Binary Context Sufficiency in Hidden States: Dissociating Sufficiency from Correctness in Fixed-Question, Changed-Context Multi-Hop QA

Ali Uyar

Abstract

Context-grounded QA failures conflate at least two cases: the model may misuse evidence that is present, or the provided context may itself be insufficient. We study this distinction in fixed-question, changed-context multi-hop QA. Each HotpotQA question is converted into four controlled context variants (sufficient, sufficient-noisy, missing, misleading), and we probe hidden states at prompt end and after a short two-step reasoning prefix. The strongest supported result is binary rather than three-way. On the primary Qwen2.5-7B-Instruct run, a linear probe separates sufficient from insufficient context with 0.952 AUROC on closed-book-negative examples, and the signal remains strong when final correctness is held fixed (0.979 AUROC on correct-only and 0.939 on wrong-only closed-book-negative subsets). Title masking preserves the effect (0.936 AUROC). At the same time, prompt-end representations are at least as informative as end-of-step-2 representations, missing-vs-misleading separation remains near chance in the original three-way analysis (0.505 AUROC), and a DeepSeek-R1-Distill-Qwen-7B replication is mixed rather than confirmatory (0.803 AUROC, effectively tied with prompt end and slightly below the strongest lexical baseline). We therefore make a narrow claim: hidden states encode a coarse, correctness-dissociated context-sufficiency signal, but the current evidence does not support a richer three-way evidence-state representation or a claim that the short reasoning prefix creates the signal.

1 Introduction

When a language model answers under retrieved or otherwise provided context, an error can arise for at least two different reasons. The model may fail to use evidence that is already present, or the available evidence may itself be incomplete, misleading, or contradictory. This distinction matters because it changes what should be diagnosed: a reasoning failure is not the same thing as an evidence failure. Recent work on context sufficiency has argued that this difference is central to understanding hallucination in context-grounded generation [4]. Yet most practical evaluations still collapse these cases into answer correctness alone.

A second line of work asks whether language models can assess their own knowledge or confidence before or during generation. Query-level uncertainty methods aim to predict whether a question is answerable before any answer tokens are produced [2]. Hidden-state probing work has shown that reasoning models encode strong signals about answer correctness during generation [6]. These are important results, but they leave open a narrower representational question: *when the question is held fixed and only the available context changes, do hidden states encode the sufficiency of that context independently of final correctness?*

Our experimental design keeps the question fixed while varying only the context. Each question is converted into four controlled variants: two sufficient variants (clean and noisy), one missing-support variant, and one misleading near-miss variant. We then prompt the model to produce a short, standardized two-step prefix and extract hidden states at three anchors: PROMPT_END, STEP1_END, and STEP2_END. The main analysis asks whether a lightweight probe can separate Sufficient from Insufficient (with Missing and Misleading merged) on the subset of examples that the model could not answer correctly from the question alone.

The completed evidence supports a narrower paper than the original three-way motivation. On the primary HotpotQA run with Qwen2.5-7B-Instruct, a linear probe reaches 0.952 AUROC for sufficient-vs-insufficient classification on the closed-book-negative subset and remains strong on correct-only and wrong-only closed-book-negative subsets. Title masking preserves the signal, suggesting that the effect is not driven only by page-title cues. But the same evidence also narrows the claim: prompt-end representations are at least as informative as end-of-step-2 representations, and the original missing-vs-misleading contrast remains weak. We therefore center the paper on a *binary* context-sufficiency dissociation result and treat the three-way analysis as secondary evidence.

The contribution is therefore controlled and conservative. We show that coarse context sufficiency is linearly decodable from hidden states under fixed-question, changed-context controls, and that this signal is not exhausted by final-answer correctness. We do *not* claim a robust three-way evidence-state representation, a new controller or retrieval method, or a causal hidden-state mechanism.

Contributions.

- We introduce a fixed-question, changed-context dissociation setup for separating context sufficiency from answer correctness in multi-hop QA.
- We show that hidden-state probes strongly separate sufficient from insufficient context on closed-book-negative examples, and that this separation survives correct-only and wrong-only controls.
- We show that title masking preserves the binary effect, while prompt-end representations remain at least as informative as end-of-step-2 representations; the main claim is therefore about decodability, not about reasoning creating the signal.
- We report the original three-way analysis and a DeepSeek reasoning-model replication as secondary evidence: the former is weak on missing-vs-misleading separation, and the latter is mixed rather than cleanly confirmatory.

2 Related Work

2.1 Context sufficiency has mostly been studied behaviorally

Our work is closest in motivation to recent papers that distinguish failures caused by insufficient context from failures caused by poor use of otherwise sufficient context. Joren et al. [4] formalize the notion of *sufficient context* and use it to reason about answering versus abstaining in context-grounded systems. We inherit that decomposition, but our contribution is different. Rather than study selective generation or abstention policies, we ask whether hidden states encode a *linearly decodable sufficiency signal* when the question is fixed and only the context changes.

