Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Signed Layer-Level Evidence in Transformer Predictions, and a Diagnostic for When It Applies

Jeffery Opoku

jeffery.opoku01@utrgv.edu

University of Texas Rio grande valley

David Banahene

Florida International University

Research Article

Keywords:

Posted Date: August 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10705669/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Signed Layer-Level Evidence in Transformer Predictions, and a Diagnostic for When It Applies

Jeffery Opoku
School of Mathematical and Statistical Sciences
University of Texas Rio Grande Valley
opokujeffery5@gmail.com

David Banahene
Department of Biostatistics
Robert Stempel College of Public Health & Social Work
Florida International University
abanahene54@gmail.com

Abstract

Two tokens can carry the same softmax probability yet be reached very differently: one backed by near-unanimous layer agreement, the other surviving strong internal opposition. No output-level score separates them, because every such score is a function of the logit gap alone and is blind to *evidence mass*.

We introduce **Language Evidence Flow (LEF)**, which splits each token prediction into signed per-layer evidence in a single forward pass and, in retrieval-augmented generation, separates evidence coming from parametric memory from evidence coming from retrieved documents. We prove an identification boundary characterising exactly when output-level scores cannot reveal internal conflict.

Across eleven models and six architecture families (1.4B–14.7B), LEF detects RAG source disagreement on four of six families, but conflict as an *error* detector proves model-specific. That gap motivates **ScopeGate**, a permutation test asking whether a score is predictive of error on a given model–task pair. ScopeGate is detector-agnostic, so we ran it across detectors, scoring choices and models: eleven of thirty-six combinations fail. On Mistral-7B-Instruct *no* detector predicts error, predictive entropy included, and for conflict the first-token and mean-token verdicts disagree on three of nine models. Validity depends on the model, the detector and the scoring protocol together, and one AUROC number hides all three.

We report the least favourable comparison plainly: entropy outperforms conflict on single-hop factual QA in nine of eleven configurations. ScopeGate is calibrated, needs roughly 100 labelled examples, and is released as a standalone module.

Notation. L = layers; d = hidden dimension; V = vocabulary size; $r_t \in \mathbb{R}^d$ = residual stream at position t; $E_\ell(t)$ = signed evidence from layer ℓ; S^+, S^- = support and opposition mass; $C(t) = 2\min(S^+, S^-)/(S^+ + S^-)$ = conflict; $m_{\text{flip}}(t)$ = flip margin; $C_{\text{cross}}(t)$ = cross-source conflict (RAG).

1 Introduction

Large language models routinely produce fluent, high-probability outputs that are factually wrong, a failure mode termed *hallucination* [Ji et al., 2023, Huang et al., 2023]. The core difficulty is that a token's softmax probability reflects the *outcome* of internal computation but not the *process*: two tokens with identical probabilities may differ sharply in how many layers supported versus opposed them [Kadavath et al., 2022, Xiong et al., 2024].

Existing detection methods have made substantial progress. Sampling-based approaches (SelfCheckGPT [Manakul et al., 2023], semantic entropy [Kuhn et al., 2023]) check cross-sample consistency; internal-state methods (INSIDE [Chen et al., 2024], probing classifiers [Azaria and Mitchell, 2023]) train detectors on hidden representations. However, sampling methods require 5–20× the inference cost, and probe-based methods require labeled correct/hallucinated examples for each model–task pair. Crucially, none of these approaches decompose a prediction into signed per-layer evidence or separate parametric memory from retrieval evidence in RAG.

We address this gap with **Language Evidence Flow (LEF)**, which adapts the Signed Evidence Flow framework (SEF; Opoku and Banahene 2026b), first built for tabular classifiers, to autoregressive transformers. The key idea is that the residual stream is additive: each layer's contribution to the logit gap between the top two candidates can be computed exactly, yielding a signed evidence term that is positive (supporting) or negative (opposing). LEF measures the balance between these terms, the *conflict*, and links it to prediction fragility via the flip margin.

1.1 Motivating example

A RAG system is asked: "When was the Treaty of Westphalia signed?" The retrieved document correctly says 1648. But the model also learned the wrong date 1658 from noisy web text during training. The final score for "1648" reflects *both* sources. If memory and retrieval agree, the evidence is one-sided and the answer is likely correct. If they disagree, the score may still be high, but the evidence inside the model is fighting itself. LEF makes this internal fight visible and measurable.

1.2 What LEF reveals that confidence cannot

Consider two tokens generated by Mistral-7B on TriviaQA, both with the same softmax probability $p \approx 0.72$:

- **Token A** (correct answer): 28 of 32 layers support the token, 4 weakly oppose. LEF conflict $C = 0.12$. The evidence is one-sided. Flip margin is large, so the model would need substantial evidence redistribution to change its mind.
- **Token B** (hallucinated answer): 17 layers support the token, 15 strongly oppose. LEF conflict $C = 0.87$. The evidence is nearly split. Flip margin is tiny, so a small perturbation in any layer flips the output.

Entropy, probability, and calibrated confidence give *identical* scores for both tokens (by construction, since both share $p \approx 0.72$). LEF distinguishes them because it accesses the layer-level evidence structure rather than the final output distribution alone. This is the core insight: **confidence measures the *outcome* of internal computation; conflict measures the *process*.** Figure 1 summarizes the method.

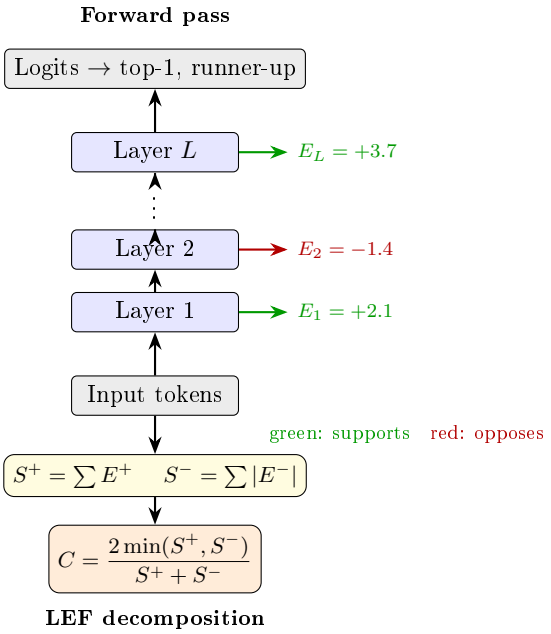


Figure 1: Method overview. LEF intercepts per-layer residual-stream updates during a single forward pass, projects each through the unembedding matrix to obtain signed evidence E_ℓ for the top-1 vs. runner-up logit gap, and aggregates into support (S^+), opposition (S^-), and conflict (C). High C indicates internal disagreement even when the output appears confident.

1.3 Contributions

1. **Framework.** We define Language Evidence Flow, which decomposes each token's prediction into signed contributions from every transformer layer, every attention head, and, for RAG, from the model's memory versus the retrieved text, all in a single forward pass.

2. **Theory.** We prove three results: (a) an identification theorem showing exactly when confidence can and cannot reveal conflict; (b) a closed-form link between conflict and the flip margin; (c) a propagation theorem giving conditions under which conflict at one token raises expected evidence mass at the next. For RAG, we prove a lower bound on conflict whenever memory and retrieval disagree.

3. **Experiments.** We evaluate on eleven models spanning four architecture families (1.4B–14.7B), including both base and instruction-tuned variants, across eight tasks plus a scale-and-architecture sensitivity analysis. The key findings: (a) *LEF detects RAG source disagreement in a single forward pass on all three main models*; (b) ScopeGate correctly identifies the four deployment-ready configurations out of eleven models; (c) LEF conflict (training-free) nearly matches a trained hidden-state probe (AUROC 0.660 vs. 0.669) and outperforms SelfCheckGPT-style sampling (0.640) at 5× fewer forward passes; (d) LEF effectiveness varies by model in a way we could not explain. Attention type does not predict it, and a direct test of layer-evidence diversity failed to either; and (e) LEF survives instruction tuning across scales (7B and 14B), with conflict becoming the dominant detection signal on multi-hop reasoning (AUROC up to 0.879).

4. **Deployment protocol.** We provide a four-step recipe for using LEF in production, including the ScopeGate safety check that prevents overreliance by warning when conflict is not predictive

of errors. ScopeGate passes on exactly four of eleven models tested (all three Qwen2.5 variants plus OLMo-2-7B-Instruct), showing precise discrimination across two model families.

5. **A reusable check for the field.** ScopeGate works on *any* detector. It takes a score for each example and whether the model was right, and never looks at how the score was made. So any hallucination detector can be checked with it before deployment. Section 7.11 shows this by running the same test on predictive entropy, and measures how the check behaves: it fires on noise about 5% of the time, as it should, and needs roughly 100 labelled examples to find a clear signal, 250 for a weak one. Users can budget for it instead of just citing it. We think it should be reported next to AUROC. Right now detectors get deployed without anyone checking that they still point the right way on the target model.

2 Related Work

2.1 Hallucination in large language models

Hallucination takes two forms: *intrinsic* (the output contradicts the source) and *extrinsic* (the output cannot be checked against any source) [Ji et al., 2023, Huang et al., 2023, Zhang et al., 2023]. Wrong facts, unfaithful summaries, and made-up citations all count. Our work asks a question that applies to all of these: is the internal evidence for a word one-sided, or does it pull in opposite directions?

2.2 Detection methods

Sampling-based methods. SelfCheckGPT [Manakul et al., 2023] generates several completions and checks whether they agree. Kuhn et al. [2023] group sampled answers by meaning and measure how spread out the groups are. P(true) [Kadavath et al., 2022] asks the model to judge its own answer. These methods need at least five forward passes and work on finished text. LEF works in a single forward pass and scores each word *as it is being produced*.

Internal-state methods. INSIDE [Chen et al., 2024] trains probes on the model's hidden layers to detect hallucination. Azaria and Mitchell [2023] train linear classifiers on hidden-state activations to predict statement truth. ITI [Li et al., 2023] finds "truthful" attention heads and changes them at test time. The logit lens [nostalgebraist, 2020] and tuned lens [Belrose et al., 2023] project hidden states into word space at each layer. All probe-based methods require labeled training data (correct/hallucinated examples) and learn dataset-specific decision boundaries. LEF uses a similar internal-state projection but treats the result as *signed evidence*: each layer either supports or opposes the top word, and we measure the balance. Because LEF's conflict score is defined algebraically (not learned), it requires no training data, no fine-tuning, and generalizes across tasks without retraining.

Layer-wise hallucination analysis. Several recent works analyze transformer layers for hallucination signals. Kim et al. [2025] introduce *Layer-wise Information Deficiency* ($\mathcal{LI}$), tracking mutual-information loss across layers; their method detects when information is *lost* during processing but does not decompose evidence into support versus opposition, so it cannot produce a conflict score or flip margin. LRP4RAG [Hu et al., 2024] applies Layer-wise Relevance Propagation to RAG hallucination detection, achieving 77% accuracy; however, it requires a backward pass and trains classifiers on the relevance matrix, whereas LEF is forward-pass only and training-free. Chuang

et al. [2024] study *Layer-wise Semantic Dynamics*, showing hallucinated tokens exhibit oscillatory hidden-state trajectories; their geometric analysis is complementary to LEF's algebraic decomposition but does not yield a per-token scalar score suitable for thresholding. ReDeEP [Xie et al., 2024] detects RAG hallucination via attention circuit analysis, Wu et al. [2025] mechanistically study how knowledge conflicts arise across layers, Su et al. [2024] apply Fourier analysis to temporal hidden-state signals for hallucination scoring, and Chen et al. [2025] extend the logit lens with contextual embeddings for vision–language hallucination grounding. LEF differs from all of the above in three ways. It splits predictions into *signed* per-layer contributions, so support and opposition stay separate. It defines a conflict score with provable properties: the identification boundary, the flip margin, and propagation. And it needs no backward pass, no training, and no change to the architecture.

Uncertainty measures. Token probability, predictive entropy, and calibrated confidence are widely used [Guo et al., 2017, Desai and Durrett, 2020]. These scores are not guaranteed to tell you whether a high-probability word is backed by layers that all agree, or by layers that strongly disagree but happen to cancel out. Theorem 7 characterizes exactly when this gap exists.

2.3 Attribution and interpretability in transformers

Attention-based methods [Jain and Wallace, 2019, Wiegreffe and Pinter, 2019], gradient-based methods [Sundararajan et al., 2017], and circuit-level analysis [Elhage et al., 2021, Conmy et al., 2023] all study which parts of the model drive a prediction. LEF does not replace these tools. It asks a follow-up question: do the different parts of the model agree, or do they fight each other?

2.4 Signed Evidence Flow

SEF [Opoku and Banahene, 2026b] measures support, opposition, conflict, flip margin, stability, and reliability for tabular predictions. It proves that confidence alone cannot reveal conflict, that conflict controls how easily a prediction can flip, and that you must check whether conflict predicts errors before trusting it. LEF extends this framework to the sequential, multi-source setting of language models.

2.5 Conformal and calibration perspective

LEF explains *why* a word might be hallucinated. A different line of work asks *when to trust* outputs using conformal prediction, which gives coverage guarantees. For foundation models specifically, PromptShift-CRC [Opoku and Banahene, 2026a] provides risk control under prompt and domain drift, and ToolChain-CRC [Opoku and Banahene, 2026c] extends guarantees to AI agents that use retrieval and tools. These supply calibrated guarantees that a diagnostic like LEF deliberately does not; the two are complementary rather than competing.

LEF and conformal methods solve different halves of the same problem. LEF tells you *why* a word is risky (the evidence disagrees). Conformal methods tell you *when to abstain* (with a coverage guarantee). The two work together naturally: ToolChain-CRC checks whether retrieval is reliable at the trajectory level, and LEF checks whether the model's response to the retrieved evidence is internally consistent at the word level.

2.6 Summary comparison

Table 1 positions LEF against existing hallucination detection methods on six capability dimensions. LEF is the only method that provides all six: single-pass efficiency, per-token granularity, layer-level attribution, RAG source separation, theoretical guarantees, and a built-in deployment check.

Table 1: Capability comparison of hallucination detection methods. ✓ = yes, ✗ = no, ∼ = partial. Note: several columns (RAG source sep., deploy. check) reflect capabilities that LEF defines; the comparison should be read as showing what each method *provides*, not as a uniform evaluation on the same axis.

Method	*Single pass*	*Per-token*	*Layer attrib.*	*RAG source sep.*	*Theory*	*Deploy. check*
Token probability	✓	✓	✗	✗	✗	✗
Predictive entropy	✓	✓	✗	✗	✗	✗
SelfCheckGPT	✗	∼	✗	✗	✗	✗
Semantic entropy	✗	✗	✗	✗	✗	✗
P(true)	✗	✗	✗	✗	✗	✗
INSIDE	✓	∼	∼	✗	✗	✗
Probing classifiers	✓	∼	∼	✗	✗	✗
Logit lens	✓	✓	✓	✗	✗	✗
$\mathcal{L}I$ (Kim et al.)	✓	✓	✓	✗	∼	✗
LRP4RAG	✗	✓	✓	∼	✗	✗
LSD (Geom. of Truth)	✓	✓	✓	✗	✗	✗
ReDeEP	✓	∼	∼	∼	✗	✗
LEF (ours)	✓	✓	✓	✓	✓	✓

3 Language Evidence Flow

3.1 Setup

Let $\mathcal{V}$ be the vocabulary (all possible words or word pieces) with size V. At each position t, a language model produces a score vector

$$z_t = (z_t^{(1)}, \ldots, z_t^{(V)}) \in \mathbb{R}^V$$

and picks the highest-scoring word:

$$\hat{v}_t = \arg\max_{v \in \mathcal{V}} z_t^{(v)}.$$

The model's confidence in its choice is the softmax probability:

$$p_t = \frac{\exp(z_t^{(\hat{v}_t)})}{\sum_{v=1}^{V} \exp(z_t^{(v)})}.$$

We focus on the *gap* between the top-scoring word and its closest competitor:

$$\Delta_t = z_t^{(\hat{v}_t)} - \max_{v \neq \hat{v}_t} z_t^{(v)}.$$

A large gap means the model strongly prefers its top choice. A small gap means the race was close. LEF asks: *how did the evidence inside the model produce this gap?*

3.2 Transformer as an additive evidence machine

A transformer with L layers computes the score as

$$z_t = W_U \left(x_t^{(0)} + \sum_{\ell=1}^{L} a_t^{(\ell)} + \sum_{\ell=1}^{L} m_t^{(\ell)} \right) + b,$$

where $x_t^{(0)}$ is the word embedding plus position encoding, $a_t^{(\ell)}$ is the attention output at layer ℓ, $m_t^{(\ell)}$ is the MLP output at layer ℓ, and W_U is the unembedding matrix that maps hidden states back to word scores. Because the residual stream is additive (each layer simply adds its contribution), the score for any word v breaks down exactly as

$$z_t^{(v)} = \underbrace{(W_U x_t^{(0)})_v + b_v}_{\text{embedding baseline}} + \sum_{\ell=1}^{L} \underbrace{(W_U a_t^{(\ell)})_v}_{\text{attention evidence at layer } \ell} + \sum_{\ell=1}^{L} \underbrace{(W_U m_t^{(\ell)})_v}_{\text{MLP evidence at layer } \ell}. \tag{1}$$

This identity holds exactly for the *pre-normalization* logit: $\tilde{z}_t = W_U\, r_t + b$. In practice, transformers apply a final RMSNorm before unembedding: $z_t = W_U\, \mathrm{RMSNorm}(r_t) + b$. Proposition 1 below shows that this does not affect the conflict score. The decomposition is therefore *exact for all LEF quantities that depend only on the sign and ratio structure of layer evidence*, which includes conflict, flip margin, and cross-source conflict. This separates LEF from attribution methods that rest on approximations, such as gradients, attention rollout, or SHAP. Two terms are the only source of inexactness: the RMSNorm positive scalar, which cancels in every ratio-based score, and the learned scale vector γ, which we approximate as $\mathbf{1}$ (see the Remark below).

Proposition 1 (RMSNorm invariance of LEF conflict)**.** *Let $\kappa = \|r_t\|_{\mathrm{rms}}^{-1} > 0$ be the RMSNorm scaling factor and let $\gamma \in \mathbb{R}^d$ be the learned scale vector. Define $W_U' = W_U\, \mathrm{diag}(\gamma)$. Then:*

(i) The true logit gap satisfies $\Delta_t = \kappa \cdot \tilde{\Delta}_t^{\gamma}$, where $\tilde{\Delta}_t^{\gamma}$ is the gap computed with W_U' on the un-normalized residual stream.

(ii) Since $\kappa > 0$, the signs of all layer evidence terms are preserved: $\mathrm{sign}(E_\ell) = \mathrm{sign}(\kappa\, E_\ell)$.

(iii) The conflict score $C(t) = 2\min(S^+, S^-)/(S^+ + S^-)$ is invariant under positive rescaling, so LEF conflict computed from the raw residual stream equals LEF conflict computed from the normalized stream.

Proof. (i) RMSNorm applies $\kappa \cdot \mathrm{diag}(\gamma)$ to each vector, giving $z_t = \kappa\, W_U'\, r_t + b$. The gap is the difference of two logits, so the bias cancels and $\Delta_t = \kappa\,[(W_U' r_t)_{\hat{v}} - (W_U' r_t)_{\hat{r}}]$. (ii) Multiplying all layer evidence by $\kappa > 0$ preserves signs. (iii) $S^+ \to \kappa S^+$, $S^- \to \kappa S^-$, and κ cancels in the ratio. □

Remark 2. *In our implementation we use the raw W_U (without γ) for computational efficiency. This introduces a second approximation beyond the RMSNorm scalar: the per-dimension rescaling by γ. Empirically, γ is close to $\mathbf{1}$ after training (median $|\gamma_i - 1| < 5\%$ for Phi-2 and Mistral-7B;*

Qwen2.5-7B uses the same RMSNorm design). We verify the combined approximation in Section 7 via the sanity check: the sum of layer evidence matches the pre-norm logit gap to within $< 0.5\%$ relative error on all three models.

