

Contradictory Evidence in LLMs: Tracing Internal Conflict Behavior with White-Box Signals

Passant Elchafei¹, Ali Monshizadehkamkar¹, Shahed Masoudian¹,
Monorama Swain¹, Yufang Hou², Markus Schedl¹

¹ Johannes Kepler University Linz, Austria

² IT:U Interdisciplinary Transformation University Austria

{passant.elchafei, alikamkar98}@gmail.com

{shahed.masoudian, monorama.swain, markus.schedl}@jku.at

yufang.hou@it-u.at

Abstract

When LLMs are given contradictory passages, the generated response may appear decisive while masking how internal signals behave under conflicting evidence. Does it favor one passage, blend both, or ignore both? We study model behavior under inter-context contradiction by introducing a white-box tracing framework that extracts internal model signals based on hidden-state shifts, counterfactual attribution via passage removal, and token-level uncertainty during generation under two single-passage prompts and one dual-passage prompt with conflicting evidence. We evaluate four open-weight LLMs from the LLaMA (8B and 70B), OLMo, and Qwen on WikiContradict, a human-annotated QA benchmark with contradictory passages. Each dual-passage answer is labeled as committing to passage A or passage B, blending both, or fabricating an answer unsupported by either passage. Comparing these labels with the extracted signals shows that larger models reduce fabrication, but hidden-state and probability-level signals can disagree with the final answer. Moreover, fabricated answers often remain high-confidence, showing that confidence alone is unreliable for identifying ungrounded behavior under contradictory evidence. We further validate on ConflictBank, a larger misinformation-conflict QA dataset, showing consistent disagreement patterns across models. Finally, we test whether pre-generation white-box signals are informative beyond interpretation by training classifiers to predict output behavior. Classifiers using white-box signals outperform probability baselines, showing these signals contain diagnostic information about model behavior under contradictory evidence beyond log probabilities.

1 Introduction

Large language models (LLMs) are increasingly used for question answering settings where the provided evidence may contain conflicting informa-

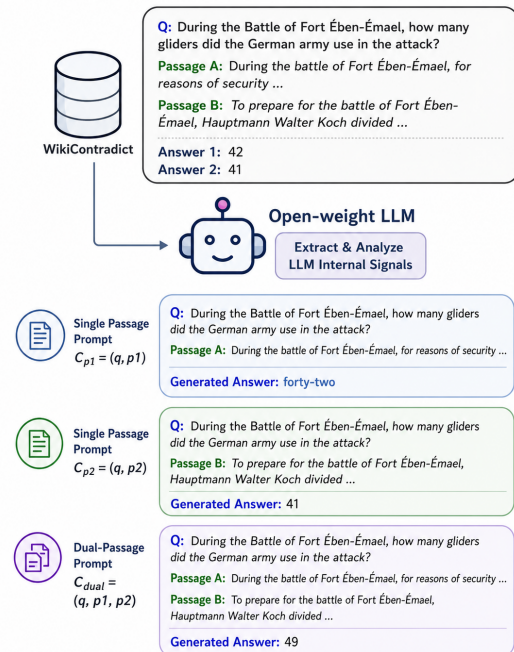


Figure 1: Evidence-conditioned prompting setup for white-box analysis of contradictory evidence. For each WikiContradict instance, we query an open-weight LLM with each passage separately (C_{p1} , C_{p2}) and with both passages together C_{dual} .

tion. Such conflicts are also common in retrieval-augmented generation (RAG) scenarios, where retrieved passages may disagree. Knowledge conflicts can arise as *context-memory conflict*, where external evidence conflicts with the model’s parametric knowledge; *inter-context conflict*, where retrieved passages conflict with each other; and *intra-memory conflict*, where conflicting knowledge is encoded within the model itself (Xu et al., 2024). This paper focuses on *inter-context conflict*: a setting where two passages provide contradictory answers to the same question, as illustrated in Figure 1. Unlike standard factual QA, there may be no single answer supported by all provided evidence. A model may commit to one

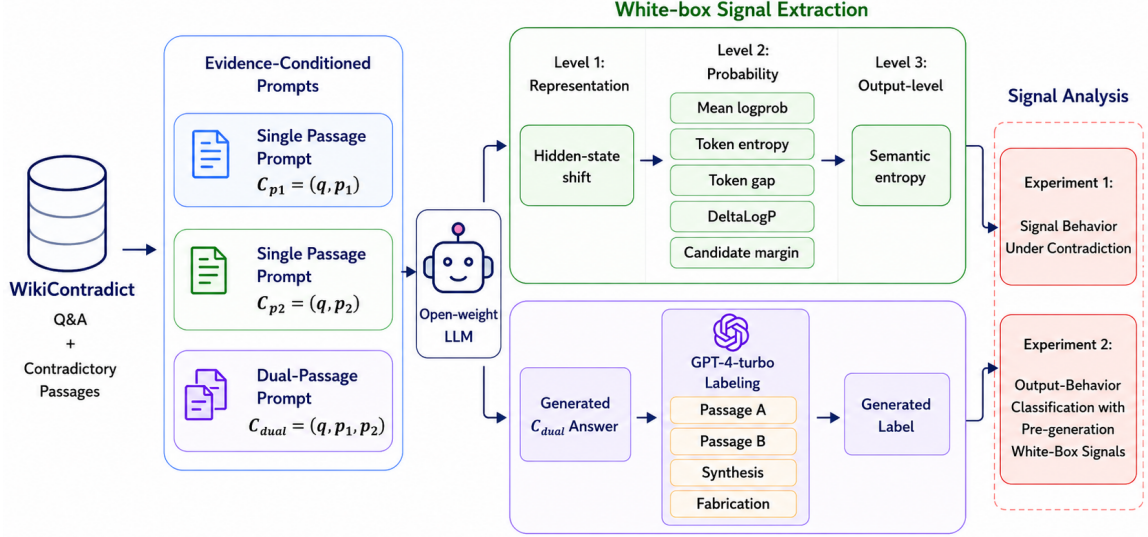


Figure 2: LLM white-box tracing pipeline. Each WikiContradict instance is converted into three evidence-conditioned prompts and evaluated with an open-weight LLM. We then extract hidden state, probability, and output level signals from the three conditions, and label the generated C_{dual} answer using GPT-4-turbo. The extracted signals and output labels are used to trace how internal behavior changes under contradictory evidence and how it relates to the model’s final answer.

passage, commit to the other, blend both passages, or generate an answer unsupported by either passage. For example, if one passage states that an event involved 42 gliders while another states 41, a model may answer “42”, “41”, blend both inconsistently, or produce an ungrounded answer such as “49”. This motivates our main research question: **how does model behavior differ between single-evidence and dual-evidence contradictory conditions?** Final-answer evaluation alone cannot answer this question, because it does not show how behavior changes relative to each passage separately. Probability-based signals are also insufficient: they indicate how likely the generated answer is, but not whether behavior follows one passage, blends both passages, or produces an answer unsupported by either passage. We therefore use single-passage conditions as reference points for analyzing the dual-passage conflict condition.