2.2 Question-only uncertainty is a deliberate contrast condition

A separate line of work studies whether a model can estimate, before generation, whether it is able to answer a query. Query-level uncertainty methods are aimed at knowledge-boundary detection rather than context-conditioned evidence analysis [2]. In our setting, that mismatch is intentional. Because all four variants of a question share the same question text, any purely question-only method is blind to the manipulated evidence state. We therefore use a locked question-only answerability self-assessment baseline as a contrast condition, not as a claim of exhaustively reproducing the query-level uncertainty literature.

2.3 Correctness probes target a different variable

Our work is also adjacent to hidden-state probing studies that focus on answer correctness rather than evidence sufficiency. Zhang et al. [6] show that reasoning-model hidden states encode whether intermediate answers are correct and can support verifier-style early exit. We ask a different question. A model can be correct despite insufficient provided context, and it can be wrong despite sufficient provided context. Our correct-only and wrong-only dissociation subsets are designed precisely to separate these variables, and the fact that the sufficiency probe remains strong on both is the core scientific result.

2.4 Retrieval control is downstream, not the present claim

Several recent systems use hidden states to decide whether or when additional retrieval is necessary. Baek et al. [1] are especially relevant because they use intermediate hidden states to guide retrieval decisions. These papers motivate why a sufficiency signal might matter, but they are not our target. We do not propose or optimize a controller, and we do not report retrieval gains as the contribution. The paper is a representational dissociation study; any downstream control interpretation is secondary.

3 Methods

3.1 Task setup and context variants

We study fixed-question, changed-context multi-hop question answering on HotpotQA bridge questions [5]. Each retained question group contains exactly four context variants, each with four paragraphs:

(1) **sufficient_clean**: both supporting paragraphs plus easy distractors;
(2) **sufficient_noisy**: both supporting paragraphs plus harder distractors;
(3) **missing_support**: one supporting paragraph removed and replaced by a distractor;
(4) **misleading_nearmiss**: one supporting paragraph replaced by a high-overlap near-miss paragraph.

The original project treated these as a three-way evidence-state problem (Sufficient, Missing, Misleading). After the primary Milestone-4 results, the project was conservatively reframed to the binary task Sufficient vs. Insufficient, with Missing and Misleading merged. We keep the three-way analysis in the appendix because it is scientifically informative but not strong enough to support the main paper.

Missing and misleading variants are constrained not to contain the normalized gold answer string in titles or paragraph text. The grouped construction is important: the question remains fixed while the evidence state changes, so the main contrast cannot be reduced to question difficulty alone.

3.2 Models and prompts

The strongest completed run in the bundle uses Qwen2.5-7B-Instruct as the primary model. The prompt surface is standardized across the main experiment: the model receives the question, four paragraphs of context, and a request to write exactly two short reasoning steps. Hidden states are extracted at three anchors:

- PROMPT_END: the last prompt token before generation;
- STEP1_END: the last token of the first reasoning sentence;
- STEP2_END: the last token of the second reasoning sentence.

The completed project state also includes a title-masked robustness run, in which page titles are removed from all contexts before the same downstream stages are rerun, and a DeepSeek-R1-Distill-Qwen-7B reasoning-model replication. For DeepSeek, the locked two-step prefix failed systematically under the original prompt format on the first seven of seven sampled groups. The replication was therefore resumed under a conservative model-specific compatibility mode: the official chat template is used, an assistant-side `<think>\n` prefix is forced, and the first two complete sentences inside the think block are parsed as step 1 and step 2 while PROMPT_END remains anchored to the last real prompt token before the forced assistant prefix. We treat this as a compatible secondary replication rather than as a prompt-identical anchor study.

3.3 Closed-book screening and evaluation subsets

To disentangle sufficiency from parametric knowledge, the pipeline first runs question-only closed-book screening. The main metric is computed on the *closed-book-negative* subset: grouped examples for which the model could not answer correctly from the question alone.

