Integrating Local and Global Entropy for Uncertainty Quantification in LLMs

Johanne Medina[1]* **Tianyi Zhou**[2]*
Keivin Isufaj[1] Aristides Gionis[2] Sanjay Chawla[1]
[1]Qatar Computing Research Institute, Hamad bin Khalifa University
[2]KTH Royal Institute of Technology and Digital Futures
{jomedina,keisufaj,schawla}@hbku.edu.qa
{tzho,argioni}@kth.se

Abstract

Large language models hallucinate confidently, making uncertainty quantification (UQ) essential for reliable deployment. Existing methods rely predominantly on token-level signals, leaving the geometric structure of intermediate hidden states underused. In this paper, we take the geometric complexity of hidden-state matrices as a measure of the *global* uncertainty of LLMs, while treating token-level uncertainty estimation as a *local* metric. We show that these signals provide complementary reliability information: token-level uncertainty provides strong overall discrimination, while hidden-state geometry further separates errors within similar local-uncertainty regimes, including responses that appear confident at the token level. Building on this complementarity, we propose **Global-Local Uncertainty (GLU)**, an unsupervised, single-pass score that fuses the two signals via a multiplicative gate. Across three model families and six benchmarks, GLU achieves the best or tied-best AUROC among the three methods in the main comparison in 13 of 18 model–dataset settings, while requiring only a single forward pass and remaining length-normalized and architecture-agnostic.

1 Introduction

Hallucination remains one of the most persistent failure modes of large language models (LLMs). Despite rapid advances in capability, frontier systems continue to produce fluent, specific, and incorrect answers. A recent cross-domain benchmark spanning 6,000 questions across 42 topics highlights a persistent gap between capability and reliability in frontier LLMs [1]. The AA-Omniscience Index awards points for correct answers, deducts points for incorrect answers, and assigns zero to abstentions. In the benchmark's original evaluation, the highest-scoring model achieved 4.8[2] on its $[-100, 100]$ scale; only three evaluated models scored above zero. This gap underscores a central challenge for reliable deployment in which models must not only produce accurate answers, but also recognize when their answers should not be trusted. Uncertainty quantification (UQ) offers a principled path forward, and a wide range of methods have been proposed, from token-level entropy [2] and sampling-based consistency [3, 4] to evidential [5, 6] and attention-based approaches [7, 8]; yet few are deployed in practice. While other methods impose at least one of the following barriers of multiple generation passes, task-specific supervision, sensitivity to response length, or architectural assumptions, we argue that a practical UQ score should be *unsupervised*, computed in a *single forward pass*, *length-normalized*, and *architecture-agnostic*.

*Co-first authors and contributed equally.

[2]This is the highest score reported in the original 2025 study. The live leaderboard changes as new models are evaluated: https://artificialanalysis.ai/evaluations/omniscience.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

The linear representation hypothesis [9] posits that high-level concepts are encoded as directions in the hidden-state space of LLMs. When a model retrieves a well-encoded fact, its hidden states move in a tight, consistent direction implying that the geometry of retrieval is compact. When it lacks reliable factual grounding, the hidden states wander through loosely related directions as the model generates together a plausible-sounding reply. Consider a model asked for the birth year of a figure it has not reliably encoded. It may produce a confident four-digit number while its representations drift through adjacent concepts rather than converging on a specific memory. We observe that this wandering is measurable. Motivated by Skean et al. [8], who show that intermediate hidden layers encode richer representations than the final layer and that information-theoretic geometry metrics effectively characterize representation quality, we quantify this drift via the entropy of eigenvalue distribution from the hidden-state trajectory. This geometric signal is computed directly from the hidden states of greedy generation with no additional forward pass.

Hidden-state representations and next-token predictions are connected through the model's output head, but they are distinct measurements. Park et al. [9] formalize related concept directions in hidden-state representation and token representation spaces using a causal inner product. This motivates examining both representations. This separation provides a principled basis for why token-level entropy and hidden-state geometric complexity measure different things. Token-level entropy operates over next-token distributions, whereas hidden-state geometric complexity operates over semantic directions. The two measures therefore need not agree.

We therefore term hidden-state geometric complexity *global* uncertainty and token-level entropy *local* uncertainty, and ask whether the two provide complementary information about response reliability. Figure 1 provides empirical evidence for this complementarity. Token-level uncertainty provides the primary separation between correct and incorrect responses, but geometry further stratifies reliability within similar local-uncertainty regimes. Among responses below the median token uncertainty, the error rate increases from 17% in the low-geometry region to 28% in the high-geometry region; among responses above the median, it similarly increases from 60% to 74%. This effect is particularly relevant for token-confident errors: within the low-token-uncertainty region shown in the right panel, responses with higher geometric complexity exhibit a 35% error rate compared with 14% at lower complexity. Thus, geometry provides additional reliability information beyond token uncertainty rather than acting as an independent substitute for it. The high geometry complexity group still contains many correct responses, so geometry alone is not a reliable rejection rule. This motivates multiplicative fusion, in which geometric complexity modulates the local uncertainty signal. We compare additive and multiplicative strategies across models and tasks and find that multiplicative fusion is consistently stronger.

This paper makes the following contributions: (1) We extract complementary uncertainty signals from two distinct representational layers: a global geometric signal from the hidden-state representations via matrix Rényi entropy of hidden-state trajectories, and a local token-level signal from the unembedding layer via Shannon and evidential epistemic entropy. Together, they provide a principled, multi-view perspective on LLM uncertainty that goes beyond logit-only approaches. (2) We show empirically that the two signals provide complementary reliability information: hidden-state geometry provides additional discrimination within similar token-uncertainty regimes, including among responses that appear confident at the token level. (3) We propose Global-Local Uncertainty (**GLU**) , a lightweight multiplicative fusion of the two signals that requires no labeled data, no additional forward passes, and is normalized for response length. (4) We validate GLU across three model families and six benchmarks, and conduct comprehensive ablation studies on the choice of local and global uncertainty estimators, demonstrating both the effectiveness of the framework and the individual contribution of each component.

Our code is publicly available at `https://github.com/qcri/GLU/tree/main`.

2 Related Work

Local and output-distributional uncertainty. A common approach estimates LLM reliability from generation-time signals such as sequence likelihood, token entropy, model-elicited confidence, or probabilistic views of next-token inference [6, 10, 11]. LogTokU is especially relevant as it argues that probabilities alone can lose evidence-strength information thus instead estimates token-level aleatoric and epistemic uncertainty directly from logits, building on the broader evidential-learning

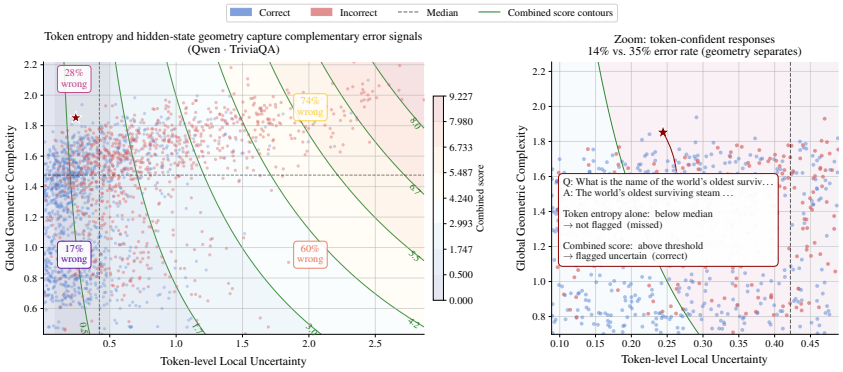

Figure 1: **Local uncertainty and hidden-state geometry provide complementary reliability signals.** Each point represents a response from Qwen2.5-7B on TriviaQA (blue = correct, red = incorrect). The x-axis shows token-level local uncertainty and the y-axis hidden-state geometric complexity; green contours denote equal values of their multiplicatively combined score. Left: Token-level uncertainty provides the primary separation between correct and incorrect responses, while geometry provides additional discrimination within similar local-uncertainty regimes. Error rates increase with geometric complexity on both sides of the median token uncertainty, indicating that geometry contributes reliability information beyond the local signal. Right: Among token-confident responses, higher geometric complexity is associated with a substantially higher error rate. The highlighted example illustrates an incorrect response that appears confident under token entropy alone but is assigned higher uncertainty after geometric modulation. Together, the panels motivate GLU's multiplicative fusion, where geometry modulates rather than replaces the local uncertainty signal.

view that uncertainty can be represented through evidence over class probabilities [5, 6]. Zhou et al. further show that token-level uncertainty is useful for confabulation detection, using it to select and aggregate hidden states for response-level reliability prediction [12]. Other methods use local distributional features during decoding, such as layer-wise logit contrast, or generate multiple samples and measure agreement or semantic dispersion [2–4, 13, 14]. These approaches are effective and often minimally intrusive, but they primarily measure uncertainty in tokens or sampled outputs rather than the global coherence of the model's internal trajectory.