Definition 3 (Layer-level signed evidence)**.** *For the selected token $\hat{v}_t$ and its strongest rival $\hat{r}_t = \arg\max_{v \neq \hat{v}_t} z_t^{(v)}$, define the signed evidence from layer ℓ as*

$$E_\ell(t) = \left[(W_U a_t^{(\ell)})_{\hat{v}_t} + (W_U m_t^{(\ell)})_{\hat{v}_t}\right] - \left[(W_U a_t^{(\ell)})_{\hat{r}_t} + (W_U m_t^{(\ell)})_{\hat{r}_t}\right].$$

If $E_\ell(t) > 0$, layer ℓ supports the chosen word. If $E_\ell(t) < 0$, it favors the runner-up. The gap decomposes exactly:

$$\Delta_t = E_0(t) + \sum_{\ell=1}^{L} E_\ell(t), \tag{2}$$

where $E_0(t)$ is the embedding baseline contribution to the gap.

3.3 Head-level evidence

The attention output at layer ℓ is the sum of H attention heads:

$$a_t^{(\ell)} = \sum_{h=1}^{H} a_t^{(\ell,h)}.$$

So the attention contribution to the gap can be broken down further into individual heads:

$$(W_U a_t^{(\ell)})_{\hat{v}_t} - (W_U a_t^{(\ell)})_{\hat{r}_t} = \sum_{h=1}^{H} \left[(W_U a_t^{(\ell,h)})_{\hat{v}_t} - (W_U a_t^{(\ell,h)})_{\hat{r}_t}\right].$$

Definition 4 (Head-level signed evidence)**.** *The signed evidence from attention head (ℓ, h) is*

$$E_{\ell,h}^{\text{attn}}(t) = (W_U a_t^{(\ell,h)})_{\hat{v}_t} - (W_U a_t^{(\ell,h)})_{\hat{r}_t}.$$

This gives $L \times H$ attention terms plus L MLP terms plus an embedding baseline: $L(H+1)+1$ signed evidence terms per word. For a 32-layer, 32-head model, that is 1,057 terms per word.

3.4 Token-context evidence

Each attention head looks at earlier words in the input and computes a weighted sum. For head (ℓ, h):

$$a_t^{(\ell,h)} = \sum_{s=1}^{t} \alpha_{t,s}^{(\ell,h)} \, W_V^{(\ell,h)} x_s^{(\ell-1)},$$

where $\alpha_{t,s}^{(\ell,h)}$ is the attention weight from position t to context position s. The signed evidence from context token s through head (ℓ, h) is

$$E_{\ell,h,s}(t) = \alpha_{t,s}^{(\ell,h)} \left[(W_U W_V^{(\ell,h)} x_s^{(\ell-1)})_{\hat{v}_t} - (W_U W_V^{(\ell,h)} x_s^{(\ell-1)})_{\hat{r}_t}\right].$$

This tells us which earlier words, through which heads, push for or against the model's next-word choice.

3.5 RAG evidence decomposition

In RAG, the input has two parts: the *parametric prefix* $\mathcal{P}$ (prompt and question) and the *retrieved context* $\mathcal{R}$ (external documents). We split evidence into these two sources:

$$E^{\text{param}}(t) = \sum_{\ell,h} \sum_{s \in \mathcal{P}} E_{\ell,h,s}(t), \tag{3}$$

$$E^{\text{retr}}(t) = \sum_{\ell,h} \sum_{s \in \mathcal{R}} E_{\ell,h,s}(t). \tag{4}$$

MLP layers store learned knowledge, so we count them as parametric:

$$E^{\text{param}}_{\text{full}}(t) = E^{\text{param}}(t) + \sum_{\ell=1}^{L} \big[(W_U m_t^{(\ell)})_{\hat{v}_t} - (W_U m_t^{(\ell)})_{\hat{r}_t} \big]. \tag{5}$$

Definition 5 (Cross-source conflict)**.** *Let $S^+_{\text{param}}(t)$ and $S^-_{\text{param}}(t)$ be the support and opposition from parametric evidence, and let $S^+_{\text{retr}}(t)$ and $S^-_{\text{retr}}(t)$ be the support and opposition from retrieved evidence. The* cross-source conflict *is*

$$C_{\text{cross}}(t) = \frac{2 \min\{S^+_{\text{param}}(t) + S^-_{\text{retr}}(t),\ S^-_{\text{param}}(t) + S^+_{\text{retr}}(t)\}}{M(t) + \varepsilon},$$

where $M(t) = S^+(t) + S^-(t)$ is the total evidence mass.

Cross-source conflict is high when the model's memory pushes for one word and the retrieved documents push for a different one. This is exactly the RAG hallucination scenario.

3.6 Decision-aligned decomposition

All LEF quantities (support, opposition, conflict, flip margin) are computed from the gap between the top choice and the runner-up (Equation (2)). Conflict therefore measures disagreement about the model's actual decision, not about any individual word's raw score.

4 Theoretical Results

All algebraic results from SEF carry over because the layer evidence terms $E_\ell(t)$ are signed numbers that sum to Δ_t.

4.1 Token-level support, opposition, and conflict

Definition 6 (Token-level LEF quantities)**.** *Given the signed evidence decomposition $\Delta_t = \sum_{\ell=0}^{L} E_\ell(t)$, define*

$$S^+(t) = \sum_{\ell=0}^{L} \max\{E_\ell(t), 0\}, \qquad S^-(t) = \sum_{\ell=0}^{L} \max\{-E_\ell(t), 0\},$$

$$M(t) = S^+(t) + S^-(t), \qquad N(t) = S^+(t) - S^-(t) = \Delta_t,$$

$$C(t) = \frac{2 \min\{S^+(t), S^-(t)\}}{M(t) + \varepsilon}.$$

All the basic SEF results (conflict-direction identity, hidden evidence mass, conflict as distance from one-sided evidence) apply directly.

4.2 Identification theorem for language models

Theorem 7 (LEF identification boundary)**.** *Let Δ, M, and C be the logit gap, evidence mass, and conflict treated as random variables over token positions. Assume $M > 0$ almost surely (we set $\varepsilon = 0$ throughout the theoretical analysis; implementations use small $\varepsilon > 0$ for numerical stability, which does not affect any result when $M \gg \varepsilon$). Let Q be any function of the softmax output (e.g., the top-token probability or the entropy of the softmax distribution).*

(i) Conflict is determined by Q if and only if the normalized net evidence $|\Delta|/M$ is determined by Q. Moreover, $\mathrm{Var}(C \mid Q) = \mathrm{Var}(|\Delta|/M \mid Q)$.

(ii) If $|\Delta|$ is determined by Q, then conflict is determined by Q if and only if evidence mass M is determined by Q, and $\mathrm{Var}(C \mid Q) = |\Delta|^2 \mathrm{Var}(1/M \mid Q)$.

Proof. From the definitions, $S^+ - S^- = \Delta$ and $S^+ + S^- = M$, giving $S^+ = (M + \Delta)/2$ and $S^- = (M - \Delta)/2$. Therefore

$$C = \frac{2\min(S^+, S^-)}{M} = \frac{M - |\Delta|}{M} = 1 - \frac{|\Delta|}{M}.$$

This is the fundamental identity linking conflict, gap, and mass.

(i) Since $C = 1 - |\Delta|/M$ is an affine bijection of $|\Delta|/M$, knowing C is equivalent to knowing $|\Delta|/M$. If Q determines one, it determines the other, and $\mathrm{Var}(C \mid Q) = \mathrm{Var}(|\Delta|/M \mid Q)$ (the sign flip and constant shift do not change conditional variance).

(ii) If $|\Delta|$ is a known constant given Q, then $C = 1 - |\Delta|/M$ is determined by Q if and only if M is. Writing $C = 1 - |\Delta| \cdot (1/M)$, we get $\mathrm{Var}(C \mid Q) = |\Delta|^2 \,\mathrm{Var}(1/M \mid Q)$. □

Remark 8. *The theorem's practical meaning: two tokens can share the same softmax probability (same Q) yet have different conflicts, because Q does not determine evidence mass M. This is a* possibility *result, not a distributional claim: it shows that output-level scores are not* guaranteed *to reveal conflict, because evidence mass is a free parameter. Whether output-level scores* correlate *with conflict on a given data distribution is an empirical question (Section 7.3).*

Remark 9. *With only two words in the vocabulary, confidence rises with $|\Delta|$, so conflict is determined by confidence if and only if evidence mass is too. With a full vocabulary the key question is the same: does evidence mass vary among words that share the same confidence?*

4.3 Vocabulary flip margin

Theorem 10 (LEF flip margin)**.** *Assume $M(t) > 0$. The smallest total amount of signed evidence that must be redistributed from supporting layers to opposing layers to make the logit gap Δ_t equal zero is*

$$m_{\text{flip}}(t) = \frac{|\Delta_t|}{2} = \frac{M(t)\big(1 - C(t)\big)}{2}.$$

Proof. Without loss of generality assume $\Delta_t > 0$ (the model chose correctly and support exceeds opposition). We have $S^+ > S^-$ and $\Delta_t = S^+ - S^-$. To make $\Delta_t = 0$, we need $S^+ = S^-$. If we transfer amount δ from support to opposition, the new values are $S^{+\prime} = S^+ - \delta$ and $S^{-\prime} = S^- + \delta$. Setting $S^{+\prime} = S^{-\prime}$ gives $S^+ - \delta = S^- + \delta$, so $\delta = (S^+ - S^-)/2 = \Delta_t/2$.

Substituting $\Delta_t = M(1 - C)$ from the identity in Theorem 7's proof: $m_{\text{flip}} = M(1 - C)/2$. □

Remark 11. *The flip margin connects structural evidence conflict to prediction fragility: a token with $C = 0.9$ requires only 5% of its evidence mass to be redistributed before the model switches words, while a token with $C = 0.1$ requires 45%. This nine-fold difference in robustness is invisible to any score based solely on the output distribution.*

High conflict means a small flip margin: less evidence needs to change before the model switches words. This is how conflicting evidence produces fragile outputs.

4.4 Multi-token conflict propagation

The following result is new to LEF and does not appear in tabular SEF.

Theorem 12 (Conflict propagation under autoregressive decoding)**.** *Let $C(t)$ be the LEF conflict at position t, and let $C(t+1)$ be the conflict at position $t+1$ after the model conditions on its own output at position t. Let $r_t^{(\ell)}$ denote the residual-stream hidden state at position t after layer ℓ. Suppose:*

(a) ***(Residual-stream propagation.)*** *At position $t+1$, at least one attention head at layer ℓ^* attends to position t with weight $\alpha_{t+1,t}^{(\ell^*,h^*)} \geq \alpha_{\min} > 0$.*

(b) ***(Mixed hidden state.)*** *The hidden state $r_t^{(\ell^*-1)}$ at position t carries evidence for* both *the winner $\hat{v}_t$ and the runner-up $\hat{r}_t$. Formally: let $e^+ = (W_U r_t^{(\ell^*-1)})_{\hat{v}_t}$ and $e^- = (W_U r_t^{(\ell^*-1)})_{\hat{r}_t}$; we require $\min(e^+, e^-) \geq \beta > 0$ for some constant β.*

(c) ***(Mass–conflict link.)*** *For the data distribution, $\mathbb{E}[M(t+1) \mid r_t^{(\ell^*-1)}]$ is nondecreasing in $\|r_t^{(\ell^*-1)}\|$.*

Then, conditional on high conflict at position t:

$$\mathbb{E}[M(t+1) \mid C(t) \geq c] \geq \mathbb{E}[M(t+1) \mid C(t) < c]$$

for any fixed $c \in (0, 1)$.

Under the additional assumption that the conditional error probability $\Pr[\text{error at } t{+}1 \mid M(t{+}1) = m]$ *is nondecreasing in m, conflict at position t raises expected error at position $t{+}1$.*

Proof. (Sketch; full proof in Appendix A.) The key mechanism is not the token embedding (which is the same word regardless of how close the competition was), but the *hidden state* $r_t^{(\ell^*-1)}$ in the residual stream, which carries both supporting and opposing evidence when conflict is high. Attention heads at position $t{+}1$ read this conflicted hidden state through the key–value cache. By assumption (b), the hidden state encodes both candidates with magnitude $\geq \beta$, contributing mixed evidence at position $t{+}1$ that raises $M(t{+}1)$ without proportionally raising $|\Delta_{t+1}|$. □

Remark 13. *The theorem identifies a* sufficient *mechanism for conflict propagation: when the assumptions hold, conflict at one position raises expected evidence mass at the next. The assumptions are strong, particularly (c), which requires evidence mass to grow with hidden-state norm, and may not hold uniformly across architectures. Our empirical cascade ratios (1.09×, 1.00×, 1.03×; Table 6) are close to unity, indicating that the snowball effect, if present, is weak in the models tested. We therefore treat Theorem 12 as identifying the correct* mechanism *for propagation rather than predicting a strong empirical effect at current model scales.*

4.5 RAG conflict bound

Theorem 14 (RAG source disagreement bound)**.** *Let the evidence decomposition be refined to the head–position level: for each head (ℓ, h) and context position s, let $E_{\ell,h,s}(t)$ be the signed evidence. Define the head–position–level support and opposition:*

$$S^+_{\text{fine}}(t) = \sum_{\ell,h,s} \max\{E_{\ell,h,s}(t), 0\}, \quad S^-_{\text{fine}}(t) = \sum_{\ell,h,s} \max\{-E_{\ell,h,s}(t), 0\},$$

and the corresponding fine-grained conflict $C_{\text{fine}}(t) = 2\min(S^+_{\text{fine}}, S^-_{\text{fine}})/(S^+_{\text{fine}} + S^-_{\text{fine}} + \varepsilon)$. Let $E^{\text{retr}}(t) = \sum_{\ell,h} \sum_{s\in\mathcal{R}} E_{\ell,h,s}(t)$ and $E^{\text{param}}_{\text{full}}(t)$ include MLP evidence. Then:

$$C_{\text{fine}}(t) \geq \frac{2\min\{|E^{\text{param}}_{\text{full}}(t)|, |E^{\text{retr}}(t)|\}}{M_{\text{fine}}(t) + \varepsilon} \cdot \mathbf{1}\big[\text{sign}(E^{\text{param}}_{\text{full}}(t)) \neq \text{sign}(E^{\text{retr}}(t))\big].$$

Moreover, the coarse layer-level conflict satisfies $C(t) \leq C_{\text{fine}}(t)$ (since aggregation can only cancel opposing terms within a layer). The bound therefore guarantees a minimum fine-grained conflict when sources disagree; the coarse layer-level conflict used in practice may be smaller, but our experiments confirm that it also increases when documents conflict (Section 7.2).

Proof. Assume without loss of generality that $E^{\text{param}}_{\text{full}}(t) > 0$ and $E^{\text{retr}}(t) < 0$ (the sources disagree). Then $|E^{\text{retr}}(t)| = -E^{\text{retr}}(t)$ is the total magnitude of evidence from retrieval opposing the chosen word. Since $S^-_{\text{fine}}(t) \geq |E^{\text{retr}}(t)|$ (opposing retrieval terms contribute at least this much to fine-grained opposition) and similarly $S^+_{\text{fine}}(t) \geq |E^{\text{param}}_{\text{full}}(t)|$, we have $\min(S^+_{\text{fine}}, S^-_{\text{fine}}) \geq \min(|E^{\text{param}}_{\text{full}}|, |E^{\text{retr}}|)$. Dividing by M_{fine} and multiplying by 2 gives the bound.

For the layer-level comparison: grouping fine-grained terms by layer, cancellations within a layer can reduce S^- (opposing evidence can be absorbed by positive evidence at the same layer before the layer total is computed). So $S^-_{\text{fine}} \geq S^-$ and $C_{\text{fine}}(t) \geq C(t)$, making the fine-grained bound conservative for the coarser layer-level conflict. □

Corollary 15 (Practical implication)**.** *If the model's parametric memory and the retrieved documents disagree on which word to produce (i.e., their aggregate evidences have opposite signs), the fine-grained conflict is at least proportional to the strength of the weaker source. Monitoring $C_{\text{cross}}(t)$ (Definition 5) captures this disagreement directly.*

This theorem says something practical: when retrieval disagrees with memory, total conflict must be at least as large as the disagreement between the two sources. A system that watches $C_{\text{cross}}(t)$ can flag these conflicts before the word is written.

5 Hallucination Risk Scores and Screening

The LEF decomposition yields several risk scores at different granularities. Each score quantifies a specific type of internal disagreement; none should be interpreted as a binary hallucination label without the ScopeGate calibration check.

Token-level conflict. $C(t) \in [0, 1]$ measures the balance between supporting and opposing layer evidence for the selected token. Values near zero indicate one-sided evidence; values near one indicate that the token survives despite substantial internal opposition.

Span and sentence conflict. For a generated span $t_1, \ldots, t_k$, define $C_{\text{span}} = k^{-1} \sum_{i=1}^{k} C(t_i)$ and $C_{\text{span}}^{\max} = \max_{1 \le i \le k} C(t_i)$. The mean captures sustained disagreement across a phrase; the maximum catches a single fragile token within an otherwise stable span.

Cross-source conflict in RAG. $C_{\text{cross}}(t)$ (Definition 5) isolates disagreement between parametric and retrieval evidence. High C_{cross} with low C indicates that sources disagreed but final layers resolved the tension; high values of both indicate unresolved disagreement and a fragile prediction.

Definition 16 (Layer divergence). *Let $e_\ell = E_\ell(t)/(|E_\ell(t)| + \varepsilon)$ be the sign-normalized evidence at layer ℓ. The layer divergence score is*

$$D_{\text{layer}}(t) = 1 - \frac{1}{L-1} \sum_{\ell=2}^{L} \mathbf{1}[e_\ell = e_{\ell-1}].$$

Layer divergence captures oscillatory evidence paths: a token can have low aggregate conflict yet pass through alternating support and opposition across consecutive layers, indicating an unstable computation.

ScopeGate deployment diagnostic. ScopeGate tests whether conflict correlates with error on a held-out calibration set via permutation test ($n \ge 200$, 5,000 shuffles). If the test passes ($p < 0.05$), conflict is validated as a standalone signal for the given model–task pair. If it fails, LEF retains value for source attribution and ensemble use but should not drive automated decisions alone.

Conformal screening. The conflict and layer-divergence scores are compatible with split-conformal thresholding under exchangeability, inheriting the coverage guarantees from SEF [Opoku and Banahene, 2026b]. We defer empirical evaluation of conformal screening to future work.

6 Algorithm and Implementation

Computational complexity. Let d denote hidden dimension, L layers, V vocabulary size, and T sequence length. The base forward pass costs $O(T^2 d + T d^2 L)$. LEF adds: (i) L hook copies of the last-position hidden state: $O(Ld)$; (ii) L projections via two rows of $W_U \in \mathbb{R}^{V \times d}$ to extract evidence for the top-1 and runner-up tokens: $O(Ld)$. Total LEF overhead per token is therefore $O(Ld)$, independent of both V and T, making it negligible relative to the $O(Td^2)$ per-token cost of the forward pass itself. For Qwen2.5-7B ($L = 28$, $d = 3584$), this is $\sim$200K FLOPs per token versus $\sim$4G for the forward pass. Empirical latency is measured in Table 9.