We use two further dissociation subsets:

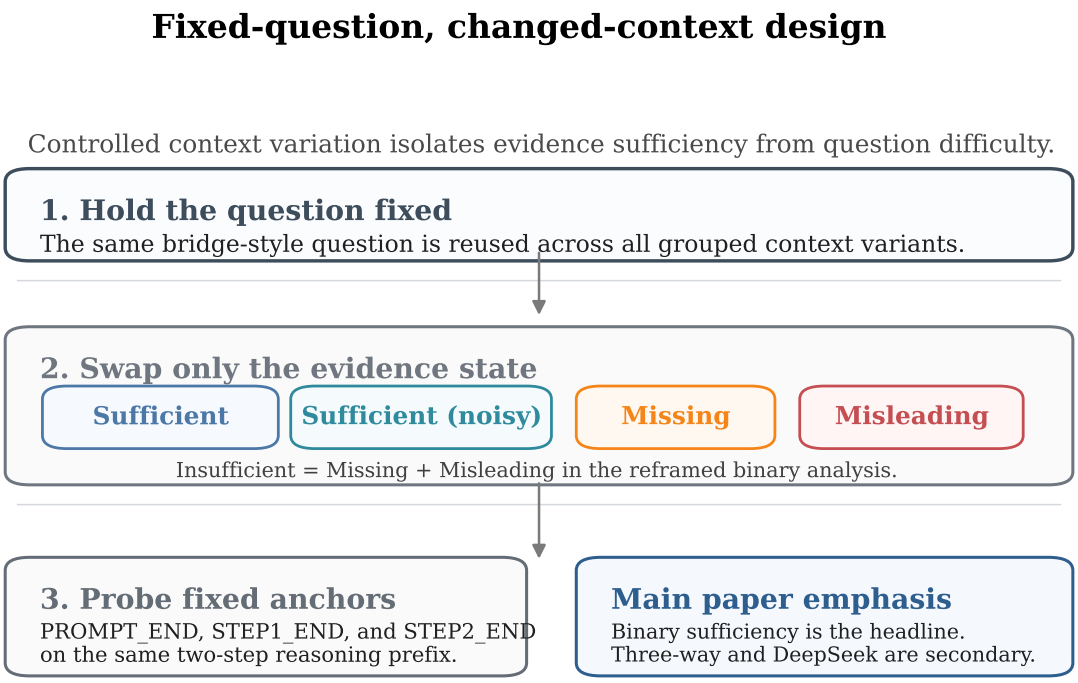

Figure 1: Fixed-question, changed-context pipeline. The question is held fixed, the context is varied across four controlled evidence states, and hidden states are probed at PROMPT_END, STEP1_END, and STEP2_END. The main paper asks whether these anchors separate sufficient from insufficient context.

- **correct-only closed-book-negative**: examples answered correctly under context despite being closed-book-negative;
- **wrong-only closed-book-negative**: examples answered incorrectly under context and also closed-book-negative.

These subsets are important because they make correctness non-diagnostic. On the primary Qwen test run, there are 87 correct predictions with insufficient context and 368 wrong predictions despite sufficient context, so correctness and context sufficiency are empirically non-identical even before probing.

3.4 Probes and controls

All language models remain frozen. The main probe is a linear logistic-regression classifier trained on hidden states. In the original three-way setting, the target is SUFFICIENT/MISSING/MISLEADING. In the reframed main analysis, the target is the binary contrast SUFFICIENT vs. INSUFFICIENT. Layers are selected on the validation split, and calibration is applied after training.

We compare the hidden-state probe against four kinds of controls:

(1) a **question-only answerability self-assessment** control, which operationalizes the pre-generation question-only contrast;
(2) a **context sufficiency self-report** baseline, in which the model explicitly labels the context as sufficient or insufficient;
(3) an **output-confidence** baseline based on final-answer token log-probabilities;
(4) a **TF-IDF lexical heuristic** baseline based only on context overlap features.

We also train a **correctness probe** on the same hidden-state family to test whether the sufficiency result is simply a relabeled correctness signal.

3.5 Metrics and retained data

The main metric of the final paper is sufficient-vs-insufficient AUROC on the closed-book-negative test subset. We also report correct-only and wrong-only closed-book-negative AUROCs. Cluster bootstrap confidence intervals are computed over question groups.

After prefix-format filtering and conservative Stage-04b QC, the primary Qwen run retains 2,006 groups (8,024 variants), split into 1,242 train groups, 376 validation groups, and 388 test groups. The title-masked robustness run retains 1,980 groups (7,920 variants). The completed DeepSeek replication retains 2,047 groups (8,188 variants) under the compatibility mode described above. These counts are summarized in Table A3.

4 Primary Results

4.1 Binary sufficiency dissociation is strong on the primary run

The primary HotpotQA/Qwen2.5-7B-Instruct result supports a binary evidence-sufficiency dissociation story. On the closed-book-negative test subset, the step-2 hidden-state probe reaches 0.952 AUROC for sufficient-vs-insufficient classification. This is substantially above the strongest non-hidden-state baseline, the TF-IDF lexical heuristic at 0.815, and also above the correctness probe at 0.809. The question-only answerability control is flat at 0.500, which is the expected failure mode in a fixed-question, changed-context design: the control never sees the manipulated context and therefore cannot distinguish among the four variants.

