Black-Box Uncertainty Quantification for Large Language Models via Ensemble-of-Ensembles

Wang Ma[1,*] Debarun Bhattacharjya[2,*] Junkyu Lee[3,*]
Nhan H. Pham[4,*] Harsha Kokel[3,*] Qiang Ji[1]
[1]Rensselaer Polytechnic Institute [2]Lirio [3]IBM Software Innovation Lab
[4]Independent Researcher
{maw6, jiq}@rpi.edu, dbhattacharjya@lirio.com, Junkyu.Lee@ibm.com,
nhan.ph0407@gmail.com, harsha.kokel@ibm.com

Abstract

Reliable use of large language models (LLMs) requires uncertainty estimates that separate ambiguity intrinsic to a prompt from genuine knowledge gaps. Bayesian and deep-ensemble methods deliver this aleatoric/epistemic decomposition in principle, but are computationally prohibitive at LLM scale and require white-box access. Existing black-box methods, in contrast, summarize to a single scalar that cannot make the distinction. We introduce a **two-level ensemble** that bridges this gap. Stochastic decoding of a fixed prompt forms the inner ensemble; meaning-preserving semantic perturbations of the prompt form the outer ensemble. Variability is measured in a continuous embedding space, and the law of total covariance gives an exact within-/between-perturbation decomposition to estimate aleatoric and epistemic uncertainty (AU/EU); it requires only black-box sampling access. On top of this estimator, we contribute three theoretical results that ground the framework: (i) a finite-sample bias identity for nested Monte Carlo with a closed-form bias-corrected epistemic estimator; (ii) a *Bayesian bridge* bounding the gap between perturbation-population AU/EU and Bayesian targets; and (iii) a *separation result* identifying a class of failure modes that any consistency-only black-box method provably misses while perturbation-based EU detects. Across five short QA benchmarks and four instruction-tuned LLMs spanning two families and two scales, our estimator is competitive with strong white-box baselines, and at a matched generation budget it matches leading black-box methods on error ranking while decomposing AU and EU; controlled paired interventions show that AU and EU respond selectively to distinct, actionable sources of failure: prompt ambiguity and model knowledge gaps; and a long-form pilot on the NQ long-answer split shows the framework extends to claim-level factuality without retraining.

1 Introduction & Related Work

Large language models (LLMs) are widely deployed in search, code assistance, and scientific workflows, where uncertainty quantification (UQ) helps detect hallucinations and guide abstention. Crucially, when an LLM is uncertain, the appropriate action depends on *why*: an ambiguous query calls for clarification, whereas a knowledge gap calls for retrieval or deferral. Existing black-box UQ methods, however, collapse output variability into a single consistency score that flags unreliability without diagnosing its source. We disentangle these two sources from black-box sampled generations via a two-level ensemble with exact finite-sample bias correction.

*This work was done while Wang Ma was an intern, and Debarun Bhattacharjya, Junkyu Lee, Nhan H. Pham, and Harsha Kokel were researchers, at IBM Research.

Code: https://github.com/wma17/Black-box_UQ_for_LLMs_Ensemble_of_Ensemble

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

From Bayesian UQ to black-box LLMs. Bayesian neural networks [12] and deep ensembles [26] give a principled aleatoric/epistemic split via the law of total variance, but exact or approximate inference over billions of parameters is intractable for modern LLMs, where deployment often offers only *black-box* sampling access.

Related work. LLM-UQ divides into two strands: *white-box* methods that exploit logits, hidden states, or trained heads to score semantic uncertainty or detect hallucinations [22, 10, 24, 11], and *black-box* methods that infer uncertainty from behavioral variability across samples without internal access [35, 13, 58, 59, 4]; surveys offer broader context [57, 33]. A growing line of *semantic-aware* black-box scorers sharpens semantic entropy via kernel-, graph-, or volume-based objects on response embeddings [38, 37, 61, 29, 41]. Concurrent AU/EU-split attempts for LLMs use trained clarification heads, spectral decompositions, variational ICL, paraphrase prompting, or conformal coverage [16, 52, 19, 6, 40]. These works share our goal but emit a single scalar or are structurally *one-level* (perturbations *or* decoding, not both), so they cannot expose the AU/EU split our two-level construction targets. The closest perturbation-based black-box method is SPUQ [13], which perturbs prompts but is one-level and reports a scalar consistency score. Outside the LLM setting, Wang and Ji [53] argued from a first-order optimization view that local input perturbations approximate parameter perturbations for pre-trained vision models; we adapt that intuition to LLMs and turn it from heuristic motivation into an explicit AU/EU approximation bound for a nested estimator (Thm. 4.2; App. A.3). Our two-level construction is what enables both the AU/EU split (Sec. 4.1) and the separation from consistency-only UQ that fixed-prompt methods provably cannot exhibit (Sec. 4.3).

Contributions. We move black-box LLM-UQ beyond a scalar consistency score: our **two-level ensemble** (Fig. 1) admits an exact within-/between-perturbation decomposition of embedded-generation covariance to estimate AU and EU, and we equip the decomposition with theory that makes it both estimable and falsifiable. Concretely:

1. **Exact within-/between-perturbation decomposition with finite-M bias correction for AU/EU estimation.** An exact within-/between-perturbation split of embedded-generation covariance, plus a closed-form bias correction that makes the epistemic estimator unbiased at any inner sample size $M \geq 2$ (Sec. 4.1).
2. **Bayesian bridge to parameter-posterior AU/EU with diagnostic conditions.** Defect bounds connecting perturbation-population AU/EU to Bayesian AU/EU through three local conditions that we estimate as diagnostics from the same outputs the estimator already collects (Sec. 4.2).
3. **Provable separation from consistency-only UQ.** A class of failures any consistency-only method provably misses while perturbation-based EU detects; in paired interventions, prompt clarification significantly reduces EU while leaving fixed-prompt baselines unchanged (Sec. 4.3, 5.4).
4. **Empirical results.** On five short-form QA benchmarks and four instruction-tuned LLMs spanning two families and two scales, the estimator is competitive with strong white-box baselines (Sec. 5.2); at a matched generation budget, it matches leading one-level black-box methods on error ranking while decomposing AU and EU (Sec. 5.3); paired interventions on AmbigQA and Natural Questions show that AU and EU respond selectively to distinct, actionable sources of failure (Sec. 5.4); the predicted finite-M bias correction is verified empirically (Sec. 5.5); a long-form pilot on the NQ long-answer split shows the framework extends to claim-level factuality without retraining (Sec. 5.6); and ablations confirm stability under embedding choice and sampling budget (Sec. 5.7).

2 Background

Bayesian uncertainty decomposition. Bayesian neural networks [14, 12] treat parameters θ as random variables with posterior $p(\theta \mid D)$, yielding a predictive $p(y \mid x, D) = \mathbb{E}_{p(\theta|D)}[p(y \mid x, \theta)]$ that decomposes total predictive uncertainty into *aleatoric* (irreducible data/decoding noise) and *epistemic* (reducible model knowledge) components [23, 18, 8]. For continuous outputs parameterized by $(\mu_\theta(x), \sigma^2_\theta(x))$ [2],

$$\mathrm{AU} = \mathbb{E}_{p(\theta|D)}[\sigma^2_\theta(x)], \quad \mathrm{EU} = \mathrm{Var}_{p(\theta|D)}[\mu_\theta(x)], \quad \mathrm{TU} = \mathrm{AU} + \mathrm{EU}.$$

This regression form is the conceptual target we adapt: we replace scalar mean/variance with embedding-space mean/covariance so the same identity applies to free-form generation, and we replace the parameter posterior (intractable for billion-parameter LLMs) with a perturbation distribution we can sample from.

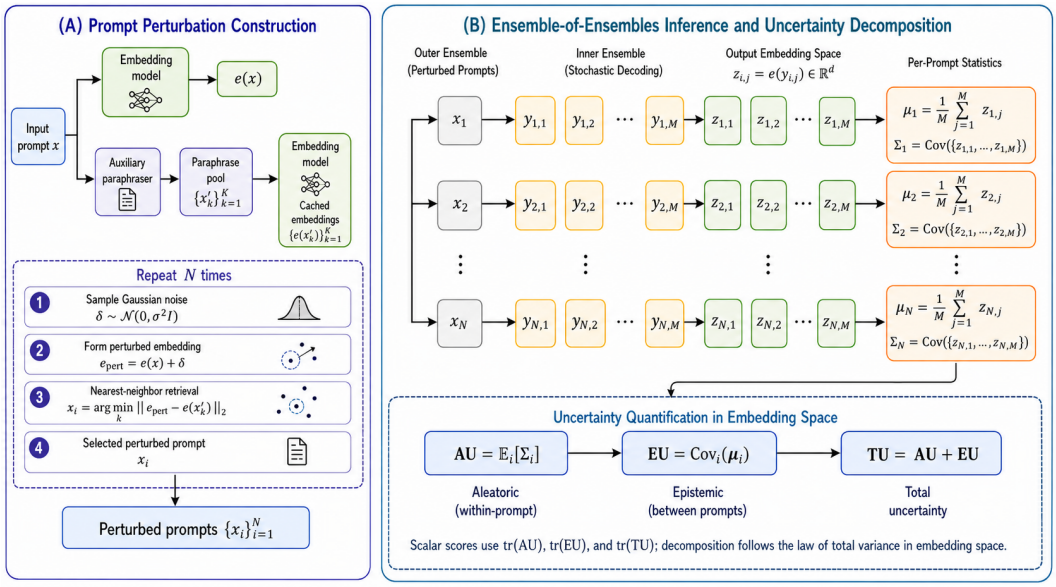

Figure 1: **Two-level ensemble pipeline. (A)** Localized semantic perturbation: snap noisy query embeddings to nearest of K_{pool} cached paraphrases (N times $\rightarrow$ outer ensemble). **(B)** Per-prompt M stochastic decodes; embedding-space statistics $(\hat{\mu}_i, \hat{\Sigma}_i)$ feed the law-of-total-covariance AU/EU/TU split.

Input perturbation as a black-box surrogate. Direct sampling from $p(\theta \mid D)$ is infeasible for pre-trained black-box LLMs. As a methodological motivation, the local equivalence of Wang and Ji [53] (Prop. 3.3) approximates small parameter variations by structured input-embedding perturbations: $f(x+\Delta x, \theta) \approx f(x, \theta+\Delta\theta)$, so input perturbations induce output changes similar to those a small parameter perturbation would induce. We treat this as a *proxy* for $p(y \mid x, D)$ at fixed parameters: input perturbations serve as a tractable surrogate for an intractable parameter posterior. The exact relationship is deferred to Thm. 4.2 (Sec. 4.2), which formalizes the local conditions and bounds the gap between perturbation-population AU/EU and Bayesian AU/EU. Throughout, we operate on three distinct objects: a Bayesian target, an operational perturbation-population quantity, and a finite-sample estimator; the estimator we build in Sec. 3 is defined in perturbation-population terms, and the Bayesian interpretation is established under the local bridge conditions of Sec. 4.2.

3 Proposed Method: Uncertainty Quantification via Ensemble-of-Ensembles

We propose a two-level black-box UQ framework for LLMs that requires neither retraining nor access to token-level logits. An *outer* ensemble draws meaning-preserving semantic perturbations $\Delta \sim p_\Delta$ of the input x, yielding $x_\Delta := T(x, \Delta)$; an *inner* ensemble draws stochastic decodes $Y \sim p(\cdot \mid x_\Delta)$ at fixed deployed parameters $\theta^\star$. We measure variability in a sentence-embedding space $Z = e(Y) \in \mathbb{R}^d$ (assume $\mathbb{E}\|Z\|^2 < \infty$), where the law of total covariance gives

$$\underbrace{\text{Cov}(Z)}_{U_T(x)} = \underbrace{\mathbb{E}_\Delta[\text{Cov}(Z \mid \Delta)]}_{U_A(x)\ \text{(within-perturbation)}} + \underbrace{\text{Cov}_\Delta(\mathbb{E}[Z \mid \Delta])}_{U_E(x)\ \text{(between-perturbation)}} . \tag{1}$$

We adopt Eq. (1) as the definition of the *perturbation-population* uncertainties $(U_A, U_E, U_T) \equiv (\mathcal{U}_A^\Delta, \mathcal{U}_E^\Delta, \mathcal{U}_T^\Delta)$. The interpretation mirrors the Bayesian split: U_A aggregates within-perturbation variability (decoding stochasticity at a fixed prompt and ambiguity in the prompt itself), while U_E measures between-perturbation disagreement of conditional means, i.e. how much the model's expected behavior changes under semantically equivalent phrasings. The corresponding Bayesian targets and finite-sample estimators are formally distinguished in Sec. 4.2 (full definitions in App. A.3.1; summary in App. Tab. 8). Fig. 1 summarizes the full pipeline; pseudocode is in App. Alg. 1.

Outer ensemble: meaning-preserving perturbations. We obtain N semantically equivalent variants $\{x_1, \ldots, x_N\}$ of x by a two-stage localized procedure adapted from Hu et al. [17]: its embedding-space nearest-neighbor step provides the locality our Bayesian bridge (Sec. 4.2) requires.

Stage 1 (offline, once per dataset). An auxiliary paraphraser produces a pool $\mathcal{P}_x = \{x'_k\}_{k=1}^{K_{\text{pool}}}$ for each input x; we cache its embeddings $\{e(x'_k)\}_{k=1}^{K_{\text{pool}}}$.

Stage 2 (online, per query). For $i = 1, \ldots, N$ we sample isotropic Gaussian noise in embedding space and snap to the nearest cached paraphrase:

$$\delta_i \sim \mathcal{N}(0, \varepsilon^2 I_d), \qquad x_i \;=\; \arg\min_{x'_k \in \mathcal{P}_x} D\big(e(x) + \delta_i,\; e(x'_k)\big), \tag{2}$$

where $D(\cdot,\cdot)$ is cosine distance. Since $N \ll K_{\text{pool}}$, this draws a localized neighborhood rather than enumerating the pool. The hyperparameters $(K_{\text{pool}}, \varepsilon)$ trade off locality (staying inside a region where the bridge linearization is plausible) and coverage (paraphrase diversity); both are ablated in App. F. The AU/EU split is robust across perturbation families: random paraphrasing and SPUQ-style noise both yield meaningful AU/EU under the two-level pipeline (App. Tab. 24).

Inner ensemble: stochastic decoding. For each x_i we draw M stochastic decodes $\{y_{i,j}\}_{j=1}^{M} \sim p(\cdot \mid x_i)$ under fixed sampling temperature, and embed them as $z_{i,j} = e(y_{i,j})$. The per-prompt sample mean and covariance are

$$\hat{\mu}_i \;=\; \tfrac{1}{M}\sum_{j=1}^{M} z_{i,j}, \qquad \hat{\Sigma}_i \;=\; \tfrac{1}{M-1}\sum_{j=1}^{M}(z_{i,j} - \hat{\mu}_i)(z_{i,j} - \hat{\mu}_i)^\top. \tag{3}$$

Two-level estimators and bias correction. Aggregating across the outer ensemble gives the empirical AU and (raw) EU estimators

$$\widehat{\mathcal{U}}_A(x) \;=\; \tfrac{1}{N}\sum_{i=1}^{N}\hat{\Sigma}_i, \qquad \widehat{\mathcal{U}}_E^{\text{raw}}(x) \;=\; \tfrac{1}{N-1}\sum_{i=1}^{N}(\hat{\mu}_i - \bar{\mu})(\hat{\mu}_i - \bar{\mu})^\top, \quad \bar{\mu} = \tfrac{1}{N}\sum_i \hat{\mu}_i. \tag{4}$$

Because $\hat{\mu}_i$ is itself an M-sample average, $\widehat{\mathcal{U}}_E^{\text{raw}}$ absorbs an $\frac{1}{M}$ leakage from $\widehat{\mathcal{U}}_A$ (Prop. 4.1); we therefore use the *bias-corrected* estimator throughout,

$$\widehat{\mathcal{U}}_E(x) \;=\; \widehat{\mathcal{U}}_E^{\text{raw}}(x) \;-\; \tfrac{1}{M}\widehat{\mathcal{U}}_A(x), \qquad \widehat{\mathcal{U}}_T(x) \;=\; \widehat{\mathcal{U}}_A(x) + \widehat{\mathcal{U}}_E(x). \tag{5}$$

For correctness ranking we reduce each $d \times d$ matrix to a scalar by the trace, $u_\bullet(x) = \operatorname{tr}\big(\widehat{\mathcal{U}}_\bullet(x)\big)$; `mean`, `max`, and `sum` reductions are ablated in App. F. Per-query cost is NM LLM decodes plus a one-time paraphrase pool of size K_{pool} amortized across the dataset, and can be reduced via per-query adaptive stopping (App. F.1.2). Throughout the paper, K_{pool} refers exclusively to this offline pool; it is distinct from the perturbation count of one-level baselines.