Representation-based and global uncertainty. A complementary line of work studies whether reliability is encoded in hidden states, layer activations, representation geometry, or attention behavior. Layer-by-Layer shows that intermediate layers can contain especially informative representations and analyzes them using information-theoretic, geometric, and invariance-based measures [8]. RAUQ similarly moves beyond output probabilities by identifying uncertainty-aware attention heads whose attention patterns correlate with incorrect generations, producing efficient sequence-level uncertainty estimates without task-specific labels [15]. More broadly, probing and representation-based methods use hidden states or activations to predict factuality, confabulation, or response reliability [7, 12, 16, 17]. These methods capture response-level failures that token probabilities can miss, but representation-only signals may overlook local knowledge gaps already visible in the output distribution.

Supervised detectors and calibration dependence. Several hallucination detectors learn reliability decision boundaries from labeled or pseudo-labeled generations [12, 16–18]. While effective in-distribution, these methods depend on calibration data and may be sensitive to shifts across datasets, model families, or generation settings.We include a supervised-detector comparison in Appendix A.

Our method instead combines local and global uncertainty without learning a detector. It uses hidden-state geometry to amplify token-level uncertainty when the response trajectory is diffuse, capturing local uncertainty and response-level drift without labeled calibration data, pseudo-labels, or retraining.

3 Preliminary

We formally introduce the notation of this paper. Let $\mathbf{x}$ denote a prompt and $\mathbf{y} = (y_1, \ldots, y_T)$ denote a generated response of length T, where each token y_t is drawn from a vocabulary $\mathcal{V}$ of size V. At generation step t and transformer layer ℓ, the language model produces a hidden-state vector $\mathbf{h}_t^{(\ell)} \in \mathbb{R}^d$. The model then autoregressively predicts the next token using the final-layer representation. Specifically, the final hidden state is mapped to the logit vector

$$\mathbf{z}_t = \mathbf{W}_U \mathbf{h}_t^{(L)} \in \mathbb{R}^V,$$

where $\mathbf{W}_U$ is the unembedding matrix (LLM head). The conditional distribution over the next token is given by

$$p(y_t \mid \mathbf{x}, \mathbf{y}_{<t}) = \mathrm{softmax}(\mathbf{z}_t).$$

Next, we collect hidden states across all T generation steps at layer ℓ into the *representation matrix*

$$\mathbf{H}^\ell = \begin{bmatrix} \mathbf{h}_1^{\ell)} \\ \vdots \\ \mathbf{h}_T^{(\ell)} \end{bmatrix} \in \mathbb{R}^{T \times d}, \tag{1}$$

where each row encodes the model's internal state at one generation step. More concretely, $\mathbf{H}$ encodes the full *trajectory* of the response in representation space, the sequence of rows traces how the model's internal representation evolves as it generates each token. We simply write $\mathbf{H}$ when there is no ambiguity.

We define the *Gram matrix* of the response as

$$\mathbf{K} = \mathbf{H}\mathbf{H}^\top \in \mathbb{R}^{T \times T}, \tag{2}$$

whose (i, j)-th entry $\mathbf{K}_{ij} = \langle \mathbf{h}_i^{(L)}, \mathbf{h}_j^{(L)} \rangle$ measures the similarity between the representations at steps i and j. Intuitively, $\mathbf{K}$ captures the geometric structure of the response trajectory: when the model generates a coherent response, the hidden states tend to concentrate along a few dominant directions, making $\mathbf{K}$ low-rank; when the response is uncertain or incoherent, the hidden states scatter across many directions, making $\mathbf{K}$ closer to a scaled identity.

4 Global and Local Uncertainty of LLMs

Uncertainty in LLM generation appears at two complementary levels. At the *token level*, the model may be uncertain over plausible next tokens, capturing local aleatoric or epistemic uncertainty. At the *response level*, the hidden-state trajectory $\mathbf{H}$ may become geometrically complex, spreading across multiple directions in latent space where this structure is captured by the eigenvalues and eigenvectors of its representation geometry [8]. These signals are not redundant; we observe that a response can have confident token predictions while its global trajectory drifts in Figure 1, producing a plausible but hallucinated answer. We therefore measure both token-level uncertainty (local) and geometric entropy (global).

4.1 Local Uncertainty: Token-level Entropy

Token-level uncertainty captures local moments where the model faces lexical ambiguity or lack of knowledge. We quantify the token-level uncertainty from both the probability distribution and the logit-induced distribution, then combine the two signals.

Following the idea from evidential deep learning (EDL) [5], we decompose predictive uncertainty into two components: *aleatoric uncertainty* (AU), reflecting irreducible ambiguity in the data, and *epistemic uncertainty* (EU), reflecting lack of evidence that could in principle be reduced with more information. EDL places a Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\alpha})$ over the token probability simplex, where the total concentration $\alpha_0 = \sum_j \alpha_j$ measures the evidence strength. The corresponding EU is then measured as $k/(\alpha_0 + k)$, which is high when evidence is weaker. For each generated token t with

logit vector $z_t \in \mathbb{R}^{|\mathcal{V}|}$, let $\tau_{(1)}, \ldots, \tau_{(k)}$ denote the top-k tokens by logit value. We define AU and EU for each token t as follows:

$$\mathrm{AU}(t) = -\sum_{j=1}^{k} p_j \log_2 p_j, \quad p_j = \frac{e^{z_{\tau_{(j)}}}}{\sum_{j'} e^{z_{\tau_{(j')}}}}, \tag{3}$$

$$\mathrm{EU}(t) = \frac{k}{\sum_{j=1}^{k}(\alpha_j + 1)}, \quad \alpha_j = \log(1 + e^{z_{\tau_{(j)}}}), \tag{4}$$

where p_j is the softmax probability over the top-k candidates, and α_j treats logits as Dirichlet concentration parameters following Ma et al. [6].

Aleatoric uncertainty is the Shannon entropy, and it is small when one candidate dominates the truncated distribution and large when probability mass is spread across several plausible candidates. The top-k restriction focuses the measure on likely next-token alternatives. AU therefore measures predictive ambiguity.

Response-level aggregation. A token is locally uncertain only when *both* the predictive distribution is spread (high AU) and the evidence base is weak (high EU). We capture this joint condition with the per-token score. Averaging the AU and EU over all tokens dilutes the signal with confident stop-words and punctuation. Instead, we aggregate over the m most uncertain tokens:

$$\bar{R} = \frac{1}{m} \sum_{t \in \mathcal{W}_m} -\mathrm{AU}(t) \cdot \mathrm{EU}(t), \tag{5}$$

where $\mathcal{W}_m$ is the index set of the m tokens with the lowest (most negative) scores and $m \leq T$. This worst-case aggregation correlates strongly with factual errors.

While token-level uncertainty provides fine-grained signals, aggregating these estimates into a single response-level score is non-trivial [6]. In the following section, we discuss a complementary approach that naturally operates at the response level, bypassing the aggregation problem entirely.

4.2 Global Uncertainty: Rényi Entropy of Hidden-State Gram Matrices

We define *geometric entropy* as the Rényi entropy of the normalized eigenspectrum of the hidden-state Gram matrix, quantifying the dispersion of representations across latent geometric directions.

More concretely, we characterize global uncertainty through the geometric complexity of hidden-state representations. As established in Section 3, the Gram matrix $\mathbf{K}^{(\ell)} = \mathbf{H}^{(\ell)}\mathbf{H}^{(\ell)^\top} \in \mathbb{R}^{T \times T}$ encodes the pairwise similarity structure of the response trajectory at layer ℓ. Its eigenvalues $\lambda_1 \geq \cdots \geq \lambda_T \geq 0$ characterize how the trajectory's variance is distributed across orthogonal directions in representation space: a sharply peaked spectrum reflects a low-rank, coherent trajectory, while a flat spectrum reflects a diffuse, uncertain one. We now convert this spectral structure into a single scalar uncertainty score.