The same signal survives when final correctness is held fixed. On correct-only closed-book-negative examples, the step-2 probe reaches 0.979 AUROC; on wrong-only closed-book-negative examples, it remains at 0.939. This is the core reason we interpret the effect as a dissociation from correctness rather than a mere re-labeling of answer success. If the probe were only learning "will the model be right?", performance should collapse once correctness is held constant. It does not.

4.2 The signal is already accessible at prompt end

The strongest honest interpretation of the anchor comparison is that coarse context sufficiency is already accessible at prompt end. On the primary run, the prompt-end probe slightly exceeds the step-2 probe on the main binary metric (0.958 vs. 0.952), while the paired comparison does not reject equality ($p = 0.201$ after correction). Step 1 and step 2 are also very close (0.951 and 0.952). We therefore do *not* claim that the short reasoning prefix creates the signal or that reasoning makes sufficiency more linearly accessible in this setup. A more conservative interpretation is that, once the full question-context prompt has been read, a coarse sufficiency judgment is already available in the hidden state and remains available through the short reasoning prefix.

This distinction matters for the framing of the paper. The main contribution is not "mid-reasoning hidden states are uniquely informative." The contribution is that hidden states carry a strong and correctness-dissociated *binary context-sufficiency signal* under controlled context manipulations.

4.3 Why the paper is binary rather than three-way

The project did begin as a three-way evidence-state study. The original three-way result is not bad in the absolute sense: step 2 reaches 0.601 macro-F1 on the closed-book-negative subset, well above the strongest non-mid baseline. But the useful signal in that setting is almost entirely coarse sufficient-vs-insufficient separation. Missing-vs-misleading AUROC is only 0.505, with a bootstrap interval spanning chance, and prompt-end representations exceed step-2 representations on both three-way macro-F1 and missing-vs-misleading AUROC. For this reason, the three-way story is kept secondary and reported honestly in the appendix rather than forced into the headline claim.

5 Robustness and Secondary Replication

5.1 Title masking preserves the binary effect

The most important robustness check in the completed bundle is title masking. Here the page titles are removed from the context, and the downstream stages are rerun on the masked grouped data. If the main signal were driven primarily by easy lexical cues from page titles, we would expect a large collapse. That is not what happens.

Under title masking, the step-2 probe still reaches 0.936 AUROC on the closed-book-negative subset, with 0.959 on the correct-only closed-book-negative subset and 0.934 on the wrong-only subset. The strongest non-hidden-state baseline drops to 0.762, so the step-2 margin over the best lexical baseline actually widens from 0.137 to 0.175. The binary dissociation therefore survives the removal of an obvious source of shallow metadata.

At the same time, title masking does not rescue a reasoning-created-signal story. Prompt end remains slightly higher than step 2 (0.949 vs. 0.936), and the paired test again fails to reject equality. Our interpretation is therefore restrained: title masking is an artifact check, not evidence that the signal becomes more reasoning-specific once titles are removed.

5.2 DeepSeek-R1-Distill-Qwen-7B gives mixed secondary evidence

The completed reasoning-model replication is informative but not cleanly confirmatory. The original locked two-step surface form failed on the initial DeepSeek sample, so the completed run used a conservative model-specific compatibility mode documented in the appendix.

Under that compatibility mode, the DeepSeek step-2 probe still separates sufficient from insufficient context with 0.803 AUROC on the closed-book-negative test subset. This is clearly above the question-only answerability control (0.500) and above the correctness probe (0.498), so the run is not a null result. However, it is effectively tied with prompt end (0.804) and slightly below the strongest non-hidden-state baseline, the TF-IDF lexical heuristic at 0.806. The run also produces only three correct examples in the closed-book-negative test subset, so the correct-only dissociation metric is not estimable. We therefore treat DeepSeek as mixed secondary evidence rather than as a successful confirmatory replication.

5.3 What these robustness results support

Taken together, the robustness results support a narrow claim. The primary binary dissociation is not driven solely by title cues, and it is not exhausted by final correctness. What they do *not* support is a stronger claim about universal model behavior. The reliable contribution is a strong controlled primary result plus a robustness check that survives title masking; the DeepSeek run is best read as informative boundary evidence rather than as the center of the paper.

6 Discussion

The central takeaway is narrow but robust: hidden states encode a *coarse binary property of the provided context*, namely whether the

Table 1: Primary HotpotQA binary results on the closed-book-negative split and correctness-controlled subsets.