Prior work has studied uncertainty (Vashurin et al., 2025), internal states (Orgad et al., 2025), and knowledge conflicts (Xu et al., 2024) in LLMs. However, it remains unclear whether white-box signals, i.e., internal model signals such as hidden-state shifts, context attribution, and token-level uncertainty, differ between single- and dual-passage contradictory evidence conditions, and whether they provide diagnostic information about output behavior beyond output log probabilities. We ad-

dress this gap through three research questions:

RQ1: How do hidden-state, context attribution, and token-level uncertainty signals differ between single-passage evidence conditions and the dual-passage contradictory condition?

RQ2: What output behaviors do LLMs produce when responding to contradictory passages: committing to passage A, committing to passage B, blending both passages, or fabricating an answer unsupported by either passage?

RQ3: Do pre-generation white-box signals provide information beyond output log probabilities for interpreting and classifying output behavior under contradictory evidence?

To answer these research questions, we introduce a white-box tracing framework for open-weight LLMs. As shown in Figure 2, we compare two single-passage prompts with one dual-passage prompt, extract these signals, and label each dual-passage answer by its output behavior. Across four open-weight LLMs, larger models reduce fabrication but do not eliminate cases where internal signals and output behavior point to different passages. Fabricated answers can remain high-confidence, showing that confidence is unreliable for detecting answers unsupported by either passage. Finally, classifiers using pre-generation white-box signals outperform probability-only baselines, showing that internal signals capture information beyond

log probabilities. We further test generalizability on ConflictBank (Su et al., 2024b), a larger misinformation-conflict QA dataset, and observe consistent cross-level disagreement patterns.

Overall, our contributions are threefold: (i) we introduce a white-box tracing framework that extracts internal signals (hidden-state shifts, counterfactual attribution, token-level uncertainty) under single-passage and dual-passage conditions, enabling contrastive analysis of how contradiction changes model behavior across four open-weight LLMs: LLaMA (8B, 70B), OLMo, and Qwen; (ii) we show that fabrication, defined as an answer unsupported by either passage, can be high-confidence under contradiction, making confidence unreliable for detecting unsupported answers; (iii) we show that pre-generation white-box signals provide diagnostic information beyond output log probabilities: they improve output-behavior classification over probability-only baselines and help interpret model behavior under contradictory evidence, achieving macro-F1 of 0.641 for Qwen and 0.460 for OLMo.

2 Related Work

Internal signals for uncertainty and factuality. Prior work uses hidden states and internal activations to study uncertainty, factuality, and hallucination-related behavior, including semantic entropy (Farquhar et al., 2024), hidden-state truthfulness signals (Orgad et al., 2025), real-time hallucination detection (Su et al., 2024a; Obeso et al., 2025), UQ benchmarking (Vashurin et al., 2025). Our work differs by studying contradictory evidence directly: we compare dual-passage behavior with single-passage baselines and analyze whether hidden state, attribution, and token-level signals align with the final answer.

Knowledge conflicts and context faithfulness. (Xie et al., 2024) study the behavior under knowledge conflicts: models are receptive to coherent evidence but exhibit confirmation bias. (Xu et al., 2024) survey three conflict categories (context-memory, inter-context, intra-memory), and define inter-context conflict as conflict among contextual knowledge sources. (Hou et al., 2024) introduce WikiContradict, a benchmark of 253 human-annotated Wikipedia contradictions. (Ming et al., 2025) benchmark faithfulness across unanswerable and inconsistent contexts. However, most work focuses on final outputs rather than comparing white-

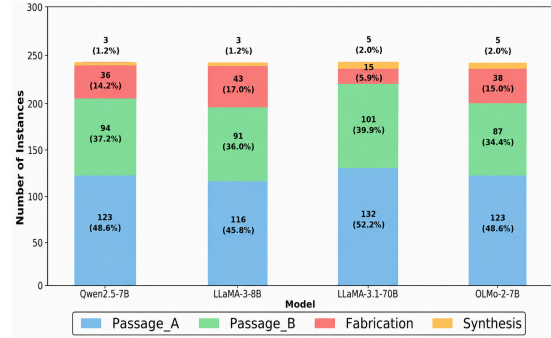


Figure 3: Output-label distribution under the dual-passage condition C_{dual} across the four evaluated models. LLaMA-3.1-70B shows fewer FABRICATION cases than the smaller models, while most outputs commit to one of the two passages.

box diagnostic signals across single-evidence and contradictory dual-evidence conditions.

Counterfactual attribution. Prior work studies context attribution by estimating how input evidence contributes to model outputs (Cohen-Wang et al., 2024). We adopt a counterfactual attribution view for contradictory evidence using leave-one-out log-probability changes: each passage is removed in turn, and the resulting change in output probability estimates its contribution to the generated answer.

3 White-Box Tracing of Contradictory Evidence

We study contradictory-evidence question answering using WikiContradict, which consists of 253 human-annotated Wikipedia contradictions. Each instance consists of a question q , two passages p_1 and p_2 , and two conflicting short answers a_1 and a_2 , where p_1 supports a_1 and p_2 supports a_2 . As illustrated in Figure 1, for each instance we construct three evidence-conditioned prompts:

$$\begin{aligned} C_{p1} &= (q, p_1), \\ C_{p2} &= (q, p_2), \\ C_{\text{dual}} &= (q, p_1, p_2). \end{aligned}$$

The single-passage prompts C_{p1} and C_{p2} provide reference behavior for each passage, while the dual-passage prompt C_{dual} exposes the model to contradictory evidence. Comparing C_{dual} with C_{p1} and C_{p2} allows us to analyze how white-box signals and output behavior differ between single-evidence and contradictory dual-evidence conditions.

