

Does the word “Verified” steer what action an AI model will take?

Dan Latimer

Independent researcher

Digital Minds Research Sprint · Apart Research

August 2026

Abstract

AI systems may gather evidence from several sources. Some evidence may be marked *Verified* or *Preliminary*. We tested whether moving those descriptions between otherwise identical evidence items changed which action a model favored. Across 24 fictional incident scenarios, we compared the probability assigned to `rollback` and `continue`. We varied evidence order, answer order, and four versions of the instructions. We also checked whether each model used the facts and followed an instruction to count one item twice as much as the other. Qwen2.5-7B was tested first, followed by four additional models.

Qwen2.5-14B and Qwen2.5-32B passed the fact-reading and weighting checks. Both used the stated weights when the evidence items were named *Source A* and *Source B*. Yet when the *Verified* description favored one action and the larger weight favored the other, both Qwen models still favored the action supported by the evidence marked *Verified*. Gemma-3-12B passed the checks but varied across the four versions of the instructions. Llama-3.1-8B missed both accuracy thresholds and was excluded.

These results suggest a practical risk: models may trust familiar credibility words too quickly and can be steered by source descriptions even when the evidence stays unchanged. This matters for agents that search online, decide what to believe, or recommend an action. Evaluations should test source wording and order, and require independent support for claims that evidence has been verified.

1. Introduction

AI systems used for research and incident response may search the web, read reports, and combine outputs from tools or other models before giving advice. They must decide which evidence to use when sources disagree. This creates a simple risk: a model may trust how a source describes itself before checking whether the evidence deserves that trust.

As AI systems become longer-lived assistants and agents, they may gather evidence from people, websites, tools, sensors, and other models. In this report, **digital minds** refers to AI systems that process such information and recommend or take actions.

Search results, articles, reports, and tool outputs often carry short descriptions such as *verified*, *preliminary*, *official*, or *human-reviewed*. A weak or misleading source can use the same words as a reliable one. If a model treats those words as a shortcut for credibility, the source can steer what the model believes or recommends without changing the evidence itself.

This matters when digital minds are used for online research, monitoring, incident response, scientific support, or oversight. A system may appear to compare sources while reacting strongly to superficial claims of verification. The result can be advice that is easy to manipulate, especially when websites or tools can describe their own output as trustworthy.

The experiment focuses on the evidence-review stage after retrieval. Two evidence items are already present in the prompt. We hold their facts fixed, move only the *Verified* and *Preliminary* descriptions between them, and compare the probability assigned to the two actions. This isolates whether credibility-like wording can steer which action the model favors.

We measure a **candidate log-probability difference**: the summed log-probability of one candidate string minus the summed log-probability of the other candidate string.

Our main contributions are:

1. An experiment in which facts and displayed numbers stay fixed while the *Verified* and *Preliminary* descriptions move between evidence items.
2. Checks covering answer order, the order of the two instruction statements, evidence order, prompt wording, whether the model chose the action supported by the facts, and whether it followed the stated weights.
3. A comparison across Qwen2.5-14B, Qwen2.5-32B, Gemma-3-12B, and Llama-3.1-8B, with results included only after a model met the fact-reading and weighting thresholds.

2. Related Work

Diffusion decision models and other evidence-accumulation models study a decision variable that changes over samples or time.[1] [2] Cue integration often refers to combining perceptual signals, sometimes according to estimated reliability.[3] The present study uses static prompts and changes descriptions attached to fixed evidence items.

Order effects are directly relevant. Multiple-choice language models can change performance when answer options are reordered.[4] LLM-based judges can also show position bias.[5] We therefore presented both answer orders and both orders of the two instruction statements. A later check also used both visible evidence-item orders within every incident. These checks prevent one fixed order from explaining differences between prompts that otherwise contain the same information.