Entropy of the normalized spectrum. Normalizing the eigenvalues as $\tilde{\lambda}_i = \lambda_i / \mathrm{tr}(\mathbf{K}^{(\ell)})$ yields a probability distribution over principal directions. We measure its dispersion via the order-q Rényi entropy of that distribution [8]:

$$S_q(\mathbf{K}) \;=\; \frac{1}{1-q} \log\Bigl(\textstyle\sum_{i=1}^{T} \tilde{\lambda}_i^{\,q}\Bigr), \qquad q > 0,\ q \neq 1. \tag{6}$$

We use $q = 2$ (collision entropy), which admits the closed form

$$S_2(\mathbf{K}) \;=\; -\log \mathrm{tr}\left(\left(\frac{\mathbf{K}}{\mathrm{tr}(\mathbf{K})}\right)^2\right) \;=\; -\log \frac{\|\mathbf{K}\|_F^2}{\mathrm{tr}(\mathbf{K})^2}, \tag{7}$$

and therefore requires only the squared Frobenius norm of $\mathbf{K}$ rather than its full eigendecomposition. Forming $\mathbf{K} = \mathbf{H}\mathbf{H}^\top$ costs $O(T^2 d)$, after which $\|\mathbf{K}\|_F^2$ is computed in $O(T^2)$; the eigendecomposition alternative would cost $O(T^3)$. This advantage is substantial in practice since $d \gg T$ for large

language models (hidden dimension d in the thousands, response length T in the tens to hundreds). Low S_2 indicates that the trajectory concentrates along a few dominant directions (coherent generation); high S_2 indicates a diffuse trajectory (geometrically incoherent generation). We give the detailed proof in Appendix B.

Normalization for cross-response and cross-model comparability. We adopt two normalizations to make $\tilde{S}$ comparable across responses and models. First, $S_2(\mathbf{K}^{(\ell)})$ grows with T and attains its maximum of $\log T$ when the spectrum is uniform ($\tilde{\lambda}_i = 1/T$); we therefore divide by $1 + \log T$ so the normalized score lies in $[0, 1)$ regardless of response length. Second, because representation quality varies with depth and no single layer is uniformly informative across architectures, we average the length-normalized entropy over all L layers to obtain

$$\tilde{S} \;=\; \frac{1}{L}\sum_{\ell=1}^{L} \frac{S_2\big(\mathbf{K}^{(\ell)}\big)}{1+\log T} \;\in\; [0,1]. \tag{8}$$

This layer-agnostic formulation avoids model-specific layer selection and yields a single, response-level uncertainty score computable in one forward pass.

4.3 Global-Local Fusion

The two signals cover distinct regimes (Fig. 1): correct responses cluster at low local uncertainty but spread broadly along $\tilde{S}$, while incorrect responses include a *confident-but-wrong* subset which has low token entropy yet elevated geometric complexity, that token-level detectors miss entirely. We fuse the two via a multiplicative gate:

$$\boxed{\mathrm{GLU} \;=\; (1+\tilde{S})\,\bar{R}} \tag{9}$$

The factor $(1 + \tilde{S})$ acts as a geometric amplifier, such that, when the trajectory is diffuse, local uncertainty is up-weighted; when compact, GLU reduces to the local signal alone. This directly targets the confident-but-wrong failure mode, which additive fusion cannot isolate because it treats the global signal as a fixed offset rather than a modulator.

5 Experiments and Ablation Studies

In this section, we evaluate GLU across six benchmarks and three model families. To empirically validate the design choices motivated in Section 4.3, we conduct comprehensive ablation studies over local and global uncertainty components, isolating the contribution of each and confirming that the multiplicative geometric–probabilistic combination outperforms alternatives.

Datasets and Models. We evaluate on six benchmarks spanning factuality, knowledge retrieval, mathematical reasoning, Arabic QA, long-form generation, and multi-turn reasoning: *TruthfulQA* [19], *TriviaQA* [20], *MATH* [21], *ArabicaQA* [22], *LongForm* [23], and the *GSM8K multi-turn* setting from Laban et al. [24]. Further information on the composition of the datasets are provided in Appendix C. We evaluate three instruction-tuned model families: *Qwen2.5-7B* [25], *Gemma3-12B* [26], and *Fanar1-9B* [27], an Arabic-centric Arabic–English model continually pretrained from Gemma-9B, allowing us to test whether GLU remains effective across different model families, languages, and latent-space configurations.

Evaluation protocol. We adopt greedy decoding to generate responses for each model–benchmark pair, and store the corresponding hidden states, logits, and probability distributions. All uncertainty estimation methods are applied post hoc to this shared set of model outputs, ensuring a fair and controlled comparison. For GLU, we use the same $k = 10$ top-logit candidates and $m = 10$ selected response tokens in every model–dataset setting. We report two complementary evaluation metrics. AUROC measures how well an uncertainty score discriminates between correct and incorrect responses (higher is better; 0.5 corresponds to random guessing). Prediction Rejection Ratio (PRR) measures how well a score puts wrong answers first in the rejection order; higher is better. A PRR of 1 is ideal, 0 is equivalent to random rejection, and a negative value is worse than random. We refer the reader to Malinin and Gales [28] for its full definition. The two metrics capture complementary

Table 1: Methods and ablations. For GLU variants, the final score combines a global geometric uncertainty term S with a local token-level uncertainty term $\bar{u}$ as $\mathrm{GLU}(x, y) = (1 + S)\bar{u}$, where larger uncertainty corresponds to a more negative score.

Method	Global	Local	Explanation
Baselines			
LogProb	–	$\frac{1}{T}\sum_{t=1}^{T} \log p(y_t \mid y_{<t}, x)$	Average token log-probability of the generated response.
LogTokU	–	$\bar{R}_{\mathrm{EDL}} = \frac{1}{m}\sum_{t\in\mathcal{W}_m} -\mathrm{AU}_{\mathrm{EDL}}(t) \cdot \mathrm{EU}(t)$	Token-level uncertainty baseline using Dirichlet aleatoric uncertainty and epistemic uncertainty [6].
P(True)	$p(\text{“True”} \mid x, y)$	–	The model self-evaluates whether its generated answer is true [10].
RAUQ	$\mathrm{RAUQ}(x, y)$	–	Recurrently fuses uncertainty-aware attention-head activations with token probabilities; peaks over a middle-layer subset [15].
Ablations on Eq.9			
GLU	$\tilde{S} = \frac{1}{L}\sum_{\ell=1}^{L} \frac{S_q^\ell}{1+\log T}$	$\bar{R} = \frac{1}{m}\sum_{t\in\mathcal{W}_m} - \mathrm{AU}(t) \cdot \mathrm{EU}(t)$	Length-normalised matrix Rényi-2 entropy averaged across all L layers, combined with the mean of the m most uncertain local scores.
GLU-$\bar{S}$	$\bar{S}_q = \frac{1}{L}\sum_{\ell=1}^{L} S_q^\ell$	$\bar{R}$	Removes the $1/(1+\log T)$ length normalisation; both $\bar{S}$ and $\bar{S}_\alpha$ average Rényi entropy across all L layers.
GLU-$\tilde{S}^*$	$\max_\ell \tilde{S}^\ell$	$\bar{R}$	Selects the single most uncertain length-normalised layer instead of averaging over all L layers as in $\tilde{S}$.
GLU-AU	$\tilde{S}$	$R = \frac{1}{m}\sum_{t\in\mathcal{W}_m} - \mathrm{AU}(t)$	Drops the epistemic factor entirely; retains only the local Shannon entropy term amplified by $\tilde{S}$.
GLU-S^*	$S_q^* = \max_\ell S_q^\ell$	$\bar{R}$	Uses the single most uncertain raw layer instead of the depth-averaged global entropy.
Additive fusion variants			
Add-S_q	$\bar{S}_q$	$\bar{R}$	Uses additive fusion instead of the multiplicative GLU interaction.
Add-$\tilde{S}$	$\tilde{S}$	$\bar{R}$	Uses additive fusion with length-normalized geometry.

aspects of performance: AUROC reflects discriminative capability, whereas PRR assesses whether the uncertainty ranking is effective for selective prediction.

Baselines and Ablations. The baselines cover the major families of unsupervised uncertainty estimation methods, including LogProb[29] and LogTokU (local token-level methods) [6], P(True) (self-evaluation) [10], and RAUQ (attention-based uncertainty estimation) [15]. The GLU ablations isolate the contribution of individual components in Eq. 9. Summarized breakdown of the ablations is in Table 1 while the full ablations descriptions are provided in Appendix D.

5.1 Main Results

In this section, we examine whether the multiplicative fusion yields more reliable uncertainty estimates across models and tasks. Table 2 reports AUROC and PRR for all methods across three models and six datasets; we discuss each metric in turn.

Among the methods displayed in Table 2, GLU achieves the best AUROC in 13 of 18 settings and ranks second in the remainder, confirming that multiplicative fusion of complementary signals consistently outperforms either component in isolation. The pattern directly validates the complementarity hypothesis from Figure 1. GLU improves over the local-only baseline LogTokU by up to 22.0% (Gemma-TriviaQA: 0.817 vs. 0.597) and over the global-only baseline RAUQ by up to 11.7% (Gemma-TriviaQA: 0.817 vs. 0.700).