As shown in Figure 2, our framework analyzes this behavior across three levels: hidden-state sig-

nals, probability-level signals, and final output behavior. We extract white-box signals from the prompt conditions, label each C_{dual} answer as following Passage A, following Passage B, combining both passages, or producing an answer unsupported by either passage, and compare these labels with the extracted signals to analyze agreement between signal-based indicators and output behavior. We evaluate four open-weight LLMs: LLaMA-3-8B, LLaMA-3.1-70B, OLMo-2-7B, and Qwen2.5-7B, using the same strict answer-only prompt format. The exact model checkpoints are listed in Appendix D.3.

3.1 Evidence-Conditioned Prompt Templates

The answer-generation prompt template is provided in Appendix A.1. In all three conditions, we use the same strict answer-only instruction so that the model returns a short answer phrase rather than an explanation. The only difference between conditions is the evidence provided: C_{p1} contains passage A only, C_{p2} contains passage B only, and C_{dual} contains both contradictory passages.

3.2 Dual-Passage Answer Labeling

For each answer generated under C_{dual} , we assign an output-behavior label based on its relation to the two conflicting reference answers from WikiContradict. GPT-4-turbo is given the question, the two reference answers, and the model-generated C_{dual} answer, and is asked to select one of four labels: PASSAGE_A, PASSAGE_B, SYNTHESIS, or FABRICATION. We use an 8-shot prompt, with two examples for each label; the full prompt is provided in Appendix A.2. To validate the labeling procedure, we manually annotated a stratified sample of 50 instances. GPT-4-turbo matched the human labels in 98.0% of validation cases, with Cohen’s $\kappa = 0.965$, indicating almost perfect agreement. We therefore apply the same labeling procedure to all generated answers in the full dataset. Further details are provided in Appendix B.4. Figure 3 summarizes the output label distribution for C_{dual} across WikiContradict for each model.

Order-swapped label distribution. To check passage-order sensitivity, we generate an order-swapped condition $C_{\text{dual-swap}} = (q, p_2, p_1)$ and report output-label distributions in Appendix B.3, Figure 4. This condition serves as a diagnostic robustness check. Because reversing passage order

Component	Value
WikiContradict instance	
Instance	wc_0020
Question	What year did Paul Fraisse become director of the Institute of Psychology of the University of Paris?
Answer 1	1965
Answer 2	1952
White-box tracing outputs	
C_{p1} answer	1965
C_{p2} answer	1952
C_{dual} answer	1961
Hidden-state Signals	Passage B
Candidate preference	Passage A
Output label	Fabrication
Disagreement	True

Table 1: Example of cross-level disagreement. The WikiContradict instance provides two conflicting answers, while the tracing outputs show that the dual-passage answer is Fabrication and the internal signals point to different passages.

may affect output-label distributions, we exclude it from signal extraction and the main experiments. The main analyses therefore use a single fixed passage order.

3.3 White-Box Signal Extraction

Building on the evidence-conditioned prompts, we extract white-box signals under C_{p1} , C_{p2} , and C_{dual} . Comparing the dual-passage condition with the two single-passage conditions lets us measure how internal signals and output behavior differ under contradiction. We analyze three signal levels: hidden-state representations, probability-level behavior, and the generated answer. In Section 4, we examine which signals differ across conditions, distinguish output labels, and provide diagnostic information for output-behavior classification beyond output log probabilities.

3.3.1 Level 1: Hidden-States Signals

We extract hidden vectors $\mathbf{h}_T^{(\ell)}$ at the last input token position T from three depth-normalized layers, corresponding to approximately 25%, 50%, and 75% of each model’s transformer depth. This choice captures early, middle, and late representations while keeping the analysis comparable across models with different numbers of layers. Prior layer-wise analyses show that transformer layers encode different linguistic and semantic information at different depths (He et al., 2024). The selected layers are $\{7, 14, 21\}$ for Qwen2.5-7B, $\{8, 16, 24\}$ for LLaMA-3-8B and OLMo-2-7B, and $\{20, 40, 60\}$ for LLaMA-3.1-70B. Hidden drift

Output Label	Qwen2.5-7B		LLaMA-3-8B		LLaMA-3.1-70B		OLMo-2-7B	
	n	Disagree	n	Disagree	n	Disagree	n	Disagree
Passage_A	123	68 (55.3%)	116	63 (54.3%)	132	68 (51.5%)	123	57 (46.3%)
Passage_B	94	31 (33.0%)	91	41 (45.1%)	101	54 (53.5%)	87	23 (26.4%)
Synthesis	0	—	3	3 (100.0%)	5	2 (40.0%)	5	4 (80.0%)
Fabrication	36	18 (50.0%)	43	20 (46.5%)	15	5 (33.3%)	38	20 (52.6%)

Table 2: Cross-level disagreement rates by output label. Disagreement occurs when internal signals do not align in selecting the same passage.

measures mean cosine distance:

$$\text{drift}(C_{\text{dual}}, C_i) = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \left(1 - \frac{\mathbf{h}_{C_{\text{dual}}}^{(\ell)} \cdot \mathbf{h}_{C_i}^{(\ell)}}{\|\mathbf{h}_{C_{\text{dual}}}^{(\ell)}\| \|\mathbf{h}_{C_i}^{(\ell)}\|} \right) \quad (1)$$

where $\mathcal{L}$ is the set of selected depth-normalized layers for each model, and $C_i \in \{C_{p1}, C_{p2}\}$. We derive a hidden state direction from the smaller drift value: the representation points to Passage A if $\text{drift}(C_{\text{dual}}, C_{p1}) < \text{drift}(C_{\text{dual}}, C_{p2})$, and to Passage B otherwise. This follows the intuition that smaller cosine drift indicates greater representational similarity between the dual-passage condition and the corresponding single-passage reference condition.

3.3.2 Level 2: Probability-Level Signals

For each generated token t_k in the output sequence, we extract three token-level signals:

$$\begin{aligned} \text{logp}_k &= \log p(t_k \mid t_{<k}, \text{prompt}), \\ \text{entropy}_k &= - \sum_{v \in \mathcal{V}} p_v \log p_v, \\ \text{gap}_k &= p_{(1)} - p_{(2)}. \end{aligned}$$

where logp_k measures token confidence, entropy_k captures probability mass dispersion over vocabulary $\mathcal{V}$, and gap_k is the margin between the top two token probabilities $p_{(1)}$ and $p_{(2)}$ at position k . These token-level signals are averaged across the generated sequence to obtain mean logprob, mean token entropy, and mean token gap.