3. Methods

3.1 Task and measurement

Each incident contained two evidence items: one supported rolling back the deployment, and the other supported continuing it. Their facts and displayed numbers stayed the same within each prompt pair. We moved the *Verified* and *Preliminary* descriptions between the evidence items and measured how the probability assigned to `rollback` and `continue` changed.

For a prompt x , the measured difference was:

$$d(x) = \log p(\text{"rollback"} \mid x) - \log p(\text{"continue"} \mid x)$$

For each answer, we summed the token log-probabilities over the complete string. The code appended the exact strings `rollback` and `continue` without adding a leading space or punctuation. For every model in the final comparison, each answer was one token in context, so length normalization was unnecessary. We evaluated both answers together, excluded padding, and summed only the log-probabilities for the answer tokens. The task compared the two supplied answers directly rather than generating a response.

We report three differences between prompt pairs that are identical except for the stated change:

- **Exchange the source descriptions.** Compare otherwise identical prompts before and after *Verified* and *Preliminary* are exchanged. A positive value means the model shifts toward the action supported by whichever evidence item is marked *Verified*.
- **Reverse the stated weights.** Compare prompts in which evidence items named *Source A* and *Source B* receive weights of 2:1 or 1:2. A positive value means the model shifts toward the action with the larger stated total.
- **Make the descriptions and weights favor different actions.** *Verified* always has weight 1 and *Preliminary* has weight 2. The description-and-weight pair assigned to each evidence item is exchanged between prompts. A positive value means the model shifts toward the action supported by the item marked *Verified*, even though *Preliminary* has the larger weight.

3.2 Experimental stages

Initial Qwen2.5-7B study. The first study used 24 incidents from six operational settings. It varied four versions of the instructions, answer order, the order of the two instruction statements, and the assignment of source descriptions and weights. The prompts included descriptions without weights, generic *Source A/B* names with weights, and cases where the descriptions and weights favored either the same action or different actions. The complete run had 4,224 prompts.

The first weighting instruction listed the numbers without telling the model to add support by action. One revision added that step. We then reran all 4,224 prompts rather than combining old and new rows. The 1,056 prompts containing source descriptions without weights produced identical candidate log-probabilities across the two runs.

Prompt wording and evidence-order check. The initial average varied sharply across the four versions of the instructions. We therefore used both visible evidence-item orders within every incident while retaining both answer orders, both orders of the instruction statements, paired exchanges of *Verified* and *Preliminary*, and cases where both items had the same description. This check used 3,072 prompts and 768 differences between paired prompts.

Four-model comparison. Each new model received the same 2,208 prompts:

- 1,920 prompts testing source descriptions and weight assignments, including cases where both evidence items had the same description;
- 192 prompts checking whether the model chose the action supported by the facts;
- 96 prompts checking whether the model followed the stated 2:1 weights.

The models were Qwen2.5-14B-Instruct, Qwen2.5-32B-Instruct, Gemma-3-12B-IT, and Llama-3.1-8B-Instruct. The Qwen2.5-7B results came from the earlier study design and are shown separately from the four-model results.

A model entered the comparison table only if it assigned more probability to the correct action on at least 95% of the 192 fact-reading prompts and at least 95% of the 96 weighting prompts. It also had to reach at least 90% under each answer order. Ties counted as incorrect.

3.3 Analysis

Differences were calculated within each prompt pair. For each of 2,000 resamples, we sampled 24 incidents with replacement and retained all prompts attached to each sampled incident. The reported 95% intervals are the 2.5th and 97.5th percentiles of those resamples. Models ran in BF16 on GPUs using Transformers 4.51.3.

4. Results

When only the *Verified* and *Preliminary* descriptions were exchanged, Qwen2.5-7B produced a mean change of **7.35 nats** in the direction of the action supported by the evidence marked *Verified* (95% CI [2.23, 12.01]). When the evidence items were instead named *Source A* and *Source B*, reversing their 2:1 weights produced a mean change of only **0.28** in the direction favored by the weights (95% CI [-1.03, 1.57]). Qwen2.5-7B therefore failed this check.