Method	Anchor	CBN AUROC	Correct-only	Wrong-only
Prompt-end probe	PROMPT_END	**0.958**	0.979	**0.956**
Step-1 probe	STEP1_END	0.951	—	—
Step-2 probe	STEP2_END	0.952	**0.979**	0.939
TF-IDF lexical baseline	—	0.815	—	—
Context sufficiency self-report	—	0.735	—	—
Output confidence	—	0.592	—	—
Question-only answerability control	—	0.500	0.500	0.500
Correctness probe	STEP2_END	0.809	0.768	0.769

CBN = closed-book-negative. Dashes indicate subset metrics that are not defined for that comparator. The key point is not that step 2 exceeds prompt end, but that hidden-state probes strongly separate sufficient from insufficient context and remain strong when correctness is held fixed.

context is sufficient to support answering the question under the controlled fixed-question, changed-context setup. This signal is not reducible to final correctness, because it remains strongly decodable on both correct-only and wrong-only closed-book-negative subsets. It is also not captured by the question-only pre-generation control, which is necessarily blind to the manipulated context.

The negative results are just as important for interpreting that signal. We do not find a robust three-way evidence-state representation: missing-vs-misleading separation is weak, and the original three-way headline does not survive contact with the data. We also do not find evidence that the short reasoning prefix creates the sufficiency signal. In the primary binary result, prompt-end representations are numerically stronger than step-2 representations, and in the title-masked and DeepSeek runs they remain at least tied. The conservative conclusion is that much of the relevant coarse sufficiency information is already available once the model has read the full prompt.

The robustness checks sharpen the scope of the claim rather than widen it. Title masking matters because a strong lexical baseline is present in the primary run, so a shallow-cue explanation is an obvious reviewer concern. Title masking does reduce performance modestly, but it hurts the lexical baseline more than the hidden-state probe. That pattern supports the claim that the probe is not reading only obvious title metadata. At the same time, title masking does not make step 2 the dominant anchor, so it should not be used rhetorically to imply a reasoning-specific signal that the numbers do not support.

The DeepSeek replication should likewise be read as a boundary result, not as a clean confirmation. It shows that the binary signal is not unique to the primary instruct model, because the DeepSeek probe remains well above the question-only and correctness controls. But it also shows that the effect is not yet a robust cross-model headline: the run is only mixed, and it required a conservative model-specific compatibility mode. The paper therefore supports a strong primary result with a meaningful robustness check, not a universal cross-model claim.

Our probe remains linear and the study remains correlational. The paper does not identify a causal circuit, and it does not show that the model robustly represents Missing and Misleading as separate internal categories.

7 Limitations

This paper should be read with several limitations in mind.

Primary evidence comes from one dataset family. All main results are on controlled HotpotQA variants. The grouped construction is carefully specified, but it is still one task family. A completed secondary dataset result is not included in the current paper package and should not be implied.

The strongest result is on the instruct control, not the reasoning-distilled model. The Qwen2.5-7B-Instruct run is the cleanest and strongest completed experiment. The DeepSeek-R1-Distill-Qwen-7B replication is useful but mixed, and it required a model-specific compatibility mode after the original stage-4 surface form failed systematically. The paper therefore cannot claim strong cross-model confirmation.

Prompt-end strength weakens any reasoning-created-signal narrative. Prompt-end representations are at least as informative as step-2 representations in the main binary analyses. We therefore interpret the result as a hidden-state sufficiency signal that is already available once the model has read the prompt, not as a signal created by the short reasoning prefix.

The question-only control is operational, not exhaustive. Our question-only answerability self-assessment baseline is a locked pre-generation control intended to capture the contrast to question-only uncertainty estimation. It is not meant as a faithful reimplementation of every query-level uncertainty method in the literature.

The probe is correlational. Linear decodability does not imply a localized or causal representation. The paper makes no claim of identifying a specific circuit, and the present evidence is not sufficient to support one.

Not all planned robustness paths are complete. The included project state indicates that other Milestone-5 paths, such as natural-CoT and shuffled-order continuations, were incomplete or not ready for scientific interpretation. We therefore do not report them here. Omitting unfinished runs is the more scientifically honest choice than presenting partial or design-ambiguous results as if they were stable evidence.