Candidate scoring. We use a common answer-reranking approach: comparing the model’s log-probability scores for the two reference answers under C_{dual} . This indicates which reference answer the model finds more probable when both contradictory passages are present, and therefore which passage dominates the model’s preference under contradiction.

$$\begin{aligned} \text{cand_score} &= \log P(a_1 \mid C_{\text{dual}}) \\ &\quad - \log P(a_2 \mid C_{\text{dual}}). \end{aligned} \quad (2)$$

Where $C_{\text{dual}} = (q, p_1, p_2)$, a_1 and a_2 are the two conflicting reference answers from WikiContradict, p_1 supports a_1 , and p_2 supports a_2 . Positive values indicate preference for a_1 , and negative values indicate preference for a_2 .

Counterfactual attribution (DeltaLogP). To estimate how each passage contributes to the generated answer, we use a leave-one-out counterfactual attribution approach inspired by context attribution work (Cohen-Wang et al., 2024). We compare the log-probability of the generated answer under the dual-passage prompt $C_{\text{dual}} = (q, p_1, p_2)$ with its log-probability when one passage is removed.

$$\Delta \log P_A = \sum_{t=1}^T \left[\log P(y_t \mid C_{\text{dual}}, y_{<t}) - \log P(y_t \mid C_{p2}, y_{<t}) \right], \quad (3)$$

$$\Delta \log P_B = \sum_{t=1}^T \left[\log P(y_t \mid C_{\text{dual}}, y_{<t}) - \log P(y_t \mid C_{p1}, y_{<t}) \right]. \quad (4)$$

Here, $y_{1:T}$ is the answer generated under C_{dual} . $\Delta \log P_A$ measures how much the probability of this generated answer changes when p_1 is removed, while $\Delta \log P_B$ measures the corresponding change when p_2 is removed. We also compute the *siding score*:

$$\text{siding_score} = \Delta \log P_A - \Delta \log P_B.$$

A positive siding score indicates stronger attribution to p_1 , while a negative score indicates stronger attribution to p_2 .

3.3.3 Level 3: Output Level Uncertainty

We compute semantic entropy (SE) as an output-level uncertainty signal. Semantic entropy estimates whether the model produces semantically diverse answers when sampled under the same

Output Label	Qwen2.5-7B		LLaMA-3-8B		LLaMA-3.1-70B		OLMo-2-7B	
	n	Higher conf.	n	Higher conf.	n	Higher conf.	n	Higher conf.
Passage_A	123	37 (30.1%)	116	30 (25.9%)	132	35 (26.5%)	123	21 (17.1%)
Passage_B	94	23 (24.5%)	91	24 (26.4%)	101	21 (20.8%)	87	19 (21.8%)
Synthesis	0	—	3	1 (33.3%)	5	1 (20.0%)	5	2 (40.0%)
Fabrication	36	13 (36.1%)	43	20 (46.5%)	15	5 (33.3%)	38	13 (34.2%)

Table 3: Higher confidence rates by output label. Higher confidence means the dual-passage conflict condition C_{dual} has higher mean log-probability than the stronger of the two single-passage conditions, $\max(C_{p1}, C_{p2})$. The n column gives the number of outputs with each behavior label.

prompt (Farquhar et al., 2024). We sample $K=10$, with decoding (temperature=0.7, top_p=0.9, max_new_tokens=64). We cluster them using NLI entailment nli-deberta-v3-small model (He et al., 2023), and compute:

$$\text{SE} = - \sum_c \frac{|c|}{K} \log \frac{|c|}{K} \quad (5)$$

where K is the number of sampled answers and $|c|$ is the size of semantic cluster c . If all sampled answers express the same meaning, they form one cluster and $\text{SE} = 0$. If the sampled answers are distributed across several semantic clusters, SE increases, indicating greater output level uncertainty.

4 Experiments

We evaluate our approach through two main experiments on 253 WikiContradict instances processed under C_{dual} , C_{p1} , and C_{p2} using four open-weight models: LLaMA-3-8B-Instruct, LLaMA-3.1-70B-Instruct, OLMo-2-7B-Instruct, and Qwen2.5-7B-Instruct. The first experiment analyzes how the extracted signals behave under contradiction and whether they differ across output-label groups (Section 4.1). The second experiment evaluates whether pre-generation signals provide diagnostic information for output-behavior classification beyond log probabilities (Section 4.2). We also include a generalization analysis on Conflict-Bank (Su et al., 2024b) in Section 4.3. Code and configurations are listed in Appendix D.

4.1 Signal Behavior Under Contradiction

Cross-level disagreement. We define cross-level disagreement when the internal signals and output behavior do not point to the same passage. For each instance, we compare which passage is indicated by the hidden-state signal (Section 3.3.1), which passage is preferred by candidate scoring (Section 3.3.2), and which passage is selected by the output label (Section 3.2). Table 1 shows an

example of cross-level disagreement. Table 2 reports disagreement rates across models and output types: 26–55% when the model follows one of the two passages and 33–53% for FABRICATION. Disagreement rates vary across models: Qwen2.5-7B shows the highest rate for PASSAGE_A, LLaMA-3.1-70B shows the highest rate for PASSAGE_B, and OLMo-2-7B shows the lowest rates for both passage-following labels. This indicates that under contradictory evidence, internal signals and output behavior often point to different passages, with different disagreement rates across models.

Confidence under contradiction. Table 3 reports how often the model is more confident under contradiction than under either single-passage condition.

Let $y_{1:T_C}^{(C)}$ be the answer generated under condition C . We define the mean token log-probability as:

$$\bar{\ell}(C) = \frac{1}{T_C} \sum_{t=1}^{T_C} \log p(y_t^{(C)} | C, y_{<t}^{(C)}).$$

We then define confidence change under contradiction as:

$$\Delta_{\text{conf}} = \bar{\ell}(C_{\text{dual}}) - \max(\bar{\ell}(C_{p1}), \bar{\ell}(C_{p2})).$$

A positive value means that the dual-passage answer has higher mean token log-probability than the stronger single-passage baseline. Across all four models, FABRICATION often appears with increased confidence: 46.5% for LLaMA-3-8B, 36.1% for Qwen2.5-7B, 34.2% for OLMo-2-7B, and 33.3% for LLaMA-3.1-70B. This shows that ungrounded answers are not reliably low-confidence under contradiction.