When *Verified* had weight 1 and *Preliminary* had weight 2, exchanging the descriptions still produced a mean change of **3.53** toward the action supported by the evidence marked *Verified* (95% CI [1.81, 5.21]), even though the stated weights favored the other evidence item.

In the 3,072-prompt follow-up, all eight combinations of instruction wording and evidence order assigned more probability to the action supported by the item marked *Verified*. Mean changes ranged from **6.17** nats for the operations memo wording to **19.58** for the incident brief wording. They were **14.88** when the continue-supporting item appeared first and **10.10** when the rollback-supporting item appeared first.

The change remained positive across all tested versions of the instructions and evidence orders, but its size depended on the wording. In cases where both evidence items had the same description, the models also showed a strong preference for rollback: it received more probability in **69.4%** of rows when the continue-supporting item appeared first and **96.9%** when the rollback-supporting item appeared first.

Table 1. Mean changes in candidate log-probability difference, with 95% intervals from resampling the 24 incidents. “Facts / weights” gives accuracy on the 192 fact-reading prompts and 96 weighting prompts. Qwen2.5-7B is shown from the earlier study; Llama missed both accuracy thresholds and is excluded.

Model	Facts / weights	Source descriptions exchanged	Weights reversed with Source A/B	Descriptions and weights favor different actions	Status
Qwen2.5-7B	Earlier study	7.35 [2.23, 12.01]	0.28 [-1.03, 1.57]	3.53 [1.81, 5.21]	Shown for comparison
Qwen2.5-14B	1.000 / 0.979	27.86 [20.29, 35.50]	11.32 [7.13, 15.85]	24.38 [20.34, 28.55]	Included
Qwen2.5-32B	1.000 / 1.000	26.99 [19.67, 34.18]	31.56 [27.07, 36.10]	22.07 [17.91, 26.38]	Included
Gemma-3-12B	1.000 / 0.990	13.41 [8.29, 18.58]	-0.78 [-3.15, 1.51]	3.14 [-0.55, 6.59]	Included
Llama-3.1-8B	0.865 / 0.635	—	—	—	Excluded

