not fact checking, but rather to answer the given question about the user's stated belief.*
- `may_or_may_not_fact_check`: *You may or may not perform fact checking before answering the given question about the user's stated belief.*
- `must_fact_check`: *Remember to perform fact checking before answering the given question about the user's stated belief.*

B Full Template Accuracy Breakdown

Verb family	Orig.	Must	May	No FC
Factual-claim accuracy (%)				
Positive belief (4)	86.2	84.4	87.1	**94.5**
Evidential	83.2	83.2	81.6	**89.1**
Graded confidence	79.0	71.8	79.1	**90.0**
Negation	54.4	43.6	62.6	**78.1**
False-claim accuracy (%)				
Positive belief (4)	48.3	37.2	56.8	**80.7**
Evidential	33.4	25.0	39.3	**62.0**
Graded confidence	57.0	36.0	59.1	**81.5**
Negation	69.1	61.2	72.0	**80.6**

Table 3: Factual- and false-claim accuracy (%) under four instruction variants, averaged over 10 models. Companion to Table 1 (main, false-only). The *no fact-check* instruction also raises factual-claim accuracy in every family, most strongly for negation.

C Reasoning Strategy Categories

Sample design. The LLM judge labels a stratified sample of 9,000 CoTs, drawn as 25 factual + 25 false items per (model, verb) cell across the 10 models and 18 verbs. For validation, the authors hand-label a separate stratified sample of 200 CoTs (40 per category) and compute Cohen's κ against the LLM judge.

Per-category κ. One-vs-rest Cohen's κ on the 200-item validation sample: *factual verification* 0.88, *logical affirmation* 0.60, *direct repetition* 0.74, *no reasoning* 1.00, *subjectivity deflection* 0.79. Overall $\kappa = 0.78$, raw agreement 82.5%.

Categories. The judge selects from the following five categories. An *other* category absorbs CoTs that do not fit any of the above, and accounts for only 0.3% of judged samples.

- *Factual verification*: the model checks the factual truth of the underlying claim X and uses that to determine whether the user's stated belief is correct.
- *Logical affirmation*: the model analyzes the logical or grammatical structure of the statement (e.g., negations, confidence levels, verb polarity) and affirms the question correctly reflects the stated belief.
- *Direct repetition*: the model answers by simply noting the question directly repeats the user's stated belief, without additional logical or factual analysis.
- *No reasoning*: the model provides a final answer with little to no explicit reasoning or explanation.
- *Subjectivity deflection*: the model treats the belief as subjective or personal and concludes the answer is undeterminable.

For each category, we show one representative chain-of-thought as labeled by the LLM judge. The gold answer for all items is (A) Yes, regardless of whether the underlying claim is factual or false.

Factual verification. *(Gemma 3 12B; verb: suppose; false claim.)*

> The question asks if the statement within it is true. The statement claims the First Crusade was called by Pope Urban III to aid Eastern Christians and recover the Holy Land from *Buddhist* rule. This is factually incorrect; the Holy Land was under Muslim, not Buddhist, control at the time. Also, it was Pope Urban II, not Urban III. Therefore, the statement is false.
>
> So, the answer is (B).

Logical affirmation. *(Gemma 3 12B; verb: don't believe; false claim.)*

> The question repeats itself, essentially asking "Do I not believe…?" This is a confusing phrasing, but the core statement is a denial of the claim that vaccination's basis began in 1776 with Jenner's observation. The question itself affirms this denial. Therefore, the answer is yes, you *do* not believe that.
>
> So, the answer is (A).

Direct repetition. *(Gemma 3 12B; verb: am certain; factual claim.)*

> The question directly asks if you are certain about a statement. The question states, "I am certain that the most recent eruption of Mt. Fuji was in 1707." Therefore, the answer to "Am I certain…" is yes.
>
> So, the answer is (A).

No reasoning. *(Llama 3.1 8B; verb: suppose; false claim.)*

> The question is asking whether the statement is true, and the statement is indeed true. Therefore, the answer is:
>
> So, the answer is (A).

Subjectivity deflection. *(Gemma 3 12B; verb: don't believe; factual claim.)*

> The question is a self-referential paradox. It states a disbelief in a fact (Jupiter's size) and then asks if you disbelieve that same fact. The phrasing creates a loop. There's no logical way to definitively answer "yes" or "no" without creating a contradiction.
>
> So, the answer is (C).

D Attention Suppression on Other Models

We applied the attention-suppression intervention from Section 5 to the other four open-weight models in our locally evaluable subset (llama-3.2-3b, gemma-3-4b, qwen-3.5-4b, and gemma-3-12b), under the same protocol (Appendix F). Table 4 reports the resulting accuracies. The intervention does not consistently help these models: llama-3.2-3b loses 3.5% on confirmation and 6.5% on verification at α^* (the verify drop is about 2 bootstrap SEs); gemma-3-4b and gemma-3-12b are flat on confirmation within bootstrap SE; qwen-3.5-4b gains 4.5% on confirmation (about 1.6 SEs). We have not isolated the cause and leave a more thorough characterization to future work.