4 Theoretical Results

We establish three theoretical results: (i) an exact finite-M bias identity and concentration bound for the perturbation-population estimators (Pillar 1, Sec. 4.1); (ii) a Bayesian bridge bounding the gap to parameter-posterior AU/EU via three data-estimable local diagnostics (Pillar 2, Sec. 4.2); and (iii) a separation theorem characterizing failure modes that fixed-prompt consistency methods provably miss (Pillar 3, Sec. 4.3); proofs in App. A.

4.1 Pillar 1: bias-corrected finite-M estimator

Estimating U_A, U_E from N outer perturbations and M inner decodes per perturbation introduces a structural bias: the naive epistemic estimator absorbs $\frac{1}{M}$ of the aleatoric covariance.

Proposition 4.1 (Finite-M bias of the epistemic estimator; informal)**.** *The naive estimator $\widehat{U}_E^{\text{raw}}$ obtained as the sample covariance of inner means is biased upward by exactly $\frac{1}{M}\mathcal{U}_A^{\Delta}(x)$ relative to the perturbation-population EU, while $\widehat{U}_A^{(M)}$ is unbiased. The corrected estimator $\widehat{U}_E^{(M)} := \widehat{U}_E^{\text{raw}} - \frac{1}{M}\widehat{U}_A^{(M)}$ is unbiased and concentrates at $O(\sqrt{\log(1/\delta)/N})$ plus lower-order $1/M$ terms. Formal statement, constants, and proof in App. A.2 (Prop. A.1, Thm. A.5).*

The leakage is empirically large at small M: $\rho(\widehat{U}_A, \widehat{U}_E^{\text{raw}})$=0.65 at M=3 on TriviaQA, and the closed-form correction—a single matrix subtraction—drives it to 0.12 at the same budget (Sec. 5.5, Tab. 5).

4.2 Pillar 2: Bayesian bridge with diagnostic conditions

The targets here are the *Bayesian* AU/EU $\mathcal{U}_A^{\text{Bayes}}(x) := \mathbb{E}_\theta[\Sigma_\theta(x)]$, $\mathcal{U}_E^{\text{Bayes}}(x) := \text{Cov}_\theta(m_\theta(x))$ defined over a parameter posterior $p(\theta \mid \mathcal{D})$: the ground truth Bayesian UQ ultimately seeks. Pillar 1 bounds the (estimator)→(perturbation-population) gap from Eq. (1); the bridge below bounds the (perturbation-population)→(Bayesian) gap; composing the two yields an end-to-end bound (formal definitions of the three flavors collected in App. A.3.1; proofs in App. A.3).

Theorem 4.2 (Bayesian bridge for perturbation AU/EU; informal)**.** *Under three local conditions on the model around x (first-order linearity of the conditional mean, locally Gaussian posterior, and bounded semantic coverage defect), the bias-corrected estimators of Pillar 1 satisfy*

$$\left\|\widehat{U}_A^{(M)}(x) - \mathcal{U}_A^{\text{Bayes}}(x)\right\|_F \;\le\; \underbrace{\textit{MC error}}_{O(\sqrt{\log(1/\delta)/N})} \;+\; \underbrace{\varepsilon_A(x)}_{\textit{bridge defect}},$$

and analogously for $\widehat{U}_E^{(M)}$. The defects $\varepsilon_A, \varepsilon_E$ shrink with the three conditions; precise statement, assumptions, and proof are in App. A.3 (Thms. A.10, A.12).

The two model-side conditions align with existing LLM theory: local first-order linearity matches the regime of linear-probing-then-fine-tuning [51], and the locally-Gaussian posterior is the Laplace approximation standard in the Bayesian-LoRA line [60]. We further *measure* each defect from outputs the estimator already collects: D1, an OLS residual of $\mu_\Delta(x)$ on Δ (local linearity); D2, a Henze–Zirkler statistic on the embedding clouds (local Gaussianity); D3, an MMD between the perturbation cloud and a held-out semantic reference (distributional coverage). Per-dataset values are in App. C (Tab. 12). Consistent with first-order scaling, sweeping $\varepsilon \in [0.01, 0.2]$ induces a nearly linear rise in EU (R^2=0.978) while AU shifts by 2.3% (App. F.2.1); instruction perturbations similarly act as implicit parameter updates under the dual gradient view of ICL [7] (App. D).

4.3 Pillar 3: provable separation from consistency-only UQ

Theorem 4.3 (Perturbation–consistency separation; informal)**.** *Let $\mathcal{F}_{\text{cons}}$ be* any *uncertainty estimator that depends only on repeated decoding at a fixed prompt (e.g., 54, 35, 24; SPUQ 13 is not in this class because it perturbs the prompt before decoding). On the class of inputs where decoding is effectively deterministic but $\mu_\Delta(x)$ varies non-trivially across semantic-preserving perturbations, $\mathcal{F}_{\text{cons}}(x) = 0$ while perturbation EU is bounded* away *from zero. The class has positive measure whenever the model has prompt-format-driven outputs. Formal statement and construction in App. A.4.*

Confident hallucinations, memorized templates, and prompt-format artifacts produce stable outputs at a fixed prompt, so any consistency-only score is exactly zero on them while perturbation EU remains positive. In Sec. 5.4, disambiguating a prompt significantly reduces EU while leaving all fixed-prompt baselines unchanged (Tab. 4).

5 Experiments

5.1 Experimental setup

Experiment roadmap. We organize the experiments around six questions. **(Q1)** Under the standard UQ protocol, do our uncertainty scores rank incorrect answers above correct ones? (§5.2, Tab. 1) **(Q2)** At a matched budget, how does two-level sampling compare with one-level designs? (§5.3, Tab. 2) **(Q3)** Beyond correctness prediction, do AU and EU respond to *different* actionable sources of failure? (§5.4, Tabs. 3, 4) **(Q4)** Does the finite-M bias correction predicted by Prop. 4.1 behave as the theory predicts? (§5.5, Tab. 5) **(Q5)** Does the framework extend beyond short-form answers to claim-level long-form factuality? (§5.6, Tab. 6) **(Q6)** Are the conclusions stable under key design choices such as the embedding backbone and the two-level sampling budget? (§5.7, Tab. 7)

Tasks and datasets. We evaluate black-box uncertainty estimation on five short-form QA benchmarks: **TriviaQA** [21] (no-context), **TruthfulQA** [31], **GPQA** [46], **SciQA** [3], and **CoQA** [44]; we also include **AmbigQA** [36] and **Natural Questions** [25] for the paired AU/EU interventions in §5.4, and the **NQ long-answer** split for the long-form pilot in §5.6. The goal is to test whether uncertainty reliably predicts answer correctness.

Table 1: **Headline correctness-prediction AUROC** across five short-form benchmarks. Cos: cosine > 0.8; R-L: ROUGE-L > 0.5. $N{=}M{=}20$. **Bold**: overall best; underlined: best black-box; gray : best white-box. Std. devs. and 70B/72B results in App. Tabs. 9,10,17.

	Method	LLaMA-3-8B										Qwen2.5-7B									
		TriviaQA		TruthfulQA		SciQA		GPQA		CoQA		TriviaQA		TruthfulQA		SciQA		GPQA		CoQA	
		Cos	R-L	Cos	R-L	Cos	R-L	Cos	R-L	Cos	R-L	Cos	R-L	Cos	R-L	Cos	R-L	Cos	R-L	Cos	R-L
Ours	AU	**0.799**	**0.845**	0.782	0.694	0.735	**0.757**	**0.770**	**0.839**	0.738	0.735	**0.773**	**0.815**	**0.812**	**0.731**	0.516	0.508	0.506	0.758	0.585	0.641
	EU	0.585	0.588	0.745	0.647	0.534	0.548	0.746	0.590	0.630	0.631	0.570	0.547	0.684	0.583	0.502	0.501	0.612	**0.949**	0.545	0.516
	TU	0.784	0.826	**0.799**	**0.715**	0.680	0.703	0.671	0.769	0.723	0.724	0.751	0.782	0.809	0.719	0.515	0.512	0.536	0.914	0.548	0.625
Black-box	D-SE	0.742	0.784	0.511	0.550	0.698	0.709	0.641	0.667	0.648	0.659	0.647	0.667	0.530	0.540	0.503	0.503	0.757	0.672	0.615	0.531
	S-Consis.	0.740	0.782	0.524	0.555	0.712	0.706	0.602	0.621	0.641	0.602	0.647	0.667	0.536	0.542	0.658	0.660	0.555	0.544	0.529	0.515
	SIMBA	0.749	0.785	0.506	0.520	0.732	0.722	0.635	0.697	0.583	0.537	0.632	0.663	0.598	0.643	**0.714**	0.695	0.546	0.513	0.566	0.610
	SelfCheckGPT	0.754	0.789	0.517	0.527	0.723	0.713	0.639	0.690	0.566	0.523	0.641	0.650	0.592	0.651	0.693	0.676	0.548	0.503	**0.651**	0.685
	SPUQ	0.615	0.641	0.602	0.583	0.659	0.651	0.554	0.510	0.505	0.505	0.616	0.640	0.544	0.590	0.646	0.634	0.592	0.512	0.506	0.513
	P(True)	0.632	0.652	0.600	0.639	0.566	0.568	0.524	0.623	**0.788**	**0.774**	0.603	0.638	0.763	0.717	0.680	0.681	0.561	0.519	0.520	0.526
White-box	G-NLL	0.786	0.798	0.546	0.549	**0.770**	0.756	0.506	0.543	0.707	0.664	0.646	0.662	0.611	0.700	0.524	0.626	**0.850**	0.748	0.501	**0.827**
	PE	0.710	0.732	0.578	0.576	0.725	0.719	0.680	0.542	0.667	0.636	0.644	0.657	0.562	0.622	0.552	0.675	0.610	0.869	0.511	0.711
	LN-PE	0.761	0.796	0.496	0.496	0.737	0.725	0.639	0.549	0.658	0.654	0.637	0.658	0.570	0.623	0.568	0.650	0.625	0.787	0.526	0.808
	SE	0.739	0.774	0.560	0.525	0.742	0.742	0.601	0.511	0.690	0.675	0.657	0.674	0.597	0.625	0.508	**0.723**	0.742	0.608	0.589	0.667
	LN-SE	0.742	0.777	0.543	0.492	0.735	0.737	0.623	0.608	0.663	0.664	0.653	0.670	0.573	0.608	0.534	0.692	0.826	0.637	0.566	0.543

Models. We use four instruction-tuned LLMs spanning two families and two scales: **LLaMA-3-70B**, **LLaMA-3-8B**, **Qwen2.5-72B**, and **Qwen2.5-7B** [15, 43]. The main paper presents the smaller two (LLaMA-3-8B, Qwen2.5-7B) in Tab. 1; 70B/72B results are in App. Tab. 17.

Correctness labeling. Two complementary automatic criteria: (1) **Cosine similarity** between sentence embeddings of generated vs. reference answers (threshold > 0.8), and (2) **ROUGE-L** overlap (threshold > 0.5) [45, 30]. Both criteria are reported to reduce metric-specific artifacts. In §5.3 and App. F.2.3, we additionally evaluate under **Exact Match (EM)** on TriviaQA as an encoder-independent correctness criterion.

Uncertainty measures. We report AUROC for correctness prediction using **aleatoric** (AU), **epistemic** (EU), and **total** uncertainty (TU) as defined in Eq. (1); following the UQ convention, the positive class is *incorrect* and the scoring signal is the (uncorrected sign) uncertainty value, so AUROC measures how well higher uncertainty ranks errors above correct answers.

Baselines. Black-box: D-SE [24], Self-Consistency [54], SIMBA [4], SelfCheckGPT [35], SPUQ [13], and a P(True) probe [22]. White-box: G-NLL [1], PE / LN-PE [34], and SE / LN-SE [11, 24]. The long-form pilot (§5.6) additionally compares against five long-form-specific scorers from Bouchard et al. [5]: avgEntail, NonContradiction, UQLM cosine, plus SelfCheckGPT and ROUGE-L (cons.) at long-form. Full baseline configurations are in App. H; SPUQ uses its three perturbation modes with $K_{\mathrm{SPUQ}}{=}10$ perturbations (SPUQ's perturbation count, distinct from our offline pool size K_{pool}; K_{SPUQ} plays the role of our outer count N) and pairwise ROUGE-L aggregation [13] (see App. D).

Sampling budget. For the headline AUROC table (Tab. 1) we use $N{=}M{=}20$ (budget ablation in App. F.1); the AU/EU validation experiments (§5.4) and all other ablations use the compute-efficient default $N{=}M{=}10$, and the long-form pilot (§5.6) uses the further-reduced $N{=}M{=}5$ for efficiency over a separately-extracted claim set (see App. E). When comparing to SPUQ in App. D, we match the *outer perturbation count* ($K_{\mathrm{SPUQ}}{=}N{=}10$); our method then incurs the additional cost of an inner ensemble (M stochastic decodes per perturbation) which is intrinsic to the two-level decomposition and not present in any one-level baseline. Per-method per-question call counts and wall-clock times are tabulated in App. Tab. 20.

Other settings. Decoding temperature T is treated as an internal property of the deployed LLM, not an estimator knob: a targeted T-sweep (App. F.7) shows EU AUROC stable across $T \in [0.3, 1.5]$. The perturbation noise scale $\varepsilon{=}0.1$ comes from the noise-level ablation (App. F.2.1); the paraphrase pool size is $K_{\mathrm{pool}}{=}500$ for short-form QA (the ablation in App. Tab. 30 shows $K_{\mathrm{pool}} \in [20, 100]$ is also competitive).

5.2 Headline AUROC: short-form correctness prediction

Tab. 1 reports correctness-prediction AUROC against strong white-box and black-box baselines under the standard protocol; the behavioral semantics of the AU/EU split are tested in §5.4.

Table 2: **Matched generation budget (B) comparison** on TriviaQA (EM label; AUROC with incorrect as positive; 400 questions per model). **Bold**: best per column; underlined: best black-box. †Requires token log-probabilities (white-box). Bootstrap 95% CIs and additional baselines: App. Tabs. 21, 22.

Method	LLaMA-3-8B			Qwen2.5-7B		
	B=10	25	100	10	25	100
SPUQ (one-level, perturbed prompts)	**0.756**	**0.769**	0.778	0.759	0.764	0.782
Embedding variance (fixed prompt)	0.754	0.762	**0.782**	**0.774**	0.801	0.816
Perturbation-only (N=B, M=1)	0.754	0.667	0.657	0.743	0.553	0.549
Decoupled (separate components, summed)	**0.756**	0.738	0.750	0.765	0.712	0.731
Semantic Density, uniform	0.600	0.594	0.586	0.721	0.776	0.765
Semantic Density, likelihood-weighted†	0.663	0.651	0.644	**0.774**	**0.821**	**0.821**
Ours: AU (within)	0.744	0.752	0.766	0.772	0.789	0.788
Ours: TU (total)	0.744	0.737	0.755	0.736	0.734	0.751

Table 3: 2×2 **paired interventions** (LLaMA-3-8B, N=M=10, 1,000 examples). All four contrasts significant at p<0.01. Δ: change in AUROC. Protocol and Qwen replication: App. B.