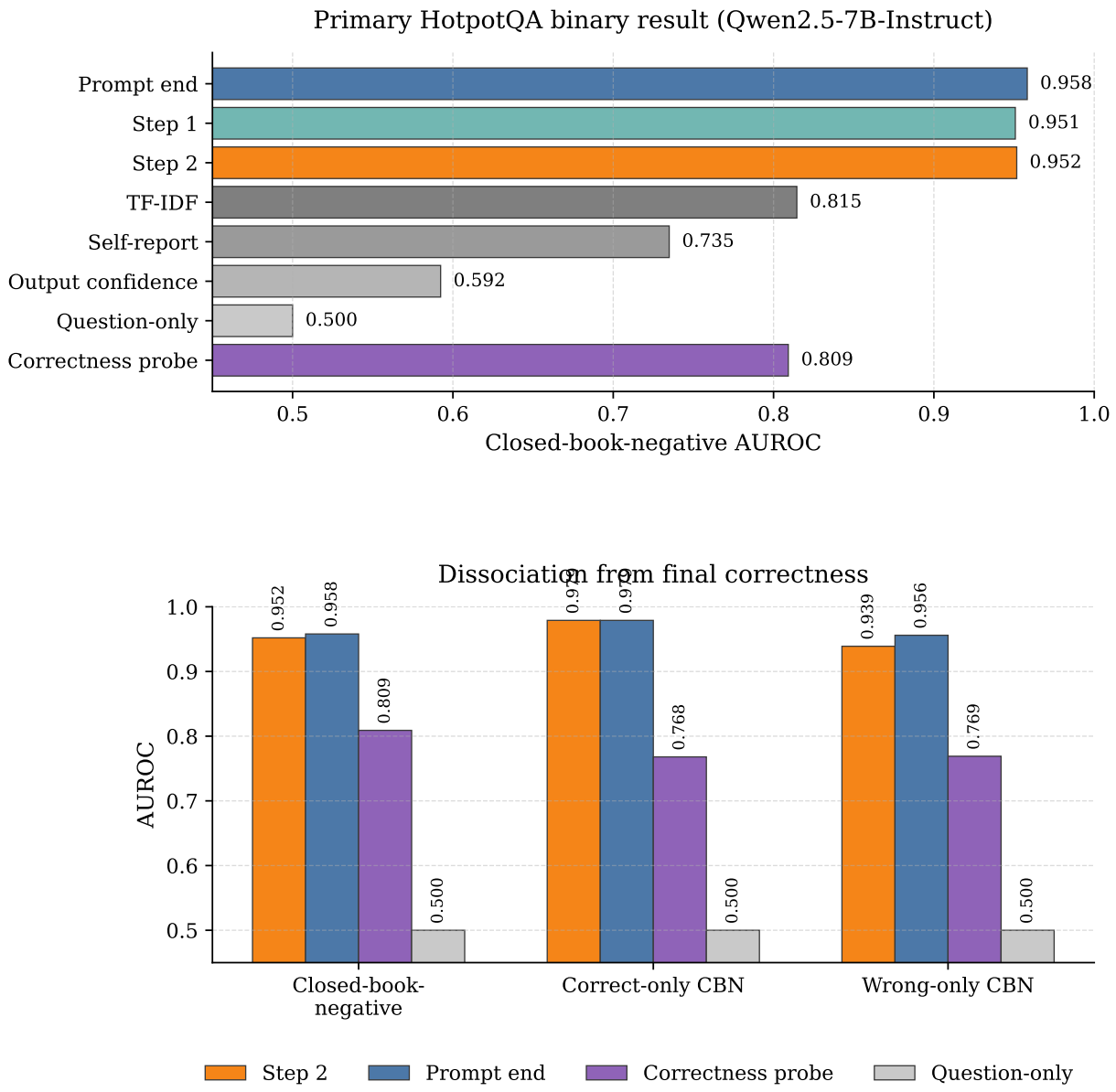

Figure 2: Primary binary results on HotpotQA with Qwen2.5-7B-Instruct. Hidden-state probes strongly separate sufficient from insufficient context on the closed-book-negative split and remain strong on correctness-controlled subsets. Prompt end is slightly stronger than step 2, so the figure supports a hidden-state sufficiency signal but not a reasoning-created-signal story.

8 Conclusion

The completed evidence supports a conservative but meaningful claim. In fixed-question, changed-context multi-hop QA, hidden-state probes linearly separate sufficient from insufficient context, and this signal remains strong even when final-answer correctness is held fixed. That is the main result.

The same evidence also tells us what *not* to claim. We do not find robust missing-vs-misleading separation, and we do not find evidence that the short reasoning prefix creates the sufficiency signal. Title masking preserves the binary effect and therefore strengthens the claim that the result is not purely title-driven, but the DeepSeek replication is mixed rather than confirmatory. The right conclusion is therefore not a stronger controller or retrieval story. It is a narrower representational one: coarse context sufficiency is linearly decodable from hidden states under controlled context manipulations, but the richer three-way evidence-state story remains unsupported in the current project state.