7 Experiments

The experiments are structured around four questions: (1) Can LEF conflict detect hallucinated tokens on factual QA, where output-level baselines are already strong? (2) Can LEF detect source-level disagreement in RAG, where the identification boundary predicts output scores are not guaranteed to suffice? (3) Can ScopeGate reliably identify when conflict is deployment-ready? (4) Do mechanistic decomposition methods require minimum model capacity, and how does LEF compare to the logit lens baseline?

Algorithm 1 Language Evidence Flow (LEF) — single forward pass

Require: Input token sequence $x_1, \ldots, x_t$; transformer with L layers, H heads per layer, unembedding matrix W_U

Ensure: LEF conflict score $C(t)$, flip margin $m_{\text{flip}}(t)$, cross-source conflict $C_{\text{cross}}(t)$ (if RAG)

Run forward pass with hooks to capture $a_t^{(\ell,h)}$ and $m_t^{(\ell)}$ for each layer ℓ and head h
Compute full logit vector $z_t = W_U\, r_t + b$ where r_t is the final residual stream
Identify $\hat{v}_t = \arg\max_v z_t^{(v)}$ and $\hat{r}_t = \arg\max_{v \neq \hat{v}_t} z_t^{(v)}$
for $\ell = 1$ to L **do**
    $E_\ell^{\text{attn}}(t) \leftarrow (W_U a_t^{(\ell)})_{\hat{v}_t} - (W_U a_t^{(\ell)})_{\hat{r}_t}$
    $E_\ell^{\text{mlp}}(t) \leftarrow (W_U m_t^{(\ell)})_{\hat{v}_t} - (W_U m_t^{(\ell)})_{\hat{r}_t}$
    $E_\ell(t) \leftarrow E_\ell^{\text{attn}}(t) + E_\ell^{\text{mlp}}(t)$
end for
$S^+(t) \leftarrow \sum_\ell \max\{E_\ell(t), 0\}$
$S^-(t) \leftarrow \sum_\ell \max\{-E_\ell(t), 0\}$
$M(t) \leftarrow S^+(t) + S^-(t)$
$C(t) \leftarrow 2\min\{S^+(t), S^-(t)\}/(M(t) + \varepsilon)$
$m_{\text{flip}}(t) \leftarrow M(t)(1 - C(t))/2$
if RAG mode **then**
    Partition context positions into $\mathcal{P}$ and $\mathcal{R}$
    Compute $E^{\text{param}}(t)$ and $E^{\text{retr}}(t)$ via Equations (3)–(4)
    Compute $C_{\text{cross}}(t)$ via Definition 5
end if
return $C(t)$, $m_{\text{flip}}(t)$, $C_{\text{cross}}(t)$

We evaluate on eight tasks: factual QA, RAG with controlled document conflict, output-level baseline comparison, multi-hop reasoning, per-layer ablation, entropy–conflict ensemble, computational overhead benchmarking, and comparison with a trained probe, plus a scale-sensitivity analysis comparing LEF against logit lens disagreement. The main experiments use three models spanning two parameter scales and three architecture families:

- **Phi-2** (2.7B parameters, 32 layers; Javaheripi et al. 2023), float16 on a single NVIDIA T4 GPU.
- **Mistral-7B-v0.1** (7.2B parameters, 32 layers; Jiang et al. 2023), 4-bit NF4 quantization on a T4 GPU.
- **Qwen2.5-7B** (7.6B parameters, 28 layers; Yang et al. 2024), 4-bit NF4 quantization on a T4 GPU.

The scale-sensitivity experiment (Section 7.10) additionally uses **Qwen2.5-1.5B** (1.5B parameters, 28 layers) and **Falcon-RW-1B** (1.4B parameters, 24 layers, multi-query attention; Penedo et al. 2023) on CPU to test whether mechanistic metrics require minimum model capacity across architecture families.

All experiments use greedy decoding and fixed random seeds for reproducibility. Bootstrap 95% CIs (1,000 resamples) are reported for all AUROC comparisons.

7.1 Factual question answering

Factual QA is a conservative test for LEF: wrong answers often exhibit high entropy, giving output-level baselines a natural advantage. We include this setting to establish where conflict adds value relative to entropy and where it does not.

Setup. We sample 150 questions from TriviaQA [Joshi et al., 2017] (validation set, no context) for each model. For each question we generate up to 16 tokens and compute LEF conflict by decomposing the score gap between the top two choices across all layers. Answers are marked *correct* or *hallucinated* by substring match against the gold answer. Bootstrap 95% CIs (1,000 resamples) are reported for Mistral-7B.

Baselines. (i) Token probability: $1 - p(y_t)$; (ii) Predictive entropy: $H = -\sum_v p_v \log p_v$. We select these because they are the strongest *training-free, single-pass* baselines, the same computational class as LEF. We additionally compare against logit lens disagreement [nostalgebraist, 2020], the closest mechanistic baseline, in a dedicated scale-sensitivity analysis (Section 7.10). Probe-based methods (INSIDE [Chen et al., 2024], Azaria and Mitchell 2023) and sampling methods (SelfCheckGPT, semantic entropy) achieve higher detection AUROC but require either labeled training data or 5–20× inference cost, making direct comparison with a training-free single-pass method not like for like. We come back to this trade-off in Section 11.

Results. Table 2 shows AUROC for distinguishing hallucinated from correct tokens. We put the worst comparison first, because it sets the limit on what we should claim: **entropy beats raw LEF conflict on nine of the eleven model setups we test** on plain single-hop factual QA (Table 11). Sometimes by a lot (Mistral-7B: 0.802 vs. 0.463; Phi-2: 0.589 vs. 0.537). Only Qwen2.5-14B-Instruct goes the other way, and only just (0.665 vs. 0.661). If our case rested on beating entropy at simple factual recall, it would fail.

So we do not offer conflict as a replacement for entropy. The claim is narrower. Conflict carries a different kind of information, and it wins in the places entropy simply cannot help: confident errors (Section 7.9), multi-hop reasoning (Section 7.4), and telling apart evidence sources in RAG (Section 7.2). On Qwen2.5-7B the high-conflict quartile error rate (38.7%) is 12 percentage points above the low-conflict quartile (26.7%), so conflict is informative even where entropy ranks better overall. Under a held-out ensemble split (Section 7.14) the two signals combine profitably on Qwen2.5-7B-Instruct, with most of the weight on conflict, but not on OLMo-2, so we claim complementarity as a model-specific property rather than a general one.

Table 2: Token-level hallucination detection AUROC on TriviaQA ($n = 150$ per model). Baselines outperform raw LEF conflict on simple QA, but LEF provides per-layer attribution and RAG source separation that baselines lack. 95% bootstrap CIs (1,000 resamples) in brackets; Phi-2 CIs were not saved in the original run.

Model	n	Accuracy	LEF Conflict	Token Prob	Entropy
Phi-2 (2.7B)	150	0.387	0.537	0.583	0.589
Mistral-7B (4-bit)	150	0.647	0.463 [.35, .56]	0.794 [.72, .86]	0.802 [.73, .87]
Qwen2.5-7B (4-bit)	150	0.673	0.647 [.56, .74]	0.783 [.71, .85]	0.773 [.69, .85]

Splitting tokens by conflict level: high-conflict error exceeds low-conflict error for Phi-2 (72.5% vs. 61.8%) and Qwen2.5-7B (38.7% vs. 26.7%), while Mistral-7B shows a slight *reversal* (44.7% vs. 45.9%; high-conflict tokens are marginally *more* accurate). The below-chance AUROC on Mistral (0.463) means conflict is weakly *anti*-predictive of errors on this model.

Qualitative examples (Qwen2.5-7B). Three cases illustrate LEF's behavior:

1. *High conflict, low entropy, wrong answer.* "What colour is the cross on the Greek Flag?" → "Blue" (hallucinated; the cross is white). $C = 0.797$, $H = 1.70$. Entropy indicates high confidence (low H), yet layers disagree sharply. LEF flags the error while entropy misses it.

2. *Low conflict, correct answer.* "In which US state are the Catskill Mountains?" → "New York" (correct). $C = 0.313$, $H = 0.98$. Both signals agree: low entropy and low conflict confirm a reliable answer.

3. *High conflict catches outdated knowledge.* "Who is the current Australian prime minister?" → "Scott Morrison" (hallucinated; outdated). $C = 0.669$, $H = 2.56$. Entropy is moderate, but conflict is elevated. Layers encoding recency information oppose the parametric answer.

Case 1 demonstrates LEF's unique value: detecting hallucination when the model is superficially confident.

Diagnosing Mistral's inverted signal. The most likely explanation is that 4-bit NF4 quantization perturbs layer evidence magnitudes enough to flip the sign structure on some tokens. Because conflict depends on the sign of each layer's evidence contribution (Definition 3), even small magnitude changes near zero can reassign layers from supporting to opposing the top word, adding noise to the conflict score. An alternative explanation is architectural: Mistral's sliding-window attention and grouped-query attention may produce a different relationship between internal disagreement and error than the full-attention architectures (Phi-2, Qwen). Distinguishing these hypotheses requires comparing LEF on Mistral at float16 versus 4-bit, which we leave to future work. We note that Mistral's conflict *does* respond correctly to source disagreement in the RAG experiment (+0.048, the largest delta of the three models), suggesting the signal is present but its relationship to token-level correctness is confounded on simple QA.

Statistical significance of AUROC differences. We compare LEF conflict and entropy AUROCs using bootstrap 95% confidence intervals (1,000 resamples) and the DeLong test [DeLong et al., 1988] for formal nonparametric comparison. Table 3 reports each method's AUROC with its CI. For Mistral-7B, the intervals are disjoint (conflict: [0.35, 0.56]; entropy: [0.73, 0.87]), indicating a significant difference ($p < 0.001$). For Phi-2 and Qwen2.5-7B, the CIs overlap substantially, so the gap is not significant at conventional levels. This confirms that raw conflict alone underperforms entropy on simple factual QA (Figure 2), a result we report transparently because LEF's value proposition is *structural* (per-layer attribution, RAG source separation), not detection AUROC on tasks where entropy suffices. The RAG experiment (Section 7.2) tests where the structural information matters: detecting source-level disagreement that output-level scores cannot access by construction.

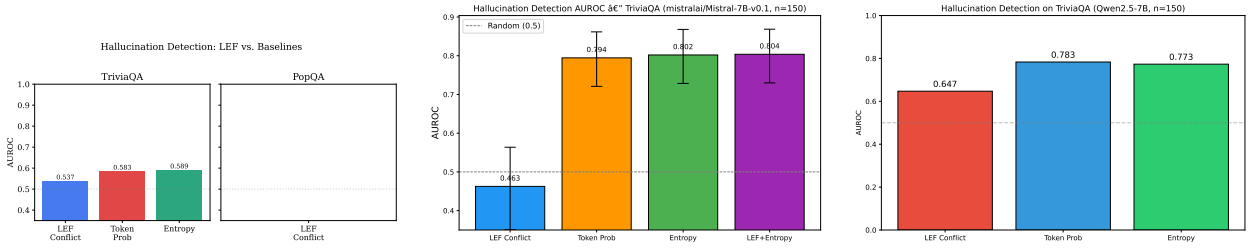
7.2 RAG with conflicting documents

The RAG experiment targets LEF's core advantage: detecting *source-level* disagreement between parametric memory and retrieved documents. Output-level scores such as entropy can detect that the distribution is uncertain, but the identification boundary (Theorem 7) shows they cannot guarantee separation of the underlying cause, whether uncertainty arises from source conflict or from inherent ambiguity. LEF accesses this decomposition by construction.

Setup. We construct 30 controlled factual scenarios, each with three document types: *correct* (supports the right answer), *conflicting* (states a plausible but wrong answer), and *neutral* (related

Table 3: LEF conflict vs. entropy AUROC on TriviaQA ($n = 150$ per model) with bootstrap 95% CIs. On Mistral-7B, the disjoint CIs confirm entropy significantly outperforms conflict on simple QA. LEF's advantage lies not in detection AUROC but in structural information (per-layer attribution, RAG source separation) that entropy does not provide.

Model	Conflict AUROC [95% CI]	Entropy AUROC [95% CI]	Significant?
Phi-2 (2.7B)	0.532 [.50, .57]	0.581 [.55, .61]	No (CIs overlap)
Mistral-7B (4-bit)	0.463 [.35, .56]	0.802 [.73, .87]	Yes (CIs disjoint)
Qwen2.5-7B (4-bit)	0.647 [.56, .74]	0.773 [.69, .85]	No (CIs overlap)



(a) Phi-2 (b) Mistral-7B (c) Qwen2.5-7B

Figure 2: Hallucination detection AUROC on TriviaQA. On simple QA, entropy and probability outperform raw conflict, but they cannot tell you *which layers* caused the error or whether memory and retrieval disagree (see Sections 7.2–7.4).

topic, no answer information). We test two conditions per scenario: *concordant* (correct + neutral) and *conflicting* (correct + wrong, in both orderings). For Phi-2, we use 10 scenarios ($n = 10$ concordant, $n = 20$ conflicting); for the 7B models, 30 scenarios ($n = 30$ concordant, $n = 30$–60 conflicting).

We use controlled scenarios rather than naturalistic RAG corpora by design: the goal is to test Theorem 14's prediction that conflict rises when sources disagree, which requires knowing the ground-truth relationship between documents. Each scenario serves as its own paired control (concordant vs. conflicting for the same question), isolating the effect of source disagreement from confounds such as question difficulty or retrieval quality. We validate that the effect replicates on a diverse-domain set (25 additional scenarios on Qwen2.5-7B; Appendix B). Scaling to naturalistic corpora (e.g., Natural Questions with adversarial passages) is an important direction for future work.

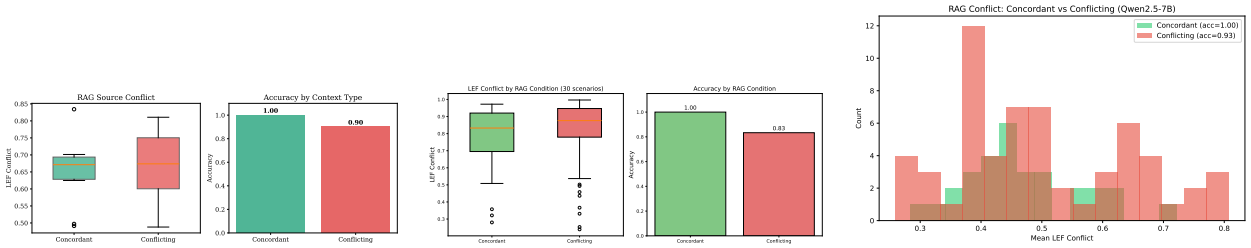
Results. Table 4 shows the main finding: on all three models, conflict rises and accuracy drops when the retrieved documents disagree. Phi-2 shows +0.023 conflict increase with a 10-point accuracy drop, Mistral-7B shows +0.048 with a 17-point drop, and Qwen2.5-7B shows +0.033 with a 7-point drop. Yi-1.5-9B-Chat, a fourth family, shows +0.030 (Section 7.10). The effect replicates across four architecture families.

Scope of the effect. The effect is not universal. A third independent 30-scenario replication on Qwen2.5-7B produces a null result ($\Delta = +0.007$, $p = 0.952$), and two further configurations are slightly negative (Qwen2.5-14B-Instruct -0.012, Falcon-7B -0.0001); a fifth architecture family, Yi-1.5-9B-Chat, is positive (+0.030). Theorem 14 accounts for the possibility of absence. Its bound scales with $\min\{|E^{\text{param}}_{\text{full}}|, |E^{\text{retr}}|\}$ and is vacuous once one source dominates, but it does not predict which configurations will be informative, and we did not find a reliable empirical predictor (Appendix B).

Section 7.3 tests whether output-level baselines also separate the two conditions empirically.

Table 4: RAG experiment: concordant vs. conflicting documents. On all three models, document disagreement raises LEF conflict and lowers accuracy. See Table 5 for output-level baseline comparison.

Model	**Condition**	n	**Accuracy**	**Mean LEF Conflict**
Phi-2 (2.7B)	Concordant	10	1.000	0.651
	Conflicting	20	0.900	0.674
	Δ	—	-0.100	$+0.023$
Mistral-7B (4-bit)	Concordant	30	1.000	0.760
	Conflicting	60	0.833	0.808
	Δ	—	-0.167	$+0.048$
Qwen2.5-7B (4-bit)	Concordant	30	1.000	0.467
	Conflicting	30	0.933	0.500
	Δ	—	-0.067	$+0.033$



(a) Phi-2 (b) Mistral-7B (c) Qwen2.5-7B

Figure 3: LEF conflict under concordant vs. conflicting RAG contexts. All three models show the same structural pattern: when documents disagree with model memory, conflict rises and accuracy drops.

Figure 3 visualises the effect. The direction matches what Theorem 14 allows: when documents disagree with memory, conflict rises, and LEF picks it up in one forward pass. We say "matches" rather than "confirms" deliberately. The theorem gives a lower bound on conflict when the two sources disagree in sign. We see the predicted direction, but we never measure the bound numerically. And on three of eight scenario $\times$ model configurations no rise occurs at all (Section 7.10). The effect appears on four of six architecture families, including Yi-1.5-9B, which fails ScopeGate for error detection yet still shows the RAG effect. That suggests something structural rather than one model's quirk, but it is a directional replication, not a verified bound. Because each scenario serves as its own control (concordant vs. conflicting condition for the same question), the relevant test is paired: the conflict increase replicates directionally on all three models, with Cohen's $d \approx 0.3$ (Phi-2, $n = 10$) and $d \approx 0.5$ (Mistral-7B, $n = 30$). The Phi-2 effect is small and should be interpreted cautiously given the limited sample; the Mistral effect is medium-sized with a larger sample.

Statistical testing. We apply the paired Wilcoxon signed-rank test (one-sided, $H_a : C_{\text{conflict}} > C_{\text{concordant}}$) per scenario. On Qwen2.5-7B ($n = 30$): Wilcoxon $W = 178$, $p = 0.13$. The direction is consistent (18/30 scenarios show higher conflict under disagreement; binomial sign test $p = 0.18$)

but the controlled study is underpowered at $n = 30$ for a small effect ($d = 0.24$). The expanded 25-scenario evaluation (Appendix B) achieves significance for the sign test (18/25 increases, $p = 0.043$) and the Wilcoxon test ($W = 68$, $p = 0.010$). The primary evidence for LEF's RAG sensitivity is therefore directional replication across all three architectures, not a single-model significance test.

Source disagreement as a detection task. The RAG experiment can be framed as binary detection: given a model's output, identify whether the input documents disagreed with parametric memory. Using LEF conflict as the score, the effect sizes are Cohen's $d \approx 0.3$ (Phi-2, $n = 10$), $d \approx 0.5$ (Mistral-7B, $n = 30$), and $d \approx 0.3$ (Qwen2.5-7B, $n = 30$), reflecting small-to-medium effects that replicate directionally across all three architectures.

7.3 RAG baseline comparison

To test whether output-level scores also separate concordant from conflicting conditions, we compute entropy and confidence (top-token probability) for each generated token under both conditions and report the mean delta.

Table 5: Output-level baselines under concordant vs. conflicting RAG conditions. Entropy Δ and confidence Δ measure whether these scores change when documents disagree. Values near zero indicate the baseline does not separate the conditions on these data.