Test	Metric	ΔAU	ΔEU	Contrast (z, p)
AmbigQA (clarify ⇒ moves AU)	Cosine	$+\mathbf{0.058}\pm0.005$	$+0.021\pm0.007$	$\Delta\text{AU}-\Delta\text{EU}=+0.037$ (z=2.74, p=0.006)
	ROUGE-L	$+\mathbf{0.044}\pm0.005$	$+0.012\pm0.007$	$\Delta\text{AU}-\Delta\text{EU}=+0.032$ (z=4.18, p<0.001)
NQ (evidence ⇒ moves EU)	Cosine	$+0.039\pm0.006$	$+\mathbf{0.069}\pm0.007$	$\Delta\text{EU}-\Delta\text{AU}=+0.030$ (z=4.28, p<0.001)
	ROUGE-L	$+0.031\pm0.006$	$+\mathbf{0.078}\pm0.007$	$\Delta\text{EU}-\Delta\text{AU}=+0.047$ (z=6.56, p<0.001)

(i) AU and TU are the strongest black-box signals overall, competitive with white-box methods that require token logits: AU reaches 0.845/0.815 under ROUGE-L on TriviaQA (LLaMA-3-8B / Qwen2.5-7B), the best black-box score in both columns, and AU or TU achieves the best black-box score in 12 and the best overall score in 11 of 20 benchmark×correctness×model cells.

(ii) AU is robust where prior black-box baselines collapse: on TruthfulQA, SPUQ (≤ 0.602) and SIMBA on LLaMA-3-8B (≤ 0.520) drift toward chance while AU remains informative.

(iii) EU contributes a complementary signal. EU targets knowledge-gap detection rather than generic correctness ranking, as validated by paired intervention in Tab. 3; on Qwen2.5-7B / GPQA / ROUGE-L it is the single best score across *all* methods (0.949), and on the three hard benchmarks (GPQA, SciQA, CoQA) AU and EU together hold the top black-box position in $5/12$ Cos/R-L columns. The 70B/72B counterpart on TriviaQA (App. Tab. 17) extends the same pattern.

5.3 Matched-budget comparison against one-level designs

Tab. 2 fixes the per-question generation budget $B \in \{10, 25, 100\}$ across 400 TriviaQA questions per model (EM labels) and compares alternative allocations: all B decodes at the original prompt (*fixed prompt*); one decode per paraphrase across B paraphrases (*perturbation-only*, N=B, M=1); an equal split between the two with summed variances (*decoupled*); SPUQ; Semantic Density [42]; and our nested estimator with $N\times M$=5×2, 5×5, and 10×10.

(i) Competitive error ranking with an explicit AU/EU split. At every budget B, SPUQ, fixed-prompt embedding variance, and our AU lie within 0.04 AUROC of one another with overlapping 95% CIs (our TU is within 0.01–0.03 on LLaMA-3-8B and 0.04–0.07 on Qwen2.5-7B of the top black-box score), matching one-level error ranking while decomposing AU and EU.

(ii) Stability across budgets. Our AU and TU vary by less than 0.03 across all budgets B. By contrast, perturbation-only (N=B, M=1) drops from 0.754 to 0.667 on LLaMA-3-8B and from 0.743 to 0.553 on Qwen2.5-7B between B=10 and 25, as M=1 lacks inner averaging and draws from an extended paraphrase pool at $B\geq$25 (App. F.1.1).

(iii) Semantic Density. Our AU outperforms uniform Semantic Density in every column, whereas the likelihood-weighted variant leads on Qwen2.5-7B at $B\geq$25 by utilizing white-box token probabilities.

Table 4: **Standardized score shifts under paired interventions** (LLaMA-3-8B; 1,000 pairs per intervention): mean of (after − before) divided by the SD of per-question changes. **Bold**: 95% bootstrap CI excludes zero.

Score	Clarify (AmbigQA)	Evidence (NQ)
Ours: AU (within)	-0.013	$\mathbf{-0.610}$
Ours: EU (between)	$\mathbf{-0.218}$	$\mathbf{-0.604}$
Ours: TU (total)	$\mathbf{-0.156}$	$\mathbf{-0.725}$
Embedding variance (fixed prompt)	-0.038	$\mathbf{-0.546}$
Self-Consistency	-0.022	$\mathbf{-0.518}$
D-SE	-0.031	$\mathbf{-0.576}$
Semantic Density, uniform	$+0.018$	$\mathbf{+0.079}$
SPUQ (perturbed prompts)	$\mathbf{-0.106}$	$\mathbf{-0.836}$

Table 5: **Bias correction validated.** Spearman ρ between AU and EU on TriviaQA (LLaMA-3-8B, N=10, Cosine).

M	3	5	10	20	30
$\rho(\widehat{U}_A, \widehat{U}_E^{\text{raw}})$	0.647	0.521	0.393	0.312	0.283
$\rho(\widehat{U}_A, \widehat{U}_E^{\text{corr}})$	0.115	0.169	0.199	0.211	0.213
Gap	0.532	0.352	0.194	0.101	0.070

Table 6: **Claim-level long-form pilot** (NQ long-answer, per-claim LLM-judge factuality). **Bold**: best per column; underlined: second-best. ROUGE-L (cons.) is the consistency-based self-similarity scorer (not the correctness criterion of Tab. 1). Pipeline, response-level companion, and discussion: App. E.

Method	**LLaMA-3-8B** (n=1,104 claims)			**Qwen2.5-7B** (n=872 claims)		
	AUROC ↑	AUPRC ↑	ECE ↓	AUROC ↑	AUPRC ↑	ECE ↓
Ours (AU)	0.698	0.940	0.045	0.659	0.920	0.034
Ours (EU)	0.663	0.928	**0.002**	0.605	0.904	**0.004**
Ours (TU)	0.696	0.940	0.029	0.652	0.918	0.037
avgEntail	**0.707**	0.933	0.044	**0.719**	**0.940**	0.028
NonContradiction	0.536	0.887	0.017	0.559	0.897	0.012
UQLM cosine	0.693	**0.944**	0.033	0.643	0.915	0.028
SelfCheckGPT	0.619	0.915	**0.002**	0.554	0.879	**0.004**
ROUGE-L (cons.)	0.548	0.905	**0.002**	0.490	0.871	**0.004**

5.4 Behavioral validation of the AU/EU split

We test the AU/EU split behaviorally with a 2×2 paired-intervention protocol that targets one source of uncertainty at a time, directly probing the actionability of the decomposition—a property correctness-prediction AUROC alone cannot reveal. All experiments use LLaMA-3-8B with N=M=10 on 1,000 examples (Qwen2.5-7B replication in App. B). The paired statistic differences the per-example AUROC contribution between treatment and control, yielding a Wald z-test for the contrast $\Delta\text{AU} - \Delta\text{EU}$ (or its twin).

Paired interventions (2×2). On AmbigQA [36], adding a gold clarification removes ambiguity; on Natural Questions [25], adding the gold evidence passage closes the knowledge gap (perturbations applied only to the question, not the passage). The framework predicts clarification moves *AU* more than EU, and evidence moves *EU* more than AU. Tab. 3 confirms both: the contrast is significant in every cell (p<0.01), and the absolute change on the predicted-target head is about 2–4× the change on the off-target head in every row.

Score-level shifts and fixed-prompt baselines. Whereas Tab. 3 evaluates discriminative error-ranking gains (ΔAUROC), Tab. 4 measures standardized shifts in raw uncertainty magnitude under both interventions across our three components, four fixed-prompt baselines, and SPUQ (LLaMA-3-8B, 10 decodes per baseline; App. B).

Supplying evidence on NQ significantly reduces uncertainty across all methods except uniform Semantic Density. By contrast, clarifying the prompt on AmbigQA significantly reduces *only* the perturbation-sensitive scores—EU (−0.218), TU (−0.156), and SPUQ (−0.106)—while all fixed-prompt baselines remain statistically unchanged, corroborating Thm. 4.3.

Because SPUQ emits a single scalar that drops under both interventions, it cannot distinguish prompt ambiguity from missing knowledge. Our decomposition separates the two: clarification resolves cross-paraphrase divergence over competing interpretations (lowering EU magnitude in Tab. 4) while sharpening within-prompt variance to better rank residual errors (ΔAU=+0.058, Tab. 3).

Table 7: **Ablations. (a)** AU AUROC under three sentence encoders; right column: min cross-encoder per-example Spearman ρ of AU. **(b)** Tied $N{=}M$ budget sweep (LLaMA-3-8B). Cosine label, $\varepsilon{=}0.1$; (a) uses $N{=}M{=}10$. **Bold**: best; underlined: second-best. Full grids: App. Tabs. 28, 29, 18.

(a) Embedding backbone

Model	Dataset	AU AUROC ↑ SBERT	MPNet	BGE	$\min_{e\neq e'} \rho^{\mathrm{AU}}_{e,e'}$ ↑
LLaMA-3-8B	TriviaQA	**0.845**	0.843	**0.845**	0.974
LLaMA-3-8B	TruthfulQA	**0.695**	0.690	0.650	0.955
Qwen2.5-7B	TriviaQA	**0.855**	0.851	0.854	0.981
Qwen2.5-7B	TruthfulQA	**0.630**	0.620	0.615	0.961

(b) Sampling budget (TruthfulQA, Cos.)

$N{=}M$	AU	EU	TU
3	0.745	0.707	0.755
5	**0.798**	0.756	0.797
10	0.792	**0.803**	**0.815**
20	0.782	0.745	0.799
30	0.768	0.746	0.791

5.5 Empirical validation of the finite-M bias correction

The behavioral test above uses the bias-corrected EU estimator. We next verify that the correction predicted by Prop. 4.1 is empirically necessary and behaves as the theory predicts (Tab. 5). The naive estimator inflates EU by $\frac{1}{M}U_A$, so the raw AU–EU correlation should start high at small M and converge from above to the corrected correlation as $1/M \to 0$.

On TriviaQA (LLaMA-3-8B, Cosine, $N{=}10$) the gap shrinks monotonically from 0.532 at $M{=}3$ to 0.070 at $M{=}30$ (Tab. 5); TruthfulQA shrinks identically ($0.474 \to 0.090$ for $M{=}3 \to 20$). The same shrinkage replicates on Qwen2.5-7B and across both datasets: at the level of the raw uncertainty scores themselves, the mean raw-vs-corrected EU gap shrinks by roughly an order of magnitude from $M{=}3$ to $M{=}30$ in every cell of the full $\{\text{LLaMA-3-8B, Qwen2.5-7B}\} \times \{\text{TriviaQA, TruthfulQA}\}$ grid (and across both Cosine and ROUGE-L correctness criteria), matching the $\frac{1}{M}$ rate predicted by Prop. 4.1 and confirming the correction is model- and dataset-invariant.

5.6 Beyond short-form: long-form pilot on NQ Long-Answer

Short-answer QA gives clean correctness labels but tests only one slice of LLM use. To check that the framework extends to long-form generation, we apply our two-level estimator to the **NQ long-answer** split [25] via segment-level decomposition: each generated long answer is split into atomic claims, each claim is scored independently with our two-level pipeline, and the resulting per-claim score is correlated with a per-claim factuality label (LLM judge). We compare against five published long-form-specific black-box baselines: **avgEntail**, **NonContradiction** [5], **UQLM cosine** [47], **SelfCheckGPT** [35], and **ROUGE-L (cons.)** [30]; baselines and our method draw from the same inner-sample pool (M stochastic decodes at the unperturbed prompt) so the consistency-only comparators see exactly the budget our inner level sees.

Tab. 6 reports the claim-level result on LLaMA-3-8B ($n{=}1{,}104$ claims) and Qwen2.5-7B ($n{=}872$ claims).

(i) Ours (AU) is the second-best AUROC method on both models, within 0.009 of avgEntail on LLaMA-3-8B and 0.060 on Qwen2.5-7B, where avgEntail leverages a pretrained NLI classifier tailored to claim entailment.

(ii) The AU $>$ EU ordering is preserved at long-form, mirroring the short-form result.

(iii) We outperform surface-similarity baselines (SelfCheckGPT, ROUGE-L (cons.)) on both models. The two-level structure transports cleanly to claim-level long-form factuality without retraining; combining our perturbation outer level with NLI-based inner aggregation is a natural extension. The response-level companion table is in App. E.

5.7 Ablation studies

We ablate five orthogonal axes; our primary findings hold across all of them. The two most central axes are summarized in Tab. 7: the embedding backbone (panel a) and the two-level sampling budget (panel b). The remaining axes (noise magnitude ε, perturbation mechanism, vector-to-scalar aggregation, and paraphrase pool size K_{pool}) are in App. F.

(1) Embedding backbone: AU is encoder-stable in absolute level and rank-stable per-example. Re-encoding the same generations under three off-the-shelf sentence encoders (Sentence-BERT [45], MPNet, BGE-base) yields AU AUROC spreads of ≤ 0.05 within every (model, dataset) cell, worst-case 0.045 on LLaMA-3-8B / TruthfulQA (Tab. 7a). The stronger statement is on *per-example* ranks: pairwise Spearman correlations of AU scores across encoders are at least 0.955 in every cell. A downstream consumer of our scores (abstention thresholds, selective evaluation) gets nearly the same ranking regardless of encoder choice; both axes extend across both correctness criteria (spread at most 0.053 under ROUGE-L; App. Tab. 28). Different encoders agree on which inputs the model is uncertain about: the AU signal lives in the LLM's behavior, not in the choice of $e(\cdot)$.

(2) Budget: gains saturate by $(N, M)=(10, 10)$. On TruthfulQA (Tab. 7b), TU AUROC climbs from 0.755 at $N=M=3$ to 0.815 at $(10, 10)$ under Cosine and then plateaus; the same saturation pattern holds on TriviaQA and under ROUGE-L (App. Tab. 18; full M and N sweeps with one dimension fixed in App. Tabs. 19, 19b). Headline runs use $(N, M)=(20, 20)$; validation experiments use $(10, 10)$ (matched to SPUQ's [13] $K_{\mathrm{SPUQ}}=10$). In practice, 3×3 or 5×5 suffices for screening and 10×10 yields stable components; adaptive stopping further cuts the median cost to 25 generations with ≤ 0.014 AUROC loss on TriviaQA (App. F.1.2).

(3) Perturbation, aggregation, and pool size are secondary. Noise scale ε has a validity–diversity sweet spot at $\varepsilon=0.1$ (AU AUROC peaks at $\varepsilon\in[0.05, 0.2]$ on TruthfulQA, drops sharply at $\varepsilon\geq0.5$; App. Tab. 23); perturbation mechanism is robust (paraphrase, distractor injection, instruction variation (strongest), and SPUQ-style noise all give non-trivial AU/EU under the two-level pipeline; App. Tab. 24); vector-to-scalar reduction (mean, sum, max, norm) shifts AU/TU AUROC by at most 0.05 (App. Tab. 26); paraphrase pool size shows a coverage–purity trade-off where $K_{\mathrm{pool}}\in[20, 100]$ is also competitive (App. Tab. 30).

Default setting. A compute-efficient default derived from these ablations is $(N, M)=(10, 10)$, $\varepsilon=0.1$, Sentence-BERT, $K_{\mathrm{pool}}\approx50$; pushing budget or pool size beyond this yields ≤ 0.02 AUROC gain. Qualitative case studies of ambiguity- and knowledge-driven uncertainty regimes, including failure cases, are in App. G.

6 Conclusion

We introduced an inference-time, black-box UQ framework for LLMs that requires neither retraining nor access to token logits. By combining stochastic decoding within a fixed prompt with meaning-preserving perturbations across prompts, the framework applies the law of total covariance in embedding space to decompose predictive variability into within-perturbation (aleatoric, AU) and between-perturbation (epistemic, EU) components.

Theoretically, we establish a finite-M bias identity with a closed-form correction, a Bayesian bridge whose conditions we measure as data-estimable diagnostics rather than assert, and a separation result identifying failures consistency-only UQ provably misses.

Empirically, the estimator is competitive with strong white-box baselines across five benchmarks and four LLMs (Tab. 1), and under a matched generation budget it matches leading black-box methods on error ranking while decomposing AU and EU (Tab. 2). Paired interventions demonstrate that AU and EU respond selectively to prompt ambiguity and knowledge gaps, capturing a disambiguation effect that fixed-prompt baselines miss (Tabs. 3, 4). Finite-M bias validation, the claim-level long-form pilot, and comprehensive ablations further confirm the robustness of the decomposition (Tabs. 5–7).