Statistical tests. We examine two questions: (1) Does contradiction change the extracted signals? (2) Do signals differ by output-label type? For question (1), we run paired Wilcoxon signed-rank tests to compare paired conditions (C_{dual} vs C_{p1} ,

Signal	Qwen2.5-7B	LLaMA-3-8B	LLaMA-3.1-70B	OLMo-2-7B
$\Delta \log P_A$	< 0.001***	< 0.001***	0.190 ns	< 0.001***
$\Delta \log P_B$	< 0.001***	< 0.001***	< 0.001***	< 0.001***
Hidden drift from C_{p1}	< 0.001***	< 0.001***	< 0.001***	< 0.001***
Candidate margin	< 0.001***	< 0.001***	< 0.001***	< 0.001***
Token entropy	0.047*	0.162 ns	0.439 ns	0.236 ns
Token gap	0.140 ns	0.067 ns	0.139 ns	0.474 ns
Hidden drift from C_{p2}	0.136 ns	0.492 ns	0.957 ns	0.073 ns

Table 4: Kruskal-Wallis test results comparing each signal across C_{dual} output label by (Passage_A, Passage_B, Synthesis, Fabrication). Significant values indicate the signal separates output types. Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ns = not significant.

Task	Feature set	Qwen2.5-7B		LLaMA-3-8B		LLaMA-3.1-70B		OLMo-2-7B	
		LogReg	GBT	LogReg	GBT	LogReg	GBT	LogReg	GBT
4-class	PreGen white-box	0.641	0.581	0.387	0.429	0.352	0.365	0.460	0.453
	Logprob-only baseline	0.497	0.569	0.364	0.366	0.295	0.288	0.398	0.409
Binary	PreGen white-box	0.614	0.532	0.525	0.576	0.427	0.473	0.478	0.503
	Logprob-only baseline	0.507	0.482	0.447	0.519	0.437	0.470	0.410	0.483

Table 5: PreGen white-box refers to the pre-generation signals set and Logprob-only baseline uses candidate margin. Performance across the four evaluated models using macro-F1 with stratified 5-fold cross-validation.

C_{dual} vs C_{p2}) in the same instances for mean token entropy, mean token gap, and mean log-probability (all token-level signals averaged over the generated answer). We also test $\Delta \log P_A$ and $\Delta \log P_B$ against zero (Appendix C.1, Table 7). Results show that contradiction produces selective rather than uniform signal shifts in most models. Token entropy and log-probability show significant changes for some single-passage comparisons, while token gap is less consistent. However, OLMo-2-7B and LLaMA-3.1-70B show significant changes across all token-level signals, suggesting more consistent token-level sensitivity to contradictory evidence. Counterfactual attribution signals are significant in most models for both passages. This indicates that removing either passage often changes the model’s probability for the generated answer.

For question (2), we run Kruskal-Wallis tests comparing each signal across the four output labels. Table 4 shows that the strongest differences across output-label groups come from attribution and preference-related signals. $\Delta \log P_A$, $\Delta \log P_B$, hidden drift from C_{p1} , and candidate margin are significant across most or all models. In contrast, token entropy, token gap, and hidden drift from C_{p2} are less consistent. This indicates that output behavior under contradiction is better separated by passage-attribution and directional signals than by generic uncertainty measures.

4.2 Classifying Output Behavior Using Pre-Generation White-Box Signals

Setup. We evaluate whether answer-independent pre-generation signals provide diagnostic information for output-behavior classification under contradiction, beyond output log probabilities. To avoid using information from the generated answer, we restrict this experiment to signals available before answer generation. We train Logistic Regression and Gradient Boosted Trees using stratified cross-validation on two feature sets: (1) PreGen white-box signals, using signals available before answer generation: hidden drift comparing C_{dual} with C_{p1} and C_{p2} (extracted at the last input token position, Section 3.3.1); and (2) logprob-only baseline, using candidate margin only. We evaluate two settings. The 4-class setting predicts the full output-behavior label: PASSAGE_A, PASSAGE_B, SYNTHESIS, or FABRICATION. The binary setting predicts whether the answer is supported by at least one passage or is FABRICATION.

Results. Table 5 shows that PreGen white-box signals outperform the logprob-only baseline in the 4-class setting across all four models, indicating that answer-independent white-box signals provide diagnostic information about output behavior beyond output log probabilities. For 4-class classification, Qwen2.5-7B improves from 0.497 to 0.641 with Logistic Regression and from 0.569 to 0.581

Output Label	Qwen2.5-7B		LLaMA-3-8B		LLaMA-3.1-70B		OLMo-2-7B	
	n	Disagree	n	Disagree	n	Disagree	n	Disagree
Passage_A	542	247 (45.6%)	714	190 (26.6%)	656	309 (47.1%)	608	53 (8.7%)
Passage_B	170	77 (45.3%)	6	1 (16.7%)	50	35 (70.0%)	83	28 (33.7%)
Synthesis	0	—	3	1 (33.3%)	30	8 (26.7%)	37	12 (32.4%)
Fabrication	38	17 (44.7%)	27	4 (14.8%)	14	6 (42.9%)	22	3 (13.6%)

Table 6: Cross-level disagreement rates by output label on ConflictBank. Disagreement occurs when internal signals do not align in selecting the same passage.

with GBT. LLaMA-3-8B improves from 0.364 to 0.387 with Logistic Regression and from 0.366 to 0.429 with GBT. LLaMA-3.1-70B improves from 0.295 to 0.352 with Logistic Regression and from 0.288 to 0.365 with GBT. OLMo-2-7B improves from 0.398 to 0.460 with Logistic Regression and from 0.409 to 0.453 with GBT. In the binary setting, PreGen white-box signals also improve over the logprob-only baseline for Qwen2.5-7B, LLaMA-3-8B, and OLMo-2-7B, while performance is comparable for LLaMA-3.1-70B. These results show that answer-independent pre-generation signals contain diagnostic information associated with output behavior under contradictory evidence, although the overall performance remains moderate and varies across models.