RAUQ overtakes GLU mainly in settings where the local token-level signal is weak: ArabicaQA for Qwen and Fanar, and multiturn MATH for Gemma. In these cases, LogTokU is near chance in AUROC and often has negative PRR, suggesting that token confidence is poorly aligned with correctness. Since GLU multiplicatively combines the local and global signals, an uninformative

Table 2: Response reliability estimation across three models and six datasets, evaluated by AUROC and PRR. **LogTokU** captures a purely *local* signal from token-level output logits; **RAUQ** captures a *global* signal from intermediate hidden-state and attention representations; **GLU** forms their *multiplicative* fusion, achieving the best score in 13/18 cells on AUROC and 12/18 on PRR while ranking second in the remaining cells. **Bold**: best per row; underline: second best.

		AUROC			PRR		
Model	**Dataset**	**LogTokU**	**RAUQ**	**GLU**	**LogTokU**	**RAUQ**	**GLU**
Qwen	TruthfulQA	**0.680**	0.582	0.616	**0.377**	0.063	0.195
	TriviaQA	0.786	0.789	**0.834**	0.606	0.581	**0.717**
	MATH	0.595	0.699	**0.767**	0.216	0.424	**0.537**
	ArabicaQA	0.521	**0.610**	0.556	−0.028	**0.186**	0.140
	Longform	0.490	0.469	**0.496**	−0.045	−0.044	**−0.040**
	Multiturn	0.523	0.644	**0.739**	0.008	0.189	**0.407**
Gemma	TruthfulQA	0.531	0.466	**0.548**	**0.115**	−0.094	0.087
	TriviaQA	0.597	0.700	**0.817**	0.224	0.416	**0.698**
	MATH	**0.801**	0.771	0.799	0.693	0.639	**0.699**
	ArabicaQA	0.538	0.642	**0.647**	0.099	0.277	**0.322**
	Longform	0.343	0.354	**0.358**	−0.249	**−0.234**	−0.242
	Multiturn	0.457	**0.734**	0.701	−0.274	**0.506**	0.423
Fanar	TruthfulQA	0.579	**0.641**	**0.641**	0.192	0.282	**0.307**
	TriviaQA	0.706	**0.821**	**0.821**	0.445	0.712	**0.730**
	MATH	0.664	0.779	**0.788**	0.324	0.624	**0.627**
	ArabicaQA	0.490	**0.613**	0.563	−0.025	**0.239**	0.143
	Longform	0.569	0.549	**0.654**	0.077	0.066	**0.200**
	Multiturn	0.463	0.804	**0.831**	−0.166	0.654	**0.687**

local component can attenuate an otherwise strong global signal, allowing the pure-global RAUQ baseline to edge ahead. However, GLU usually remains close to RAUQ in these cases, indicating that the fusion degrades gracefully rather than failing outright.

The PRR results show that GLU's advantage carries over to selective prediction. GLU achieves the best PRR in 12 of 18 settings and the highest mean PRR overall, indicating that its uncertainty rankings place unreliable responses closer to the rejection tail more effectively than either LogTokU or RAUQ alone. This is important because the practical goal is not only to distinguish correct from incorrect responses, but also to prioritize which outputs should be deferred, flagged, or verified first.

5.2 Ablation Study

To determine which design choices account for GLU's performance, Table 3 aggregates performance across all 18 settings and reveals three findings.

First, GLU ranks first on both mean AUROC (0.676) and mean PRR (0.369), confirming that multiplicative fusion of global and local signals consistently outperforms any individual component: RAUQ ranks 12th and LogTokU 14th, showing that neither signal alone is sufficient. Second, all multiplicative ablations cluster tightly at ranks 2–9, indicating that the gain is robust to the choice of local and global uncertainty estimator. The performance gain is driven by the fusion principle rather than the specific choice of estimator. Third, and most decisively, the additive variant Add-$\tilde{S}$ drops to rank 13 (0.621 AUROC, 0.259 PRR), falling below all baselines except LogTokU. This sharp degradation isolates multiplicative fusion as the source of the gain: treating the global signal as a fixed offset rather than a modulator of local uncertainty fails to capture the interaction between the two signals that GLU is designed to exploit.

5.3 GLU on TriviaQA

Beyond aggregate discrimination performance, we examine whether GLU provides a graded signal of response reliability on TriviaQA. For each model, we sort responses by increasing GLU uncertainty, partition them into deciles, and compute the empirical incorrect rate within each bin.

Figure 2 shows a consistent monotonic trend across all three models: responses assigned higher GLU uncertainty are more likely to be incorrect. The gap between the lowest- and highest-uncertainty bins reaches approximately 84 percentage points for Qwen, 83 percentage points for Gemma, and 66

Table 3: Mean performance across all 18 model–dataset settings, ranked among 14 methods. The **Space** column denotes the signal each method uses: L (local, token-level) and G (global, hidden-state geometry); L $\times$ G marks multiplicative fusion and L+G additive fusion. GLU ranks first on both mean AUROC (0.676) and mean PRR (0.369), above the global-signal baseline RAUQ (rank 12) and the local-signal baseline LogTokU (rank 14). The multiplicative ablations cluster just below it (ranks 2–9), while the additive variant Add-$\tilde{S}$ falls to rank 13, isolating the multiplicative fusion as the source of the gain. Full results and rankings are in Appendix D.

Group	Method	Space	AUROC	Rank	PRR	Rank
Proposed	blue!8**GLU**	L $\times$ G	0.676	1	0.369	1
Baseline	LogProb	L	0.655	10	0.326	10
Baseline	LogTokU	L	0.574	14	0.144	14
Baseline	P(true)	G	0.654	11	0.311	11
Baseline	RAUQ	G	0.648	12	0.305	12
Ablation	GLU-SP	L $\times$ G	0.675	3	0.367	2
Ablation	GLU-$\tilde{S}^*$	L $\times$ G	0.674	4	0.366	3
Ablation	GLU-DK	L $\times$ G	0.675	2	0.366	4
Ablation	GLU-$\tilde{S}$	L $\times$ G	0.674	5	0.365	5
Ablation	GLU-S_q-SP	L $\times$ G	0.673	6	0.363	6
Ablation	GLU-AU	L $\times$ G	0.671	7	0.362	7
Ablation	GLU-S_q-AU	L $\times$ G	0.667	8	0.357	8
Ablation	GLU-S^*	L $\times$ G	0.665	9	0.351	9
Ablation	Add-$\tilde{S}$	L+G	0.621	13	0.259	13

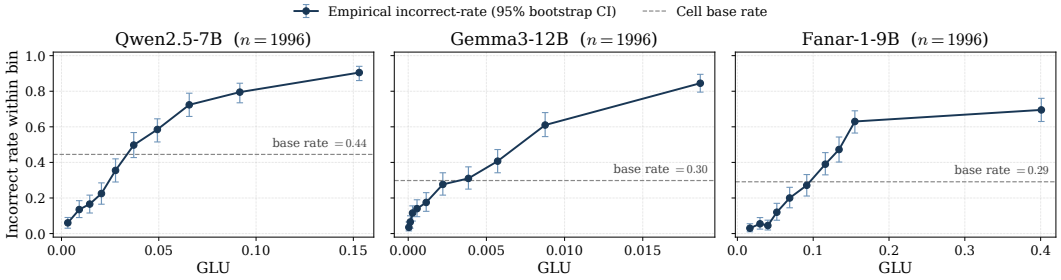

Figure 2: Binned reliability of GLU on TriviaQA. Responses are sorted by $-$GLU and partitioned into deciles; markers report the empirical incorrect-rate per bin with 95% bootstrap CIs (1,000 resamples). Dashed line: cell-level base incorrect-rate. GLU yields a monotonic increase in incorrect-rate across confidence deciles for all three models, indicating a graded reliability signal rather than only separating extreme cases.

percentage points for Fanar. These ranges are above the model-level base incorrect rates, indicating that GLU does not simply reflect dataset-level difficulty or a binary correct/incorrect separation.

This result suggests that GLU captures a graded notion of answer reliability. Low-GLU responses are mostly correct, while high-GLU responses concentrate a large fraction of errors. Therefore, GLU can support reliability-aware generation settings in which answers are selectively trusted, flagged, or routed for additional verification.