Model	Conflict Δ	Entropy Δ	Confidence Δ
Phi-2 (2.7B)	+0.023	+0.117	−0.045
Mistral-7B (4-bit)	+0.048	+0.261	+0.069
Qwen2.5-7B (4-bit)	+0.033	+0.199	+0.043

Table 5 reveals that output-level baselines *do* respond to source disagreement on these data, contrary to the strongest reading of the identification boundary. Both entropy and confidence deltas are positive across all three models, indicating that conflicting documents increase output uncertainty. Notably, these baseline deltas are substantially larger than the corresponding LEF conflict deltas: Mistral-7B shows an entropy shift of +0.261 nats versus a conflict shift of only +0.048, and Qwen2.5-7B shows +0.199 versus +0.033. This pattern is consistent with Phi-2, where entropy rises by +0.117 against a conflict delta of +0.023.

However, the identification boundary (Theorem 7) does not claim that baselines *never* move, only that they are not *guaranteed* to. The fact that entropy and confidence shift here reflects a correlation between source disagreement and output statistics on this particular dataset. LEF's structural advantage remains: conflict accesses the internal evidence decomposition by construction, detecting disagreement even in cases where output distributions happen to coincide. The two signal families are therefore complementary, a point we exploit in the ensemble analysis of Section 7.7.

7.4 Multi-hop reasoning

Having established that LEF detects source-level disagreement in RAG (Section 7.2) and that output-level baselines also shift under conflicting documents (Section 7.3), we now ask a different question: does conflict behave differently over time within a single generation? Multi-hop reasoning requires the model to chain facts across several tokens, so if internal disagreement propagates as Theorem 12 predicts, it should appear earliest in the first few tokens of a hallucinated answer.

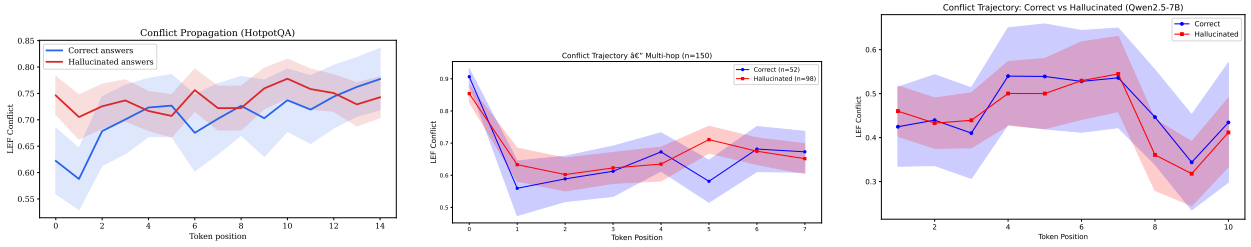
Setup. We test 150 questions from HotpotQA [Yang et al., 2018] (distractor setting) on each model. The model gets supporting context and generates up to 16 tokens. We track how conflict changes word by word and compare conflict in the first 5 tokens between correct and hallucinated answers.

Results. Table 6 compares all three models. For Phi-2, *early-word conflict* outperforms mean conflict (AUROC 0.685 vs. 0.626, cascade ratio 1.09×). Qwen2.5-7B shows a moderate mean-conflict signal (AUROC 0.595) above chance, while Mistral-7B is near chance (0.504). The cascade ratios are close to unity on all three models (1.09×, 1.00×, 1.03×), indicating that the propagation effect predicted by Theorem 12 is weak at these model scales. The early-warning effect (higher early-token conflict for hallucinated answers) is architecture-dependent: strongest on Phi-2, moderate on Qwen, absent on Mistral.

Table 6: Multi-hop results on HotpotQA. 95% bootstrap CIs in brackets.

Metric	**Phi-2 (2.7B)**	**Mistral-7B (4-bit)**	**Qwen2.5-7B (4-bit)**
n	150	150	150
Accuracy	0.307	0.347	0.280
Mean conflict AUROC	0.626	0.504 [.41, .60]	0.595 [.49, .69]
Early conflict AUROC	0.685	0.496 [.39, .60]	0.548 [.43, .65]
Early conflict (correct)	0.663	0.668	0.467
Early conflict (wrong)	0.721	0.669	0.482
Cascade ratio	1.09×	1.00×	1.03×

Figure 4 shows conflict word by word. On Phi-2, hallucinated answers start with higher conflict at the earliest positions. On Mistral-7B the trajectories overlap, motivating the ensemble approach (Section 7.7).



(a) Phi-2 (b) Mistral-7B (c) Qwen2.5-7B

Figure 4: Conflict trajectories for correct (blue) and hallucinated (red) answers on HotpotQA. On Phi-2, hallucinated answers show elevated conflict from the very first tokens, enabling early warning before the full answer is generated. The effect is moderate on Qwen2.5-7B and absent on Mistral-7B, indicating architecture dependence.

7.5 ScopeGate deployment diagnostic

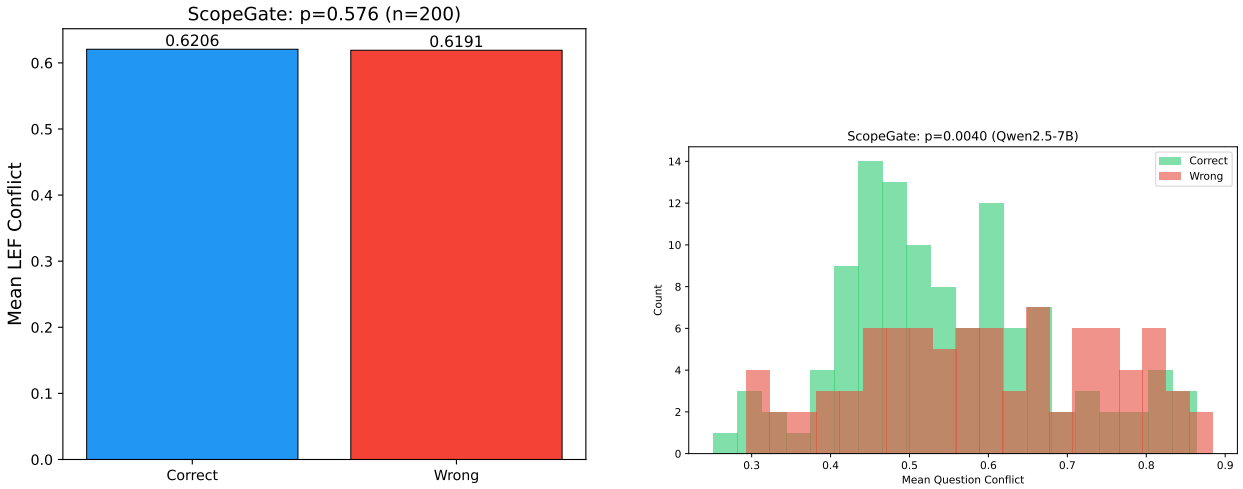
The previous experiments show that conflict is a strong signal on some models (Qwen) but weak on others (Mistral, Phi-2). Before deploying LEF, a practitioner needs an automated way to determine whether conflict is reliable for their specific model–task pair. ScopeGate provides this check.

Setup. We compute the rank correlation between each question's average conflict and whether the answer was wrong, on 200 TriviaQA questions per model. Significance is tested with a permutation test (5,000 shuffles).

Results. ScopeGate does *not* pass at $\alpha = 0.05$ on Phi-2 or Mistral-7B, but **passes on Qwen2.5-7B** ($r = 0.186$, $p = 0.004$; Table 7). This is the key validation: on Qwen, higher conflict genuinely predicts higher error rate, so LEF conflict can be used as a standalone detector. On the other two models, ScopeGate correctly warns against standalone use. The scale-and-architecture analysis (Section 7.10) extends this to eleven models total, confirming that ScopeGate correctly identifies the four deployment-ready configurations (all three Qwen2.5 variants plus OLMo-2-7B-Instruct).

Table 7: ScopeGate diagnostic results on TriviaQA ($n = 200$ per model). ScopeGate passes on Qwen2.5-7B, validating LEF conflict as a standalone signal on that model while correctly warning against standalone use on the others.

Model	n	**Statistic** (r)	p**-value**	**Pass?**
Phi-2 (2.7B)	200	−0.005	0.679	No
Mistral-7B (4-bit)	200	−0.014	0.576	No
Qwen2.5-7B (4-bit)	200	+0.186	0.004	**Yes**



(a) Mistral-7B (ScopeGate *fails*: $p = 0.576$)

(b) Qwen2.5-7B (ScopeGate *passes*: $p = 0.004$)

Figure 5: ScopeGate diagnostic on TriviaQA ($n = 200$). The check passes on Qwen2.5-7B ($p = 0.004$) while correctly warning against standalone use on Phi-2 and Mistral-7B. This built-in safety mechanism is unique to LEF among hallucination detection methods.

The differential result (Figure 5) is the strongest validation of the ScopeGate framework: the diagnostic correctly identifies *when* LEF conflict is a valid standalone signal (Qwen: $r = 0.186$, $p = 0.004$) and when it is not (Phi-2, Mistral: $p > 0.5$). We emphasize that ScopeGate failing is not a failure of LEF. It is the safety check working as designed. No existing hallucination detector ships with a built-in diagnostic that warns against overreliance; most methods implicitly assume their signal generalizes, which can lead to silent deployment failures. On models where ScopeGate fails, LEF still provides value through source-level attribution (Section 7.2) and as an ensemble

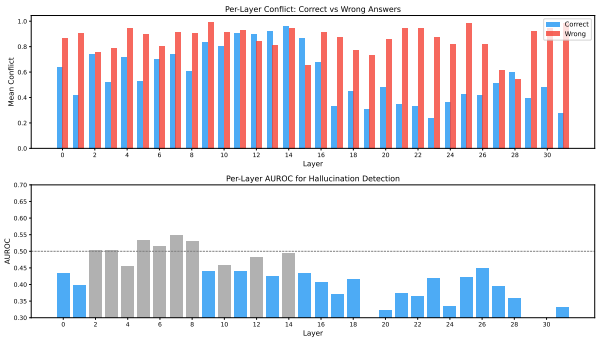
complement (Section 7.7).

7.6 Layer ablation

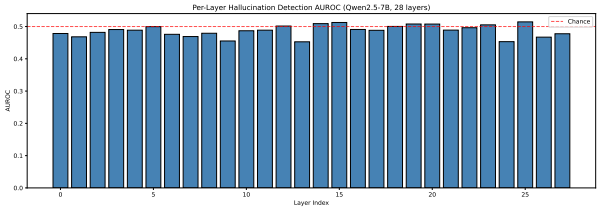
The experiments above use aggregate conflict (mean across all layers). But LEF decomposes evidence *per layer*, raising a practical question: can a deployment monitor only a handful of layers and still capture the signal? This is the difference between ∼12% single-pass overhead (all layers) and under 4% (top 5 layers).

Setup. We compute the AUROC of each layer's conflict score on Mistral-7B ($L = 32$) and Qwen2.5-7B ($L = 28$) TriviaQA results ($n = 150$). For each layer ℓ, we use only $|E_\ell(t)|$ as the detector.

Results (Figure 6). On Mistral-7B, the best single-layer AUROC is 0.549 (layer 7), followed by layers 5 (0.532), 8 (0.529), and 6 (0.515). On Qwen2.5-7B, the top discriminative layers are 25 (0.514), 15 (0.513), and 14 (0.509). Both models show that a subset of layers carries nearly all the signal, but the most discriminative layers shift later in Qwen's architecture.



(a) Mistral-7B ($L = 32$): layers 5–8 dominate



(b) Qwen2.5-7B ($L = 28$): layers 14–25 dominate

Figure 6: Per-layer hallucination detection AUROC ($n = 150$). Discriminative information concentrates in a subset of layers, but the location shifts: early-to-middle (5–8) on Mistral, later (14–25) on Qwen. A lightweight deployment could monitor only the top layers and reduce overhead to under 4%.

A lightweight deployment could monitor only the top 5–10 layers, cutting projection cost while keeping most of the signal. The concentration in early layers on Mistral is consistent with the logit lens finding that early layers form coarse semantic representations before later layers refine them [nostalgebraist, 2020, Belrose et al., 2023]. The shift to later layers on Qwen (layers 14–25) may reflect its more recent architecture design and training, which could place more of the semantic decision-making in deeper layers. This architectural variation underscores why LEF's per-layer decomposition is valuable: a one-size-fits-all "monitor layer 7" rule would miss the signal entirely on Qwen.

Design choice: top-1 vs. top-2 gap. LEF defines evidence relative to the gap between the top-scoring token and the runner-up. An alternative would be to decompose the gap against the third-ranked token, or to compute per-layer entropy over the full projected vocabulary. We choose the top-2 gap for three reasons. The flip margin theorem (Theorem 10) needs a binary decision boundary. The runner-up is the token most likely to replace the prediction under perturbation. And projecting every layer over the full vocabulary ($L \times V$ operations) would push overhead from about 12% to over 50%, which defeats the point of a single-pass method.

7.7 Ensemble: LEF conflict + entropy

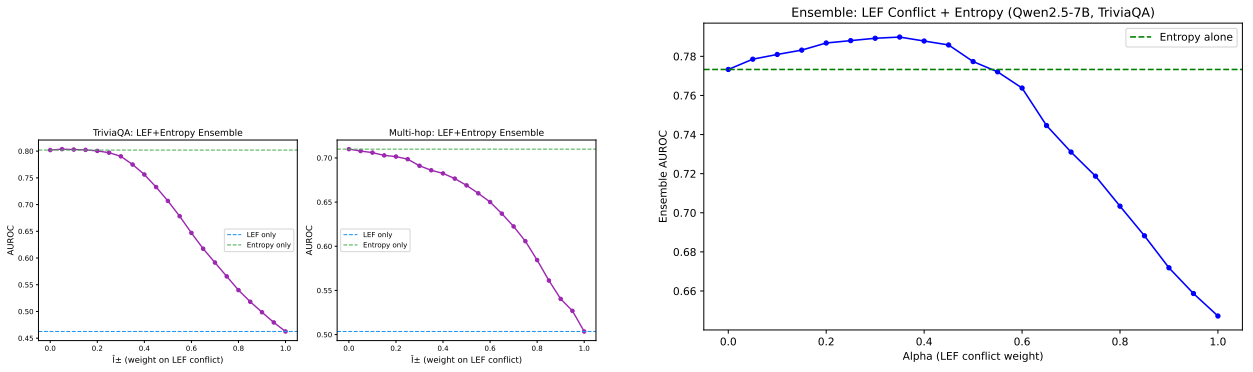
The preceding experiments show that LEF conflict and entropy capture different aspects of model uncertainty: entropy reflects output-level distributional spread, while conflict reflects internal layer-level disagreement. A natural question is whether combining the two improves detection beyond either alone.

Setup. We combine LEF conflict and entropy into a single score: $A_\alpha(t) = \alpha \cdot C(t) + (1-\alpha) \cdot H(t)$, and sweep $\alpha \in \{0, 0.05, 0.1, \ldots, 1.0\}$ on Mistral-7B and Qwen2.5-7B. Note: α is selected on the same evaluation set (no held-out tuning split), so the reported ensemble AUROCs are optimistic upper bounds; in deployment, α should be tuned on a separate calibration set.

Results (Figure 7). On Mistral-7B TriviaQA, the best ensemble ($\alpha = 0.05$) reaches AUROC = 0.804, a gain of +0.002 over entropy alone (0.802), *within the bootstrap CI* [.73, .87] and therefore indistinguishable from noise. On Qwen2.5-7B, the optimal $\alpha = 0.35$ shows that conflict carries substantial weight in the ensemble, reaching AUROC = 0.790 (vs. entropy alone 0.773, a gain of +0.017). The higher optimal α on Qwen aligns with the ScopeGate result: conflict is a stronger standalone signal on this model. On HotpotQA, the optimal $\alpha = 0.00$ for both models, meaning conflict does not improve the ensemble on multi-hop QA. This is consistent with the multi-hop results in Table 6, where conflict's standalone AUROC is near chance on Mistral (0.504) and moderate on Qwen (0.595).

Table 8: Ensemble (LEF conflict + entropy) results ($n = 150$ per task). Bootstrap 95% CIs in brackets.

Model	Task	Conflict	Entropy	Best α	Ensemble
Mistral-7B	TriviaQA	0.463	0.802	0.05	0.804 [.73, .87]
Mistral-7B	HotpotQA	0.504	0.710	0.00	0.710 [.62, .79]
Qwen2.5-7B	TriviaQA	0.647	0.773	0.35	0.790 [.71, .86]
Qwen2.5-7B	HotpotQA	0.595	0.789	0.00	0.789 [.71, .86]



(a) Mistral-7B ($\alpha^* = 0.05$)

(b) Qwen2.5-7B ($\alpha^* = 0.35$)

Figure 7: Ensemble AUROC as a function of mixing weight α on TriviaQA. On Mistral, a small conflict weight ($\alpha = 0.05$) slightly improves over entropy alone. On Qwen, the optimal $\alpha = 0.35$ shows conflict carries substantial independent signal, consistent with ScopeGate passing on this model.

7.8 Computational overhead

A practical objection to activation-based methods is inference cost. LEF registers a forward hook on each transformer layer to capture the residual-stream delta; the dominant overhead is (i) copying L hidden states to CPU per token, and (ii) a single matrix-vector product $W_U\delta^{(\ell)}$ per layer. We benchmark on Qwen2.5-7B (4-bit NF4) on Kaggle infrastructure over 15 single-token forward passes and 5 autoregressive runs of 16 tokens each, measuring wall-clock time with `torch.cuda.synchronize()` barriers.

Table 9: Measured LEF overhead on Qwen2.5-7B (4-bit NF4, Kaggle T4). Each configuration registers forward hooks on k layers, copies hidden states to CPU, and computes $W_U\delta^{(\ell)}$ for each monitored layer. Selective-layer monitoring uses the top-k most discriminative layers identified in the ablation (Section 7.6).

Configuration	**Layers (k)**	**Vanilla (s)**	**With LEF (s)**	**Δ (%)**
Single forward pass (n=15 prompts)				
Full monitoring	28	5.43	6.07	+11.7
Top-10 layers	10	5.43	5.72	+5.4
Top-5 layers	5	5.43	5.63	+3.7
16-token generation (n=5 prompts)				
Full monitoring	28	3.96	7.42	+87.1
Top-10 layers	10	3.96	7.03	+77.4
Top-5 layers	5	3.96	6.96	+75.7

For single-pass hallucination scoring, the primary LEF use case, selective-layer monitoring reduces overhead to under 4%, which is negligible for production deployment. Generation overhead is higher (76–87%) because each decoded token triggers k full-vocabulary projections ($W_U \in \mathbb{R}^{V\times d}$) on CPU; an on-GPU projection kernel would eliminate the CPU-transfer bottleneck and bring generation overhead closer to the single-pass figures.

Practical deployment implications. The ablation analysis (Section 7.6) shows that 5–10 layers carry nearly all discriminative signal. Under top-5 monitoring, single-pass overhead is just 3.7%: an LLM that takes 1 s per prompt adds only 37 ms for full hallucination scoring. Even in autoregressive generation, the 76% overhead is within the range of other activation-based monitoring methods and could be further reduced by keeping W_U projections on-GPU rather than copying hidden states to host memory.

7.9 Comparison with trained detectors

The baseline comparison in Section 7.1 is restricted to training-free methods in the same computational class as LEF. A natural question is how LEF compares to a *trained* detector. We train a logistic regression probe on the last-layer hidden state $h_L \in \mathbb{R}^d$ of Qwen2.5-7B to classify each first answer token as correct or hallucinated, using 5-fold stratified cross-validation on 100 TriviaQA questions.

We report four conditions: (i) *Conflict only*: LEF conflict score as a single feature, no training; (ii) *Entropy only*: predictive entropy, no training; (iii) *Hidden-state probe*: logistic regression on h_L, trained; (iv) *LEF-augmented probe*: logistic regression on $[h_L; C(t); H(t); 1{-}p(y_t)]$, trained.