4.3 Generalization to ConflictBank

To test whether the observed cross-level disagreement patterns generalize beyond WikiContradict, we evaluate our framework on ConflictBank (Su et al., 2024b), a large knowledge-conflict benchmark covering misinformation, temporal, and semantic conflicts. We use 750 instances from the misinformation-conflict QA subset. In this setting, each instance pairs a default factual evidence context with a synthetically generated misinformation context, creating an asymmetric conflict between factual and misleading evidence. We apply the same prompting conditions, white-box signal extraction pipeline, output taxonomy, and disagreement analysis used for WikiContradict. Table 6 reports the corresponding cross-level disagreement rates on ConflictBank. The results show that cross-level disagreement persists on ConflictBank across all four models. Although the output distribution is skewed toward PASSAGE_A, especially for LLaMA-3-8B, disagreement remains common even when the model selects the factual evidence side. For example, LLaMA-3-8B selects PASSAGE_A in 714 cases, of which 190 show internal disagreement. Qwen2.5-7B and LLaMA-

3.1-70B show even higher disagreement rates for PASSAGE_A, at 45.6% and 47.1%, respectively. These results support the main finding that final answers can mask disagreement among internal signals.

5 Discussion

Our experiments reveal how LLMs behave under contradictory evidence.

Contradictory evidence creates different output behaviors. Under C_{dual} , a model can commit to the answer supported by p_1 , commit to the answer supported by p_2 , blend information from both passages, or generate an answer unsupported by either passage. These behaviors are not captured by a simple faithful/unfaithful distinction, because an answer may be supported by one passage while ignoring the conflict introduced by the other. We therefore analyze behavior under contradictory evidence rather than treating the task as standard hallucination detection.

Final answers can mask disagreement among internal signals. The cross-level disagreement results show that internal signals and output behavior do not point to the same passage. This occurs not only for FABRICATION, but when the final answer commits to one of the two passages. Thus, the final answer does not capture how the model responds to competing evidence sources.

Sampling uncertainty under contradiction. Semantic entropy is useful, but for short-answer contradictions it is often less discriminative than attribution and hidden-state signals. Models may generate stable answers even when passages conflict; in these cases, sampled-output uncertainty can remain low, while internal signals show disagreement or sensitivity to competing passages.

Confidence alone is unreliable under contradiction. A finding is that fabricated answers are not always low-confidence. Across the four

models, FABRICATION shows higher confidence under contradiction in 46.5% of LLaMA-3-8B, 36.1% of Qwen2.5-7B, 34.2% of OLMo-2-7B, and 33.3% of LLaMA-3.1-70B cases. This means that confidence-based filtering would miss a fraction of unsupported answers. The result is important for RAG-style settings, where retrieved passages may conflict and a model may produce a fluent, high-confidence FABRICATION answer, i.e., an answer unsupported by either passage.

Pre-generation White-box signals provide diagnostic information. Across models, classifiers using pre-generation white-box signals outperform logprob-only baseline. This indicates that pre-generation white-box signals contain diagnostic information about output behavior that is not captured by log probabilities alone.

ConflictBank provides a complementary setting to WikiContradict: while WikiContradict contains human-authored contradictions without a single designated gold passage, ConflictBank is asymmetric and includes misinformation evidence. The consistent disagreement patterns across both datasets suggest that the proposed white-box signals capture conflict behavior beyond the specific WikiContradict setting.

6 Conclusion

We introduced a white-box tracing framework for analyzing how LLMs behave under contradictory evidence. Using WikiContradict, we compare single-passage baselines, C_{p1} and C_{p2} , with a dual-passage conflict condition, C_{dual} , across LLaMA-3-8B, LLaMA-3.1-70B, OLMo-2-7B, and Qwen2.5-7B. We label dual-passage answers as PASSAGE_A, PASSAGE_B, SYNTHESIS, or FABRICATION, and trace how these behaviors relate to hidden-state drift, probability-level signals, counterfactual attribution, and output-level uncertainty.

Three findings stand out. First, larger models reduce fabrication (LLaMA-3.1-70B: 6% vs. 14–17% for smaller models) but do not eliminate cross-level disagreement between internal signals and output behavior. Second, fabricated answers show higher confidence in 33–47% of cases, making confidence-based filtering unreliable for detecting answers unsupported by either passage under contradiction. Third, pre-generation white-box signals provide diagnostic information about output behavior: they improve 4-class output-behavior classification over the logprob-only baseline across

all four models. These results show that analyzing output behavior under contradictory evidence requires more than final answer analysis. It benefits from combining diagnostic generation-time analysis with pre-generation classification signals. A further evaluation on ConflictBank, a larger misinformation-conflict QA dataset, shows that cross-level disagreement persists beyond WikiContradict. Future work should evaluate whether these signals generalize to long-form, multi-hop, and naturally retrieved RAG settings.

7 Acknowledgments

This research was funded in whole or in part by the Austrian Science Fund (FWF): [COE12](#), [DFH23](#), [P36413](#); and by the State of Upper Austria and the Federal Ministry of Education, Science, and Research: LIT-2024-13-SEE-111.

8 Limitations

Our main experiments are conducted on WikiContradict, a controlled benchmark of short-answer question-answering instances with paired contradictory passages. We additionally evaluate on ConflictBank to test generalizability in a larger misinformation-conflict setting. However, both datasets remain controlled benchmarks, and the findings may not directly generalize to long-form generation, multi-hop reasoning, or naturally retrieved RAG settings where conflicts are less controlled. Our analysis is also limited to four open-weight instruction-tuned models. Although these models cover different families and sizes, additional models are needed to assess how broadly the observed signal patterns generalize. Finally, our experiments are intended to test whether white-box signals contain useful information about output behavior, not to claim a deployable detector for unsupported answers. Explaining the causes of model-specific differences, such as architecture, training data, or instruction tuning effects, is left for future work.

9 Ethical considerations

This work aims to analyze output behavior under contradictory evidence, not to provide a deployable safety detector. The proposed white-box signals should not be interpreted as guarantees of factual correctness. We use controlled benchmarks of contradictory and misinformation-conflict evidence, so

the findings may not fully generalize to noisier real-world RAG settings. Output labels are obtained with LLM-assisted annotation and manually validated on a stratified sample; although agreement is high, some labeling errors may remain. We used generative AI tools to assist with language polishing and clarity of presentation. All scientific content, experimental design, analysis, and conclusions were developed and verified by the authors.