Looking ahead, the AU/EU split can guide active data acquisition for high-EU queries [28, 39], selective retrieval in RAG [20, 50], adaptive sampling or tool invocation [48, 47], and conformal prediction in the black-box regime [40].

Limitations. (**L1**) Our estimators target perturbation-population AU/EU at fixed model weights, whose correspondence to Bayesian parameter uncertainty relies on the local regularity conditions diagnosed by D1–D3 (Sec. 4.2). (**L2**) Automated paraphrasing can introduce semantic drift that inflates between-prompt variance (App. F.2.3). (**L3**) Uncertainty estimates inherit the representational geometry and potential blind spots of the sentence encoder $e(\cdot)$ (Tab. 7a). (**L4**) Nested sampling requires NM decodes per query to separate AU and EU, incurring higher inference cost than single-level baselines (Sec. 5.3). (**L5**) Our empirical and interventional evaluations focus primarily on short-form QA with controlled single-factor interventions. See App. I for a detailed discussion.

Acknowledgments and Disclosure of Funding

This work was supported by IBM through the IBM-Rensselaer Future of Computing Research Collaboration. The authors declare no competing interests.

References

[1] Lukas Aichberger, Kajetan Schweighofer, and Sepp Hochreiter. Rethinking uncertainty estimation in natural language generation. *arXiv preprint arXiv:2412.15176*, 2024.

[2] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. *Advances in neural information processing systems*, 33:14927–14937, 2020.

[3] Sören Auer, Dante AC Barone, Cassiano Bartz, Eduardo G Cortes, Mohamad Yaser Jaradeh, Oliver Karras, Manolis Koubarakis, Dmitry Mouromtsev, Dmitrii Pliukhin, Daniil Radyush, et al. The sciqa scientific question answering benchmark for scholarly knowledge. *Scientific Reports*, 13(1):7240, 2023.

[4] Debarun Bhattacharjya, Balaji Ganesan, Junkyu Lee, Radu Marinescu, Katsiaryna Mirylenka, Michael Glass, and Xiao Shou. SIMBA UQ: Similarity-based aggregation for uncertainty quantification in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15880–15894, 2025.

[5] Dylan Bouchard, Mohit Singh Chauhan, Viren Bajaj, and David Skarbrevik. Fine-grained uncertainty quantification for long-form language model outputs: A comparative study. *arXiv preprint arXiv:2602.17431*, 2026. Code: `https://github.com/cvs-health/uqlm`.

[6] Kyle Cox, Jiawei Xu, Yikun Han, Rong Xu, Tianhao Li, Chi-Yang Hsu, Tianlong Chen, Walter Gerych, and Ying Ding. Mapping from meaning: Addressing the miscalibration of prompt-sensitive language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23696–23703, 2025.

[7] Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why can GPT learn in-context? Language Models secretly perform gradient descent as meta-optimizers. In *Findings of ACL*, 2023.

[8] Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International conference on machine learning*, pages 1184–1193. PMLR, 2018.

[9] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? Does it matter? *Structural Safety*, 31(2):105–112, 2009.

[10] Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. Lm-polygraph: Uncertainty estimation for language models. In *EMNLP (Demos)*, pages 446–461. Association for Computational Linguistics, 2023.

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

[12] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[13] Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. SPUQ: perturbation-based uncertainty quantification for large language models. In *EACL (1)*, pages 2336–2346. Association for Computational Linguistics, 2024.

[14] Ethan Goan and Clinton Fookes. Bayesian neural networks: An introduction and survey. In *Case Studies in Applied Bayesian Data Science: CIRM Jean-Morlet Chair, Fall 2018*, pages 45–87. Springer, 2020.

[15] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[16] Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. Decomposing uncertainty for large language models through input clarification ensembling. In *ICML*. PMLR, 2024.

[17] Wenyang Hu, Yao Shu, Zongmin Yu, Zhaoxuan Wu, Xiaoqiang Lin, Zhongxiang Dai, See-Kiong Ng, and Bryan Kian Hsiang Low. Localized zeroth-order prompt optimization. *Advances in Neural Information Processing Systems*, 37:86309–86345, 2024.

[18] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506, 2021.

[19] I. Shavindra Jayasekera, Jacob Si, Filippo Valdettaro, Wenlong Chen, A. Aldo Faisal, and Yingzhen Li. Variational uncertainty decomposition for in-context learning. In *Advances in Neural Information Processing Systems*, 2025.

[20] Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, 2023.

[21] Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL (1)*, pages 1601–1611. Association for Computational Linguistics, 2017.

[22] 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.

[23] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[24] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *ICLR*. OpenReview.net, 2023.

[25] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019.

[26] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30, 2017.

[27] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7871–7880, 2020.

[28] Jiaqi Li, Yixuan Tang, and Yi Yang. Know the unknown: An uncertainty-sensitive method for llm instruction tuning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2972–2989, 2025.

[29] Xiaomin Li, Zhou Yu, Ziji Zhang, Yingying Zhuang, Swair Shah, Narayanan Sadagopan, and Anurag Beniwal. Semantic volume: Quantifying and detecting both external and internal uncertainty in LLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

[30] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[31] 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.

[32] Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Trans. Mach. Learn. Res.*, 2022, 2022.

[33] Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. Uncertainty quantification and confidence calibration in large language models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 6107–6117, 2025.

[34] Andrey Malinin and Mark J. F. Gales. Uncertainty estimation in autoregressive structured prediction. In *ICLR*. OpenReview.net, 2021.

[35] Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9004–9017, 2023.

[36] Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of EMNLP*, 2020.

[37] Dang Nguyen, Ali Payani, and Baharan Mirzasoleiman. Beyond semantic entropy: Boosting LLM uncertainty quantification with pairwise semantic similarity. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

[38] Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[39] Ruijia Niu, Dongxia Wu, Rose Yu, and Yi-An Ma. Functional-level uncertainty quantification for calibrated fine-tuning on llms. *arXiv preprint arXiv:2410.06431*, 2024.

[40] Sima Noorani, Shayan Kiyani, George J. Pappas, and Hamed Hassani. Conformal prediction beyond the seen: A missing mass perspective for uncertainty quantification in generative models. In *Advances in Neural Information Processing Systems*, 2025.

[41] Edward Phillips, Sean Wu, Soheila Molaei, Danielle Belgrave, Anshul Thakur, and David A. Clifton. Geometric uncertainty for detecting and correcting hallucinations in LLMs. *arXiv preprint arXiv:2509.13813*, 2025.

[42] Xin Qiu and Risto Miikkulainen. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. In *Advances in Neural Information Processing Systems*, volume 37, pages 134507–134533, 2024.

[43] 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, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[44] Siva Reddy, Danqi Chen, and Christopher D Manning. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019.

[45] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP/IJCNLP (1)*, pages 3980–3990. Association for Computational Linguistics, 2019.

[46] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

[47] Brendan Leigh Ross, Noël Vouitsis, Atiyeh Ashari Ghomi, Rasa Hosseinzadeh, Ji Xin, Zhaoyan Liu, Yi Sui, Shiyi Hou, Kin Kwan Leung, Gabriel Loaiza-Ganem, et al. Textual bayes: Quantifying uncertainty in llm-based systems. *arXiv preprint arXiv:2506.10060*, 2025.

[48] Zhenning Shi, Yijia Zhu, Yi Xie, Junhan Shi, Guorui Xie, Haotian Zhang, Yong Jiang, Congcong Miao, and Qing Li. Reasoning under uncertainty: Efficient llm inference via unsupervised confidence dilution and convergent adaptive sampling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32192–32206, 2025.

[49] Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018.

[50] Heydar Soudani, Hamed Zamani, and Faegheh Hasibi. Uncertainty quantification for retrieval-augmented reasoning. *arXiv preprint arXiv:2510.11483*, 2025.

[51] Akiyoshi Tomihari and Issei Sato. Understanding linear probing then fine-tuning language models from NTK perspective. In *Advances in Neural Information Processing Systems*, volume 37, pages 139786–139822, 2024.

[52] Nassim Walha, Sebastian G. Gruber, Thomas Decker, Yinchong Yang, Alireza Javanmardi, Eyke Hüllermeier, and Florian Buettner. Fine-grained uncertainty decomposition in large language models: A spectral approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

[53] Hanjing Wang and Qiang Ji. Epistemic uncertainty quantification for pre-trained neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11052–11061, 2024.

[54] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR*. OpenReview.net, 2023.

[55] Joseph Wilson, Chris van der Heide, Liam Hodgkinson, and Fred Roosta. Uncertainty quantification with the empirical neural tangent kernel. *arXiv preprint arXiv:2502.02870*, 2025.

[56] Lisa Wimmer, Yusuf Sale, Paul Hofman, Bernd Bischl, and Eyke Hüllermeier. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 2282–2292. PMLR, 2023.

[57] Zhiqiu Xia, Jinxuan Xu, Yuqian Zhang, and Hang Liu. A survey of uncertainty estimation methods on large language models. In *ACL (Findings)*, volume ACL 2025 of *Findings of ACL*, pages 21381–21396. Association for Computational Linguistics, 2025.

[58] Quan Xiao, Debarun Bhattacharjya, Balaji Ganesan, Radu Marinescu, Katsiaryna Mirylenka, Nhan H. Pham, Michael R. Glass, and Junkyu Lee. The consistency hypothesis in uncertainty quantification for large language models. In *UAI*, volume 286 of *Proceedings of Machine Learning Research*, pages 4636–4651. PMLR, 2025.

[59] Duygu Nur Yaldiz, Yavuz Faruk Bakman, Baturalp Buyukates, Chenyang Tao, Anil Ramakrishna, Dimitrios Dimitriadis, Jieyu Zhao, and Salman Avestimehr. Do not design, learn: A trainable scoring function for uncertainty estimation in generative llms. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 691–713, 2025.

[60] Adam X. Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison. Bayesian low-rank adaptation for large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

[61] Xingtao Zhao, Hao Peng, Dingli Su, Xianghua Zeng, Chunyang Liu, Jinzhi Liao, and Philip S. Yu. SeSE: Black-box uncertainty quantification for large language models based on structural information theory. *arXiv preprint arXiv:2511.16275*, 2025.

A Additional Theoretical Details

Notation. We adopt the notation of Sec. 3: $Z = e(Y) \in \mathbb{R}^d$ is the embedded output, $\mu(\Delta) = \mathbb{E}[Z \mid \Delta]$ and $\Sigma(\Delta) = \mathrm{Cov}(Z \mid \Delta)$ the conditional moments, and (U_A, U_E, U_T) the population AU/EU/TU defined in Eq. (1). Throughout we assume $\mathbb{E}\|Z\|^2 < \infty$. The identity $U_T = U_A + U_E$ is the law of total covariance and is taken as given. Standard tools (matrix-valued LLN, Hoeffding's inequality, entrywise-to-Frobenius conversion) are used without re-derivation.

Roadmap and connection to the main paper. Each appendix subsection backs up a specific main-paper claim. The dependency graph is:

Main-paper claim	Backed by	Purpose
Pillar 1: bias-corrected estimator (Prop. 4.1, main)	App. A.1	finite-M bias identity
	Prop. A.1	exact $1/M$ inflation
	Thm. A.2	a.s. consistency, raw & corrected
Pillar 1 rate bounds	App. A.2	finite-sample concentration
	Thm. A.3	population-level outer bound
	Thm. A.5	full two-level (N, M) bound
Pillar 2: Bayesian bridge (Thm. 4.2, main)	App. A.3	defect bounds for $\varepsilon_A, \varepsilon_E$
	Thm. A.10	EU-side bridge bound
	Thm. A.12	AU-side bridge bound
	App. C	data-estimable diagnostics
Pillar 3: separation result (Thm. 4.3, main)	App. A.4	explicit construction
	Thm. A.13	$\mathcal{F}_{\mathrm{cons}}{=}0$, $U_E \succ 0$

How to read the proofs. Within each subsection we follow the same pattern: state the assumptions, give the formal claim, prove it, and close with a one-line "*what this gives the main paper.*" Steps inside proofs are explicitly named ("Step 1: $\hat{\mu}_i$ is unbiased…") so the reader can match the algebra to its intended role.

A.1 Two-Level Estimation with Finite Inner Sampling

Purpose of this subsection. The main paper's Pillar 1 claims that the naive sample-covariance estimator of EU *inflates* the true EU by $\frac{1}{M}U_A$ and that subtracting $\frac{1}{M}\hat{U}_A$ removes this inflation. This subsection establishes both: (i) the exact bias identity (Prop. A.1) that motivates the corrected estimator $\hat{U}_E$, and (ii) the almost-sure consistency of $\hat{U}_A, \hat{U}_E^{\mathrm{raw}}, \hat{U}_E$ as $N \to \infty$ even with M fixed (Thm. A.2). Finite-sample rates are deferred to App. A.2.

Sampling assumptions. $\Delta_1, \ldots, \Delta_N \overset{iid}{\sim} p_\Delta$; conditional on Δ_i, $Z_{i,1}, \ldots, Z_{i,M}$ are i.i.d. with the same distribution as $Z \mid \Delta = \Delta_i$; the inner draws are mutually independent across i; and $\mathbb{E}\|Z\|^4 < \infty$. We label these (A1)–(A4) below and use them directly in proofs.

Estimators. The two-level estimators $\hat{\mu}_i, \hat{\Sigma}_i, \hat{U}_A, \hat{U}_E^{\mathrm{raw}}$ are defined as in Alg. 1, and the bias-corrected epistemic estimator is

$$\hat{U}_E \;:=\; \hat{U}_E^{\mathrm{raw}} - \tfrac{1}{M}\hat{U}_A. \tag{6}$$

The next subsection establishes the finite-M identity that motivates this correction.

A.1.1 Unbiasedness and finite-M bias identity

What we will show. Each inner-sample mean $\hat{\mu}_i$ is unbiased for the conditional mean $\mu(\Delta_i)$, and each inner-sample covariance $\hat{\Sigma}_i$ is unbiased for the conditional covariance $\Sigma(\Delta_i)$. However, the *unconditional* covariance of $\hat{\mu}_i$ inherits a $\frac{1}{M}\Sigma(\Delta_i)$ inflation from inner sampling, which propagates through the law of total covariance into a $\frac{1}{M}U_A$ inflation of the population covariance of $\hat{\mu}_i$. This is exactly the leakage the main-paper Pillar 1 references; the corrected estimator (6) subtracts it.

Proposition A.1 (Unbiasedness of $\hat{\Sigma}_i$ and finite-M covariance inflation)**.** *Under (A2) and* $\mathbb{E}\|Z\|^2 < \infty$,

$$\mathbb{E}[\hat{\Sigma}_i \mid \Delta_i] = \Sigma(\Delta_i), \qquad \mathrm{Cov}(\hat{\mu}_i \mid \Delta_i) = \frac{1}{M}\Sigma(\Delta_i), \qquad \mathbb{E}[\hat{\mu}_i \mid \Delta_i] = \mu(\Delta_i). \tag{7}$$

Consequently, under (A1)–(A3),

$$\mathrm{Cov}(\hat{\mu}_i) = U_E + \frac{1}{M} U_A. \tag{8}$$

Proof. Fix i and condition on Δ_i.