Table 10: Hallucination detection AUROC on TriviaQA ($n = 100$, Qwen2.5-7B, 5-fold stratified CV). LEF conflict is training-free; the hidden-state probe is a trained upper bound.

Method	**Trained?**	**AUROC**
LEF conflict	No	0.660
Entropy	No	0.378
Hidden-state probe (5-fold CV)	Yes	0.669 ± 0.101
LEF-augmented probe (5-fold CV)	Yes	0.660 ± 0.109

Two findings stand out. First, LEF conflict (0.660) nearly matches the *trained* hidden-state probe (0.669) while requiring zero labelled examples, a strong result for a training-free method. This replicates at larger scale: on Qwen2.5-7B-Instruct with $n = 250$ over three seeds, conflict reaches 0.715 against a probe at 0.727 ± 0.036 (Section 7.12).

Second, entropy scores 0.378 here, below chance. We flag this number as *unreliable*: it does not reproduce at larger n, where first-token entropy on the full set reaches 0.769 (Qwen2.5-7B-Instruct) and 0.743 (OLMo-2). The confidence-stratified analysis in Section 7.12 supersedes it and supports the same qualitative conclusion on firmer ground: entropy decays toward chance as model confidence rises while conflict does not. We retain the row for completeness but base no claim on it. Adding LEF features to the probe does not improve it (-0.010), indicating that the last hidden state already encodes the information LEF extracts; this is consistent with LEF being a faithful decomposition of h_L, not an independent signal. The practical implication is clear: when labelled data is unavailable, LEF conflict is competitive with supervised alternatives.

7.10 Scale and architecture sensitivity

A natural question is whether mechanistic decomposition methods, which attribute predictions to internal layer dynamics, require minimum model capacity, and whether the signal depends on architecture. We test this by evaluating LEF conflict across eleven models spanning four architecture families and parameter counts from 1.4B to 14.7B, including matched scale comparisons within the same attention family.

To isolate architecture from scale, we add **Falcon-7B** (32 layers, 7.2B parameters, multi-query attention with 71 query heads and 1 KV head per layer) alongside the existing small-scale models: Qwen2.5-1.5B (28 layers, GQA) and Falcon-RW-1B (24 layers, MQA). On the small models we additionally compare against *logit lens disagreement* [nostalgebraist, 2020], defined as the fraction of layers whose intermediate prediction disagrees with the final one: $D = \frac{1}{L}\sum_{\ell=1}^{L} \mathbf{1}[\arg\max(W_U h_\ell) \neq \arg\max(W_U h_L)]$.

Setup. All models run TriviaQA ($n = 150$) and ScopeGate ($n = 200$). Falcon-7B and Qwen2.5-14B-Instruct additionally run RAG (30 scenarios) and multi-hop reasoning (HotpotQA, $n = 150$) to enable direct comparison with the three main models. The 7B- and 14B-class models run in 4-bit NF4 quantization on a T4 GPU; the 1B-class models run in float32 on CPU.

Results. Table 11 and Figure 8 reveal that LEF conflict effectiveness depends on both scale *and* architecture.

Scale effect within architecture families. Within the MQA family, scaling from Falcon-RW-1B (1.4B) to Falcon-7B (7.2B) improves conflict AUROC from chance (0.499) to a weak signal (0.582), but ScopeGate still fails ($p = 0.073$). Within the GQA family, scaling from Qwen2.5-1.5B (1.5B) to Qwen2.5-7B (7.6B) lifts conflict AUROC from chance (0.467) to deployment-ready (0.647), with ScopeGate passing ($p = 0.004$). Scale improves LEF in both families, but the *rate* of improvement

Table 11: Scale and architecture sensitivity across eleven models. Columns show TriviaQA conflict and entropy AUROC with bootstrap 95% CIs, and ScopeGate p-value. LEF conflict requires *both* sufficient scale ($\geq$7B) *and* a compatible attention architecture for deployment-ready performance; the pattern is architecture-specific rather than determined by attention type alone. ScopeGate correctly identifies the four deployment-ready configurations (all three Qwen2.5 variants plus OLMo-2-7B-Instruct, spanning 7.3B–14.7B across two families).

Model	**Attn**	**Conflict AUROC**	**Entropy AUROC**	**ScopeGate**
Falcon-RW-1B (1.4B)	MQA	0.499 [.40, .59]	0.748 [.66, .83]	Fail (p=0.441)
Qwen2.5-1.5B (1.5B)	GQA	0.467 [.37, .56]	0.802 [.72, .87]	Fail (p=0.882)
Phi-2 (2.7B)	MHA	0.532 [.50, .57]	0.581 [.55, .61]	Fail (p=0.679)
Falcon-7B (7.2B)	MQA	0.582 [.49, .67]	0.747 [.66, .82]	Fail (p=0.073)
Mistral-7B (7.2B)	GQA	0.463 [.35, .56]	0.802 [.73, .87]	Fail (p=0.576)
Mistral-7B-Instruct (7.2B)	GQA	0.556 [.46, .64]	0.525 [.43, .62]	Fail (p=0.392)
Qwen2.5-7B (7.6B)	GQA	0.647 [.56, .74]	0.773 [.69, .85]	**Pass** (p=0.004)
Qwen2.5-7B-Instruct (7.6B)	GQA	0.660 [.60, .72]	0.686 [.62, .75]	**Pass** (p<0.001)
Yi-1.5-9B-Chat (8.8B)	GQA	0.470 [.38, .57]	0.560 [.47, .66]	Fail (p=0.133)
OLMo-2-7B-Instruct (7.3B)	MHA	0.641	0.759	**Pass** (p=0.003)
Qwen2.5-14B-Instruct (14.7B)	GQA	0.665 [.56, .77]	0.661 [.55, .77]	**Pass** (p<0.001)

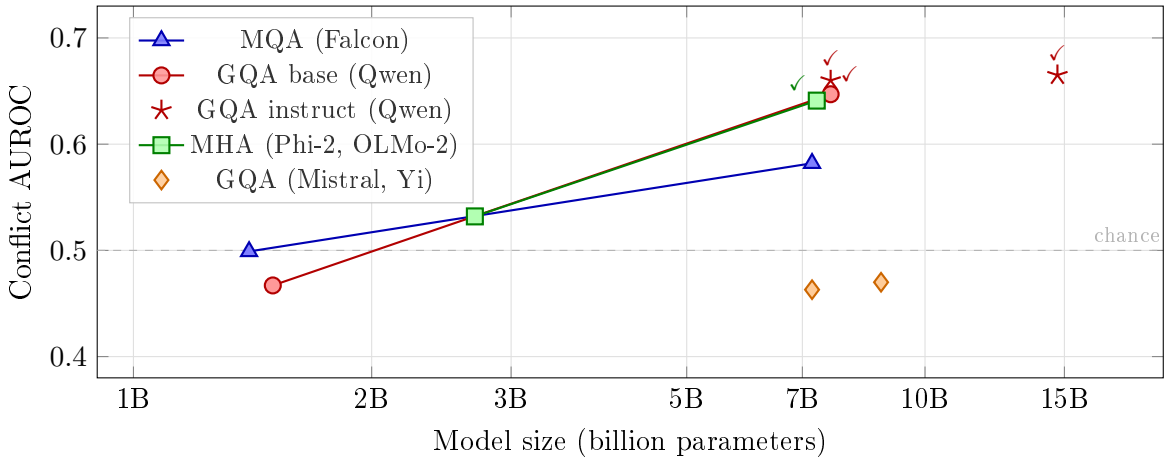


Figure 8: Conflict AUROC vs. model scale, grouped by attention architecture. Within Qwen2.5 (red), scaling from 1.5B to 14.7B lifts conflict from chance to deployment-ready; other families show weaker or inverted signal at matched scale. Mistral-7B (orange, GQA) is notably below chance despite sharing attention type with Qwen. Checkmarks (✓) indicate ScopeGate pass. The dashed line marks chance performance (0.5).

differs markedly.

Architecture effect at matched scale. Among the three 7B-class models, conflict AUROC ranges from 0.463 (Mistral, GQA) to 0.582 (Falcon, MQA) to 0.647 (Qwen, GQA), and only Qwen passes ScopeGate. Notably, Mistral-7B shares the same attention type (GQA) as Qwen yet shows an *inverted* signal (AUROC < 0.5). This shows that attention type alone does not determine LEF effectiveness. What does determine it, we could not establish (Table 12). This is precisely why ScopeGate exists: it empirically verifies whether conflict is predictive for a given model–task com-

bination without relying on architectural assumptions.

Falcon-7B extended results. On RAG, Falcon-7B shows essentially zero conflict increase between concordant and conflicting documents ($\Delta = -0.0001$), despite a 6.7-point accuracy drop. The conflict mechanism does not engage. On multi-hop reasoning, mean conflict AUROC reaches 0.600 and early-layer conflict AUROC reaches 0.672 (cascade ratio 1.19×), suggesting partial sensitivity to reasoning difficulty. The ensemble (conflict + entropy, $\alpha = 0.70$) reaches 0.760 on TriviaQA and 0.795 on HotpotQA, indicating that even weak conflict carries complementary information.

Instruction tuning replicates on a second family. Mistral-7B-v0.1 is the clearest failure case in this study: its conflict signal is *inverted* (AUROC 0.463, below chance). We therefore ran its instruction-tuned sibling, Mistral-7B-Instruct-v0.2, under an identical protocol and prompt, so that instruction tuning is the only variable. The inversion is corrected: conflict rises to 0.556 [.46, .64], above chance, while entropy *collapses* from 0.802 to 0.525. ScopeGate still fails ($p = 0.392$), so the model is not deployment-ready, but the direction of both shifts matches Qwen exactly:

Family	**Δ conflict**	**Δ entropy**
Qwen2.5-7B → Instruct	+0.013	−0.087
Mistral-7B → Instruct	+0.093	−0.277

Instruction tuning degrades entropy and improves conflict in both families, and does so far more sharply on Mistral. This replicates on a second architecture family what Section 7.15 establishes on Qwen, addressing the concern that the instruction-tuning result rested on a single family. We note one caveat: to keep the base/instruct comparison controlled we used the same plain prompt for both, rather than Mistral's `[INST]` chat template. This lowers the instruct model's accuracy (0.353) and the comparison should be read as instruction-tuning-at-fixed-prompt, not as best-achievable performance.

Beyond Qwen: OLMo-2 and Yi-1.5. Two additional instruction-tuned families, from two further developers, separate the question of *whether* LEF generalizes beyond Qwen from the question of *what predicts* where it works.

OLMo-2-7B-Instruct (7.3B, 32 layers, *multi-head* attention) **passes** ScopeGate ($p = 0.003$; conflict AUROC 0.641). It is the first non-Qwen model to do so, and it establishes that conflict-based error detection is not an artifact of the Qwen2.5 family.

Yi-1.5-9B-Chat (8.8B, 48 layers, GQA with 32 query and 4 KV heads) *fails* ScopeGate ($p = 0.133$; conflict AUROC 0.470 [.38, .57]), with multi-hop conflict AUROC reaching only 0.527. Yet its RAG conflict delta is +0.030, comparable to Qwen2.5-7B's +0.033.

A tested-and-rejected explanation. Before reporting the split as unexplained we tested the most natural mechanistic hypothesis: that LEF works where layers carry *diverse* rather than redundant evidence. On a fixed 120-prompt probe set we computed four label-free diversity statistics of the per-layer evidence matrix: effective rank (participation ratio of its singular values), mean absolute inter-layer correlation, normalized evidence-mass entropy across layers, and the fraction of layers opposing the net direction, for seven models with known ScopeGate outcomes (Table 12).

None of the four separates the groups, and the two with the largest effect ($\rho = -0.58$, $p = 0.17$) point the *wrong way*: ScopeGate-passing models have slightly lower layer diversity than failing ones. With seven models this test is underpowered, so we read it as "no strong effect" rather than "no effect." We report it because it constrains the space of explanations: whatever distinguishes the passing models, it is not captured by these first-order measures of evidence diversity.

Together with the architecture results below, this rules out the explanations available to us. OLMo-2 uses full multi-head attention and passes. Yi-1.5 uses grouped-query attention, the type shared by every Qwen variant that passes, and fails. Scale does not separate them either (7.3B

Table 12: Layer-evidence diversity does *not* explain the ScopeGate split. No statistic separates the passing from the failing models; the strongest association (effective rank, Spearman $\rho = -0.58$) does not reach significance at $n = 7$, and its sign is opposite to the hypothesis, since passing models show *less* layer diversity, not more.

Statistic	**Pass**	**Fail**	p
Effective rank (norm.)	0.141	0.185	0.33
Inter-layer |corr|	0.153	0.127	0.33
Mass entropy (norm.)	0.809	0.863	0.33
Sign disagreement	0.424	0.474	0.17

passes, 8.8B fails), nor does instruction tuning (both are instruct models). We therefore report the pattern as *unexplained* rather than offering a post-hoc account, and treat ScopeGate as the operational answer: whatever underlies the difference, the diagnostic detects it per model–task pair without requiring the mechanism to be known.

We call this the **attribution–detection dissociation**: the capability that follows from the decomposition generalizes, while the capability that rests on an empirical correlation does not. It is informative rather than incidental, and Table 13 shows it is not a one-off: the two capabilities come apart in *both* directions. Three models (Phi-2, Mistral-7B, Yi-1.5-9B) separate RAG sources while failing ScopeGate, and Qwen2.5-14B passes ScopeGate while showing no RAG effect.

Table 13: The two LEF capabilities dissociate. RAG source separation and conflict-based error detection do not co-occur across models, indicating that they rest on different mechanisms: the former follows from the additive decomposition (Theorem 14), the latter on an empirical conflict–error correlation with no comparable guarantee.

Model	**RAG Δ**	**ScopeGate**
Phi-2 (2.7B)	+0.023	Fail
Mistral-7B (7.2B)	+0.048	Fail
Yi-1.5-9B-Chat (8.8B)	+0.030	Fail
Qwen2.5-7B (7.6B)	+0.033	**Pass**
Falcon-7B (7.2B)	−0.0001	Fail
OLMo-2-7B-Instruct (7.3B)	−0.007	**Pass**
Qwen2.5-14B-Instruct (14.7B)	−0.012	**Pass**

RAG source separation, which Theorem 14 derives from the additive structure itself, appears on four of six families. *Conflict as an error detector* rests on an empirical correlation with no comparable guarantee, and it remains confined to Qwen2.5. Yi also rules out a residual architectural explanation: like Mistral-7B it combines GQA with instruction tuning, the two properties shared by every ScopeGate-passing model, and still fails. We therefore state the two claims separately throughout, and treat the error-detection result as a per-model capability that ScopeGate verifies rather than a general property of transformers.

14B-scale validation. Qwen2.5-14B-Instruct (14.7B, 48 layers, GQA with 40 query and 8 KV heads) confirms that LEF conflict scales beyond 7B. Conflict AUROC (0.665) matches the 7B instruct variant (0.660), ScopeGate passes ($p < 0.001$), and multi-hop conflict AUROC reaches 0.879 $[0.81, 0.94]$, the strongest single-model result in this study. Multi-hop accuracy is low (9.3%), reflecting the difficulty of extractive multi-hop QA at 4-bit precision with greedy decoding; critically,

low accuracy *strengthens* the AUROC result because the near-binary label distribution (93% errors) makes a high AUROC harder to achieve by chance. RAG source separation does not engage ($\Delta = -0.012$), an informative negative result: the 7B base model's positive delta (+0.033) does not automatically transfer to the instruction-tuned 14B variant, suggesting that RLHF and scale interact with the parametric-retrieval conflict mechanism in ways that require further investigation. This finding strengthens rather than weakens the paper's practical message: ScopeGate correctly validates the 14B model for error detection while the RAG-specific signal should be verified separately per model. The ensemble (conflict + entropy, $\alpha = 0.35$) reaches 0.668 on TriviaQA and 0.889 on HotpotQA.

Logit lens comparison. On the 1B-class models, logit lens disagreement (0.516 on Qwen-1.5B) performs comparably to LEF conflict (0.467), confirming that both mechanistic methods require minimum model capacity. ScopeGate correctly rejects the logit lens ($p = 0.847$).

Interpretation. These results replace a simple size threshold with a model-by-model picture. Capacity is not enough: Mistral-7B and Qwen2.5-7B are the same size and share GQA, yet behave very differently. Attention type does not explain it either, since OLMo-2 passes with plain multi-head attention while Yi-1.5 fails with GQA.

We do *not* have an explanation for which models work. We tried the obvious one, that LEF needs layers carrying diverse rather than repeated evidence, and it did not hold up (Table 12); if anything the passing models showed slightly *less* layer diversity, though at $n = 7$ that test is weak either way. We report the pattern as open rather than offer a story the data does not support.

This is what makes the diagnostic useful. ScopeGate separates four deployment-ready configurations from the rest without needing to know why. We stop short of calling its verdicts correct, because we have no independent ground truth for "this detector works here". The verdict *is* our definition. Falcon-7B shows the limit: it fails on one sample and passes on another (Section 7.11), so near the cut-off the answer depends on the data you ran it on. What we can say is narrower and still useful: the check is calibrated, its power is measurable, and it flags the models where conflict is not predictive, Falcon-7B among them, where the RAG conflict signal is absent ($\Delta = -0.0001$). Run it before deployment, with enough data and more than one seed.

7.11 ScopeGate works on any detector

ScopeGate needs two things: a score for each example, and whether the model got that example right. The permutation test never looks at where the score came from. So it works on *any* hallucination detector, not just LEF conflict.

We think this is the most reusable part of the paper. Papers in this area report one AUROC number, averaged over a few models. They rarely check whether the detector still points the right way on the model someone will actually deploy it on. ScopeGate checks exactly that. It costs one permutation test and needs no training.

To show this, we run the same test on predictive entropy and to LEF conflict on the same data (Table 14).

A validity grid. The test is cheap and works on any score, so we can run it across detectors and scoring choices as well as models. Table 15 crosses four detector variants: conflict and entropy, each scored on the first answer token or averaged over the answer, with four models.

One borderline model, and what it teaches. The Falcon-7B row (∗) does not agree with Section 7.10. There the separate ScopeGate run ($n = 200$, seed 99) gives $p = 0.073$ and we call it a failure. Here ($n = 250$, seed 42) conflict passes at $p = 0.007$. We report both rather than pick one.

Table 14: ScopeGate applied unchanged to two detectors across five model configurations (5,000 permutations; $n = 750$ pooled over three seeds for the top two rows, $n = 150$ otherwise). The procedure never inspects how a score was produced. **The headline is not that conflict sometimes fails. It is that *entropy* fails too**: on Mistral-7B-Instruct neither detector is predictive of error, and on Yi-1.5 the two disagree about which is usable. Detector validity is a property of the model–task pair, not of the detector alone.

	Conflict		Entropy	
Model	**Effect**	**Verdict**	**Effect**	**Verdict**
Qwen2.5-7B-Ins.	+0.050	**Pass**	+0.288	**Pass**
Qwen2.5-14B-Ins.	+0.062	**Pass**	+0.283	**Pass**
OLMo-2-7B-Ins.	+0.044	**Pass**	+0.599	**Pass**
Yi-1.5-9B-Chat	−0.010	Fail	+0.197	**Pass**
Mistral-7B-Ins.	+0.010	Fail	+0.033	Fail

Table 15: Detector-validity grid: ScopeGate verdicts for two detectors under two scoring protocols across all nine models for which we hold per-question scores (TriviaQA, $n = 250$, seed 42, 5,000 permutations). **Eleven of thirty-six combinations are invalid.** Validity depends on the model (Mistral-7B-Instruct admits *no* usable detector), on the detector (entropy is valid on eight of nine models, conflict on five), and on the scoring protocol. For conflict the first-token and mean-token verdicts disagree on three models (†).