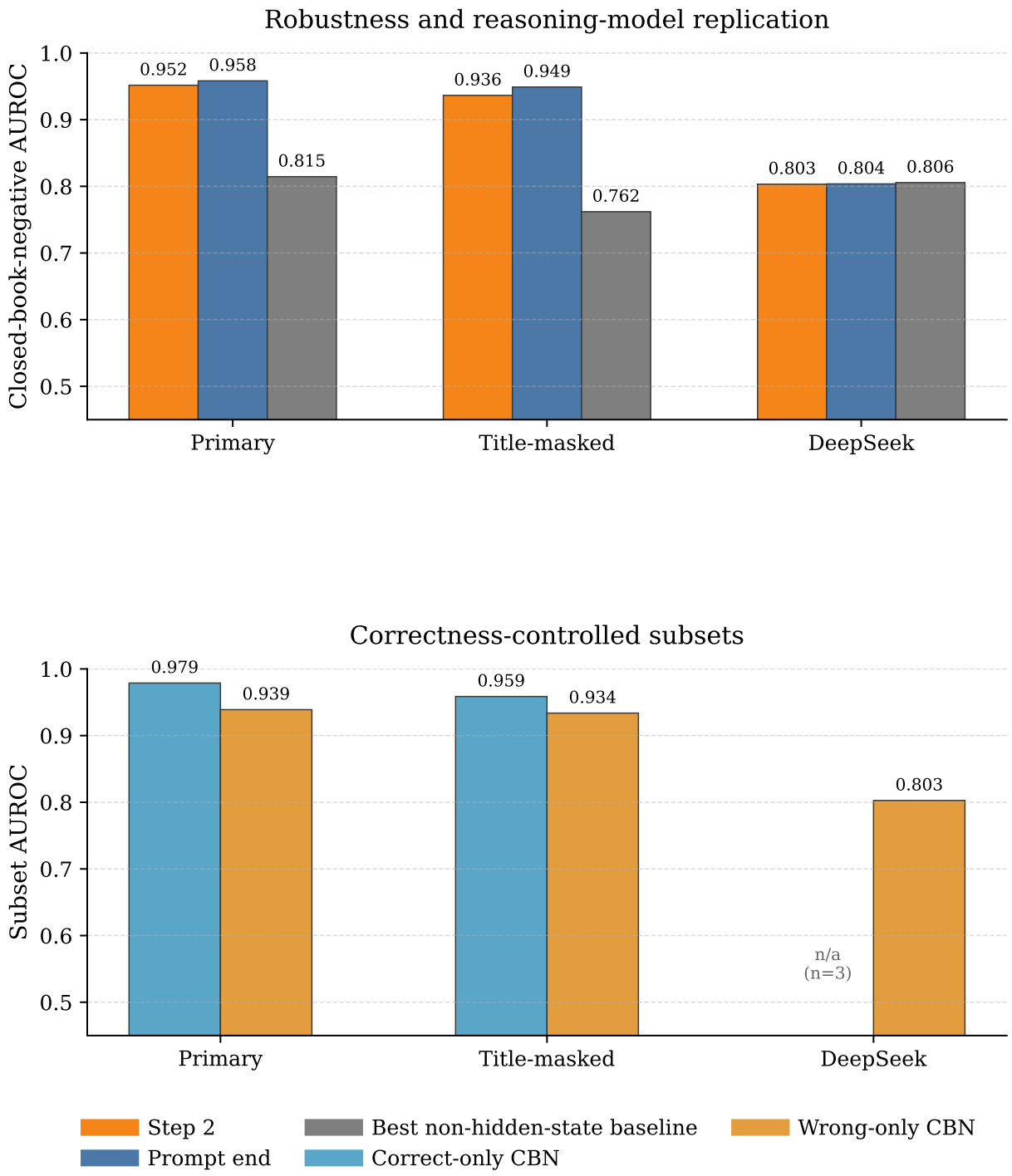

Figure 3: Robustness and secondary replication. Title masking preserves the main binary effect and the correctness-controlled pattern. DeepSeek remains above the question-only and correctness controls, but is effectively tied with prompt end and slightly below the strongest lexical baseline, so it should be read as mixed secondary evidence rather than as a confirmatory replication.

References

[1] Ingeol Baek, Hwan Chang, ByeongJeong Kim, Jimin Lee, and Hwanhee Lee. 2025. Probing-RAG: Self-Probing to Guide Language Models in Selective Document Retrieval. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3287–3304.

[2] Lihu Chen and Gaël Varoquaux. 2025. Query-Level Uncertainty in Large Language Models. *arXiv preprint arXiv:2506.09669*.

[3] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*.

[4] Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. 2025. Sufficient Context: A New Lens on Retrieval Augmented Generation Systems. In *International Conference on Learning Representations*.

[5] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

[6] Anqi Zhang, Yulin Chen, Jane Pan, Chen Zhao, Aurojit Panda, Jinyang Li, and He He. 2025. Reasoning Models Know When They Are Right: Probing Hidden States for Self-Verification. In *Conference on Language Modeling*.

A Secondary Three-Way Analysis

The original project framing treated evidence state as a three-way label over SUFFICIENT, MISSING, and MISLEADING. We report those results here because they explain why the final paper is framed around binary sufficiency instead.

Table A1: Secondary three-way evidence-state results on the primary HotpotQA run.