6 Conclusion

We proposed **GLU**, a lightweight unsupervised method that fuses two complementary uncertainty signals: global geometric uncertainty from hidden-state trajectories, quantified via matrix Rényi entropy, and local token-level uncertainty from the output distribution, based on Shannon and evidential epistemic uncertainty. The two signals capture distinct failure regimes, with the global signal providing additional discrimination among confident-but-wrong responses that token-level

methods can miss. GLU combines these signals through multiplicative fusion while requiring no labeled data, no additional forward passes, and remaining normalized for response length. Across three model families and six benchmarks, GLU matches or outperforms all unsupervised baselines and remains competitive with supervised methods that lack cross-dataset generalization. Ablation studies further show that the gains are robust to the choice of local and global uncertainty estimators, supporting the complementarity of the two signals across models and tasks.

Limitations. GLU requires access to hidden states and token-level output distributions, making it most directly applicable to open-weight models. While GLU achieves the best mean AUROC and PRR across our evaluation, the relative strength of global and local signals varies by task; in some settings, RAUQ remains highly competitive, suggesting that adaptive fusion could further improve robustness. Future work should explore principled layer selection, adaptive weighting, and validation on larger models, additional languages, and domain-specific deployments.

Acknowledgements

We thank all anonymous reviewers and the area chair for their valuable feedback and constructive suggestions. This research was supported by the Ministry of Transport, Qatar, under grant number OVPR-SRO-MOT-2025-001, the Swedish Research Council (VR) [2024-05603], the European Commission MSCA DN [101168951], and the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

AI Use Statement

Generative AI tools were used during the development of this manuscript. Their use included drafting and debugging portions of experimental code, assisting with the interpretation and presentation of experimental results, and drafting and editing portions of the manuscript. All implementations, experimental protocols, and numerical results included in the final paper were reviewed and verified by the authors. Experiments reported in the paper were executed by the authors on the stated models and datasets. The authors take full responsibility for the correctness and originality of the paper and for parts of the code produced with the assistance of generative AI.

References

[1] Declan Jackson, William Keating, George Cameron, and Micah Hill-Smith. AA-Omniscience: Evaluating cross-domain knowledge reliability in large language models. *arXiv preprint arXiv:2511.13029*, 2025.

[2] Jianyi Zhang, Da-Cheng Juan, Cyrus Rashtchian, Chun-Sung Ferng, Heinrich Jiang, and Yiran Chen. Sled: self logits evolution decoding for improving factuality in large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

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

[4] Yasin Abbasi Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, Ali Taylan Cemgil, and Nenad Tomasev. Mitigating llm hallucinations via conformal abstention. (arXiv:2405.01563), April 2024. doi: 10.48550/arXiv.2405.01563. URL `http://arxiv.org/abs/2405.01563`. arXiv:2405.01563 [cs].

[5] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 3183–3193, Red Hook, NY, USA, 2018. Curran Associates Inc.

[6] Huan Ma, Jingdong Chen, Joey Tianyi Zhou, Guangyu Wang, and Changqing Zhang. Estimating llm uncertainty with evidence. *arXiv preprint arXiv:2502.00290*, 2025.

[7] Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. Llm-check: Investigating detection of hallucinations in large language models. *Advances in Neural Information Processing Systems*, 37:34188–34216, 2024.

[8] Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *Proceedings of the 42nd International Conference on Machine Learning*, PMLR 267, 2025.

[9] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[10] 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 preprint arXiv:2207.05221*, 2022.

[11] Siddhartha Dalal and Vishal Misra. Beyond the black box: A statistical model for llm reasoning and inference. *arXiv preprint arXiv:2402.03175*, 2024.

[12] Tianyi Zhou, Johanne Medina, and Sanjay Chawla. Can llms detect their confabulations? estimating reliability in uncertainty-aware language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 38164–38172, 2026.

[13] Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9004–9017. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.557. URL `https://doi.org/10.18653/v1/2023.emnlp-main.557`.

[14] Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *arXiv preprint arXiv:2405.20003*, 2024.

[15] Artem Vazhentsev, Lyudmila Rvanova, Gleb Kuzmin, Ekaterina Fadeeva, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin, Mrinmaya Sachan, Preslav Nakov, and Artem Shelmanov. Efficient hallucination detection for LLMs using uncertainty-aware attention heads, 2026. URL https://openreview.net/forum?id=FSOoR1ZFtf.

[16] Loka Li, Zhenhao Chen, Guangyi Chen, Yixuan Zhang, Yusheng Su, Eric Xing, and Kun Zhang. Confidence matters: Revisiting intrinsic self-correction capabilities of large language models. 2024. URL https://arxiv.org/abs/2402.12563.

[17] Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. Llms know more than they show: On the intrinsic representation of llm hallucinations. International Conference on Learning Representations (ICLR), 2024.

[18] Xuefeng Du, Chaowei Xiao, and Yixuan Li. Haloscope: Harnessing unlabeled llm generations for hallucination detection. In *Advances in Neural Information Processing Systems*, 2024.

[19] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 3214–3252, 2022.

[20] Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1601–1611. Association for Computational Linguistics, 2017.

[21] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021.

[22] Abdelrahman Abdallah, Mahmoud Kasem, Mahmoud Abdalla, Mohamed Mahmoud, Mohamed Elkasaby, Yasser Elbendary, and Adam Jatowt. Arabicaqa: A comprehensive dataset for arabic question answering, 2024.

[23] Abdullatif Köksal, Timo Schick, Anna Korhonen, and Hinrich Schütze. Longform: Effective instruction tuning with reverse instructions, 2023.

[24] Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*, 2025.

[25] Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025.

[26] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, and Anton Tsitsulin. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[27] Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, Majd Hawasly, Mus'ab Husaini, Soon-Gyo Jung, Ji Kim Lucas, Walid Magdy, Safa Messaoud, Abubakr Mohamed, Tasnim Mohiuddin, Basel Mousi, Hamdy Mubarak, Ahmad Musleh, Zan Naeem, Mourad Ouzzani, Dorde Popovic, Amin Sadeghi, Husrev Taha Sencar, Mohammed Shinoy, Omar Sinan, Yifan Zhang, Ahmed Ali, Yassine El Kheir, Xiaosong Ma, and Chaoyi Ruan. Fanar: An arabic-centric multimodal generative ai platform. *arXiv preprint arXiv:2501.13944*, 2025.

[28] Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jN5y-zb5Q7m.

[29] Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. To believe or not to believe your llm: Iterative prompting for estimating epistemic uncertainty. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Table 4: **Classifier-based:** AUROC for response reliability estimation across three benchmarks and three model families. **Bold**: best per column per model. Underline: second best. Although classifier-based methods output better scores, they are not generalizable. *Italic*: third best.

Model	Method	TruthfulQA	TriviaQA	Math
Qwen	**Prob(Exact)**	0.758	0.781	0.627
	Probe(EOS)	**0.794**	0.812	0.703
	Probe(EU)	0.759	0.754	0.646
	Probe(AVG)	0.761	0.786	0.699
	RespEntropy	0.705	0.793	**0.888**
	GLU	0.616	**0.834**	0.767
Gemma	**Prob(Exact)**	0.728	0.796	0.773
	Probe(EOS)	0.728	0.810	0.834
	Probe(EU)	0.687	0.751	0.669
	Probe(AVG)	**0.733**	**0.818**	0.786
	RespEntropy	0.708	0.775	**0.867**
	GLU	0.548	0.817	0.799
Fanar	**Prob(Exact)**	0.711	0.783	0.827
	Probe(EOS)	0.706	0.739	0.790
	Probe(EU)	0.709	0.751	0.794
	Probe(AVG)	0.734	0.765	0.833
	RespEntropy	**0.753**	**0.885**	**0.842**
	GLU	0.641	0.821	0.788

A Supervised vs Unsupervised

GLU requires no labeled data at test time, which is its primary practical advantage. To give a sense of what that constraint costs, we include here results from classifier-based methods that *do* have access to ground-truth correctness labels during training. The comparison is deliberately asymmetric, we are not claiming parity, but it lets readers calibrate how much headroom remains above our unsupervised scores and confirms that GLU is already competitive with several supervised probes despite operating in a fundamentally harder setting.

B Proof of the Collision-Entropy Closed Form

We prove that the order-2 Rényi entropy of the normalised eigenspectrum of a Gram matrix $\mathbf{K} \in \mathbb{R}^{T\times T}$ admits the two equivalent closed forms stated in (7).