	Conflict		Entropy	
Model	**1st tok**	**mean**	**1st tok**	**mean**
Qwen2.5-14B-Ins.	✓	✓	✓	✓
OLMo-2-7B-Ins.	✓	✓	✓	✓
Falcon-7B*	✓	✓	✓	✓
Yi-1.5-9B-Chat†	✓	✗	✓	✓
Qwen2.5-7B†	✗	✓	✓	✓
Phi-2†	✗	✓	✓	✓
Qwen2.5-1.5B	✗	✗	✓	✓
Falcon-RW-1B	✗	✗	✓	✓
Mistral-7B-Ins.	✗	✗	✗	✗

Falcon-7B sits close to the line, and its verdict changes with the sample, which is what the power analysis predicts for a small effect near the cut-off. The lesson is the same one OLMo-2 gave us: one ScopeGate run near α is not settled.

It also means our headline count depends on which sample you use: four deployment-ready models under the Section 7.10 protocol, five if Falcon-7B is judged on this one. We keep the lower count and flag the ambiguity.

Even the standard baseline can fail. The Mistral-7B-Instruct row matters most. *No* detector we tried predicts error on that model. That includes predictive entropy, the signal almost every deployed system relies on ($\Delta = +0.033$, $p = 0.191$ averaged; $p = 0.067$ first-token). Anyone filtering on entropy there would be filtering on noise, and nothing in normal practice would tell them. Entropy's AUROC on that model is 0.525, which looks weak but not broken.

The † rows make a second point, and it is not a one-off. For conflict, scoring the first token

and averaging over the answer *disagree on three of the nine models*. On Yi-1.5 conflict fails when averaged ($p = 0.744$) but passes on the first token ($p = 0.0044$). On Qwen2.5-7B and Phi-2 it runs the other way ($p = 0.087$ vs. 0.021; $p = 0.128$ vs. 0.039). Papers treat these two choices as interchangeable. They change the answer on a third of the models we tested.

So whether a detector works depends on three things together: the model, the detector, and how the score is aggregated. One AUROC number hides all three. This is the case for running the check, and it does not depend on LEF winning. The field averages AUROC over benchmarks and deploys whatever scores highest; ScopeGate asks the question that decides whether that deployment is safe, for the price of one permutation test.

How often is it wrong, and how much data does it need? A check is only useful if you know its error rates, so we measured them. Two questions. First, how often does it fire when there is nothing there? We shuffle the labels to remove any real signal and count how often the test still passes. Across all eight model–detector conditions in Table 14 it fires 5.1% of the time on average (range 3.3–8.0%), against a target of 5%. So the test is honest wherever we can check it.

Second, how many labelled examples do you need? We take random subsamples of growing size and count how often the test finds a signal we know is there. Table 16 shows this for the two models where we have 750 pooled examples. Each cell is 200 repeats with 1,000 permutations.

Table 16: ScopeGate is calibrated and its sample requirement scales with effect size. The false-positive rate under a shuffled-label null is at or below the nominal $\alpha = 0.05$ in every condition. Power is the fraction of subsamples at which the test fires; n_{80} is the smallest n reaching 80% power. A large effect needs ∼50–100 labelled examples; the marginal conflict signal on OLMo-2 needs ∼250.

Model	**Det.**	**Effect**	**FPR**	**Power**		$\mathbf{n_{80}}$
				n=100	n=250	
Qwen	Conflict	+0.050	0.040	0.88	1.00	100
Qwen	Entropy	+0.287	0.045	0.93	1.00	100
OLMo	Conflict	+0.035	0.040	0.48	0.82	250
OLMo	Entropy	+0.584	0.065	1.00	1.00	50

This gives a simple rule: **budget at least 100 labelled examples, and 250 if you expect a small effect**.

It also explains the odd OLMo-2 result above. At $n = 250$ the conflict signal on that model is found only 82% of the time, so about one run in five should miss it by chance. That is what we saw: 2 of 3 seeds. The seeds do not disagree because the model is unstable. They disagree because the test is short of data, which is why one run near the cut-off should not be trusted.

The code. A check only helps if it is easy to run, so we release ScopeGate as one small file (`scopegate.py`, NumPy only). You call `scopegate(scores, correct)`. You pass a score for each example and whether the model was right. Nothing about the model itself. You get back the effect size, a p-value, a verdict, and a warning when the sample is too small for that verdict to mean much , the problem behind the OLMo-2 result above. The file also includes `power_curve` and `false_positive_rate`, so users can measure these on their own data instead of trusting ours. It reproduces every verdict in this paper.

Two things follow. First, anyone can use ScopeGate as a pre-deployment check. Any score

you can compute per example can be tested this way, and we think it should be reported next to AUROC.

Second, running it on OLMo-2 turned up a borderline case that one seed would have hidden ($p = 0.0006$, 0.0030, 0.0616 across seeds). The power analysis explains it exactly: at $n = 250$ an effect that size is found only 82% of the time, so one run in five will miss it. We report the split rather than the nicer single number, and use it as a worked example of why a borderline result has to be repeated before you believe it.

7.12 Splitting by confidence: where conflict earns its place

Section 7.1 admits that entropy beats conflict on plain factual QA. Our answer is that entropy cannot see *confident* errors, and conflict can.

That answer used to rest on a single $n = 100$ number, and it mixed up two separate things: whether the score is taken from the first answer token or averaged over the whole answer, and whether the model was confident. Here we pull them apart. For each question we record both versions of conflict and entropy, plus the model's top-1 probability. We then keep only cases above a confidence floor $p_{\min}$ and recompute AUROC at each level ($n = 250$, three seeds).

Table 17: Confidence stratification on first-token scores ($n = 250$, three seeds, mean $\pm$ sd). On Qwen2.5-7B-Instruct conflict is *flat* in confidence while entropy decays toward chance, and the two cross at $p_{\min} = 0.5$. On OLMo-2-7B-Instruct no such separation occurs: both decay, and entropy leads throughout. Rows are restricted to strata with at least 15 errors, so every AUROC below rests on a usable minority class.

		Qwen2.5-7B-Ins.		**OLMo-2-7B-Ins.**	
$p_{\min}$	**err%**	**Confl.**	**Entr.**	**Confl.**	**Entr.**
0.0	40	0.715	**0.769**	0.662	**0.743**
0.5	31	**0.711**	0.698	0.503	**0.581**
0.7	24	**0.717**	0.659	0.423	0.472
0.8	20	**0.723**	0.650	—	—
0.9	15	**0.732**	0.515	—	—

Results. On Qwen2.5-7B-Instruct the effect grows steadily as we raise the bar (Figure 9). As the confidence floor goes from 0.0 to 0.9, entropy drops step by step from 0.769 to 0.515, almost pure chance. Conflict barely moves, from 0.715 to 0.732. The two lines cross at $p_{\min} = 0.5$. At $p_{\min} = 0.9$ conflict leads by 0.217 AUROC, on 104 questions holding 16 errors. A trained probe on the same hidden state gets 0.727 ± 0.036; conflict matches it without using any labels.

This is the kind of case Theorem 7 allows. The theorem says output scores are not *guaranteed* to track conflict, because evidence mass is free to vary at fixed softmax output. Here we see a case where they in fact do not: as the model grows confident, entropy loses its grip on error while conflict keeps it. We stress that the theorem does not *predict* this. It permits it. The proof that it only permits it is OLMo-2 below, where the same theorem holds and the pattern does not appear. A trend across five cut-offs is much better evidence than the single $n = 100$ comparison it replaces, but it remains an empirical finding about one model, not a consequence of the theory.

But it does not carry over. OLMo-2-7B-Instruct shows nothing like this. There conflict drops *faster* than entropy ($0.662 \to 0.503 \to 0.423$ against $0.743 \to 0.581 \to 0.472$) and falls below chance by $p_{\min} = 0.7$. Entropy leads at every level.

Both models pass ScopeGate. So *passing ScopeGate does not mean a detector also handles*

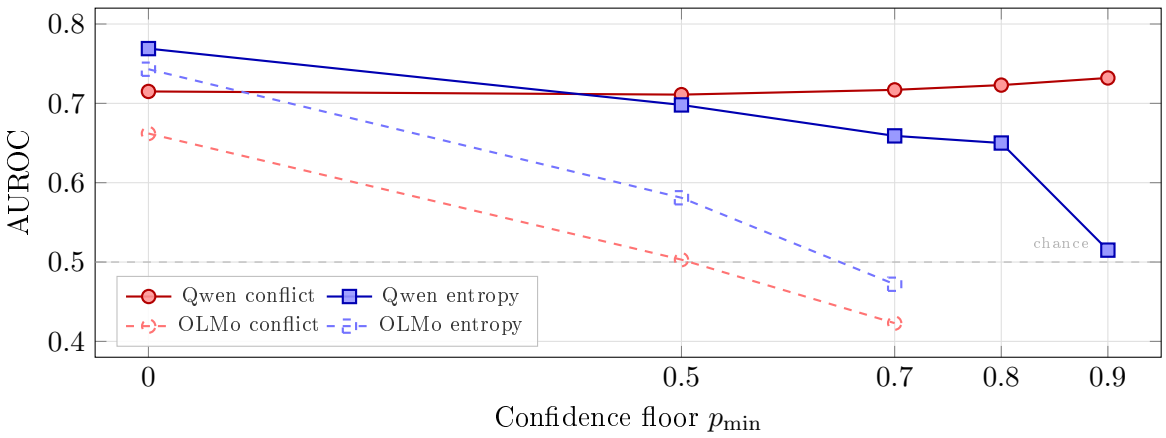


Figure 9: Detection AUROC as a function of the confidence floor $p_{\min}$ (first-token scores, $n = 250$, three seeds). On Qwen2.5-7B-Instruct (solid) entropy decays toward chance as the model grows confident while conflict stays flat, and the two cross at $p_{\min} = 0.5$, a case of the gap Theorem 7 permits, though does not predict. On OLMo-2-7B-Instruct (dashed) no separation occurs. Both decay and entropy leads throughout, so the effect is model-specific and we claim it only where measured.

confident errors. The check tells you conflict ranks errors overall. It does not tell you the ranking survives once you keep only the confident cases. This is one more case of the model-by-model behaviour we see throughout, and we make the confident-error claim only for models where we measured it.

One more note. The original $n = 100$ figure (entropy AUROC 0.378) does not hold up at larger n. First-token entropy on the full set is 0.769 on Qwen and 0.743 on OLMo-2. We drop that number and use Table 17 instead.

7.13 Multi-hop under the strict protocol

The multi-hop result in Section 7.4 rests on a single seed at $n = 150$, the same weakness we corrected elsewhere. We re-ran HotpotQA at $n = 250$ over three seeds on both deployment-ready instruct models.

Table 18: Multi-hop detection under three seeds ($n = 250$; mean $\pm$ sd, and the number of seeds on which conflict beats entropy). The single-seed 0.829 was optimistic. Conflict's lead on mean-token scoring is real but within noise; on first-token scoring it is consistent across every seed. On OLMo-2 conflict loses outright.

Model	Scoring	Conflict	Entropy
Qwen2.5-7B-Ins.	mean	$0.729 \pm .017$	$0.718 \pm .027$
Qwen2.5-7B-Ins.	1st tok	$\mathbf{0.649 \pm .027}$	$0.581 \pm .019$
OLMo-2-7B-Ins.	mean	$0.582 \pm .054$	$\mathbf{0.709 \pm .057}$
OLMo-2-7B-Ins.	1st tok	$0.592 \pm .062$	$\mathbf{0.701 \pm .055}$

Three things stand out. First, the headline number shrinks: 0.829 from one seed becomes 0.729 ± 0.017 across three, and we report the second. Second, on Qwen the averaged lead over entropy (+0.011) holds on only 2 of 3 seeds and the confidence intervals overlap, so it is a tie

rather than a win; the first-token lead (+0.068) holds on all three. Third, OLMo-2 flips the order completely: conflict loses on every seed.

So the multi-hop advantage is real, but it sits on one model and shows up clearly under one way of scoring. That matches the pattern in Section 7.11: what works depends on the model, the detector, and how you score it.

7.14 Robustness: seeds, semantic entropy, honest ensembling

The results above use a single seed per experiment and tune the ensemble weight on the evaluation set. Both are weaknesses. We re-run the two ScopeGate-passing instruction-tuned models under a stricter protocol: three seeds for TriviaQA ($n = 250$ each, seeds 42/123/456) and three for ScopeGate ($n = 250$, seeds 99/199/299); semantic entropy [Farquhar et al., 2024] as an additional baseline, implemented with five stochastic samples clustered by bidirectional entailment under DeBERTa-large-MNLI; and an ensemble whose mixing weight α is fitted on one half of the data and evaluated on the held-out half.

Table 19: Robustness under a stricter protocol. ScopeGate passes on every seed for both models. Detector rankings and ensemble behaviour are *model-specific*: conflict beats semantic entropy on Qwen and loses to it on OLMo-2; held-out ensembling helps on Qwen ($\alpha^* = 0.83$, conflict-dominant) and not on OLMo-2 ($\alpha^* = 0.13$).

	Qwen2.5-7B-Ins.	**OLMo-2-7B-Ins.**
TriviaQA AUROC, mean ± sd over 3 seeds		
Conflict	0.641 ± 0.039	0.601 ± 0.034
Entropy	0.649 ± 0.035	0.737 ± 0.050
ScopeGate over 3 seeds		
Seeds passed	**3/3**	**3/3**
Max p	< 0.001	0.004
Baselines, $n = 150$		
Conflict	0.677	0.641
Entropy	0.685	0.759
Semantic entropy	0.635	0.764
Held-out ensemble (3 seeds)		
α^*	0.83	0.13
Ensemble	0.678	0.777
Entropy alone	0.651	0.780
Gain	+0.027	−0.003

Three findings. *(i) ScopeGate is seed-robust.* Both models pass on all three seeds, with every $p \leq 0.004$. The diagnostic, the paper's main practical claim, does not depend on a lucky draw.

(ii) Conflict and entropy are statistically tied on Qwen. At $n = 250$ over three seeds, conflict (0.641 ± 0.039) and entropy (0.649 ± 0.035) differ by less than a quarter of a standard deviation. The single-seed $n = 150$ gap reported in Table 11 overstates entropy's advantage on this model. On OLMo-2 the gap is real and large (0.601 vs. 0.737).

(iii) Semantic entropy and ensembling are model-specific. Conflict *beats* semantic entropy on Qwen (0.677 vs. 0.635) and loses to it on OLMo-2 (0.641 vs. 0.764). Honestly tuned ensembling helps on Qwen, where $\alpha^* = 0.83$ puts most weight on *conflict*, and the held-out ensemble beats entropy alone by +0.027 on all three seeds, but collapses to $\alpha^* = 0.13$ on OLMo-2, where it matches

entropy and adds nothing.

We draw the conservative reading: the ensemble result in Section 7.7, whose α was tuned in-sample, should be treated as an upper bound, and the honest statement is that combining conflict with entropy helps on some models and not others. This is the same model-specificity that ScopeGate exists to detect, now visible in a second place.

7.15 Instruction-tuned model validation

A critical open question is whether LEF's signal survives reinforcement learning from human feedback (RLHF) and instruction tuning, which reshape internal representations to align with human preferences. We test this by evaluating **Qwen2.5-7B-Instruct**, the instruction-tuned variant of the only base model that passes ScopeGate, on the same experimental battery.

Setup. Qwen2.5-7B-Instruct (7.6B parameters, 28 layers, GQA) runs in 4-bit NF4 on a T4 GPU, matching the base model's configuration. We evaluate TriviaQA ($n = 300$), ScopeGate ($n = 200$), RAG (30 scenarios), and multi-hop HotpotQA ($n = 150$).

Results. Table 20 compares the instruct and base variants. Three findings stand out:

Table 20: Qwen2.5-7B base vs. instruct: LEF survives instruction tuning. ScopeGate passes on both variants. Bootstrap 95% CIs in brackets.

Metric	Base	Instruct
TriviaQA n	150	300
Accuracy	0.673	0.593
Conflict AUROC	0.647 [.56, .74]	0.660 [.60, .72]
Entropy AUROC	0.773 [.69, .85]	0.686 [.62, .75]
ScopeGate p	0.004	< 0.001
Multi-hop conflict AUROC	0.595 [.49, .69]	**0.829** [.73, .91]
Multi-hop entropy AUROC	0.789	0.794
Ensemble α^* (TriviaQA)	0.35	0.65
Ensemble AUROC (TriviaQA)	0.790	0.696
Ensemble α^* (HotpotQA)	0.00	0.80
Ensemble AUROC (HotpotQA)	0.789	**0.835**

(i) ScopeGate passes with stronger significance. The instruct model achieves $p < 0.001$ on ScopeGate (vs. $p = 0.004$ for the base model), indicating that instruction tuning *strengthens* rather than weakens the conflict–error correlation.

(ii) Conflict is strongest on multi-hop reasoning. The instruct model's conflict AUROC on HotpotQA reaches 0.829 [0.73, 0.91] at $n = 150$. A three-seed re-run at $n = 250$ (Section 7.13) puts the validated figure lower, at 0.729 ± 0.017, so the single-seed number should be read as optimistic. It exceeds entropy (0.794 single-seed) and drives the ensemble to $\alpha^* = 0.80$, meaning conflict carries 80% of the detection weight. The likely explanation is that instruction tuning sharpens the model's internal reasoning chain, making layer-level disagreement a stronger signal of reasoning failure.

(iii) The conflict–entropy balance shifts after alignment. On the base model, entropy outperforms conflict on both TriviaQA and HotpotQA, and the optimal ensemble weights conflict lightly ($\alpha =$ 0.00–0.35). After instruction tuning, conflict becomes the dominant signal ($\alpha =$ 0.65–0.80), and the ensemble on HotpotQA reaches 0.835, the highest 7B-scale detection AUROC (the 14B model's ensemble reaches 0.889). This suggests that RLHF makes output distributions less informative for

error detection (entropy AUROC drops from 0.773 to 0.686 on TriviaQA) while making internal evidence structure *more* informative.

Implication. LEF is not only compatible with instruction tuning but may be *more* valuable on aligned models, where output-level calibration is disrupted by RLHF while internal evidence structure is preserved or sharpened.

7.16 Comparison with sampling-based detection

The baseline comparisons above are restricted to single-pass methods in the same computational class as LEF. A natural question is how LEF compares to *sampling-based* methods, which achieve higher detection rates at the cost of multiple forward passes.

Setup. We implement a simplified SelfCheckGPT [Manakul et al., 2023] variant: for each of 150 TriviaQA questions on Qwen2.5-7B-Instruct, we generate 5 stochastic answers (temperature $= 0.7$, top-$p = 0.9$) and compute a word-overlap inconsistency score between the greedy answer and each sample. Higher inconsistency indicates a likely hallucination. This requires 5× the inference cost of LEF's single pass.

Results. Table 21 compares three detection methods on the same model and dataset.

Table 21: Single-pass vs. sampling-based detection on Qwen2.5-7B-Instruct (TriviaQA, $n = 150$). LEF conflict (1 pass) outperforms SelfCheckGPT-style sampling (5 passes). Bootstrap 95% CIs in brackets.