Step 1: $\mathbb{E}[\hat{\mu}_i \mid \Delta_i] = \mu(\Delta_i)$. This is immediate from linearity of conditional expectation:

$$\mathbb{E}[\hat{\mu}_i \mid \Delta_i] = \frac{1}{M} \sum_{j=1}^{M} \mathbb{E}[Z_{i,j} \mid \Delta_i] = \mu(\Delta_i).$$

Step 2: $\mathrm{Cov}(\hat{\mu}_i \mid \Delta_i) = \Sigma(\Delta_i)/M$. Using conditional i.i.d. and $\mathrm{Cov}(aX) = a^2\mathrm{Cov}(X)$,

$$\mathrm{Cov}(\hat{\mu}_i \mid \Delta_i) = \mathrm{Cov}\left(\frac{1}{M} \sum_{j=1}^{M} Z_{i,j} \,\middle|\, \Delta_i \right) = \frac{1}{M^2} \sum_{j=1}^{M} \mathrm{Cov}(Z_{i,j} \mid \Delta_i) = \frac{1}{M} \Sigma(\Delta_i).$$

Step 3: $\mathbb{E}[\hat{\Sigma}_i \mid \Delta_i] = \Sigma(\Delta_i)$. This is the standard unbiasedness of the sample covariance. A direct derivation: let $\tilde{Z}_{i,j} := Z_{i,j} - \mu(\Delta_i)$ and $\overline{\tilde{Z}}_i := \frac{1}{M} \sum_j \tilde{Z}_{i,j} = \hat{\mu}_i - \mu(\Delta_i)$. Then

$$\sum_{j=1}^{M} (Z_{i,j} - \hat{\mu}_i)(Z_{i,j} - \hat{\mu}_i)^\top = \sum_{j=1}^{M} (\tilde{Z}_{i,j} - \overline{\tilde{Z}}_i)(\tilde{Z}_{i,j} - \overline{\tilde{Z}}_i)^\top = \sum_{j=1}^{M} \tilde{Z}_{i,j} \tilde{Z}_{i,j}^\top - M\, \overline{\tilde{Z}}_i\, \overline{\tilde{Z}}_i^\top.$$

Taking conditional expectation and using conditional i.i.d. with $\mathbb{E}[\tilde{Z}_{i,j} \mid \Delta_i] = 0$,

$$\mathbb{E}\left[\sum_{j=1}^{M} \tilde{Z}_{i,j} \tilde{Z}_{i,j}^\top \,\middle|\, \Delta_i \right] = M\, \mathbb{E}[\tilde{Z}_{i,1} \tilde{Z}_{i,1}^\top \mid \Delta_i] = M\, \Sigma(\Delta_i),$$

and

$$\mathbb{E}\left[M\, \overline{\tilde{Z}}_i\, \overline{\tilde{Z}}_i^\top \,\middle|\, \Delta_i \right] = M\, \mathrm{Cov}(\overline{\tilde{Z}}_i \mid \Delta_i) = M \cdot \frac{1}{M} \Sigma(\Delta_i) = \Sigma(\Delta_i).$$

Therefore,

$$\mathbb{E}\left[\sum_{j=1}^{M} (Z_{i,j} - \hat{\mu}_i)(Z_{i,j} - \hat{\mu}_i)^\top \,\middle|\, \Delta_i \right] = (M-1)\Sigma(\Delta_i),$$

and dividing by $(M-1)$ gives $\mathbb{E}[\hat{\Sigma}_i \mid \Delta_i] = \Sigma(\Delta_i)$.

Step 4: unconditional covariance inflation. Apply the law of total covariance to $\hat{\mu}_i$ with conditioning variable Δ_i:

$$\mathrm{Cov}(\hat{\mu}_i) = \mathbb{E}\big[\mathrm{Cov}(\hat{\mu}_i \mid \Delta_i)\big] + \mathrm{Cov}\big(\mathbb{E}[\hat{\mu}_i \mid \Delta_i]\big) = \mathbb{E}\Big[\frac{1}{M}\Sigma(\Delta_i)\Big] + \mathrm{Cov}(\mu(\Delta_i)) = \frac{1}{M} U_A + U_E.$$

This yields (8). □

What this gives the main paper. The identity $\mathrm{Cov}(\hat{\mu}_i) = U_E + \frac{1}{M} U_A$ *is* the inflation we corrected for in Eq. (6) and that the main-paper Prop. 4.1 reports informally. Because the inflation is exact (not asymptotic), the corrected estimator $\hat{U}_E$ has the right target at every $M \geq 2$, which is what enables AU/EU separation at the small inner budgets a black-box API forces.

A.1.2 Almost sure consistency (raw and corrected forms)

Theorem A.2 (Almost sure consistency of two-level estimators)**.** *Under (A1)–(A4): (i) if $M \to \infty$ and $N \to \infty$, then $\hat{U}_A \xrightarrow{a.s.} U_A$ and $\hat{U}_E^{\mathrm{raw}} \xrightarrow{a.s.} U_E$; (ii) if $N \to \infty$ while $M \geq 2$ is fixed, then $\hat{U}_A \xrightarrow{a.s.} U_A$, $\hat{U}_E^{\mathrm{raw}} \xrightarrow{a.s.} U_E + \frac{1}{M} U_A$, and $\hat{U}_E \xrightarrow{a.s.} U_E$.*

Proof sketch. $\hat{U}_A \xrightarrow{a.s.} U_A$ is the matrix-valued SLLN applied to i.i.d. $\{\hat{\Sigma}_i\}$ with $\mathbb{E}[\hat{\Sigma}_1] = U_A$ by Prop. A.1; $\hat{U}_E^{\text{raw}} \xrightarrow{a.s.} U_E + \frac{1}{M}U_A$ is the sample-covariance SLLN on i.i.d. $\{\hat{\mu}_i\}$, with population value identified by Prop. A.1. Combining gives $\hat{U}_E = \hat{U}_E^{\text{raw}} - \frac{1}{M}\hat{U}_A \to U_E$, i.e. (ii); (i) follows by additionally letting $M \to \infty$. □

The corrected estimator $\hat{U}_E$ defined in Alg. 1 is therefore unbiased for U_E at any $M \geq 2$ and converges to it as $N \to \infty$. The rate is in App. A.2.

A.2 Finite-Sample Concentration and Two-Level Monte Carlo Bounds

Purpose of this subsection. App. A.1 established *whether* $\hat{U}_A, \hat{U}_E$ converge to their population targets; this subsection establishes *how fast*. The result the main-paper Pillar 1 quotes informally ($O(\sqrt{\log(1/\delta)/N})$ plus lower-order $1/M$ terms) is Theorem A.5. We get there in two steps: first an *outer-only* bound that pretends conditional moments $\mu(\Delta), \Sigma(\Delta)$ are known exactly (Thm. A.3), then the full two-level bound that adds the inner-sampling fluctuation back in.

The proofs combine standard tools (Hoeffding's inequality for bounded centered i.i.d. scalars and the entrywise-to-Frobenius conversion $\|A\|_F \leq d\|A\|_{\max}$ for $A \in \mathbb{R}^{d\times d}$) with the two-level sampling decomposition specific to our estimator. We refer to either as "standard" below.

A.2.1 Outer concentration with population conditional moments

As a stepping stone to the full two-level bound, we first treat $\mu(\Delta), \Sigma(\Delta)$ as known (infinite inner samples). Let $\mu_i := \mu(\Delta_i)$, $\Sigma_i := \Sigma(\Delta_i)$, $\widehat{U}_A := \frac{1}{N}\sum_i \Sigma_i$, $\widehat{U}_E := \frac{1}{N-1}\sum_i (\mu_i - \bar{\mu}^\star)(\mu_i - \bar{\mu}^\star)^\top$ with $\bar{\mu}^\star := \frac{1}{N}\sum_i \mu_i$.

Theorem A.3 (High-probability outer concentration)**.** *Assume $\Delta_1, \ldots, \Delta_N$ i.i.d. with $\|\Sigma(\Delta)\|_F \leq B_\Sigma$ and $\|\mu(\Delta)\|_2 \leq B_\mu$ a.s. For any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\|\widehat{U}_A - U_A\|_F \leq d\, B_\Sigma \sqrt{2\log(2d^2/\delta)/N}, \tag{9}$$

$$\|\widehat{U}_E - U_E\|_F \leq \tfrac{N}{N-1}\left[d\, B_\mu^2 \sqrt{2\log(2d^2/\delta)/N} + 2d\, B_\mu^2 \sqrt{2\log(2d/\delta)/N}\right]. \tag{10}$$

Proof sketch. For (9): entrywise Hoeffding on the centered, bounded scalars $(\Sigma_i)_{k\ell} - (U_A)_{k\ell}$ ($b = 2B_\Sigma$), union bound over d^2 entries, and the entrywise-to-Frobenius conversion $\|A\|_F \leq d\|A\|_{\max}$. For (10): write $\widehat{U}_E - U_E = \frac{N}{N-1}[(\bar{Q} - Q) - (\bar{M} - mm^\top)]$ with $Q = \mathbb{E}[\mu\mu^\top]$, $m = \mathbb{E}[\mu]$, and the sample analogues. Apply (i) entrywise Hoeffding+union to $\bar{Q} - Q$, (ii) coordinatewise Hoeffding+union to $\bar{\mu}^\star - m$, and (iii) the identity $\bar{M} - mm^\top = \bar{\mu}^\star(\bar{\mu}^\star - m)^\top + (\bar{\mu}^\star - m)m^\top$ with $\|\bar{\mu}^\star\|_2, \|m\|_2 \leq B_\mu$ to get $\|\bar{M} - mm^\top\|_F \leq 2B_\mu\|\bar{\mu}^\star - m\|_2$. Triangle inequality and constant-simplification on the intersection event give (10). □

The leading $\sqrt{\log(1/\delta)/N}$ rate matches what main-paper Pillar 1 quotes; d is the embedding dimension, entering both as a linear prefactor d and logarithmically inside $\sqrt{\log(2d^2/\delta)/N}$.

A.2.2 Two-level Monte Carlo concentration (outer N, inner M)

Why this step. The previous subsubsection assumed exact conditional moments, which we never have in practice; we estimate them from M inner draws. This subsection adds the inner-sampling slack and gives the rate the main paper actually relies on, with both N and M explicit.

To obtain explicit finite-sample bounds, we impose an a.s. boundedness assumption on embedded outputs.

Assumption A.4 (Bounded embedded outputs)**.** Almost surely, $\|Z\|_2 \leq B_Z$. Equivalently, for all (i,j), $\|Z_{i,j}\|_2 \leq B_Z$ a.s.

Under Assumption A.4, we have $\|\mu(\Delta)\|_2 \leq B_Z$ and $\|\Sigma(\Delta)\|_F \leq 2B_Z^2$ a.s. (the latter follows from $\|Z - \mu\|_2 \leq 2B_Z$ and $\|vv^\top\|_F = \|v\|_2^2$).

Plan. We bound (i) $\|\hat{U}_A - U_A\|_F$ and (ii) $\|\hat{U}_E - U_E\|_F$ with explicit dependence on N and M. The key is to decompose each estimator into an *outer* fluctuation term (from sampling Δ_i) plus an *inner* fluctuation term (from estimating $\mu(\Delta_i), \Sigma(\Delta_i)$ with M samples).

Inner-sample concentration (used below). For any fixed i and $\delta \in (0,1)$, conditional on Δ_i, standard Hoeffding gives $\|\hat{\mu}_i - \mu(\Delta_i)\|_2 \le 2B_Z\sqrt{2d\log(2d/\delta)/M}$ w.p. $\ge 1-\delta$ (coordinate-wise + $\|v\|_2 \le \sqrt{d}\|v\|_\infty$), and $\|\hat{\Sigma}_i - \Sigma(\Delta_i)\|_F \le C\, B_Z^2\, d\,\sqrt{\log(2d^2/\delta)/M}$ w.p. $\ge 1-\delta$ for a universal $C > 0$ (entrywise Hoeffding on $\frac{1}{M}\sum_j Z_{i,j}Z_{i,j}^\top$ plus the rank-1 perturbation $\hat{\mu}_i\hat{\mu}_i^\top - \mu_i\mu_i^\top$ controlled by the mean bound).

Theorem A.5 (Two-level Monte Carlo concentration (explicit N, M rates))**.** *Assume (A1)–(A3) and Assumption A.4. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$:*

$$\|\hat{U}_A - U_A\|_F \le d\,(2B_Z^2)\sqrt{\frac{2\log(4d^2/\delta)}{N}} \;+\; C_1\, B_Z^2\, d\,\sqrt{\frac{\log(4d^2/\delta)}{M}} \;+\; C_1\, B_Z^2\, d\,\sqrt{\frac{\log(4d^2/\delta)}{NM}}, \tag{11}$$

$$\|\hat{U}_E - U_E\|_F \le C_2\, B_Z^2\, d\,\sqrt{\frac{\log(4d^2/\delta)}{N}} \;+\; C_3\, B_Z^2\, d\,\sqrt{\frac{\log(4d^2/\delta)}{NM}} \;+\; \frac{1}{M}\,\|\hat{U}_A - U_A\|_F, \tag{12}$$

for universal constants $C_1, C_2, C_3 > 0$.

Proof sketch. Decompose $\hat{U}_A - U_A = (\frac{1}{N}\sum_i \Sigma(\Delta_i) - U_A) + \frac{1}{N}\sum_i(\hat{\Sigma}_i - \Sigma(\Delta_i))$. The first term is bounded by Thm. A.3 with $B_\Sigma \le 2B_Z^2$, giving the $\sqrt{\log/N}$ term; the per-i inner Hoeffding above and a union bound over i give the $\sqrt{\log/M}$ term; exploiting cross-outer independence (A3) when union-bounding entrywise yields the sharper $\sqrt{\log/(NM)}$ refinement. For $\hat{U}_E$, write $\hat{U}_E - U_E = (\hat{U}_E^{\mathrm{raw}} - (U_E + \frac{1}{M}U_A)) + \frac{1}{M}(U_A - \hat{U}_A)$; the first piece is the sample-covariance error of $\{\hat{\mu}_i\}$ around their population covariance $U_E + \frac{1}{M}U_A$ (Prop. A.1) and is bounded by Thm. A.3 with $\|\hat{\mu}\|_2 \le B_Z$, while the second is $1/M$ times the bound from $\hat{U}_A$. □

Remark. Assumption A.4 can be relaxed to sub-Gaussian or bounded-moment conditions by replacing Hoeffding by Bernstein; the (N, M) rates are unchanged.

What this gives the main paper. Pillar 1 in Sec. 4.1 states informally that the bias-corrected estimator concentrates at $O(\sqrt{\log(1/\delta)/N})$ plus lower-order $1/M$ terms. (11)–(12) are the precise version. The $1/(NM)$ refinement matters in practice: it tells us that doubling the inner budget M buys us a comparable rate improvement to doubling the outer budget N, which is what justifies the equal-budget default $N{=}M{\approx}10$ in the experiments.

A.3 Explicit Sufficient Conditions for the Bayesian Bridge

Purpose of this subsection. The main paper's Pillar 2 (Thm. 4.2) states that the perturbation-population uncertainties (U_A^Δ, U_E^Δ) approximate the Bayesian targets $(U_A^{\mathrm{Bayes}}, U_E^{\mathrm{Bayes}})$ up to two defect terms $\varepsilon_A, \varepsilon_E$. Here we make these defects explicit by introducing local conditions on the model and bounding $\varepsilon_A, \varepsilon_E$ in those terms.

Structure of this subsection. We treat the EU- and AU-side bridges separately. The EU bridge (§A.3.2) needs (i) local linearity of the conditional mean in both parameters and inputs (Ass. A.6), (ii) a local Gaussian posterior (Ass. A.7), and (iii) a semantic coverage condition saying the perturbation distribution spans the directions of Bayesian mean variation (Ass. A.8). The AU bridge (§A.3.3) needs only one extra piece (a Lipschitz aleatoric map; Ass. A.11) saying decoding stochasticity varies smoothly with the predictive state. Each assumption gets a *Discussion* block addressing its plausibility and how to measure its tightness from data (see also App. C for the diagnostics).

Relation to Wang and Ji [53]. Our outer ensemble builds on the local input–parameter correspondence of Wang and Ji [53], while our bridge theorems below quantify its approximation error at the distributional level. Specifically, Prop. 3.3 of Wang and Ji [53] is a pointwise, first-order existence result in a continuous input space: for a small parameter shift $\delta\theta$, there exists a (generally non-unique) continuous input perturbation δx inducing the same first-order output change, and vice versa via a first-layer construction. However, pointwise existence neither provides a mechanism for generating

meaning-preserving perturbations of discrete text nor guarantees that the chosen input-perturbation distribution matches the pushforward of the parameter posterior—a prerequisite for covariance-level AU/EU approximation. Our two-stage procedure (Sec. 3) instantiates a localized perturbation distribution over a finite pool of fluent paraphrases: isotropic Gaussian noise in embedding space serves purely as a stochastic selector over cached candidates, ensuring the LLM only processes natural text. The remaining approximation gaps are explicitly tracked by our diagnostics: D1 for local linearization, D3 for distributional coverage, and our validity audit (App. F.2.3) for semantic preservation. Beyond Wang and Ji [53], our technical contributions encompass the nested two-level estimator, the exact finite-M bias identity and correction, and the conditional AU/EU bridge bounds with data-estimable defect diagnostics.