Method	3-way F1	Suff./insuff.	Missing/misleading
Prompt-end probe	**0.637**	**0.952**	0.576
Step-1 probe	0.611	0.951	0.543
Step-2 probe	0.601	0.941	0.505
TF-IDF lexical baseline	0.511	0.815	**0.580**
Context sufficiency self-report	0.414	0.735	0.482
Question-only answerability control	0.222	0.500	0.500

The useful three-way signal is largely coarse sufficient-vs-insufficient separation. Missing-vs-misleading separation is weak at step 2 and does not support the stronger original headline.

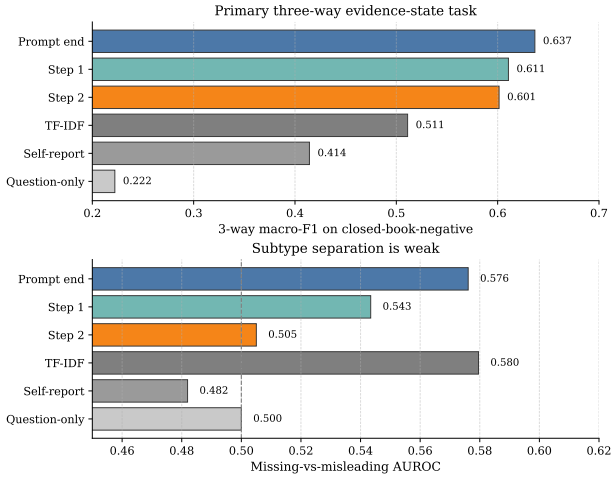

Figure A1: Secondary three-way evidence-state results on the primary HotpotQA run. The top panel shows that prompt end exceeds step 2 on three-way macro-F1. The bottom panel shows that missing-vs-misleading separation remains weak and near chance at step 2.

The three-way result is *borderline* rather than useless. Step 2 reaches 0.601 macro-F1 on the closed-book-negative subset, above the strongest non-mid baseline. But the part that matters for the original evidence-state claim is weak: missing-vs-misleading AUROC is 0.505, with a bootstrap interval of [0.477, 0.532], and prompt end exceeds step 2 on both macro-F1 and subtype AUROC. We therefore regard the three-way analysis as a useful negative result rather than as the main paper.

B Robustness and Replication Table

This summary table is deferred to the appendix because it supports interpretation rather than the main binary claim. The title-masked result is the central completed robustness check; the DeepSeek result is informative but mixed.

C Run Counts and Secondary Procedural Details

Table A3: Retained grouped examples after structured-prefix filtering and Stage-04b QC.

Run	Train groups	Val. groups	Test groups	Total groups
Primary Qwen2.5-7B-Instruct	1242	376	388	2006
Title-masked Qwen2.5-7B-Instruct	1226	370	384	1980
DeepSeek-R1-Distill-Qwen-7B	1266	390	391	2047

Each retained group contributes four context variants. The appendix reports these counts to make the paper's completed evidence boundary explicit.

The primary Qwen2.5-7B-Instruct path retains 2,006 grouped examples after structured-prefix filtering and the conservative post-prefix QC pass. The title-masked continuation retains 1,980 groups. The DeepSeek replication retains 2,047 groups under the compatibility mode described in the main text.

The compatibility mode matters because the original locked two-step prompt failed systematically on the first seven sampled DeepSeek groups, which were all rejected as malformed. The resumed replication followed a conservative rule: force only an assistant-side `<think>\n` prefix, parse the first two complete sentences inside the think block as steps, and keep PROMPT_END fixed at the last real prompt token. This preserves the project framing while acknowledging that the resulting run is not a prompt-identical reproduction of the primary stage-4 setup.

D Unreported and Incomplete Paths

The project bundle documents additional Milestone-5 work, including shuffled-order scaffolding and a natural-CoT design question. These are not included in the paper because the current bundle does not support a stable scientific interpretation of them. In particular, the natural-CoT path remained blocked on anchor-definition ambiguity in the project notes. We therefore omit these paths entirely from the empirical narrative rather than present them as if they were completed results.

Table A2: Robustness and secondary replication summary for binary sufficient-vs-insufficient AUROC on the closed-book-negative split.

Condition	Step-2	95% CI	Prompt end	TF-IDF	Correct-only	Wrong-only
Primary	0.952	[0.939, 0.964]	0.958	0.815	0.979	0.939
Title-masked	0.936	[0.920, 0.952]	0.949	0.762	0.959	0.934
DeepSeek	0.803	[0.783, 0.823]	0.804	0.806	—	0.803

The strongest non-hidden-state baseline is TF-IDF in all three settings. The DeepSeek replication is mixed rather than confirmatory: under the compatibility mode, step 2 is effectively tied with prompt end and slightly below TF-IDF; the correct-only subset is too small to estimate reliably.