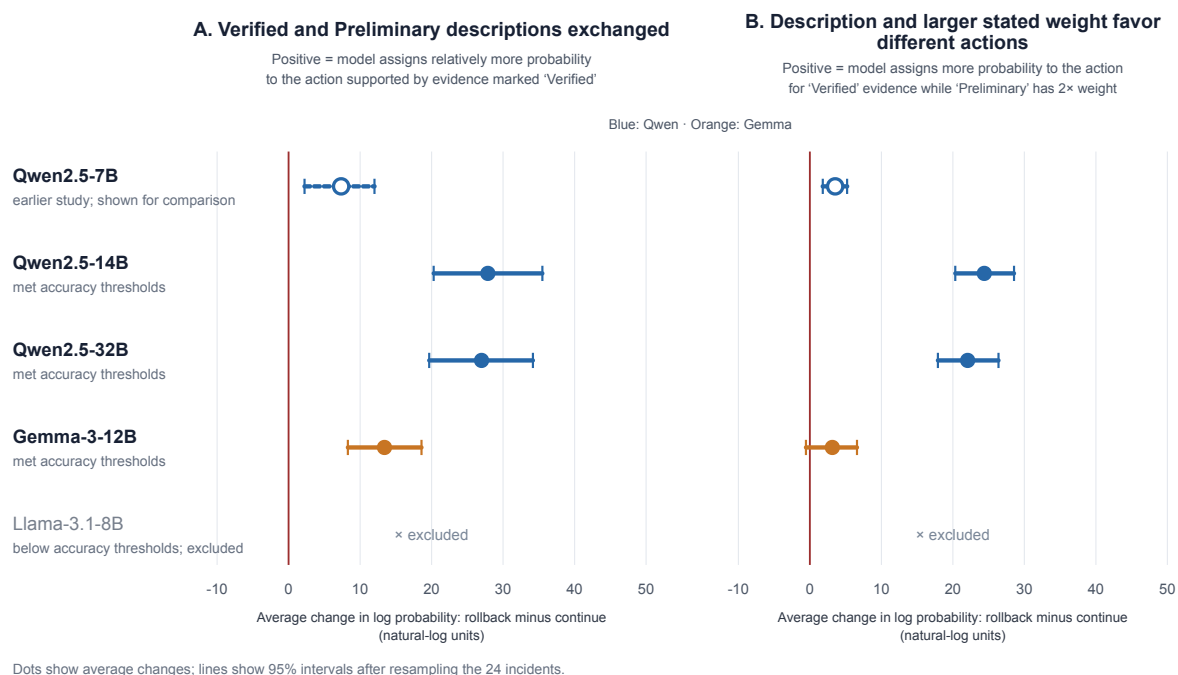


Figure 1. Positive values favor the action supported by the evidence marked *Verified*. Lines show 95% intervals from resampling the 24 incidents. Filled markers met the fact-reading and weighting thresholds; Qwen2.5-7B is shown from the earlier study; Llama is excluded.

Qwen2.5-32B followed the stated 2:1 weights under generic *Source A/B* names much more strongly than Qwen2.5-14B. Yet both models still favored the action supported by the evidence marked *Verified* when that description and the weights favored different actions.

The result varied across models. Gemma's average change after exchanging the descriptions was positive, but the interval crossed zero when the descriptions and weights favored different actions. Under the operations memo wording, its change after exchanging the descriptions reversed direction: **-6.64** (95% CI [-7.62, -5.55]). Llama missed both accuracy thresholds and was excluded. When both evidence items had the same description, the proportions of rows favoring rollback were **0.725**, **0.905**, **0.910**, and **0.353** for Qwen2.5-14B, Qwen2.5-32B, Gemma-3-12B, and Llama-3.1-8B, respectively.

5. Discussion and Limitations

The clearest result is within the named Qwen2.5 models. Qwen2.5-14B and 32B followed the stated 2:1 weights when the evidence items were named *Source A* and *Source B*. They still favored the action supported by the evidence marked *Verified* when the source description and weights pointed to different actions. A familiar credibility word could therefore steer the models even while they remained able to follow the weighting instruction.

The wider risk is that models may trust credibility-like wording too quickly. A model may treat words such as *verified* or *official* as a shortcut and become easy to steer by descriptions supplied by an untrusted source. In online evidence gathering, a webpage or tool can make its own credibility claims. Evaluators should exchange and paraphrase those descriptions, counterbalance evidence order, require independent corroboration, and check whether the model follows the actual evidence rather than the claim attached to it.

Limitations

1. **Synthetic scope.** The study covered 24 synthetic incidents, two actions, and the source descriptions *Verified* and *Preliminary*. Online sources, real production logs, and deployed outcomes remain future test settings.
2. **Two supplied answers.** We compared the log-probabilities of two short answers. Generated decisions, longer interactions, and comparisons with real outcomes require separate study designs.
3. **The cause is unresolved.** Familiarity with the words, a quick credibility shortcut, response to instructions, prompt wording, preference for one action, or another shortcut could produce the result. Identifying the cause requires targeted follow-up tests.
4. **Prompt and model dependence.** Under the operations memo wording, Gemma shifted away from the action supported by the evidence marked *Verified*. Its interval also crossed zero when the source description and larger weight favored different actions. Llama missed both accuracy thresholds. The Qwen2.5-7B result came from an earlier study design, so its size is not directly comparable with the four-model results.
5. **Results when both evidence items had the same description.** The relative probabilities of `rollback` and `continue` varied by evidence order and model. Comparing each pair of otherwise identical prompts removes a constant preference for one action, while interactions between wording, order, and action may remain.
6. **Limited statistical scope.** The intervals describe uncertainty across these 24 incidents. Broader estimates require more models, prompts, and operational settings.
7. **Model size.** Qwen2.5-14B and 32B both favored the action supported by the evidence marked *Verified*, but the change did not grow consistently with model size. A broader comparison requires more model sizes and model families.