A.3.1 Notation recap and three formally distinct flavors of (U_A, U_E, U_T)

Fix an input x and trained parameters $\theta^\star$. Let $\Delta \sim p_\Delta$ be a semantic perturbation and $x_\Delta := T(x, \Delta)$. For the stochastic model output representation $Z_\theta(x, \xi)$ (decoding noise ξ; embedding $Z = e(Y)$, main paper), define the conditional mean and covariance

$$m_\theta(x) := \mathbb{E}_\xi[Z_\theta(x, \xi) \mid x, \theta], \qquad \Sigma_\theta(x) := \mathrm{Cov}_\xi(Z_\theta(x, \xi) \mid x, \theta).$$

The paper uses three formally distinct flavors of (U_A, U_E, U_T) that are easy to conflate; we lay them out once here for reference, and the two pillars (Sec. 4.1 and 4.2) bound the gaps between them.

(1) Bayesian (ground-truth) AU/EU/TU. Defined over the parameter posterior $p(\theta \mid \mathcal{D})$ as in the canonical Bayesian decomposition [23, 8]:

$$\mathcal{U}_A^{\mathrm{Bayes}}(x) := \mathbb{E}_{\theta\sim p(\theta|\mathcal{D})}\big[\Sigma_\theta(x)\big], \quad \mathcal{U}_E^{\mathrm{Bayes}}(x) := \mathrm{Cov}_{\theta\sim p(\theta|\mathcal{D})}\big(m_\theta(x)\big), \quad \mathcal{U}_T^{\mathrm{Bayes}}(x) := \mathcal{U}_A^{\mathrm{Bayes}}(x) + \mathcal{U}_E^{\mathrm{Bayes}}(x). \tag{13}$$

This is the object Bayesian UQ ultimately seeks. We never sample from $p(\theta \mid \mathcal{D})$ directly: the posterior over an LLM's parameters is intractable.

(2) Perturbation-population AU/EU/TU (what our estimator targets). At the *fixed deployed model* $\theta^\star$, replace the parameter average by an outer-perturbation average over p_Δ:

$$\mathcal{U}_A^{\Delta}(x) := \mathbb{E}_{\Delta\sim p_\Delta}\big[\Sigma_{\theta^\star}(x_\Delta)\big], \quad \mathcal{U}_E^{\Delta}(x) := \mathrm{Cov}_{\Delta\sim p_\Delta}\big(m_{\theta^\star}(x_\Delta)\big), \quad \mathcal{U}_T^{\Delta}(x) := \mathcal{U}_A^{\Delta}(x) + \mathcal{U}_E^{\Delta}(x). \tag{14}$$

By the law of total covariance applied to the joint distribution of (Δ, ξ) at $\theta = \theta^\star$, the $\mathcal{U}_T^{\Delta}(x)$ defined above coincides with the embedding-space total covariance $\mathrm{Cov}(Z) = U_T(x)$ written in Eq. (1) of the main paper, so $(U_A, U_E, U_T) \equiv (\mathcal{U}_A^{\Delta}, \mathcal{U}_E^{\Delta}, \mathcal{U}_T^{\Delta})$ in that equation. These quantities are well-defined for any LLM with black-box sampling access; sampling from p_Δ replaces sampling from the parameter posterior.

(3) Empirical estimators (what we actually compute at run time). The two-level Monte Carlo estimators built from N outer perturbations $\{x_i\}_{i=1}^N$ and M inner decodes $\{Z_{i,j}\}_{j=1}^M$ per perturbation:

$$\widehat{U}_A^{(M)} := \tfrac{1}{N}\sum_{i=1}^{N} \widehat{\Sigma}_i, \quad \widehat{U}_E^{\mathrm{raw}} := \widehat{\mathrm{Cov}}\big(\{\hat{\mu}_i\}_{i=1}^N\big), \quad \widehat{U}_E^{(M)} := \widehat{U}_E^{\mathrm{raw}} - \tfrac{1}{M}\widehat{U}_A^{(M)}, \tag{15}$$

where $\hat{\mu}_i$ and $\widehat{\Sigma}_i$ are the sample mean and sample covariance of $\{Z_{i,j}\}_{j=1}^M$. The bias-corrected $\widehat{U}_E^{(M)}$ is the estimator we use throughout the paper; we drop the (M) superscript when there is no ambiguity.

How the three flavors relate. Pillar 1 (App. A.1 and A.2) bounds the gap $(3) \to (2)$: $\widehat{U}_A^{(M)} \to \mathcal{U}_A^{\Delta}$ and $\widehat{U}_E^{(M)} \to \mathcal{U}_E^{\Delta}$ almost surely as $N \to \infty$ at any fixed $M \geq 2$, with the $\frac{1}{M}$ leakage from inner-sampling fluctuation removed by the bias correction. Pillar 2 (this subsection) bounds the gap $(2) \to (1)$: under three local conditions on the model around x, $\|\mathcal{U}_\bullet^{\Delta}(x) - \mathcal{U}_\bullet^{\mathrm{Bayes}}(x)\|_F$ is upper-bounded by defect terms $\varepsilon_A(x), \varepsilon_E(x)$ that we estimate empirically as diagnostics from the outputs the estimator already collects (App. C). The composition of the two pillars yields the end-to-end bound $\|\widehat{U}_\bullet^{(M)} - \mathcal{U}_\bullet^{\mathrm{Bayes}}\|_F \leq$ MC error + bridge defect stated informally in Theorem 4.2 of the main paper.

Table 8: **Summary of the three formal uncertainty objects** across our theoretical hierarchy (Sec. 3–4).

	Object	Symbols	Requirements & Properties
(1)	Bayesian target	$\mathcal{U}_A^{\text{Bayes}}$, $\mathcal{U}_E^{\text{Bayes}}$	Parameter posterior $p(\theta \mid \mathcal{D})$ (intractable at LLM scale)
(2)	Perturbation population	$\mathcal{U}_A^{\Delta} \equiv U_A$, $\mathcal{U}_E^{\Delta} \equiv U_E$	Black-box sampling access; exact decomposition in Eq. (1)
(3)	Empirical estimator	$\widehat{U}_A$, $\widehat{U}_E$ (or $\widehat{U}_\bullet^{(M)}$)	$N \times M$ generations; unbiased for (2) at any $M \geq 2$

A.3.2 Epistemic bridge via linearization and semantic coverage

Assumption A.6 (Local bi-linearization of the conditional mean)**.** There exist Jacobians $J_\theta(x) \in \mathbb{R}^{d \times P}$ and $J_x(x) \in \mathbb{R}^{d \times q}$ and constants $c_\theta, c_x > 0$ such that for small $\delta\theta$ and δx,

$$m_{\theta^\star + \delta\theta}(x) = m_{\theta^\star}(x) + J_\theta(x)\,\delta\theta + r_\theta, \qquad \|r_\theta\| \leq c_\theta \|\delta\theta\|^2, \tag{16}$$

$$m_{\theta^\star}(x + \delta x) = m_{\theta^\star}(x) + J_x(x)\,\delta x + r_x, \qquad \|r_x\| \leq c_x \|\delta x\|^2. \tag{17}$$

Assumption A.7 (Local Gaussian posterior approximation)**.** In a neighborhood of $\theta^\star$, the Bayesian posterior admits a local approximation $\delta\theta := \theta - \theta^\star \sim \mathcal{N}(0, \Sigma_\theta)$.

Discussion of Assumptions A.6–A.7. Two readings of these assumptions deserve direct engagement: that NTK-style linearization fits modern LLMs poorly at the global level, and that RLHF/DPO breaks classical Bayesian-posterior assumptions. Both turn on the distinction between global and local regimes, which we now make explicit.

(i) Linearization is local, not global. Assumption A.6 only requires a local first-order Taylor expansion of $m_\theta(x)$ around the fixed reference $(\theta^\star, x)$, with second-order remainder $\|r\| \leq c\|\delta\|^2$. This is strictly weaker than the global NTK regime that underlies infinite-width arguments for full networks [55]. Local first-order linearization for language models has been studied directly [51], where it provides a useful regime for analyzing fine-tuning trajectories. In our usage we need linearity only inside the moderate-ε semantic neighborhood the perturbation distribution actually populates, and the residual is empirically measurable via the OLS diagnostic D1 in App. C.

(ii) Locally Gaussian posterior is the Laplace assumption. Assumption A.7 is the standard Laplace approximation to the posterior in a neighborhood of a twice-differentiable local minimum, applied around $\theta^\star$. Recent work has applied this exact approximation to large language models in the Bayesian-LoRA line [60], where it gives competitive posterior surrogates for instruction-tuned LLMs. The training procedure (RLHF, DPO, or standard MLE) only determines *which* local minimum $\theta^\star$ sits at, not whether the posterior shape near $\theta^\star$ is locally Gaussian; any twice-differentiable loss admits a local quadratic and hence a local Gaussian posterior. Our diagnostic D2 (Henze–Zirkler test on *embedding* samples; specifically the inner-decode and outer-mean clouds in $\mathbb{R}^d$, not on uncertainty scores; App. C) flags inputs where the embedding distribution induced by Assumption A.7 (under the linearity in Assumption A.6) visibly fails to be Gaussian.

(iii) Coverage and stability are measurable. Assumptions A.8 and A.11 are not assertions of fact but *defect terms* that appear in our bounds: the bridge gets tighter as $\varepsilon_{\text{cov}} \to 0$ and as the aleatoric Lipschitz constant decreases. Both are upper-bounded by computable quantities (D3 MMD upper bound; spread of $\hat{\Sigma}_i$ across i). The bounds remain meaningful with non-zero defects; they quantify, rather than presuppose, how well the perturbation mechanism approximates Bayesian posterior sampling.

(iv) Existing UQ methods make analogous local assumptions. MC-Dropout assumes the dropout-induced distribution approximates a Bayesian posterior locally; Deep Ensembles assume independently trained models sample meaningfully from the posterior; Laplace approximations assume local Gaussianity. Our assumptions are at the same level of mathematical commitment, with the additional property that we *measure* their tightness on each dataset rather than asserting them globally.

Define the Bayesian mean-variation operator

$$A(x) := J_\theta(x)\,\Sigma_\theta^{1/2} \in \mathbb{R}^{d \times P}. \tag{18}$$

Under Assumptions A.6–A.7, the linearized Bayesian epistemic covariance of the mean is

$$\mathcal{U}_E^{\mathrm{Bayes}}(x) = \mathrm{Cov}_\theta(m_\theta(x)) = A(x)A(x)^\top + \mathrm{Rem}_\theta, \tag{19}$$

where Rem_θ aggregates second-order terms.

Assumption A.8 (Semantic coverage defect)**.** Let P_x be the orthogonal projector onto $\mathrm{Im}(J_x(x))$. Define the semantic coverage defect

$$\varepsilon_{\mathrm{cov}}(x) := \|(I - P_x)A(x)\|_F. \tag{20}$$

Lemma A.9 (Best-match perturbation covariance for epistemic variation)**.** *There exists $\Sigma_\Delta^\star \succeq 0$ such that*

$$\left\|J_x(x)\Sigma_\Delta^\star J_x(x)^\top - A(x)A(x)^\top\right\|_F \;\le\; 2\|A(x)\|_F\,\varepsilon_{\mathrm{cov}}(x) + \varepsilon_{\mathrm{cov}}(x)^2. \tag{21}$$

If $\varepsilon_{\mathrm{cov}}(x) = 0$, equality holds.

Proof. Let P_x be the projector onto $\mathrm{Im}(J_x)$ and decompose $A = A_\parallel + A_\perp$ with $A_\parallel := P_x A$ and $A_\perp := (I - P_x)A$. Since $A_\parallel \in \mathrm{Im}(J_x)$, there exists B such that $A_\parallel = J_x B$. Set $\Sigma_\Delta^\star := BB^\top \succeq 0$, giving $J_x\Sigma_\Delta^\star J_x^\top = A_\parallel A_\parallel^\top$. Now expand

$$AA^\top - A_\parallel A_\parallel^\top = A_\perp A_\perp^\top + A_\parallel A_\perp^\top + A_\perp A_\parallel^\top,$$

and bound Frobenius norms using $\|A_\perp\|_F = \varepsilon_{\mathrm{cov}}(x)$:

$$\|AA^\top - A_\parallel A_\parallel^\top\|_F \le \|A_\perp A_\perp^\top\|_F + \|A_\parallel A_\perp^\top\|_F + \|A_\perp A_\parallel^\top\|_F \le \|A_\perp\|_F^2 + 2\|A_\parallel\|_F\|A_\perp\|_F \le \varepsilon_{\mathrm{cov}}(x)^2 + 2\|A\|_F\,\varepsilon_{\mathrm{cov}}(x).$$

This proves (21). □

Theorem A.10 (Epistemic bridge bound)**.** *Under Assumptions A.6–A.8, the perturbation-population epistemic covariance satisfies*

$$\|\mathcal{U}_E^\Delta(x) - \mathcal{U}_E^{\mathrm{Bayes}}(x)\|_F \;\le\; \underbrace{2\|A(x)\|_F\,\varepsilon_{\mathrm{cov}}(x) + \varepsilon_{\mathrm{cov}}(x)^2}_{\text{coverage}} + \underbrace{C_\theta\,\mathbb{E}\|r_\theta\| + C_x\,\mathbb{E}\|r_x\| + \|\mathrm{Rem}_\theta\|_F}_{\text{linearization remainders}}, \tag{22}$$

where expectations are taken over $\delta\theta \sim \mathcal{N}(0, \Sigma_\theta)$ and $\delta x := x_\Delta - x$, and C_θ, C_x are absolute constants (depending only on the chosen norms).

Proof. We provide a detailed proof with explicit decompositions.

Step 1: Bayesian epistemic covariance under local linearization. Under (16),

$$m_{\theta^\star+\delta\theta}(x) = m_{\theta^\star}(x) + J_\theta(x)\delta\theta + r_\theta.$$

Taking covariance over $\delta\theta$,

$$\mathrm{Cov}_\theta(m_\theta(x)) = \mathrm{Cov}\big(J_\theta\delta\theta + r_\theta\big) = J_\theta\,\mathrm{Cov}(\delta\theta)\,J_\theta^\top + \mathrm{Cov}(r_\theta) + \mathrm{Cov}(J_\theta\delta\theta, r_\theta) + \mathrm{Cov}(r_\theta, J_\theta\delta\theta).$$

Under Assumption A.7, $\mathrm{Cov}(\delta\theta) = \Sigma_\theta$ and thus the leading term is $J_\theta\Sigma_\theta J_\theta^\top = A(x)A(x)^\top$ by definition of $A(x)$. Collect the remaining terms into Rem_θ:

$$\mathcal{U}_E^{\mathrm{Bayes}}(x) = A(x)A(x)^\top + \mathrm{Rem}_\theta.$$

A standard norm bound yields $\|\mathrm{Rem}_\theta\|_F \le C_\theta\,\mathbb{E}\|r_\theta\| +$ (higher-order terms), which is the term displayed in (22).

Step 2: Perturbation-population epistemic covariance under input linearization. Under (17),

$$m_{\theta^\star}(x_\Delta) = m_{\theta^\star}(x + \delta x) = m_{\theta^\star}(x) + J_x(x)\delta x + r_x.$$

Taking covariance over Δ (equivalently over δx),

$$\mathcal{U}_E^\Delta(x) = \mathrm{Cov}_\Delta\big(m_{\theta^\star}(x_\Delta)\big) = \mathrm{Cov}_\Delta\big(J_x\delta x + r_x\big) = J_x\,\mathrm{Cov}(\delta x)\,J_x^\top + \mathrm{Rem}_x,$$

where Rem_x collects $\mathrm{Cov}(r_x)$ and cross-covariance terms between $J_x\delta x$ and r_x. Using triangle inequality and standard covariance bounds, one obtains $\|\mathrm{Rem}_x\|_F \le C_x\,\mathbb{E}\|r_x\|$ for an absolute constant C_x.