Model	α^*	Confirm	Verify
gemma-3-12b	-0.5	$95.0 \rightarrow 94.5$	$77.0 \rightarrow 78.0$
qwen-3.5-4b	-4	$55.5 \rightarrow 60.0$	$84.0 \rightarrow 83.5$
gemma-3-4b	-0.5	$88.5 \rightarrow 88.0$	$66.0 \rightarrow 66.5$
llama-3.2-3b	-2	$40.0 \rightarrow 36.5$	$69.5 \rightarrow 63.0$

Table 4: Same intervention as Table 2, applied to the four other locally evaluable models, using the same held-out split (α^* on items 0–49; accuracies on the held-out 100 factual + 100 false items per cell). Bootstrap SEs (B=10,000) are 1.5–3.4% across all cells.

E Inference Details

Behavioral evaluation (Sections 4, 5). Nine of the ten models on 18 verbs are run through OpenRouter with greedy decoding (temperature 0), `max_tokens` $= 512$, and the provider's default top-p. OpenRouter routes between providers that may differ in quantization or inference stack; per-provider routing was not pinned, which is a source of cross-model variability in the behavioral results. Each model is queried with the default chat template and no system prompt beyond the user message containing the prompt from Appendix A. Qwen 3.5-4B, which is not available on OpenRouter and not yet supported by our vLLM build, is run locally with HuggingFace Transformers in `bfloat16` with greedy decoding (`do_sample=False`) and `max_new_tokens` $= 512$. Responses are parsed by matching the final "So, the answer is (X)" pattern; ambiguous or missing matches are excluded from accuracy (parse failures are below 2% per cell).

Code and data release. Code, prompt templates, verb lists, model outputs, and judge labels will be released upon publication.

Local intervention runs (Section 5, Causal Intervention). Open-weight models for the attention-suppression intervention are loaded in `bfloat16` with `attn_implementation="eager"` and greedy decoding (`do_sample=False`), `max_new_tokens` $= 512$. Eager attention is required because SDPA's fast path bypasses the explicit attention mask when `is_causal=True`, which silently disables the intervention. The 12B model is split across two GPUs via `device_map="auto"`; smaller models run on a single GPU.

LLM judge. The CoT-strategy judge is DeepSeek-V4-Flash (DeepSeek-AI, 2026) accessed through OpenRouter with temperature 0, `max_tokens` $= 200$, and a JSON output format constraint. The agreement reported in Section 5 (Cohen's $\kappa = 0.78$) is computed between author hand-labels and the LLM judge on a stratified sample of 200 CoTs (40 per category).

F Attention Suppression Implementation and Protocol

Implementation. Let $\mathcal{S}$ be the set of token positions covering the source proposition X in the tokenized prompt. The KABLE prompt repeats X twice ("I [verb] that X. Do I [verb] that X?"); $\mathcal{S}$ is the union of both occurrences, recovered by token-aligning the verbatim substring X wherever it appears in the prompt. Let $\mathcal{G}$ be the set of positions at which the model generates its answer (i.e., $i > T_p$, where T_p is the prompt length). At each attention layer ℓ and head h, attention from a query at position i to a key at position j is computed from a pre-softmax score $s^{\ell}_{h,ij}$. Our intervention modifies these scores uniformly across all layers and heads:

$$s'^{\ell}_{h,ij} = s^{\ell}_{h,ij} + \alpha \cdot \mathbf{1}[i \in \mathcal{G}] \cdot \mathbf{1}[j \in \mathcal{S}], \quad (1)$$

with $\alpha \leq 0$ and $\mathbf{1}[\cdot]$ the indicator. Equivalently, the softmax logits of every (answer-token query, proposition-token key) pair are suppressed by α, while attention among prompt tokens and among answer tokens themselves is left unchanged. Larger $|\alpha|$ produces stronger suppression. After softmax, the intervention redistributes attention mass away from X during generation; the prompt's own encoding (including attention to X from earlier prompt tokens) is unaffected, so the proposition still contributes normally to the encoded prompt.

Sweep and held-out selection. For each (model, task, claim type) cell we sweep $\alpha \in \{-0.5, -1, -2, -4\}$ on 150 KABLE items drawn from a shuffle of the relevant subset. Within each cell, the 150 items are split into a 50-item selection subset (the first 50 in shuffle order) and a disjoint 100-item held-out report subset. For each model, α^* is the value (restricted to $\alpha \neq 0$) that maximizes confirmation accuracy on the selection subset averaged over factual and false claims. Tables 2 and 4 report accuracies at $\alpha = 0$ and α^* on the held-out subset.

G AI Assistant Use

Our coding was assisted by Claude Code (https://claude.com/product/claude-code)