[Collision-entropy closed form] Let $\mathbf{K} = \mathbf{H}\mathbf{H}^\top$ be the Gram matrix of hidden states $\mathbf{H} \in \mathbb{R}^{T\times d}$, with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_T \geq 0$. Define normalised eigenvalues $\tilde{\lambda}_i = \lambda_i/\mathrm{tr}(\mathbf{K})$. Then

$$S_2(\mathbf{K}) \;=\; -\log\Big(\textstyle\sum_{i=1}^{T} \tilde{\lambda}_i^2\Big) \;=\; -\log\,\mathrm{tr}\left(\left(\frac{\mathbf{K}}{\mathrm{tr}(\mathbf{K})}\right)^2\right) \;=\; -\log\frac{\|\mathbf{K}\|_F^2}{\mathrm{tr}(\mathbf{K})^2}. \tag{10}$$

Proof. We establish the two equalities in turn. Setting $q = 2$ in the general order-q Rényi entropy gives

$$S_2(\mathbf{K}) = \frac{1}{1-2}\log\Big(\sum_i \tilde{\lambda}_i^2\Big) = -\log\Big(\sum_i \tilde{\lambda}_i^2\Big). \tag{11}$$

Since $\tilde{\lambda}_i = \lambda_i/\mathrm{tr}(\mathbf{K})$ and $\mathrm{tr}(\mathbf{K})$ is a positive scalar,

$$\sum_{i=1}^{T} \tilde{\lambda}_i^2 = \sum_{i=1}^{T} \frac{\lambda_i^2}{\mathrm{tr}(\mathbf{K})^2} = \frac{\sum_{i=1}^{T}\lambda_i^2}{\mathrm{tr}(\mathbf{K})^2}. \tag{12}$$

Because $\mathbf{K}$ is real symmetric and positive semidefinite, its eigenvalues are real and non-negative. The eigenvalues of $\mathbf{K}^2$ are λ_i^2, so by the trace–eigenvalue identity,

$$\mathrm{tr}(\mathbf{K}^2) = \sum_{i=1}^{T} \lambda_i^2. \tag{13}$$

Moreover, the Frobenius norm satisfies $\|\mathbf{K}\|_F^2 = \mathrm{tr}(\mathbf{K}^\top \mathbf{K}) = \mathrm{tr}(\mathbf{K}^2)$, where the last step uses symmetry $\mathbf{K}^\top = \mathbf{K}$. Hence

$$\sum_{i=1}^{T} \lambda_i^2 = \mathrm{tr}(\mathbf{K}^2) = \|\mathbf{K}\|_F^2. \tag{14}$$

Since the trace is linear,

$$\frac{\mathrm{tr}(\mathbf{K}^2)}{\mathrm{tr}(\mathbf{K})^2} = \mathrm{tr}\left(\frac{\mathbf{K}^2}{\mathrm{tr}(\mathbf{K})^2}\right) = \mathrm{tr}\left(\left(\frac{\mathbf{K}}{\mathrm{tr}(\mathbf{K})}\right)^2\right). \tag{15}$$

Substituting (12)–(15) into (11) yields the first equality in (10).

Replacing $\mathrm{tr}(\mathbf{K}^2)$ by $\|\mathbf{K}\|_F^2$ via (14) gives

$$\mathrm{tr}\left(\left(\frac{\mathbf{K}}{\mathrm{tr}(\mathbf{K})}\right)^2\right) = \frac{\|\mathbf{K}\|_F^2}{\mathrm{tr}(\mathbf{K})^2}, \tag{16}$$

which is the second equality in (10). □

C Dataset Details

This appendix details the composition and preprocessing of the six benchmarks used in our evaluation. Where we subsample, we fix the random seed for reproducibility. Table 5 summarizes the final sizes.

- **TruthfulQA** [19] probes whether a model reproduces common human misconceptions. We use the full dataset of 817 questions without subsampling.
- **TriviaQA** [20] is a large-scale reading-comprehension benchmark for factual knowledge retrieval. We draw $2{,}000$ samples from the reading-comprehension (`rc`) training split.
- **MATH** [21] consists of competition mathematics problems spanning a range of difficulty levels and subjects. We evaluate on a subset of $1{,}000$ problems.
- **ArabicaQA** [22] is a native Arabic question-answering benchmark, included to test reliability estimation outside English. We draw $2{,}000$ samples from the machine-reading-comprehension (MRC) test split.
- **LongForm** [23] targets long-form generation. To keep responses within a tractable length and restrict to well-structured sources, we take the test split and retain only examples drawn from StackExchange, Wikipedia, and BigBench whose reference output is at most 512 words, yielding a subset of 731 examples.
- **GSM8K multi-turn** [24] adapts grade-school math word problems into a sharded, multi-turn conversational setting in which the problem is revealed incrementally across turns. We use the `math` subset of the released `lost_in_conversation` data, comprising 103 problems decomposed into multiple turns each.

D Methods and Ablations

This appendix provides the complete picture behind the main-text results. We show the mean performance of all 19 methods ranked across the 18 model–dataset settings (Table 7), the full per-setting AUROC and PRR for the baselines and GLU including the strongest ablation per cell (Tables 8 and 9), and the complete per-setting AUROC and PRR for all GLU ablations (Tables 10–13). These tables let readers verify that the headline results are not artefacts of aggregation and identify which variants are most robust across model families and task types.

Summary of findings. Three patterns hold across the full results. First, GLU and its multiplicative ablations occupy the top nine ranks on both mean AUROC and mean PRR (Table 7), well above every single-signal baseline, confirming that the gain comes from the multiplicative global–local fusion rather than any specific estimator. Second, the per-setting tables show the gain is broad rather

than driven by a few cells: GLU is best or second on AUROC in 13 of 18 settings and on PRR in a comparable majority, with the clearest margins on TriviaQA, MATH, and Multiturn, and ties or close seconds where the local signal collapses (ArabicaQA, Gemma-Multiturn). Third, fusion variants that drop one signal entirely (GLU-EU, GLU-EU-SP) or replace the multiplicative interaction with an additive one (Add-$\tilde{S}$, Add-S_q) fall to the bottom of Table 7, isolating both the presence of two signals and their multiplicative combination as the sources of the improvement.

Table 5: Composition of the six evaluation benchmarks.

Benchmark	Task	Size	Split / selection
TruthfulQA	Factuality / misconceptions	817	full dataset
TriviaQA	Knowledge retrieval	2,000	`rc` train, subsampled
MATH	Mathematical reasoning	1,000	subsampled
ArabicaQA	Arabic QA	2,000	MRC test, subsampled
LongForm	Long-form generation	731	SE/Wiki/BigBench, ≤ 512 words
GSM8K multi-turn	Multi-turn reasoning	103	`math` subset, sharded turns

Table 6: Methods and ablations. For GLU variants, the final score combines a global geometric uncertainty term S with a local token-level uncertainty term $\bar{u}$ as $\mathrm{GLU}(x, y) = (1 + S)\bar{u}$, where larger uncertainty corresponds to a more negative score.