Step 3: Coverage approximation. Lemma A.9 provides $\Sigma_\Delta^\star \succeq 0$ such that

$$\|J_x \Sigma_\Delta^\star J_x^\top - AA^\top\|_F \le 2\|A\|_F \varepsilon_{\mathrm{cov}} + \varepsilon_{\mathrm{cov}}^2.$$

Interpret $\Sigma_\Delta^\star$ as the "best-match" covariance in input-perturbation space that aligns perturbation-induced mean variation with Bayesian mean-variation directions (up to the defect).

Step 4: Combine all terms. Combine the decompositions:

$$\mathcal{U}_E^\Delta = J_x \mathrm{Cov}(\delta x) J_x^\top + \mathrm{Rem}_x, \qquad \mathcal{U}_E^{\mathrm{Bayes}} = AA^\top + \mathrm{Rem}_\theta.$$

Add and subtract $J_x \Sigma_\Delta^\star J_x^\top$:

$$\|\mathcal{U}_E^\Delta - \mathcal{U}_E^{\mathrm{Bayes}}\|_F \le \underbrace{\|J_x \mathrm{Cov}(\delta x) J_x^\top - J_x \Sigma_\Delta^\star J_x^\top\|_F}_{(\star)} + \underbrace{\|J_x \Sigma_\Delta^\star J_x^\top - AA^\top\|_F}_{\text{Lemma A.9}} + \|\mathrm{Rem}_x\|_F + \|\mathrm{Rem}_\theta\|_F.$$

The second term is bounded by Lemma A.9. The remainder terms are bounded as in Steps 1–2. The term $(\star)$ vanishes if $\mathrm{Cov}(\delta x) = \Sigma_\Delta^\star$; more generally it is absorbed into the linearization/remainder group (it quantifies mismatch between the actual perturbation covariance and the best-match covariance). Collecting yields (22). □

What this gives the main paper. (22) *is* the $\varepsilon_E(x)$ defect bound used in the informal Pillar 2 statement (Thm. 4.2). Reading the right-hand side: the first two terms shrink with the semantic-coverage diagnostic D3 (App. C), while the last remainder shrinks with the linearity diagnostic D1; the bridge is tight precisely when both diagnostics report small values.

A.3.3 Aleatoric bridge via stability of conditional covariance

Assumption A.11 (Lipschitz aleatoric map)**.** There exist a mapping $G : \mathbb{R}^d \to \mathbb{S}_+^d$ and constants $L_G, \rho_G \ge 0$ such that

$$\|\Sigma_\theta(x) - G(m_\theta(x))\|_F \le \rho_G \quad \text{for all } (x, \theta) \text{ in the local neighborhood,} \tag{23}$$

and G is Lipschitz:

$$\|G(u) - G(v)\|_F \le L_G \|u - v\| \quad \forall u, v \in \mathbb{R}^d. \tag{24}$$

Discussion. Assumption A.11 formalizes the empirical observation that the *within*-perturbation dispersion of LLM outputs (decoding stochasticity at a fixed prompt) varies smoothly with the predictive state $m_\theta(x)$ rather than discontinuously. This is consistent with the observation that decoding noise is primarily a function of local entropy of the next-token distribution, which is itself a smooth function of the hidden representation. Both L_G and ρ_G are defect terms: the bridge is tightest when decoding stochasticity is well-approximated by a Lipschitz function of the predictive mean, and the bound (25) degrades gracefully with ρ_G rather than breaking. We do not need $\rho_G = 0$.

Theorem A.12 (Aleatoric bridge bound)**.** *Under Assumption A.11 and the setup of Section 4.2,*

$$\|\mathcal{U}_A^\Delta(x) - \mathcal{U}_A^{\mathrm{Bayes}}(x)\|_F \;\le\; L_G \, \mathbb{E}\Big[\|m_\theta(x) - m_{\theta^\star}(x_\Delta)\|\Big] \;+\; 2\rho_G, \tag{25}$$

where the expectation is under any coupling of (θ, Δ). Moreover, under Assumptions A.6–A.8, one can choose a coupling such that

$$\mathbb{E}\Big[\|m_\theta(x) - m_{\theta^\star}(x_\Delta)\|\Big] \;\lesssim\; \|A(x)\|_F \, \varepsilon_{\mathrm{cov}}(x) + \mathbb{E}\|r_\theta\| + \mathbb{E}\|r_x\|. \tag{26}$$

Proof. We provide a detailed proof of (25); the refinement (26) follows by the same best-match argument used in the epistemic bridge.

Step 1: pointwise bound on conditional covariances. Fix (θ, Δ) and write $x_\Delta = T(x, \Delta)$. By triangle inequality,

$$\|\Sigma_\theta(x) - \Sigma_{\theta^\star}(x_\Delta)\|_F \le \|\Sigma_\theta(x) - G(m_\theta(x))\|_F + \|G(m_\theta(x)) - G(m_{\theta^\star}(x_\Delta))\|_F + \|G(m_{\theta^\star}(x_\Delta)) - \Sigma_{\theta^\star}(x_\Delta)\|_F. \tag{27}$$

Apply Assumption A.11 to the first and third terms to get $\le \rho_G$ each. Apply Lipschitzness (24) to the middle term:

$$\|G(m_\theta(x)) - G(m_{\theta^\star}(x_\Delta))\|_F \le L_G \|m_\theta(x) - m_{\theta^\star}(x_\Delta)\|.$$

Thus, for all (θ, Δ) in the local neighborhood,

$$\|\Sigma_\theta(x) - \Sigma_{\theta^\star}(x_\Delta)\|_F \le L_G \|m_\theta(x) - m_{\theta^\star}(x_\Delta)\| + 2\rho_G.$$

Step 2: take expectations under a coupling. Let (θ, Δ) be jointly distributed under any coupling with marginals $\theta \sim p(\theta \mid \mathcal{D})$ and $\Delta \sim p_\Delta$. Take expectation of the pointwise inequality:

$$\mathbb{E}\|\Sigma_\theta(x) - \Sigma_{\theta^\star}(x_\Delta)\|_F \le L_G \, \mathbb{E}\|m_\theta(x) - m_{\theta^\star}(x_\Delta)\| + 2\rho_G.$$

Step 3: relate to $\mathcal{U}_A^{\mathrm{Bayes}}$ and $\mathcal{U}_A^{\Delta}$. By definition,

$$\mathcal{U}_A^{\mathrm{Bayes}}(x) = \mathbb{E}_\theta[\Sigma_\theta(x)], \qquad \mathcal{U}_A^{\Delta}(x) = \mathbb{E}_\Delta[\Sigma_{\theta^\star}(x_\Delta)].$$

By Jensen and triangle inequality,

$$\|\mathcal{U}_A^{\mathrm{Bayes}}(x) - \mathcal{U}_A^{\Delta}(x)\|_F = \|\mathbb{E}_\theta[\Sigma_\theta(x)] - \mathbb{E}_\Delta[\Sigma_{\theta^\star}(x_\Delta)]\|_F \le \mathbb{E}\|\Sigma_\theta(x) - \Sigma_{\theta^\star}(x_\Delta)\|_F.$$

Combine with Step 2 to obtain (25).

Step 4: refinement via semantic coverage (sketch). Under Assumption A.6, local linearization provides $m_\theta(x) \approx m_{\theta^\star}(x) + J_\theta \delta\theta$ and $m_{\theta^\star}(x_\Delta) \approx m_{\theta^\star}(x) + J_x \delta x$. Choosing a coupling that aligns $J_x \delta x$ with $J_\theta \delta\theta$ in the sense of Lemma A.9 yields a mean-gap controlled by $\|A\|_F \varepsilon_{\mathrm{cov}}$ plus remainder terms, giving (26). □

What this gives the main paper. (25) *is* the $\varepsilon_A(x)$ defect bound used in the informal Pillar 2 statement (Thm. 4.2). Together with (22), the two bounds furnish the right-hand sides of the Pillar 2 inequalities, completing the bridge defect side of the main-paper claim. The MC error side of the same claim was supplied by App. A.2.

A.4 Spurious Determinism: When Consistency-Only UQ Can Fail

Purpose of this subsection. The main paper's Pillar 3 (Thm. 4.3) claims a class of inputs on which any consistency-only uncertainty estimator $\mathcal{F}_{\mathrm{cons}}$ is exactly 0 while perturbation EU is bounded *away* from 0. Here we (i) name the class formally: the "spurious determinism" regime where decoding is deterministic given the prompt but the conditional mean varies across semantic-preserving perturbations; and (ii) construct a concrete instance and prove the separation. The argument is intentionally elementary: the value of the result is in isolating a regime, not in the algebra.

This section formalizes a simple failure mode of "consistency-only" uncertainty heuristics that probe only within-prompt variability under repeated decoding. The key phenomenon is *spurious determinism*: for a fixed prompt instance, the model may be nearly deterministic (hence self-consistent) yet wrong, while semantically equivalent perturbations expose strong sensitivity.

A.4.1 A minimal construction

Let Δ be a random perturbation and consider $Z = e(Y) \in \mathbb{R}^d$ under the conditional model $p(\cdot \mid x_\Delta)$. Assume that for each fixed Δ, decoding is deterministic:

$$Z = \mu(\Delta) \quad \text{a.s. given } \Delta, \tag{28}$$

so that $\Sigma(\Delta) = \mathrm{Cov}(Z \mid \Delta) = 0$.

Now assume $\mu(\Delta)$ varies across Δ due to a spurious prompt feature. A concrete example: let $u, v \in \mathbb{R}^d$ be orthonormal and define

$$\mu(\Delta) = a\,u + s(\Delta)\,v, \qquad s(\Delta) \in \{-1, +1\}, \qquad \mathbb{P}(s = +1) = \mathbb{P}(s = -1) = \tfrac{1}{2}, \tag{29}$$

where a is a fixed semantic component and $s(\Delta)v$ is a spurious component that flips across perturbations.

Theorem A.13 (Spurious determinism implies $U_A = 0$ but $U_E \succ 0$)**.** *Assume* (28)*. Then* $U_A = \mathbb{E}_\Delta[\Sigma(\Delta)] = 0$*. If* $\mu(\Delta)$ *is non-constant with* $\mathrm{Cov}_\Delta(\mu(\Delta)) \neq 0$*, then*

$$U_E = \mathrm{Cov}_\Delta(\mu(\Delta)) \succcurlyeq 0 \quad \textit{and} \quad U_E \neq 0.$$

In particular, under the flip model (29)*,*

$$U_E = vv^\top, \qquad \mathrm{tr}(U_E) = \|v\|_2^2 = 1.$$

Table 9: Comparison of uncertainty-correctness correlation (AUROC, mean ± std over the bootstrap distribution) across models and methods on TriviaQA (no-context) and TruthfulQA. Numbers in **bold** represent the overall best performance, underlined numbers indicate the top-performing black-box method (including ours), and numbers with a gray background indicate the top-performing white-box method. This is the deviation-augmented companion to the main-paper headline Tab. 1; the GPQA/SciQA/CoQA companion is Tab. 10.

Method	LLaMA-3-8B				Qwen2.5-7B			
	TriviaQA		TruthfulQA		TriviaQA		TruthfulQA	
	Cosine	ROUGE-L	Cosine	ROUGE-L	Cosine	ROUGE-L	Cosine	ROUGE-L
Ours								
AU	**0.799** ± 0.009	**0.845** ± 0.008	0.782 ± 0.008	0.694 ± 0.019	**0.773** ± 0.009	**0.815** ± 0.002	**0.812** ± 0.034	**0.731** ± 0.005
EU	0.585 ± 0.007	0.588 ± 0.005	0.745 ± 0.029	0.647 ± 0.032	0.570 ± 0.007	0.547 ± 0.009	0.684 ± 0.017	0.583 ± 0.019
TU	0.784 ± 0.009	0.826 ± 0.008	**0.799** ± 0.088	**0.715** ± 0.023	0.751 ± 0.011	0.782 ± 0.004	0.809 ± 0.025	0.719 ± 0.009
Black-box								
D-SE	0.742 ± 0.004	0.784 ± 0.005	0.511 ± 0.026	0.550 ± 0.018	0.647 ± 0.012	0.667 ± 0.010	0.530 ± 0.030	0.540 ± 0.021
Self-Consistency	0.740 ± 0.004	0.782 ± 0.005	0.524 ± 0.026	0.555 ± 0.019	0.647 ± 0.011	0.667 ± 0.010	0.536 ± 0.031	0.542 ± 0.022
SIMBA	0.749 ± 0.005	0.785 ± 0.008	0.506 ± 0.015	0.520 ± 0.022	0.632 ± 0.012	0.663 ± 0.014	0.598 ± 0.039	0.643 ± 0.024
SelfCheckGPT	0.754 ± 0.006	0.789 ± 0.008	0.517 ± 0.021	0.527 ± 0.022	0.641 ± 0.013	0.650 ± 0.013	0.592 ± 0.033	0.651 ± 0.017
SPUQ	0.615 ± 0.003	0.641 ± 0.007	0.602 ± 0.028	0.583 ± 0.029	0.616 ± 0.015	0.640 ± 0.007	0.544 ± 0.039	0.590 ± 0.031
P(True)	0.632 ± 0.003	0.652 ± 0.005	0.600 ± 0.021	0.639 ± 0.015	0.603 ± 0.008	0.638 ± 0.004	0.763 ± 0.027	0.717 ± 0.014
White-box								
G-NLL	0.786 ± 0.008	0.798 ± 0.007	0.546 ± 0.029	0.549 ± 0.024	0.646 ± 0.013	0.662 ± 0.009	0.611 ± 0.023	0.700 ± 0.012
PE	0.710 ± 0.003	0.732 ± 0.007	0.578 ± 0.010	0.576 ± 0.008	0.644 ± 0.014	0.657 ± 0.014	0.562 ± 0.036	0.622 ± 0.026
LN-PE	0.761 ± 0.005	0.796 ± 0.007	0.496 ± 0.014	0.496 ± 0.021	0.637 ± 0.013	0.658 ± 0.008	0.570 ± 0.019	0.623 ± 0.012
SE	0.739 ± 0.005	0.774 ± 0.007	0.560 ± 0.016	0.525 ± 0.015	0.657 ± 0.011	0.674 ± 0.012	0.597 ± 0.042	0.625 ± 0.029
LN-SE	0.742 ± 0.005	0.777 ± 0.006	0.543 ± 0.019	0.492 ± 0.014	0.653 ± 0.013	0.670 ± 0.013	0.573 ± 0.031	0.608 ± 0.024

Proof. If $Z = \mu(\Delta)$ a.s. given Δ, then $\Sigma(\Delta) = \mathrm{Cov}(Z \mid \Delta) = 0$ for every Δ, so $U_A = \mathbb{E}_\Delta[\Sigma(\Delta)] = 0$.

By definition, $U_E = \mathrm{Cov}_\Delta(\mu(\Delta))$ is always positive semidefinite. Moreover, $U_E = 0$ iff $\mu(\Delta)$ is a.s. constant; thus if $\mu(\Delta)$ is non-constant, then $U_E \neq 0$.

For the flip model (29), note $\mathbb{E}[s(\Delta)] = 0$ and $\mathrm{Var}(s) = 1$. Then $\mathbb{E}[\mu(\Delta)] = a\,u$ and

$$\mu(\Delta) - \mathbb{E}[\mu(\Delta)] = s(\Delta)\,v.$$

Therefore,

$$U_E = \mathbb{E}\big[(\mu - \mathbb{E}\mu)(\mu - \mathbb{E}\mu)^\top\big] = \mathbb{E}\big[s(\Delta)^2\, vv^\top\big] = vv^\top.$$

Taking trace yields $\mathrm{tr}(U_E) = \mathrm{tr}(vv^\top) = \|v\|_2^2 = 1$. □

Interpretation. Under spurious determinism, any uncertainty method that only probes within-prompt decoding variability (i.e., only $\Sigma(\Delta)$ for a fixed prompt instance) will report near-zero uncertainty, even if the output is highly sensitive to semantic-preserving perturbations. Our perturbation-based epistemic term U_E is designed to detect precisely this sensitivity.