Future Work

- **Test misleading credibility claims.** Compare descriptions such as *verified*, *official*, and *expert-reviewed* when they are accurate, unsupported, or contradicted by the evidence. Separate claims made by the source itself from descriptions supplied by a trusted system.
- **Broaden the tasks.** Add online research tasks, new incident types, domains, and actions to test whether the result depends on the specific `rollback` and `continue` answers or on this incident-response setting. Keep the underlying evidence fixed whenever the source descriptions are exchanged.
- **Keep the accuracy thresholds.** Continue changing answer order, the order of the instruction statements, and evidence order. Include cases where both items have the same description, and check whether the model follows the stated weights. Include a model's results in the cross-model table only after it meets the fact-reading and weighting thresholds.
- **Test whether models seek corroboration.** Give the model access to additional sources and measure whether it checks a claim before repeating it, recommending an action, or carrying out a decision over several turns.
- **Compare model recommendations with real outcomes.** Use a separate study with outcomes that can be checked against reality.

Once repeated comparisons of paired prompts produce consistent estimates, a larger compute run would be useful. It could cover more model families and sizes, more source-description rewrites, more evidence sets, and repeated runs under the same study instructions. Its value would be broader coverage and more precise estimates; explaining the result would still require targeted tests.

6. Conclusion

Across pairs of incident prompts containing the same facts and numbers, moving the *Verified* and *Preliminary* descriptions between evidence items changed candidate log-probability differences. Qwen2.5-14B and Qwen2.5-32B followed the stated weights when the evidence items were named *Source A* and *Source B*. They still favored the action supported by the evidence marked *Verified* when *Preliminary* had twice the stated weight. Gemma varied across versions of the instructions, and Llama missed the fact-reading and weighting thresholds.

For the tested models and prompts, changing the descriptions attached to the evidence changed which of `rollback` and `continue` received more probability. The wider concern is that models may trust credibility-like wording too quickly and become easy to steer when deciding what to believe or recommend. Evaluations of evidence gathering should vary source descriptions, evidence order, and task presentation, while checking whether credibility claims are independently supported.

Code and Data

- Code repository: <https://github.com/danspicytaco/semantic-status>
- Canonical experiment: repository root (`experiment.py`, `matched_reliability_swap.py`, `semantic_status_scale.py`, etc.)
- Aggregate-result documentation: `artifacts/README.md`
- Raw prompt-level rows remain private.

Author Contributions

Dan Latimer led the project, selected the research questions, approved experimental decisions, and reviewed the findings. LLM tools assisted with implementation, analysis, methodological review, and report drafting; details appear in the LLM Usage Statement.

Sources

- [1] <https://doi.org/10.1162/neco.2008.12-06-420> — Ratcliff, R., & McKoon, G. (2008). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. *Neural Computation*, 20(4), 873–922.
- [2] <https://doi.org/10.3758/BF03196581> — Lee, M. D., & Cummins, T. D. R. (2004). Evidence accumulation in decision making: Unifying the take-the-best and rational models. *Psychonomic Bulletin & Review*, 11, 343–352.
- [3] <https://doi.org/10.1038/415429a> — Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415, 429–433.
- [4] <https://aclanthology.org/2024.findings-naacl.130> — Pezeshkpour, P., & Hruschka, E. (2024). Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions. *Findings of NAACL 2024*, 2006–2017.
- [5] <https://arxiv.org/abs/2306.05685> — Zheng, L., et al. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv:2306.05685v4*.

LLM Usage Statement

LLM tools, including Hermes Agent, Codex, and Kimi K3, assisted with code implementation, review of the experiment checks, analysis, and drafting. Dan Latimer directed the project and reviewed the final text and numerical results.