References

- Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. 2024. Contextcite: attributing model generation to context. In *Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS ’24*.
- Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.
- Linyang He, Peili Chen, Ercong Nie, Yuanning Li, and Jonathan R. Brennan. 2024. *Decoding probing: Revealing internal linguistic structures in neural language models using minimal pairs*. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4488–4497, Torino, Italia. ELRA and ICCL.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. *DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing*. In *The Eleventh International Conference on Learning Representations*.
- Yufang Hou, Alessandra Pascale, Javier Carnerero-Cano, Tigran Tchrakian, Radu Marinescu, Elizabeth Daly, Inkit Padhi, and Prasanna Sattigeri. 2024. *Wikicontradict: A benchmark for evaluating llms on real-world knowledge conflicts from wikipedia*. In *Advances in Neural Information Processing Systems*, volume 37, pages 109701–109747. Curran Associates, Inc.
- Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2025. *Faitheval: Can your language model stay faithful to context, even if “the moon is made of marshmallows”*. In *The Thirteenth International Conference on Learning Representations*.
- Oscar Obeso, Andy Arditi, Javier Ferrando, Joshua Freeman, Cameron Holmes, and Neel Nanda. 2025. Real-time detection of hallucinated entities in long-form generation. *arXiv preprint arXiv:2509.03531*.
- Hadas Orgad, Michael Tokar, Zorik Gekhman, Roi Reichart, Idan Szepktor, Hadas Kotek, and Yonatan Belinkov. 2025. *LLMs know more than they show: On the intrinsic representation of LLM hallucinations*. In *The Thirteenth International Conference on Learning Representations*.
- Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024a. *Un-supervised real-time hallucination detection based on the internal states of large language models*. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14379–14391, Bangkok, Thailand. Association for Computational Linguistics.
- Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yan-shu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024b. *Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llms*. In *Advances in Neural Information Processing Systems*, volume 37, pages 103242–103268. Curran Associates, Inc.
- Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. 2025. *Benchmarking uncertainty quantification methods for large language models with LM-polygraph*. *Transactions of the Association for Computational Linguistics*, 13:220–248.
- Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. *Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts*. In *The Twelfth International Conference on Learning Representations*.
- Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. *Knowledge conflicts for LLMs: A survey*. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8541–8565, Miami, Florida, USA. Association for Computational Linguistics.

A Prompt Templates

A.1 Answer Generation Prompt

For each WikiContradict instance, we instantiate the answer-generation prompt separately under three evidence conditions: C_{p1} , C_{p2} , and C_{dual} .

Answer Generation Prompt: Single-Passage Condition C_{p1}

System: You are given a question and optionally one or two context passages. Return ONLY the shortest answer phrase. Do not explain. Do not repeat the question. Do not mention conflict. Do not use full sentences unless the answer itself is a sentence. Do not add prefixes like “Answer:”. Output only the answer text.

User:

Context: [PASSAGE_A]

Question: [QUESTION]

Answer:

Answer Generation Prompt: Dual-Passage Condition C_{dual}

System: You are given a question and optionally one or two context passages. Return *ONLY* the shortest answer phrase. Do not explain. Do not repeat the question. Do not mention conflict. Do not use full sentences unless the answer itself is a sentence. Do not add prefixes like “Answer:”. Output only the answer text.

User:

Context 1: [PASSAGE_A]
Context 2: [PASSAGE_B]
Question: [QUESTION]
Answer:

A.2 Output Labeling Prompt

Output Labeling Prompt

Task: Given a question, two conflicting reference answers, and a model-generated answer, assign exactly one label describing how the generated answer relates to the two reference answers.

Question: [QUESTION]

Answer 1: [ANSWER_1 from WikiContradict]

Answer 2: [ANSWER_2 from WikiContradict]

Model answer: [MODEL_ANSWER]

Labels:

- **Passage_A:** the model answer commits to Answer 1.
- **Passage_B:** the model answer commits to Answer 2.
- **Synthesis:** the model answer blends incompatible evidence from both answers.
- **Fabrication:** the model answer is unsupported by either answer.

Few-shot examples: The prompt includes eight examples, with two examples for each label.

Output format: Return only one label: *Passage_A*, *Passage_B*, *Synthesis*, or *Fabrication*.

B Output Labeling and Label Distribution

B.1 Output Label Definitions

Each C_{dual} answer is assigned one of four labels: *PASSAGE_A* if it follows the answer supported by p_1 , *PASSAGE_B* if it follows p_2 , *SYNTHESIS* refers to answers that combine incompatible information from both reference answers into a single response, such as returning a range, conjunction, or hybrid answer that is not itself supported by either passage alone. For example, if the two reference answers are “41” and “42”, an answer such as “41–42” would be labeled as *SYNTHESIS*. *FABRICATION* refers to an answer that is unsupported by either reference answer, such as returning “49” in the same example.

B.2 LLM-Assisted Labeling

We use GPT-4-turbo (gpt-4-turbo-2024-04-09) to assign output-behavior labels. It receives the question, the two conflicting reference answers from WikiContradict, and the answer generated under C_{dual} . We use an 8-shot in-context prompt, with two examples for each label. The prompt is shown in Appendix A.2.

B.3 Order-Swapped Dual-Passage Label Distribution

To examine passage-order sensitivity in output behavior, we evaluate an order-swapped dual-passage condition:

$$C_{\text{dual-swap}} = (q, p_2, p_1).$$

This condition uses the same two contradictory passages as C_{dual} , but reverses their order. We use this analysis only for output-label distributions. It is not used in the hidden-state, attribution, or other experiments. Figure 4 shows the output-label distributions.

B.4 Labeling Validation

We manually verified a stratified sample of 50 instances to validate the GPT-4-turbo labeling procedure.

The automatic labels matched the manually verified labels in 98.0% of the validation cases, with Cohen’s $\kappa = 0.965$, indicating almost perfect agreement. Only one disagreement was observed. The validation was performed by the authors; no external participants or annotators were recruited.

Labeling disagreement example. We observed one disagreement between GPT-4-turbo and the manually verified label:

- Question: Was the total distance of the 1920 Volta a Catalunya less than 615 km (382 mi)?
- WikiContradict answer 1: Yes
- WikiContradict answer 2: No
- C_{dual} generated answer: 607 km (377 mi)
- Labels: GPT-4-turbo = *FABRICATION*; manual = *PASSAGE_A*.