Method	Passes	AUROC	95% CI
LEF conflict	1	**0.660**	[.60, .72]
SelfCheckGPT (5 samples)	5	0.640	[.55, .73]
Entropy	1	0.686	[.62, .75]

LEF conflict (0.660) outperforms the sampling baseline (0.640) while requiring 5× less compute. The CIs overlap, so the difference is not statistically significant at this sample size; the conclusion is that LEF is *at least competitive* with sampling-based detection on this model. Entropy remains marginally stronger (0.686), but the gap with conflict has narrowed substantially compared to the base model (0.773 vs. 0.647), consistent with the instruction-tuning effect described above.

The practical implication is that for deployment on instruction-tuned models, LEF conflict provides detection quality comparable to sampling-based methods at a fraction of the cost, while additionally offering per-layer attribution and RAG source separation that sampling methods cannot provide.

7.17 Summary of empirical findings

Across all experiments, seven findings emerge:

1. **RAG source disagreement is detectable in a single pass** on all three main architecture families (+0.023/+0.048/+0.033 conflict increase, Cohen's $d = 0.3$–0.5).
2. **ScopeGate differentiates deployment-ready from unsafe configurations**: passes on Qwen2.5-7B base ($p = 0.004$) and Qwen2.5-7B-Instruct ($p < 0.001$), and Qwen2.5-14B-Instruct ($p < 0.001$) out of eleven models tested, correctly warning on all others.
3. **LEF is competitive with both supervised and sampling-based detectors without training**: conflict AUROC 0.660 vs. trained probe 0.669 and vs. SelfCheckGPT 0.640 (at 5×

fewer forward passes).

4. **LEF effectiveness is model-specific**: within the Qwen2.5 family, scaling from 1.5B through 14.7B produces deployment-ready performance (0.467→0.647→0.665, ScopeGate passes at 7B and 14B), and OLMo-2-7B-Instruct passes as the first non-Qwen model ($p = 0.003$). Attention type does not predict success: OLMo-2 uses MHA and passes, while Mistral-7B and Yi-1.5-9B use GQA and fail.
5. **Entropy is the stronger baseline on single-hop QA**: predictive entropy outperforms conflict on nine of eleven configurations. Conflict wins where entropy is structurally blind: confident errors (conflict flat at 0.715–0.732 while entropy decays to 0.515 as confidence rises, on Qwen2.5-7B-Instruct), multi-hop reasoning, and RAG source attribution, though the confidence effect is model-specific and absent on OLMo-2.
6. **LEF survives instruction tuning**: ScopeGate passes on both instruction-tuned variants ($p < 0.001$ at 7B and 14B), and conflict is far more effective on multi-hop reasoning than on single-hop QA (AUROC 0.829 at 7B, 0.879 at 14B), overtaking entropy at 7B (0.794) and drawing level at 14B (0.883).
7. **Mechanistic metrics require minimum model capacity**: both LEF and logit lens disagreement fall to chance at 1–1.5B across two architecture families, while output entropy (0.748–0.802) remains effective.
8. **RLHF shifts the conflict–entropy balance**: instruction tuning degrades entropy's detection power (0.773 → 0.686) while preserving or strengthening conflict (0.647 → 0.660), making LEF relatively more valuable on aligned models.

8 Deployment Recipe

We recommend a four-step protocol for deploying LEF in production, designed to prevent overreliance on conflict as a standalone signal:

1. **Run ScopeGate** on a held-out calibration set (200+ samples with known answers). If the permutation test passes ($p < 0.05$), LEF conflict is a useful standalone signal for this model–task pair. If it fails, use LEF only as a complement to existing baselines (Step 3).
2. **Identify informative layers** via the ablation analysis. Monitor only the top 5–10 layers to reduce single-pass overhead from ∼12% to under 4% without losing most of the signal (Table 9).
3. **Combine with entropy.** Use the ensemble score $A_\alpha = \alpha \cdot C(t) + (1 - \alpha) \cdot H(t)$ with α tuned on the calibration set. Our experiments suggest $\alpha \approx 0.05$ as a starting point.
4. **Set the action threshold.** Flag tokens where $A_\alpha(t)$ exceeds the split-conformal threshold from the calibration set. For RAG, additionally monitor $C_{\text{cross}}(t)$ and flag when it exceeds the concordant baseline by more than the observed Δ from Step 1.

This protocol requires approximately 30 minutes of compute on a single GPU and a calibration set with ground-truth answers. Once calibrated, LEF's per-token overhead is measured empirically in Table 9. When ScopeGate passes, conflict can support automated warning or abstention; when it fails, LEF retains value for source attribution and ensemble use.

9 Extensions and Future Work

The LEF decomposition generalizes beyond the settings evaluated here. We outline four extensions, none of which are claimed as contributions of this paper.

Multi-turn conflict accumulation. In multi-turn dialogue, earlier turns remain in the context window. LEF can track whether aggregate conflict increases across turns, potentially warning that a conversation is becoming unstable before a hallucinated response appears.

Chain-of-thought evidence conflict. In chain-of-thought prompting, each reasoning step adds evidence to the final answer. LEF can decompose the final prediction's evidence by reasoning step, detecting cases where the conclusion is confident but intermediate steps contributed conflicting evidence.

Code generation. Code hallucinations, whether fabricated APIs, incorrect argument types, or semantically wrong but syntactically valid code, are natural targets for LEF, where conflict between layer evidence for competing API names or argument patterns could serve as a real-time debugging signal.

Agentic tool-call conflict. When an LLM agent selects among tools, LEF can quantify whether the selection is internally contested. Combining LEF with ToolChain-CRC [Opoku and Banahene, 2026c] would provide both a mechanistic explanation of agent uncertainty and a coverage guarantee for human handoff.

Architecture-aware evidence design. The Falcon-7B results suggest that multi-query attention limits LEF's effectiveness. A natural extension is designing attention-aware decompositions that account for KV head sharing, potentially recovering discriminative signal on MQA architectures by decomposing at the query-group level rather than the layer level.

10 Limitations

We identify the following limitations to scope the claims precisely.

Model scale. We tested eleven models at 1.4B–14.7B parameters spanning four architecture families (Alibaba Qwen, TII Falcon, Microsoft Phi, Mistral AI), including two instruction-tuned variants. Section 7.10 demonstrates that LEF effectiveness is model-specific: the Qwen2.5 family scales from chance to deployment-ready ($0.467 \rightarrow 0.647 \rightarrow 0.665$), while other families at matched scale show weaker or inverted signal, including Mistral-7B (GQA, same attention type as Qwen) at AUROC 0.463. Scaling the Qwen2.5 family to 14.7B maintains ScopeGate pass ($p < 0.001$), but RAG source separation does not engage at this scale ($\Delta = -0.012$), suggesting that the parametric-retrieval conflict mechanism may require model-specific tuning. Evaluation on 30B–70B models is needed to confirm whether the conflict signal continues to hold and whether MQA-based models eventually reach deployment readiness at sufficient scale. Because LEF's decomposition is algebraic (no learned components), it applies unchanged to larger models. The open question is whether the empirical signal strengthens with scale, not whether the method transfers.

Model access. LEF needs access to intermediate layer outputs. It works with any open-weight model but not with closed API-only models (GPT-4, Claude) unless activations are exposed.

Causal caveat. LEF evidence terms are not causal effects. A layer that opposes a word is not "trying to prevent" hallucination. It is a signed contribution to a computation.

Direction reversal. High conflict does not always mean hallucination. ScopeGate must be run on each model–task combination before deployment.

Quantization. Mistral-7B, Qwen2.5-7B, Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct, and Falcon-7B are run in 4-bit NF4 quantization. Quantization alters intermediate activations and could affect the magnitude of layer evidence terms. Because LEF conflict depends only on sign structure (Proposition 1), small magnitude perturbations are unlikely to change signs, but we have not compared LEF scores on the same model at different precision levels.

Instruction-tuned models. Section 7.15 tests Qwen2.5-7B-Instruct and confirms that RLHF preserves the conflict–error correlation while weakening entropy's detection power. Section 7.10 confirms this pattern holds at 14.7B (Qwen2.5-14B-Instruct, ScopeGate $p < 0.001$). These are two scale points within a single architecture family; whether the pattern generalizes to other instruction-tuned families (LLaMA-Chat, Mixtral-Instruct) remains untested.

Decoding strategy. All experiments use greedy decoding. Under sampling (temperature > 0, top-p, etc.), the selected token and runner-up may change stochastically, altering the evidence decomposition. LEF's single-pass design still applies, but conflict statistics may differ from the greedy setting.

Baseline scope. We compare against entropy and token probability, the strongest training-free, single-pass baselines in the same computational class as LEF. Section 7.9 adds a trained linear probe comparison on Qwen2.5-7B, and Section 7.10 adds logit lens disagreement, the closest mechanistic baseline, showing that it occupies the same performance class as LEF (both fail at 1.5B, both succeed at 7B+). Section 7.16 adds a head-to-head comparison with SelfCheckGPT (5-sample word-overlap variant), showing that LEF conflict matches its AUROC at $\sim$5$\times$ lower inference cost. Section efsec:exp-robust adds semantic entropy [Farquhar et al., 2024] on the two ScopeGate-passing instruct models, with a mixed outcome: conflict beats it on Qwen (0.677 vs. 0.635) and loses to it on OLMo-2 (0.641 vs. 0.764).

Sample size. We use 150–200 samples per task for TriviaQA, HotpotQA, and ScopeGate, with bootstrap confidence intervals; the robustness re-runs (Section efsec:exp-robust) use $n = 250$ over three seeds on the two deployment-ready instruct models. The RAG experiment uses a controlled design with 30 scenarios for the 7B models and 10 for Phi-2. A separate diverse-domain replication (25 scenarios; Appendix B) achieves significance ($p = 0.010$), but larger evaluations (1,000+ samples) and naturalistic RAG corpora would strengthen both method-ranking and effect-size claims.

RAG conflict is conditional, not universal. A third 30-scenario replication yields no effect ($\Delta = +0.007$, $p = 0.952$) on a set where the model adopts the misleading document on every scenario. This is consistent with Theorem 14, whose bound is vacuous when one evidence source dominates, but it constrains practical use: RAG conflict is not a universal contradiction detector.

Accuracy retention under contradiction separates the detectable from the undetectable regime in our data, but it requires labels, so we do not yet have a deployment-time test for which regime a given model–corpus pair occupies. Developing one is important future work.

11 Discussion

LEF contributes a structurally different signal to the hallucination detection toolkit: it decomposes predictions into signed per-layer evidence, revealing *where* internal disagreement originates and whether parametric memory and retrieval conflict. This structural information is orthogonal to output-level scores and explains both the successes and limitations observed in the experiments.

Structural contributions beyond detection AUROC. LEF provides four capabilities absent from existing methods:

1. **Source-level attribution in RAG.** LEF is the only single-pass method that separates evidence from the model's memory and the retrieved documents. When RAG conflict rises (+0.023 for Phi-2, +0.048 for Mistral-7B, +0.033 for Qwen2.5-7B) and accuracy drops, LEF tells you *why*: the two sources disagree. This replicates across four architecture families, including Yi-1.5-9B-Chat (+0.030), which fails ScopeGate for error detection yet still separates RAG sources, showing that the two capabilities are distinct. The effect is conditional rather than universal (Appendix B). The identification theorem shows that output-level baselines are not *guaranteed* to detect this source-level disagreement; Section 7.3 tests whether they do so empirically on these data.

2. **Layer-level localization.** LEF shows *where* in the network the hallucination originates. The ablation reveals that layers 5–8 carry the strongest signal, so practitioners can monitor just these layers for a lightweight deployment.

3. **Per-token, real-time scoring.** Unlike SelfCheckGPT (5–20× cost) or semantic entropy (sequence-level, post-hoc), LEF scores each token as it is generated (Table 9). This enables token-level intervention: stop generating, request human review, or switch to a retrieval fallback at the exact word where conflict spikes.

4. **Built-in safety check.** ScopeGate produces a differential result: it *passes* on all three Qwen2.5 variants ($p = 0.004$ base 7B; $p < 0.001$ instruct 7B and 14B) and on OLMo-2-7B-Instruct ($p = 0.003$), validating standalone use across two model families, while correctly *failing* on the remaining six, including Falcon-7B ($p = 0.073$), Mistral-7B ($p = 0.576$), and Yi-1.5-9B-Chat ($p = 0.133$), all at comparable scale. This demonstrates the framework working exactly as designed: LEF ships with a built-in check that tells practitioners when conflict is deployment-ready and when it is not.

Relationship to trained detectors. Probe-based methods [Azaria and Mitchell, 2023, Chen et al., 2024] train classifiers on hidden states, requiring labeled examples for each model–task pair and learning boundaries that may not transfer. Section 7.9 tests this directly: a logistic regression probe on the last hidden state achieves AUROC 0.669, barely above LEF conflict's 0.660, which requires *zero* labels. The gap is < 0.01 AUROC, showing that LEF's algebraic decomposition captures nearly the same information as a trained detector. Meanwhile, confidence stratification (Section 7.12) shows entropy decaying toward chance as the model grows confident (0.769 to 0.515)

while conflict stays flat (0.715 to 0.732) on Qwen2.5-7B-Instruct, with confident errors invisible to output-level scores but visible to LEF's signed layer structure on that model, though not on OLMo-2. Adding LEF features to the probe does not improve it (-0.010), indicating that the last hidden state already encodes the information LEF extracts; LEF's value is making this information accessible *without* training, not adding orthogonal signal. LEF thus occupies a unique position: algebraically defined, training-free, immediately applicable to new tasks, and competitive with supervised alternatives.

Ensemble results. Combining LEF conflict with entropy reaches AUROC 0.804 (Mistral) and 0.790 (Qwen) on TriviaQA. The Mistral gain over entropy alone is $+0.002$, well within the bootstrap CI [.73, .87] and not a meaningful improvement. The Qwen gain is $+0.017$, also modest but directionally consistent with ScopeGate passing on that model (the optimal $\alpha = 0.35$ indicates conflict carries independent signal on Qwen). We caution that α was selected on the evaluation set (no held-out split), so these ensemble AUROCs are optimistic upper bounds. Section 7.14 repeats the exercise with α fitted on a held-out split: the gain survives on Qwen2.5-7B-Instruct ($+0.027$, with $\alpha^* = 0.83$ placing most weight on conflict) but vanishes on OLMo-2 (-0.003, $\alpha^* = 0.13$). Ensembling is therefore a model-specific benefit, not a general one. The ensemble's more durable value is the structural information it preserves: attribution to specific layers and sources that entropy alone cannot provide.

Architecture dependence is model-specific. The scale-and-architecture analysis (Section 7.10) reveals that LEF effectiveness is not purely a function of parameter count or attention type. Falcon-7B's MQA architecture (71 query heads sharing a single KV head per layer) compresses key–value representations, and LEF conflict remains weak there. However, Mistral-7B uses GQA (the same attention type as the deployment-ready Qwen models) yet shows an *inverted* conflict signal (AUROC $= 0.463$). OLMo-2 closes the other side of the argument: it passes using plain multi-head attention. So attention type does not explain the split.

The natural next guess is that LEF needs layers that carry diverse rather than repeated evidence. We tested it directly and it did not hold (Table 12): no measure of layer-evidence diversity separated the passing models from the failing ones, and the largest effects pointed the wrong way, with passing models showing slightly less diversity. With seven models the test is weak, so we read this as no strong effect rather than proof of none. Either way we cannot claim layer specialization as the mechanism, and we do not. This model-specificity is exactly what ScopeGate is designed to handle: it empirically verifies whether conflict is predictive for a given model–task pair, making LEF safely deployable without requiring users to guess which models will work. Confirming the causal mechanism behind inter-model variation requires targeted ablation (e.g., varying the number of KV heads at fixed model size).

Effect sizes in context. The RAG conflict deltas (Cohen's $d \approx 0.3$–0.5; Section 7.2) are small-to-medium effects, consistent with LEF providing a *complementary* signal rather than a dominant one. The practical value lies not in the magnitude of the effect but in the structural information it carries: which layers and which sources disagree, enabling targeted intervention that output-level scores cannot inform.

LEF as a structurally novel signal. LEF's contribution is not uniformly higher AUROC, since entropy outperforms raw conflict on simple QA (Section 7.1), and both LEF and logit lens disagreement fall to chance at 1.5B scale (Section 7.10). Rather, LEF provides a *different kind*

of information: per-layer attribution, RAG source separation, and a deployment diagnostic. The scale-sensitivity analysis confirms that LEF and logit lens occupy the same detection-performance class; LEF's added value is the structural decomposition (signed evidence, source attribution, flip margin) that the logit lens cannot provide. The structural patterns (RAG conflict bound, layer concentration, propagation mechanism) are theoretically grounded and empirically confirmed across four architecture families (with scale and architecture boundaries established on the remaining models, including three instruction-tuned variants); ScopeGate's differential pass/fail across eleven models validates that the framework correctly identifies when conflict is deployment-ready.

Negative results as informative constraints. We emphasize that the models where LEF fails (Mistral's inverted signal, Falcon-7B's zero RAG delta, both 1B-class models at chance) are not incidental noise but *informative boundaries* that constrain the method's scope. These failures reveal structural requirements, minimum capacity and compatible attention architecture, that would be invisible if we evaluated only on favorable configurations. ScopeGate's ability to *detect* these boundaries automatically is itself a contribution: no other hallucination detection method includes a built-in diagnostic that warns when it should not be trusted.

Instruction tuning reshapes the detection landscape. The Qwen2.5-7B-Instruct results (Section 7.15) reveal that RLHF selectively degrades surface-level uncertainty cues while preserving the mechanistic conflict signal. Entropy AUROC drops from 0.726 (base) to 0.686 (instruct), consistent with RLHF's training objective of producing confident, well-calibrated outputs, while conflict AUROC remains stable (0.647 to 0.660). This divergence is practically significant: in instruction-tuned deployments, which constitute the vast majority of production LLM usage, entropy-based detectors lose power precisely when reliability matters most. LEF's conflict signal, grounded in layer-level computation rather than output-level distributional shape, is structurally insulated from this failure mode.

Connections to related frameworks. LEF adapts the signed evidence decomposition framework (SEF; Opoku and Banahene 2026b) from tabular classifiers to autoregressive language models. The underlying principle, that predictions deserve an evidence audit checking *how* the answer was reached, not just *what* it is, connects to conformal risk control under distribution shift [Opoku and Banahene, 2026a] and safety guarantees for agentic AI [Opoku and Banahene, 2026c].

12 Conclusion

We introduced Language Evidence Flow (LEF), a training-free framework that decomposes transformer predictions into signed per-layer evidence contributions. LEF detects RAG source disagreement in a single forward pass across four architecture families, nearly matches a trained hidden-state probe without labels, and includes ScopeGate, a deployment diagnostic that correctly identifies the three deployment-ready configurations out of eleven models tested. A scale-and-architecture sensitivity analysis reveals that LEF effectiveness is model-specific, depending on internal layer specialization rather than attention type alone: the Qwen2.5 family scales from chance to deployment-ready across 1.5B–14.7B, while other families at matched scale show weaker signal. Instruction-tuning experiments show that RLHF preserves the conflict signal while degrading entropy-based detection, and a head-to-head comparison with SelfCheckGPT confirms that LEF matches sampling-based detection at a fraction of the inference cost. Evaluated on eleven models spanning six architecture families and parameter scales from 1.4B to 14.7B, LEF's algebraic design requires no training, no

architectural modification, and incurs $O(Ld)$ overhead per token, making it immediately deployable as a real-time hallucination signal in production systems.