Method	Global	Local	Explanation
Baselines			
LogProb	–	$\frac{1}{T}\sum_{t=1}^{T} \log p(y_t \mid y_{<t}, x)$	Average token log-probability of the generated response.
LogTokU	–	$R_{\mathrm{EDL}} = \frac{1}{m}\sum_{t\in\mathcal{W}_m} -\mathrm{AU}_{\mathrm{EDL}}(t) \cdot \mathrm{EU}(t)$	Token-level uncertainty baseline using Dirichlet aleatoric uncertainty and epistemic uncertainty [6].
P(True)	$p(\text{“True”} \mid x, y)$	–	The model self-evaluates whether its generated answer is true [10].
RAUQ	$\mathrm{RAUQ}(x, y)$	–	Recurrently fuses uncertainty-aware attention-head activations with token probabilities; peaks over a middle-layer subset [15].
Ablations on Eq.9			
GLU	$\tilde{S} = \frac{1}{L}\sum_{\ell=1}^{L} \frac{S_q^\ell}{1+\log T}$	$\bar{R} = \frac{1}{m}\sum_{t\in\mathcal{W}_m} -\mathrm{AU}(t) \cdot \mathrm{EU}(t)$	Length-normalised matrix Rényi-2 entropy averaged across all L layers, combined with the mean of the m most uncertain local scores.
GLU-EDL	$\tilde{S}$	R_{EDL}	Replaces the local ShETokU term with the LogTokU-style Dirichlet token uncertainty.
GLU-$\bar{S}$	$\bar{S}_q = \frac{1}{L}\sum_{\ell=1}^{L} S_q^\ell$	$\bar{R}$	Removes the $1/(1+\log T)$ length normalisation; both $\tilde{S}$ and $\bar{S}_q$ average Rényi entropy across all L layers.
GLU-$\tilde{S}^*$	$\max_\ell \tilde{S}^\ell$	$\bar{R}$	Selects the single most uncertain length-normalised layer instead of averaging over all L layers as in $\tilde{S}$.
GLU-SP	$\tilde{S}$	$\bar{R}_{\mathrm{EU}^{\mathrm{sp}}} = \frac{1}{k}\sum_{t\in\mathcal{K}} -AU(t) \cdot \mathrm{EU}^{\mathrm{sp}}(t)$	Uses softplus evidence $q_k = \mathrm{softplus}(\ell_k) + \varepsilon$ instead of $\mathrm{ReLU}(\ell_k) + 1$ in the Dirichlet epistemic uncertainty term.
GLU-DK	$\tilde{S}$	$\bar{R}_{\mathrm{EU}}^{k'},\ k' = \left\lfloor k/(1+\tilde{S}) \right\rfloor$	Uses an adaptive local window size, shrinking the number of selected uncertain tokens as global geometric uncertainty increases.
GLU-AU	$\tilde{S}$	$R_{AU} = \frac{1}{m}\sum_{t\in\mathcal{W}_m} -\mathrm{AU}(t)$	Drops the epistemic factor entirely; retains only the local Shannon entropy term amplified by $\tilde{S}$.
GLU-EU	$\tilde{S}$	$R_{EU} = \frac{1}{m}\sum_{t\in\mathcal{W}_m} -\mathrm{EU}(t)$	Drops Shannon entropy entirely; retains only the Dirichlet epistemic uncertainty term amplified by $\tilde{S}$.
GLU-S_q-AU	$\bar{S}_q$	R_{AU}	Jointly ablates length normalisation and the epistemic factor; tests whether raw multi-layer geometry pairs best with Shannon-only local uncertainty.
GLU-S_q-SP	$\bar{S}_q$	$\bar{R}_{\mathrm{EU}^{\mathrm{sp}}}$	Tests whether softplus evidence pairs better with unnormalised depth-averaged entropy than with $\tilde{S}$.
GLU-S^*	$S_q^* = \max_\ell S_q^\ell$	$\bar{R}$	Uses the single most uncertain raw layer instead of the depth-averaged global entropy.
GLU-EU-SP	$\tilde{S}$	$\bar{R}_{\mathrm{EU}^{sp}} = \frac{1}{k}\sum_{t\in\mathcal{K}} -\mathrm{EU}^{sp}(t)$	Uses only the softplus-based evidential uncertainty term, dropping the Shannon entropy factor.
GLU-S_q-SP-EU	$\bar{S}_q$	$\bar{R}_{\mathrm{EU}^{sp}} = \frac{1}{k}\sum_{t\in\mathcal{K}} -\mathrm{EU}^{sp}(t)$	Combines raw depth-averaged matrix Rényi entropy with the softplus-based evidential-only local score.
Additive fusion variants			
Add-S_q	$\bar{S}_q$	$\bar{R}$	Uses additive fusion instead of the multiplicative GLU interaction.
Add-$\tilde{S}$	$\tilde{S}$	$\bar{R}$	Uses additive fusion with length-normalized geometry.

Table 7: Mean performance across all 18 model–dataset settings. Methods are grouped into baselines, the proposed GLU method, and GLU ablations.

Group	Method	Mean AUROC	AUROC rank	Mean PRR	PRR rank
Proposed	blue!8**GLU**	0.676	1	0.369	1
Baseline	LogProb	0.655	10	0.326	10
Baseline	RAUQ	0.648	12	0.305	12
Baseline	$P(\mathrm{true})$	0.654	11	0.311	11
Baseline	LogTokU	0.574	14	0.144	14
Ablation	GLU-SP	0.675	3	0.367	2
Ablation	GLU-$\tilde{S}^*$	0.674	4	0.366	3
Ablation	GLU-DK	0.675	2	0.366	4
Ablation	GLU-$\bar{S}$	0.674	5	0.365	5
Ablation	GLU-S_q-SP	0.673	6	0.363	6
Ablation	GLU-AU	0.671	7	0.362	7
Ablation	GLU-S_q-AU	0.667	8	0.357	8
Ablation	GLU-S^*	0.665	9	0.351	9
Ablation	Add-$\tilde{S}$	0.621	13	0.259	13
Ablation	GLU-EDL	0.574	15	0.140	15
Ablation	GLU-S_q-SP-EU	0.557	16	0.103	16
Ablation	GLU-EU	0.544	17	0.079	17
Ablation	GLU-EU-SP	0.543	18	0.077	18
Ablation	Add-S_q	0.454	19	-0.118	19

Table 8: Full AUROC comparison. Ranking emphasis is applied only across the baselines and GLU. The best ablation is reported for context but is not included in the row-wise ranking.

Model	Dataset	LogTokU	RAUQ	LogProb	$P(\mathrm{true})$	blue!8**GLU**	Best ablation	
Qwen	TruthfulQA	**0.680**	0.582	*0.615*	0.556	blue!8<u>0.616</u>	GLU-EDL	0.680
	TriviaQA	0.786	*0.789*	0.775	<u>0.805</u>	blue!8**0.834**	GLU-S_q-SP	0.837
	MATH	0.595	*0.699*	0.684	<u>0.759</u>	blue!8**0.767**	GLU-S_q-SP	0.772
	ArabicaQA	0.521	*0.610*	<u>0.625</u>	**0.711**	blue!80.556	GLU-AU	0.562
	Longform	<u>0.490</u>	0.469	*0.486*	0.389	blue!8**0.496**	GLU-EDL	0.558
	Multiturn	0.523	*0.644*	0.635	**0.800**	blue!8<u>0.739</u>	GLU-$\bar{S}$	0.758
Gemma	TruthfulQA	0.531	0.466	<u>0.553</u>	**0.629**	blue!8*0.548*	Add-S_q	0.589
	TriviaQA	0.597	*0.700*	<u>0.804</u>	0.671	blue!8**0.817**	GLU-S_q-AU	0.819
	MATH	<u>0.801</u>	0.771	0.683	**0.817**	blue!8*0.799*	Add-$\tilde{S}$	0.801
	ArabicaQA	0.538	*0.642*	**0.689**	0.597	blue!8<u>0.647</u>	GLU-AU	0.650
	Longform	0.343	0.354	<u>0.373</u>	**0.472**	blue!8*0.358*	GLU-S_q-AU	0.390
	Multiturn	0.457	<u>0.734</u>	*0.730*	**0.810**	blue!80.701	GLU-S^*	0.713
Fanar	TruthfulQA	0.579	**0.641**	*0.633*	0.543	blue!8**0.641**	GLU-AU	0.661
	TriviaQA	0.706	**0.821**	*0.797*	0.729	blue!8**0.821**	GLU-S_q-AU	0.828
	MATH	0.664	<u>0.779</u>	*0.746*	0.666	blue!8**0.788**	GLU-AU	0.801
	ArabicaQA	0.490	<u>0.613</u>	**0.619**	*0.594*	blue!80.563	GLU-AU	0.586
	Longform	<u>0.569</u>	0.549	*0.558*	0.525	blue!8**0.654**	GLU-$\tilde{S}^*$	0.659
	Multiturn	0.463	<u>0.804</u>	*0.780*	0.691	blue!8**0.831**	GLU-AU	0.831

Table 9: Full PRR comparison. Ranking emphasis is applied only across the baselines and GLU. The best ablation is reported for context but is not included in the row-wise ranking.

Model	Dataset	LogTokU	RAUQ	LogProb	$P(\text{true})$	blue!8GLU	Best ablation	
Qwen	TruthfulQA	**0.377**	0.063	*0.185*	0.102	blue!80.195	GLU-EDL	0.392
	TriviaQA	*0.606*	0.581	0.627	0.598	blue!8**0.717**	GLU-S_q-SP	0.723
	MATH	0.216	*0.424*	0.381	**0.574**	blue!80.537	GLU-S_q-SP	0.545
	ArabicaQA	-0.028	*0.186*	0.201	**0.379**	blue!80.140	GLU-AU	0.160
	Longform	-0.045	*-0.044*	**-0.035**	-0.216	blue!8-0.040	GLU-EDL	0.047
	Multiturn	0.008	*0.189*	0.131	**0.777**	blue!80.407	GLU-$\bar{S}$	0.444
Gemma	TruthfulQA	0.115	-0.094	*0.091*	**0.267**	blue!80.087	Add-S_q	0.154
	TriviaQA	0.224	*0.416*	0.686	0.268	blue!8**0.698**	GLU-S_q-AU	0.706
	MATH	0.693	*0.639*	0.438	0.636	blue!8**0.699**	GLU-S_q-SP	0.700
	ArabicaQA	0.099	*0.277*	**0.343**	0.152	blue!80.322	GLU-AU	0.323
	Longform	-0.249	*-0.234*	-0.189	**0.004**	blue!8-0.242	GLU-S_q-AU	-0.201
	Multiturn	-0.274	*0.506*	0.524	**0.723**	blue!80.423	GLU-DK	0.451
Fanar	TruthfulQA	0.192	*0.282*	**0.310**	0.181	blue!80.307	GLU-AU	0.347
	TriviaQA	0.445	0.712	*0.688*	0.448	blue!8**0.730**	GLU-$\bar{S}$	0.732
	MATH	0.324	0.624	*0.520*	0.325	blue!8**0.627**	GLU-AU	0.648
	ArabicaQA	-0.025	0.239	**0.254**	*0.148*	blue!80.143	GLU-AU	0.194
	Longform	*0.077*	0.066	0.083	-0.011	blue!8**0.200**	GLU-$\tilde{S}^*$	0.221
	Multiturn	-0.166	0.654	*0.622*	0.252	blue!8**0.687**	GLU-AU	0.691

Table 10: AUROC for GLU ablations, part 1: EDL, global geometric variants, and local uncertainty variants.