C Extended Experimental Results

C.1 Dual-Passage vs. Single-Passage Signal Comparison

For token-level signals, we report the mean value under C_{dual} and the corresponding single-passage reference condition. These means are descriptive

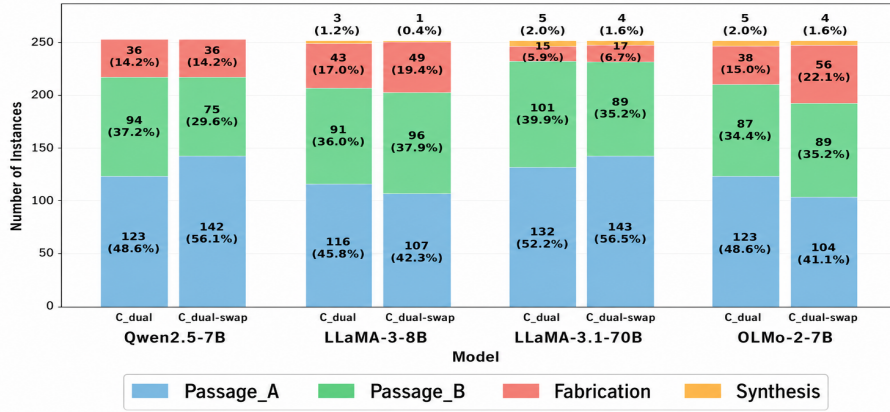


Figure 4: Comparison of output-label distributions for the dual-passage condition $C_{\text{dual}} = (q, p_1, p_2)$ and the order-swapped dual condition $C_{\text{dual-swap}} = (q, p_2, p_1)$. Each model has two stacked bars showing how reversing passage order changes generated answer labels. Across both conditions, LLaMA-3.1-70B produces fewer Fabrication cases than the smaller models.

summaries across instances. Statistical significance is assessed using Wilcoxon signed-rank tests over paired instance-level differences. Table 7 reports complete Wilcoxon signed-rank test results from Experiment 1.

D Reproducibility Details

D.1 Code and Data Availability

To support reproducibility, we provide a public repository:

<https://github.com/pelchafei/whitebox-contradictory-evidence/>.

D.2 Hyperparameters

Table 8 shows all settings and experimental hyperparameters.

Parameter	Value
Signal extraction	
do_sample	False
Sampling semantic entropy	
do_sample	True
num_return_sequences	10
temperature	0.7
top_p	0.9
NLI clustering	
Model	cross-encoder/ nli-deberta-v3-small
Threshold	0.7
Hidden states	
Qwen2.5-7B layers	{7, 14, 21}
LLaMA-3-8B layers	{8, 16, 24}
LLaMA-3.1-70B layers	{20, 40, 60}
OLMo-2-7B layers	{8, 16, 24}
Classifiers	
Logistic Regression	
Penalty	L2
max_iter	1000
class_weight	balanced
Gradient Boosted Trees	
n_estimators	100
random_state	42
Cross-validation	Stratified 5-fold
Feature scaling	StandardScaler

Table 8: Experimental hyperparameters.

D.3 Model Checkpoints

We use the following open-weight model versions:

- Qwen2.5-7B: Qwen/Qwen2.5-7B-Instruct
- LLaMA-3-8B: meta-llama/Meta-Llama-3-8B-Instruct
- LLaMA-3.1-70B: hugging-quants/Meta-Llama-3.1-70B-Instruct-AWQ-INT4
- OLMo-2-7B: allenai/OLMo-2-1124-7B-Instruct

Model	Signal	vs.	Mean C_{dual}	Mean Reference (C_{p1} or C_{p2})	Wilcoxon p_{adj}
Qwen2.5-7B	Token entropy	C_{p1}	0.230	0.186	0.002**
	Token entropy	C_{p2}	0.230	0.197	0.014*
	Token gap	C_{p1}	0.844	0.867	0.117 ns
	Token gap	C_{p2}	0.844	0.857	1.000 ns
	Mean logprob	C_{p1}	-0.110	-0.090	0.065 ns
	Mean logprob	C_{p2}	-0.110	-0.101	1.000 ns
	$\Delta \log P_A$	0	6.645	—	< 0.001***
	$\Delta \log P_B$	0	7.047	—	< 0.001***
LLaMA-3-8B	Token entropy	C_{p1}	0.438	0.404	0.322 ns
	Token entropy	C_{p2}	0.438	0.398	0.004**
	Token gap	C_{p1}	0.715	0.734	1.000 ns
	Token gap	C_{p2}	0.715	0.743	0.106 ns
	Mean logprob	C_{p1}	-0.210	-0.197	1.000 ns
	Mean logprob	C_{p2}	-0.210	-0.191	0.044*
	$\Delta \log P_A$	0	5.385	—	< 0.001***
	$\Delta \log P_B$	0	4.300	—	< 0.001***
LLaMA-3.1-70B	Token entropy	C_{p1}	0.651	0.576	0.007**
	Token entropy	C_{p2}	0.651	0.560	< 0.001***
	Token gap	C_{p1}	0.655	0.681	0.037*
	Token gap	C_{p2}	0.655	0.691	< 0.001***
	Mean logprob	C_{p1}	-0.279	-0.250	0.008**
	Mean logprob	C_{p2}	-0.279	-0.240	< 0.001***
	$\Delta \log P_A$	0	2.807	—	< 0.001***
	$\Delta \log P_B$	0	3.097	—	< 0.001***
OLMo-2-7B	Token entropy	C_{p1}	0.558	0.442	< 0.001***
	Token entropy	C_{p2}	0.558	0.463	< 0.001***
	Token gap	C_{p1}	0.736	0.783	< 0.001***
	Token gap	C_{p2}	0.736	0.781	< 0.001***
	Mean logprob	C_{p1}	-0.220	-0.171	< 0.001***
	Mean logprob	C_{p2}	-0.220	-0.173	< 0.001***
	$\Delta \log P_A$	0	5.669	—	< 0.001***
	$\Delta \log P_B$	0	3.777	—	< 0.001***

Table 7: Wilcoxon signed-rank results comparing C_{dual} with single-passage references for token-level signals, where Wilcoxon p_{adj} is computed from paired instance-level differences and Bonferroni-adjusted. $\Delta \log P_A$ and $\Delta \log P_B$ are tested against zero. Significance codes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ns = not significant. Mean columns are descriptive summaries across instances for the corresponding conditions and are not used directly to compute the Wilcoxon p_{adj} values.