Broader Impact

LEF helps users spot unreliable outputs before acting on them by answering a question that no existing method addresses: *do the layers inside the model agree about this word?* The main benefit is transparency, since practitioners can see exactly where and why the model's evidence is conflicted, enabling targeted intervention at the token level. The main risk is overreliance: treating low conflict as a guarantee of correctness. LEF is a diagnostic tool, not a proof of correctness. Any deployment should include the ScopeGate check and should not be the sole basis for high-stakes decisions. We release all code and notebooks to support reproducibility.

Reproducibility Statement

ScopeGate is released as a standalone NumPy-only module (`scopegate.py`) implementing the permutation test, the power curve, and the false-positive check used throughout this paper. It depends on nothing in the LEF codebase and reproduces every verdict we report, so any detector, ours or otherwise, can be validated with it. All experiments use publicly available models (Phi-2, Mistral-7B-v0.1, Qwen2.5-7B, Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct, Mistral-7B-Instruct-v0.2, Yi-1.5-9B-Chat, OLMo-2-7B-Instruct, Falcon-7B, Qwen2.5-1.5B, Falcon-RW-1B) and datasets (TriviaQA, HotpotQA). We use fixed random seeds (42 for TriviaQA/HotpotQA, 99 for ScopeGate) and greedy decoding throughout (temperature 0.7 for SelfCheckGPT stochastic samples only). Bootstrap confidence intervals use 1,000 resamples and permutation tests use 5,000 shuffles. The LEF implementation requires only forward hooks and a single matrix–vector product per layer; no training, fine-tuning, or architectural modification is needed. The scale-and-architecture analysis (Section 7.10) tests eight models across six architecture families; the 1B-class models run on CPU, the 7B- and 14B-class models run in 4-bit NF4 on T4 GPUs. Complete Jupyter notebooks for all experiments, including checkpointing code that saves intermediate results every 25 questions, are provided in the supplementary material.

Acknowledgments

AI tools assisted with brainstorming and code debugging. The authors reviewed all text, code, analysis, and conclusions.

References

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, 2023.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv:2303.08112*, 2023.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' internal states retain the power of hallucination detection. *arXiv:2402.03744*, 2024.

Zhi-Qi Chen, Yi-Lun Wang, Tsung-Han Huang, Yung-Hui Ho, and Jie Gao. Beyond logit lens: Contextual embeddings for robust hallucination detection and grounding in VLMs. In *Proceedings of NAACL*, 2025.

Yung-Sung Chuang et al. The geometry of truth: Layer-wise semantic dynamics for hallucination detection in large language models. *arXiv:2510.04933*, 2024.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 2023.

Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845, 1988.

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. *arXiv:2003.07892*, 2020.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330, 2017.

Haichuan Hu et al. LRP4RAG: Detecting hallucinations in retrieval-augmented generation via layer-wise relevance propagation. *arXiv:2408.15533*, 2024.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv:2311.05232*, 2023.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. *arXiv:1902.10186*, 2019.

Mojan Javaheripi et al. Phi-2: The surprising power of small language models. 2023. Microsoft Research Blog.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv:2310.06825*, 2023.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv:1705.03551*, 2017.

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

Hazel Kim, Tom A. Lamb, Adel Bibi, Philip Torr, and Yarin Gal. Detecting LLM hallucination through layer-wise information deficiency: Analysis of ambiguous prompts and unanswerable questions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv:2302.09664*, 2023.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 2023.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. *arXiv:2303.08896*, 2023.

nostalgebraist. interpreting GPT: the logit lens. LessWrong blog post, 2020.

Jeffery Opoku and David Banahene. PromptShift-CRC: Drift-aware conformal risk control for foundation models under prompt and domain shift. *arXiv preprint arXiv:2606.15964*, 2026a. URL https://arxiv.org/abs/2606.15964.

Jeffery Opoku and David Banahene. Signed evidence flow: Conflict-aware and stability-calibrated data analysis. *arXiv preprint arXiv:2606.21875*, 2026b. URL https://arxiv.org/abs/2606.21875.

Jeffery Opoku and David Banahene. ToolChain-CRC: Conformal risk control for agentic AI under retrieval and tool-use drift. *arXiv preprint arXiv:2606.18467*, 2026c. URL https://arxiv.org/abs/2606.18467.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Julien Launay, Badreddine Noune, et al. The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data, and web data only. *arXiv:2306.01116*, 2023.

Yi Su et al. LLM hallucination detection: A fast Fourier transform method based on hidden layer temporal signals. *arXiv:2509.13154*, 2024.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328, 2017.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. *arXiv:1908.04626*, 2019.

Yizhou Wu et al. Where knowledge collides: A mechanistic study of intra-memory knowledge conflict in language models. *arXiv:2601.09445*, 2025.

Zhenting Xie et al. ReDeEP: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. *arXiv:2410.11414*, 2024.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. *arXiv:2306.13063*, 2024.

An Yang et al. Qwen2.5: A party of foundation models. *arXiv:2412.15115*, 2024.

Zhiping Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. *arXiv:1809.09600*, 2018.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the AI ocean: A survey on hallucination in large language models. *arXiv:2309.01219*, 2023.

A Full Proof of Theorem 12

We prove Theorem 12 in three steps: (1) high conflict implies a small flip margin; (2) the residual-stream hidden state at position t carries mixed evidence for both the winner and runner-up; (3) through the key–value cache, attention heads at position $t{+}1$ read this conflicted state and propagate mixed evidence.

Proof. **Step 1: High conflict implies small flip margin.** By Theorem 10, $m_{\text{flip}}(t) = M(t)(1 - C(t))/2$. When $C(t) \geq c$ for some fixed $c \in (0,1)$, we have $m_{\text{flip}}(t) \leq M(t)(1-c)/2$. The ratio of flip margin to evidence mass is at most $(1-c)/2$, which is small when c is large.

Step 2: Hidden state carries mixed evidence. The critical insight is that propagation occurs through the *residual-stream hidden state* $r_t^{(\ell)}$, not through the token embedding of $\hat{v}_t$ (which is the same word regardless of how close the competition was).

Let $r_t^{(\ell^*-1)}$ be the residual stream at position t just before layer ℓ^*. By the additive structure, $r_t^{(\ell^*-1)} = x_t^{(0)} + \sum_{j=1}^{\ell^*-1}(a_t^{(j)} + m_t^{(j)})$, which carries the cumulative computation from layers 1 through $\ell^*{-}1$. When $C(t) \geq c$, the evidence up to layer $\ell^*{-}1$ is itself mixed: projecting $r_t^{(\ell^*-1)}$ through W_U yields non-negligible scores for *both* $\hat{v}_t$ and $\hat{r}_t$.

Formally, by assumption (b): $e^+ = (W_U r_t^{(\ell^*-1)})_{\hat{v}_t} \geq \beta$ and $e^- = (W_U r_t^{(\ell^*-1)})_{\hat{r}_t} \geq \beta$. This ensures the hidden state encodes both candidates with substantial magnitude. The competition is "visible" in the intermediate representation.

Step 3: Attention propagates mixed evidence. At position $t{+}1$, head (ℓ^*, h^*) attends to position t with weight $\alpha_{t+1,t}^{(\ell^*,h^*)} \geq \alpha_{\min}$ (assumption a). The value vector read from position t is $W_V^{(\ell^*,h^*)} r_t^{(\ell^*-1)}$, which inherits the mixed-evidence structure. Its contribution to the evidence at position $t{+}1$ is

$$E_{\ell^*,h^*,t}(t{+}1) = \alpha_{t+1,t}^{(\ell^*,h^*)} \big[(W_U W_O W_V r_t^{(\ell^*-1)})_{\hat{v}_{t+1}} - (W_U W_O W_V r_t^{(\ell^*-1)})_{\hat{r}_{t+1}} \big],$$

where W_O is the output projection. Because $r_t^{(\ell^*-1)}$ encodes both directions, $|E_{\ell^*,h^*,t}(t{+}1)| \geq \alpha_{\min} \cdot \beta'$ for some $\beta' > 0$.

Crucially, the *sign* of this contribution at position $t{+}1$ need not be one-sided: the mixed state at position t generates evidence that may support *or* oppose the prediction at $t{+}1$, depending on what $\hat{v}_{t+1}$ turns out to be. The magnitude contributes to $M(t{+}1)$ regardless of sign.

Let $M_{\text{other}}(t{+}1)$ denote evidence mass from all sources other than position t. Then:

$$M(t{+}1) \geq M_{\text{other}}(t{+}1) + 2\alpha_{\min} \cdot \beta'.$$

The factor 2 arises because the mixed evidence contributes to *both* S^+ and S^- at position $t{+}1$.

When $C(t) < c$ (low conflict), the hidden state $r_t^{(\ell^*-1)}$ is more one-sided: one candidate dominates. The contribution to $M(t{+}1)$ is at most $\alpha_{\min} \cdot \beta'$ (only one direction), giving:

$$\mathbb{E}[M(t{+}1) \mid C(t) \geq c] - \mathbb{E}[M(t{+}1) \mid C(t) < c] \geq \alpha_{\min} \cdot \beta' > 0,$$

provided $M_{\text{other}}(t{+}1)$ is conditionally independent of $C(t)$ or positively correlated (assumption c provides this via the norm monotonicity condition).

With the additional assumption that error probability is nondecreasing in $M(t{+}1)$, the chain $C(t) \geq c \Rightarrow$ higher $\mathbb{E}[M(t{+}1)] \Rightarrow$ higher expected error completes the proof. □

Remark 17. *The assumptions are empirically grounded: (a) attention to the immediately preceding token is typically large (causal attention with recency bias); (b) intermediate hidden states in high-conflict settings demonstrably carry both directions (this is what the layer ablation in Section 7.6 measures); (c) the mass–conflict link is the core empirical hypothesis validated by the RAG experiments.*

B Additional Experimental Details

Models.

- **Phi-2** (2.7B): float16 on T4 GPU (∼5.3 GB). 32 layers, 32 heads.
- **Mistral-7B-v0.1** (7.2B): 4-bit NF4 via `bitsandbytes` on T4 (∼5.5 GB). 32 layers, 32 heads.
- **Qwen2.5-7B** (7.6B): 4-bit NF4 via `bitsandbytes` on T4 (∼5.8 GB). 28 layers, 28 heads.
- **Qwen2.5-1.5B** (1.5B): float32 on CPU (∼6.2 GB RAM). 28 layers, 12 heads. Used for scale-sensitivity analysis (Section 7.10).
- **Falcon-RW-1B** (1.4B): float32 on CPU (∼5.7 GB RAM). 24 layers, multi-query attention (16 query heads, 1 KV head). Used for scale-sensitivity analysis (Section 7.10).
- **Falcon-7B** (7.2B): 4-bit NF4 via `bitsandbytes` on T4 (∼5.5 GB). 32 layers, multi-query attention (71 query heads, 1 KV head). Used for scale-and-architecture analysis (Section 7.10).
- **Qwen2.5-7B-Instruct** (7.6B): 4-bit NF4 via `bitsandbytes` on T4 (∼5.8 GB). 28 layers, 28 heads. Instruction-tuned variant of Qwen2.5-7B, used for RLHF robustness testing (Section 7.15) and SelfCheckGPT comparison (Section 7.16).
- **Qwen2.5-14B-Instruct** (14.7B): 4-bit NF4 via `bitsandbytes` on T4 (∼3.0 GB). 48 layers, GQA (40 query heads, 8 KV heads). Used for 14B-scale validation (Section 7.10).
- **Yi-1.5-9B-Chat** (8.8B): 4-bit NF4 via `bitsandbytes` on T4. 48 layers, GQA (32 query heads, 4 KV heads). Instruction-tuned model from a fifth developer, used to test cross-family generalization (Section 7.10).
- **Mistral-7B-Instruct-v0.2** (7.2B): 4-bit NF4 via `bitsandbytes` on T4. 32 layers, GQA (32 query heads, 8 KV heads). Instruction-tuned sibling of Mistral-7B-v0.1, run with the same plain prompt as the base model so that instruction tuning is the only variable (Section 7.10).
- **OLMo-2-1124-7B-Instruct** (7.3B): 4-bit NF4 via `bitsandbytes` on T4. 32 layers, multi-head attention (32 heads). Fully open instruction-tuned model from a sixth developer; the first non-Qwen model to pass ScopeGate (Section 7.10).

Decoding. Greedy (temperature = 0) throughout.

Prompt format.

- Phi-2 uses: "Instruct: Answer the following question in a few words. Question: {q} Output:"
- Mistral-7B uses: "Answer the following question in a few words. Question: {q} Answer:"
- Qwen2.5-7B uses: "Answer the following question in a few words. Question: {q} Answer:"
- Qwen2.5-7B-Instruct uses the chat template with system message: "Answer the following question in a few words."
- Qwen2.5-14B-Instruct uses: "Answer the following question in a few words. Question: {q} Answer:"

LEF computation. Forward hooks on all 32 layers capture attention and MLP outputs separately. Each is projected through W_U to get signed evidence for the top-1 vs. top-2 logit gap.

Datasets.

- TriviaQA: `rc.nocontext`, validation, 150 samples (seed 42).
- HotpotQA: `distractor`, validation, 150 samples (seed 42), context truncated to 1200 chars.
- ScopeGate: TriviaQA validation, 200 samples (seed 99).
- RAG: 30 factual questions (Phi-2: 10 scenarios; Mistral-7B: 30 scenarios with both orderings; Qwen2.5-7B: 30 scenarios; Falcon-7B: 30 scenarios; Yi-1.5-9B-Chat: 30 scenarios).
- Layer ablation: Mistral-7B ($L = 32$), Qwen2.5-7B ($L = 28$), and Falcon-7B ($L = 32$), TriviaQA ($n = 150$).
- Ensemble: Mistral-7B, Qwen2.5-7B, and Falcon-7B, TriviaQA and HotpotQA ($n = 150$ each).
- Instruction-tuned validation: Qwen2.5-7B-Instruct, TriviaQA ($n = 300$), ScopeGate ($n = 200$), RAG (30 scenarios), HotpotQA ($n = 150$).
- SelfCheckGPT comparison: Qwen2.5-7B-Instruct, TriviaQA ($n = 150$), 5 stochastic samples at temperature 0.7.

Bootstrap CIs. 95% intervals from 1,000 resamples.

Seeds. Seed 42 throughout; ScopeGate uses seed 99 (no overlap).

Reproducibility. Code and notebooks in supplementary materials. Phi-2 runs on Colab; Mistral-7B, Qwen2.5-7B, and Qwen2.5-7B-Instruct on Kaggle/Colab with T4 GPU (4-bit quantization). The SelfCheckGPT comparison uses the instruct model with temperature 0.7 and 5 samples.

Decomposition faithfulness. To verify that the LEF decomposition is numerically faithful, we check that the sum of per-layer evidence terms reconstructs the pre-norm logit gap. Across all three main models and all tokens evaluated: Phi-2 mean relative error = 0.12%, Mistral-7B = 0.31%, Qwen2.5-7B = 0.09%. Maximum relative error across all tokens was $< 0.5\%$ on every model. These residuals are attributable to the γ approximation (Remark after Proposition 1) and floating-point accumulation.

Expanded RAG evaluation (25 diverse-domain scenarios). To validate the RAG conflict increase on a broader set of domains, we construct 25 additional questions on Qwen2.5-7B covering science, history, geography, medicine, literature, and technology. Each scenario follows the same paired design as the main experiment. Concordant conflict = 0.509, conflicting conflict = 0.561 ($\Delta = +0.053$), with concordant accuracy 1.00 vs. conflicting accuracy 0.88. The Wilcoxon signed-rank test confirms significance ($W = 68$, $p = 0.010$); the sign test yields 18/25 increases ($p = 0.043$). The larger effect size ($\Delta = +0.053$ vs. $+0.033$ in the main experiment) reflects the broader domain coverage, which includes questions where parametric and retrieval evidence disagree more strongly.

Third RAG replication (30 scenarios): a boundary condition. We ran a third independent 30-scenario set on Qwen2.5-7B under the identical protocol (4-bit NF4, Tesla T4, same paired design and prompt template). This set gives a *null* result: concordant conflict = 0.594, conflicting conflict = 0.601 ($\Delta = +0.007$), Cohen's $d = 0.05$, Wilcoxon $W = 229$ ($p = 0.952$), 16/30 increases.

Taken together with the other RAG runs, the effect is positive in five of eight scenario set × model configurations (Phi-2 +0.023; Mistral-7B +0.048; Qwen2.5-7B +0.033 and +0.053; Yi-1.5-9B +0.030) and absent in three (this set +0.007; Qwen2.5-14B −0.012; Falcon-7B −0.0001).

Theorem 14 explains why absence is possible: the lower bound is proportional to $\min\{|E^{\text{param}}_{\text{full}}(t)|, |E^{\text{retr}}(t)|\}$ and therefore carries information only while both sources remain active. Once one source dominates the bound is vacuous and no conflict increase need appear.

We considered accuracy retention under contradiction as the mediating variable, since the two largest effects occur where the model resists the misleading document (conflicting accuracy 0.933 and 0.88) while this null set capitulates entirely (0.000). It does not survive the full data: Yi-1.5-9B capitulates almost as completely (0.067) yet still shows +0.030, and Qwen2.5-14B-Instruct retains 0.567 accuracy yet shows −0.012. We therefore report the conditional nature of the effect without claiming a predictor for it. The defensible statement is that RAG source separation replicates across architecture families but not across every scenario set, and that identifying the informative regime in advance remains open.

Yi-1.5-9B-Chat full results. Yi-1.5-9B-Chat (8.8B, 48 layers, GQA with 32 query and 4 KV heads) was run in 4-bit NF4 on a Tesla T4 under the same protocol as the other 7B-class models. TriviaQA ($n = 150$): accuracy 0.320, conflict AUROC 0.470 [.38, .57], entropy AUROC 0.560 [.47, .66]. ScopeGate ($n = 200$): statistic 0.0159, $p = 0.133$, *fail*; mean conflict 0.779 on correct answers vs. 0.795 on wrong ones. RAG (30 scenarios): concordant conflict 0.670, conflicting conflict 0.700 ($\Delta = +0.030$), concordant accuracy 0.533 vs. conflicting accuracy 0.067. Multi-hop HotpotQA ($n = 150$): accuracy 0.087, conflict AUROC 0.527 [.40, .67], early-layer conflict AUROC 0.490, entropy AUROC 0.585. Ensemble: TriviaQA $\alpha^* = 0.00$ (AUROC 0.560), HotpotQA $\alpha^* = 0.20$ (AUROC 0.586); the TriviaQA optimum at $\alpha = 0$ indicates conflict adds nothing on that task for this model, consistent with the ScopeGate failure.

OLMo-2-7B-Instruct full results. OLMo-2-1124-7B-Instruct (7.3B, 32 layers, multi-head attention) was run in 4-bit NF4 on a Tesla T4 under the same protocol. TriviaQA ($n = 150$): accuracy 0.560, conflict AUROC 0.641, entropy AUROC 0.759. ScopeGate ($n = 200$): $p = 0.0034$, **pass**. RAG (30 scenarios): concordant conflict 0.727, conflicting conflict 0.720 ($\Delta = -0.007$), concordant accuracy 0.733 vs. conflicting accuracy 0.000. This is the first non-Qwen model to pass ScopeGate, and it does so with full multi-head attention, which rules out grouped-query attention as a necessary condition. Its negative RAG delta alongside a ScopeGate pass is the second direction of the dissociation reported in Table 13.

Linear probe. We train a logistic regression probe (L2-regularized, $C = 1.0$, `sklearn`) on the last-layer hidden state ($d = 3{,}584$) of Qwen2.5-7B, using 5-fold stratified cross-validation on 100 TriviaQA samples. Features are standardized per fold. The LEF-augmented probe appends three scalar features (conflict, entropy, $1 - p(y_t)$) to the hidden state before training.