Model	Dataset	GLU-EDL	GLU-$\bar{S}$	GLU-S^*	GLU-$\tilde{S}^*$	GLU-SP	GLU-DK	GLU-AU
Qwen	TruthfulQA	0.680	0.609	0.602	0.612	0.616	0.615	0.596
	TriviaQA	0.788	0.837	0.836	0.835	0.835	0.837	0.824
	MATH	0.559	0.772	0.755	0.759	0.768	0.759	0.755
	ArabicaQA	0.523	0.542	0.535	0.551	0.555	0.557	0.562
	Longform	0.558	0.473	0.420	0.483	0.497	0.515	0.455
	Multiturn	0.485	0.758	0.740	0.734	0.742	0.744	0.726
Gemma	TruthfulQA	0.533	0.545	0.527	0.546	0.549	0.543	0.531
	TriviaQA	0.605	0.817	0.816	0.817	0.817	0.817	0.819
	MATH	0.799	0.799	0.795	0.795	0.799	0.798	0.789
	ArabicaQA	0.539	0.644	0.623	0.639	0.647	0.642	0.650
	Longform	0.346	0.363	0.350	0.358	0.357	0.356	0.379
	Multiturn	0.458	0.703	0.713	0.713	0.700	0.704	0.713
Fanar	TruthfulQA	0.578	0.645	0.637	0.637	0.632	0.632	0.661
	TriviaQA	0.695	0.824	0.822	0.819	0.811	0.813	0.827
	MATH	0.661	0.793	0.796	0.788	0.786	0.788	0.801
	ArabicaQA	0.487	0.564	0.558	0.559	0.546	0.555	0.586
	Longform	0.578	0.617	0.612	0.659	0.658	0.655	0.573
	Multiturn	0.461	0.828	0.825	0.828	0.829	0.828	0.831

Table 11: AUROC for GLU ablations, part 2: evidence-only, combined S_q variants, and additive fusion baselines.

Model	Dataset	GLU-EU	GLU-EU-SP	GLU-S_q-AU	GLU-S_q-SP	GLU-S_q-SP-EU	Add-S_q	Add-$\tilde{S}$
Qwen	TruthfulQA	0.678	0.678	0.593	0.610	0.609	0.454	0.562
	TriviaQA	0.782	0.782	0.828	0.837	0.749	0.320	0.625
	MATH	0.495	0.497	0.760	0.772	0.638	0.317	0.731
	ArabicaQA	0.522	0.522	0.547	0.542	0.499	0.518	0.535
	Longform	0.550	0.546	0.422	0.475	0.493	0.428	0.374
	Multiturn	0.450	0.450	0.739	0.758	0.635	0.256	0.679
Gemma	TruthfulQA	0.531	0.531	0.527	0.546	0.510	0.589	0.516
	TriviaQA	0.594	0.592	0.819	0.817	0.618	0.424	0.615
	MATH	0.782	0.782	0.789	0.799	0.798	0.730	0.801
	ArabicaQA	0.536	0.536	0.646	0.644	0.532	0.592	0.640
	Longform	0.345	0.345	0.390	0.362	0.357	0.327	0.344
	Multiturn	0.441	0.443	0.711	0.703	0.440	0.628	0.647
Fanar	TruthfulQA	0.557	0.555	0.661	0.637	0.561	0.400	0.616
	TriviaQA	0.614	0.607	0.828	0.814	0.630	0.370	0.824
	MATH	0.528	0.521	0.800	0.792	0.545	0.377	0.780
	ArabicaQA	0.494	0.495	0.585	0.547	0.498	0.498	0.570
	Longform	0.507	0.517	0.531	0.624	0.516	0.574	0.529
	Multiturn	0.386	0.377	0.825	0.828	0.392	0.377	0.785

Table 12: PRR for GLU ablations, part 1: EDL, global geometric variants, and local uncertainty variants.

Model	Dataset	GLU-EDL	GLU-$\bar{S}$	GLU-S^*	GLU-$\tilde{S}^*$	GLU-SP	GLU-DK	GLU-AU
Qwen	TruthfulQA	0.392	0.176	0.159	0.188	0.197	0.179	0.163
	TriviaQA	0.618	0.723	0.722	0.719	0.718	0.723	0.704
	MATH	0.127	0.544	0.520	0.525	0.538	0.521	0.516
	ArabicaQA	-0.018	0.129	0.118	0.133	0.139	0.138	0.160
	Longform	0.047	-0.063	-0.122	-0.054	-0.038	-0.012	-0.092
	Multiturn	-0.077	0.444	0.408	0.396	0.412	0.384	0.400
Gemma	TruthfulQA	0.115	0.081	0.049	0.082	0.087	0.076	0.057
	TriviaQA	0.237	0.698	0.697	0.698	0.697	0.697	0.705
	MATH	0.688	0.700	0.695	0.694	0.700	0.698	0.675
	ArabicaQA	0.099	0.319	0.299	0.314	0.321	0.319	0.323
	Longform	-0.245	-0.236	-0.249	-0.238	-0.243	-0.244	-0.213
	Multiturn	-0.267	0.426	0.439	0.443	0.423	0.451	0.442
Fanar	TruthfulQA	0.185	0.309	0.294	0.299	0.297	0.281	0.347
	TriviaQA	0.422	0.732	0.730	0.728	0.718	0.722	0.730
	MATH	0.319	0.634	0.642	0.629	0.626	0.632	0.648
	ArabicaQA	-0.030	0.141	0.136	0.137	0.122	0.137	0.194
	Longform	0.093	0.130	0.127	0.221	0.206	0.205	0.074
	Multiturn	-0.185	0.678	0.658	0.679	0.682	0.681	0.691

Table 13: PRR for GLU ablations, part 2: evidence-only, combined S_q variants, and additive fusion baselines.

Model	Dataset	GLU-EU	GLU-EU-SP	GLU-S_q-AU	GLU-S_q-SP	GLU-S_q-SP-EU	Add-S_q	Add-$\tilde{S}$
Qwen	TruthfulQA	0.387	0.387	0.147	0.178	0.157	-0.161	0.121
	TriviaQA	0.607	0.606	0.710	0.723	0.552	-0.463	0.267
	MATH	-0.008	-0.003	0.524	0.545	0.295	-0.391	0.462
	ArabicaQA	-0.019	-0.019	0.147	0.128	0.018	0.047	0.067
	Longform	0.036	0.030	-0.121	-0.060	-0.025	-0.089	-0.185
	Multiturn	-0.122	-0.116	0.434	0.443	0.198	-0.299	0.283
Gemma	TruthfulQA	0.113	0.112	0.051	0.082	0.049	0.154	0.025
	TriviaQA	0.211	0.205	0.706	0.698	0.257	-0.236	0.228
	MATH	0.663	0.663	0.675	0.700	0.688	0.502	0.692
	ArabicaQA	0.095	0.095	0.321	0.319	0.094	0.154	0.280
	Longform	-0.248	-0.248	-0.201	-0.237	-0.232	-0.340	-0.270
	Multiturn	-0.294	-0.291	0.430	0.426	-0.304	0.303	0.355
Fanar	TruthfulQA	0.174	0.171	0.343	0.298	0.153	-0.206	0.258
	TriviaQA	0.247	0.229	0.732	0.721	0.285	-0.359	0.718
	MATH	-0.005	-0.023	0.646	0.633	0.037	-0.364	0.587
	ArabicaQA	-0.013	-0.013	0.187	0.121	-0.006	-0.026	0.151
	Longform	0.001	0.012	0.011	0.140	0.006	0.076	-0.010
	Multiturn	-0.405	-0.412	0.686	0.677	-0.368	-0.423	0.636