What this gives the main paper. Theorem A.13 *is* the formalization of the main paper's Pillar 3 statement (Thm. 4.3). The class "$\mathcal{F}_{\mathrm{cons}}$=0 but U_E bounded away from 0" is exactly the spurious-determinism regime constructed above; it has positive measure under any base distribution placing mass on prompt-format-driven outputs, since the flip model is realizable whenever the LLM has a deterministic mode that depends on a non-semantic prompt feature. Empirically, in Sec. 5.4, disambiguating a prompt significantly reduces EU while leaving all fixed-prompt consistency baselines statistically unchanged (Tab. 4).

B Validating the AU/EU Decomposition: Additional Details

This appendix provides additional setup detail and supplementary results for the controlled diagnostics in Sec. 5.4 of the main paper.

Calibration metric (ECE): definition used throughout the paper. A raw uncertainty score is not a probability, so a calibration metric requires a mapping to $[0, 1]$. We use the same protocol in every ECE column reported anywhere in the paper (Tab. 6 main, Tab. 16, Tab. 13, Tab. 15): (i) split

Table 10: AUROC (mean ± std) on GPQA, SciQA, and CoQA: the deviation-augmented companion to the main-paper headline Tab. 1. The TriviaQA/TruthfulQA companion is Tab. 9. Numbers in **bold** represent the overall best performance; underlined numbers indicate the top-performing black-box method (including ours); numbers with a gray background indicate the top-performing white-box method.

Method	GPQA		SciQA		CoQA	
	Cosine	ROUGE-L	Cosine	ROUGE-L	Cosine	ROUGE-L
LLaMA-3-8B						
Ours						
AU	**0.770** ± 0.127	**0.839** ± 0.031	0.735 ± 0.006	**0.757** ± 0.003	0.738 ± 0.007	0.735 ± 0.008
EU	0.746 ± 0.076	0.590 ± 0.043	0.534 ± 0.008	0.548 ± 0.009	0.630 ± 0.008	0.631 ± 0.007
TU	0.671 ± 0.158	0.769 ± 0.030	0.680 ± 0.007	0.703 ± 0.006	0.723 ± 0.009	0.724 ± 0.010
Black-box						
D-SE	0.641 ± 0.013	0.667 ± 0.012	0.698 ± 0.005	0.709 ± 0.003	0.648 ± 0.008	0.659 ± 0.009
S-consistency	0.602 ± 0.030	0.621 ± 0.010	0.712 ± 0.011	0.706 ± 0.008	0.641 ± 0.017	0.602 ± 0.026
SIMBA	0.635 ± 0.014	0.697 ± 0.022	0.732 ± 0.009	0.722 ± 0.009	0.583 ± 0.013	0.537 ± 0.022
SelfCheckGPT	0.639 ± 0.016	0.690 ± 0.018	0.723 ± 0.009	0.713 ± 0.009	0.566 ± 0.017	0.523 ± 0.025
SPUQ	0.554 ± 0.036	0.510 ± 0.026	0.659 ± 0.008	0.651 ± 0.008	0.505 ± 0.018	0.505 ± 0.017
P(True)	0.524 ± 0.021	0.623 ± 0.014	0.566 ± 0.003	0.568 ± 0.004	**0.788** ± 0.018	**0.774** ± 0.021
White-box						
G-NLL	0.506 ± 0.053	0.543 ± 0.034	**0.770** ± 0.005	0.756 ± 0.005	0.707 ± 0.005	0.664 ± 0.007
PE	0.680 ± 0.008	0.542 ± 0.027	0.725 ± 0.008	0.719 ± 0.008	0.667 ± 0.002	0.636 ± 0.005
LN-PE	0.639 ± 0.009	0.549 ± 0.011	0.737 ± 0.007	0.725 ± 0.003	0.658 ± 0.004	0.654 ± 0.003
SE	0.601 ± 0.062	0.511 ± 0.031	0.742 ± 0.007	0.742 ± 0.002	0.690 ± 0.006	0.675 ± 0.007
LN-SE	0.623 ± 0.017	0.608 ± 0.018	0.735 ± 0.006	0.737 ± 0.003	0.663 ± 0.005	0.664 ± 0.007
Qwen2.5-7B						
Ours						
AU	0.506 ± 0.108	0.758 ± 0.000	0.516 ± 0.022	0.508 ± 0.020	0.585 ± 0.018	0.641 ± 0.054
EU	0.612 ± 0.216	**0.949** ± 0.000	0.502 ± 0.016	0.501 ± 0.013	0.545 ± 0.028	0.516 ± 0.024
TU	0.536 ± 0.186	0.914 ± 0.005	0.515 ± 0.013	0.512 ± 0.013	0.548 ± 0.019	0.625 ± 0.051
Black-box						
D-SE	0.757 ± 0.169	0.672 ± 0.005	0.503 ± 0.000	0.503 ± 0.000	0.615 ± 0.041	0.531 ± 0.021
S-consistency	0.555 ± 0.019	0.544 ± 0.029	0.658 ± 0.014	0.660 ± 0.013	0.529 ± 0.013	0.515 ± 0.013
SIMBA	0.546 ± 0.023	0.513 ± 0.034	**0.714** ± 0.008	0.695 ± 0.010	0.566 ± 0.029	0.610 ± 0.024
SelfCheckGPT	0.548 ± 0.017	0.503 ± 0.035	0.693 ± 0.008	0.676 ± 0.009	**0.651** ± 0.021	0.685 ± 0.021
SPUQ	0.592 ± 0.059	0.512 ± 0.035	0.646 ± 0.007	0.634 ± 0.009	0.506 ± 0.040	0.513 ± 0.037
P(True)	0.561 ± 0.028	0.519 ± 0.013	0.680 ± 0.009	0.681 ± 0.008	0.520 ± 0.019	0.526 ± 0.018
White-box						
G-NLL	**0.850** ± 0.041	0.748 ± 0.023	0.524 ± 0.067	0.626 ± 0.064	0.501 ± 0.051	**0.827** ± 0.005
PE	0.610 ± 0.098	0.869 ± 0.005	0.552 ± 0.051	0.675 ± 0.095	0.511 ± 0.016	0.711 ± 0.003
LN-PE	0.625 ± 0.056	0.787 ± 0.014	0.568 ± 0.052	0.650 ± 0.028	0.526 ± 0.020	0.808 ± 0.011
SE	0.742 ± 0.057	0.608 ± 0.005	0.508 ± 0.046	**0.723** ± 0.022	0.589 ± 0.036	0.667 ± 0.038
LN-SE	0.826 ± 0.076	0.637 ± 0.009	0.534 ± 0.048	0.692 ± 0.083	0.566 ± 0.029	0.543 ± 0.050

the (uncertainty, error-label) pairs $70/30$ into evaluation and calibration folds with a fixed seed; (ii) fit a single-feature logistic regression (scikit-learn `LogisticRegression()` defaults; equivalent to a sigmoid / Platt calibrator) on the calibration fold mapping uncertainty to $\hat{p}(\text{error}) \in [0, 1]$; (iii) apply this map to the evaluation fold; (iv) compute ECE with 15 equal-width bins on $[0, 1]$, with a left-closed, right-open bin convention except for the final bin which is closed at both ends. The error label is $1-$correctness, where correctness is the same Cosine or ROUGE-L criterion (or LLM-judge `lenient` label, for long-form) used in the corresponding AUROC column. We return `NaN` on degenerate splits (single-class fold). This single procedure yields all reported ECE values.

Datasets and pairing. For the AmbigQA [36] test we use the public AmbigQA development split, restricted to questions whose ambiguity is annotated and whose gold clarification is a one-sentence rewrite of the question (this excludes free-form clarifying paragraphs). For the NQ evidence test we use the Natural Questions [25] short-answer subset and pair each closed-book question with its gold long answer passage; the long answer is concatenated as additional context but *outer perturbations are applied only to the question span*, so the perturbation distribution is held fixed across the closed/open conditions.

Statistical procedure. Per-example AUROC contributions are computed via a Hanley–McNeil-style pairwise formulation: each correct/incorrect pair contributes $\mathbb{1}[u_- > u_+] + \frac{1}{2}\mathbb{1}[u_- = u_+]$ to the AUROC (with u_+/u_- the uncertainty of the correct/incorrect answer under the incorrect-as-positive convention). The paired statistic $\Delta\mathrm{AU} - \Delta\mathrm{EU}$ averages these contributions per example before differencing, which removes between-example variance. The z-statistic in Table 3 is a Wald test on the per-example paired differences using their per-example sample variance; the BCa intervals reported in Table 11 are computed from a paired bootstrap with 1,000 resamples, and the same procedure also gives the standard errors in Table 3.

Replication on Qwen2.5-7B. Table 11 replicates the full 2×2 paired-intervention protocol of Table 3 on Qwen2.5-7B under Cosine and ROUGE-L correctness. The predicted decomposition pattern transfers cleanly: *clarification on AmbigQA selectively moves AU* (ROUGE-L: z=2.44, p=0.015; Cosine directional, p=0.125), and *evidence on NQ selectively moves EU* (z=9.33 and z=12.13, both $p<0.001$). The NQ contrasts are substantially larger than the LLaMA-3-8B headline ($\Delta\mathrm{EU}-\Delta\mathrm{AU}= + 0.085$ and $+0.110$, vs. $+0.030$ and $+0.047$ on LLaMA-3-8B), consistent with Qwen's larger closed-book knowledge gap on NQ being more dramatically closed by evidence injection.

Table 11: **Qwen2.5-7B replication of the 2×2 paired-intervention test.** $\Delta\mathrm{AU}$ and $\Delta\mathrm{EU}$ are paired AUROC differences (treatment minus control). **Bold** marks the winning side of the contrast per row (clarification predicts AU>EU; evidence predicts EU>AU); **bold** p marks $p < 0.05$. Same protocol as Table 3.

Dataset	Metric	$\Delta\mathrm{AU}$	$\Delta\mathrm{EU}$	Contrast (predicted-direction Δ, z, p)
AmbigQA (clarification ⇒ moves AU)	Cosine	$\mathbf{+0.030}\pm0.006$	$+0.013\pm0.009$	$\Delta\mathrm{AU}-\Delta\mathrm{EU}= + 0.017$ (z=1.54, p=0.125)
	ROUGE-L	$\mathbf{+0.025}\pm0.005$	-0.003 ± 0.009	$\Delta\mathrm{AU}-\Delta\mathrm{EU}= + 0.027$ (z=2.44, $\mathbf{p=0.015}$)
NQ (evidence ⇒ moves EU)	Cosine	$+0.013\pm0.008$	$\mathbf{+0.099}\pm0.009$	$\Delta\mathrm{EU}-\Delta\mathrm{AU}= + 0.085$ (z=9.33, $\mathbf{p<0.001}$)
	ROUGE-L	-0.039 ± 0.008	$\mathbf{+0.071}\pm0.008$	$\Delta\mathrm{EU}-\Delta\mathrm{AU}= + 0.110$ (z=12.13, $\mathbf{p<0.001}$)

Why bootstrapping rather than reporting raw AUROC differences? The two intervention conditions share the same examples and only differ in how the input is augmented. Reporting raw between-condition AUROC differences would treat them as independent samples and inflate variance. The paired contrast we report removes this and is valid even though AU and EU AUROCs differ in absolute level.

Score-level shifts and baseline comparisons. Tab. 4 evaluates raw score changes across the exact question pairs of Tab. 3. Fixed-prompt baselines (embedding variance, self-consistency, discrete semantic entropy, and uniform Semantic Density) are computed from 10 stochastic decodes of the prompt per condition, SPUQ is computed from 10 single-decode perturbations per condition (temperature 0.9, top-p 0.99), and our AU/EU/TU components use the N=M=10 runs. For each question pair, we compute the paired difference $\delta = \mathrm{score}_{\mathrm{after}} - \mathrm{score}_{\mathrm{before}}$ and report the standardized shift $\bar{\delta}/\mathrm{sd}(\delta)$; statistical significance (bold in Tab. 4) means that the 95% percentile bootstrap confidence interval of $\bar{\delta}$ (over 5,000 resamples of question pairs) excludes zero. On an absolute scale, our EU decreases by -7.6×10^{-5} (95% CI $[-9.8, -5.5]\times10^{-5}$) under clarification, relative to a pre-intervention mean of 5.5×10^{-4}.

Bias correction. The bias-correction validation is at the *score* level (AU–EU correlation ρ) and is reported in main-paper Table 5 on the TriviaQA/LLaMA-3-8B/Cosine slice. We do not include an additional appendix table because ρ is the right object for the finite-M inflation predicted by Prop. 4.1: the correction acts on the uncertainty score itself, not on downstream AUROC.

C Bridge Diagnostics: Estimating Assumption Quality from Data

The Bayesian bridge in Sec. 4 closes the gap between perturbation-based AU/EU and Bayesian targets only under Assumption A.6 (local first-order linearization of the conditional mean), Assumption A.7 (locally Gaussian posterior), and Assumption A.8 (bounded semantic coverage defect). This appendix turns each assumption into a *diagnostic* that is estimable from data, and reports its value on our datasets.

Local linearity defect (Assumption A.6 diagnostic). For a perturbation neighborhood $\mathcal{N}_\varepsilon(x)$, define the linearity defect $\mathrm{LD}(x;\varepsilon) := \|f(x+\Delta) - f(x) - J_x\Delta\| / \|\Delta\|$, where J_x is the embedding-space slope of f estimated by least squares on $\{(\Delta_i, f(x+\Delta_i))\}_{i=1}^N$. The bridge bound depends on $\mathbb{E}[\mathrm{LD}^2]$, which is directly measurable from the same outer-perturbation samples already collected for the AU/EU estimator. We report per-input mean, standard deviation, and 95th percentile of $\mathrm{LD}(x;\varepsilon)$.

Posterior Gaussianity (Assumption A.7 diagnostic). Assumption A.7 is a Laplace approximation on *parameters* ($\delta\theta \sim \mathcal{N}(0, \Sigma_\theta)$); under the local linearity in Assumption A.6, this induces an approximately Gaussian distribution on *embeddings*. We therefore test the induced embedding distribution directly. (We do not test scalar uncertainty scores; HZ on a one-dimensional summary would be a different test of a different object and is not what the bridge requires.) Two multivariate clouds in $\mathbb{R}^d$ (d=384) are available per input: (a) the *outer-mean* cloud $\{\hat{\mu}_i(x)\}_{i=1}^N$ across the N=10 perturbations: this is the cloud most directly tied to Assumption A.7, since under linearity it is the pushforward of $\delta\theta$ via J_θ; and (b) the *inner-sample* cloud $\{z_{i,j}(x)\}_{j=1}^M$ across the M=10 stochastic decodes at fixed perturbation, which probes within-perturbation aleatoric Gaussianity (a side condition used in the AU bridge). The HZ column in Table 12 reports the rejection rate of test (b), which is the better-powered of the two: with only N=M=10 samples in d=384, both forms of the test are deeply underpowered, and on the per-input outer-mean cloud (test (a)) the average p-value collapses to numerical zero ($\sim 10^{-14}$ on all four cells), reflecting the test's high-dimensional small-sample behavior under our budget; the HZ statistic is reported as a Gaussianity diagnostic alongside LD and MMD, and the load-bearing diagnostic for the bridge bound itself is LD.

Coverage diagnostic (perturbation–reference MMD). The semantic coverage defect that enters the bridge bound is, in principle, a Wasserstein distance between the empirical perturbation distribution and a Bayesian posterior over inputs the model treats equivalently. The latter is not directly observable; we report instead the maximum mean discrepancy (MMD) in embedding space, computed with a Gaussian RBF kernel and median-heuristic bandwidth, between the embedded perturbation cloud and a held-out 1,000-prompt reference distribution from the same dataset. This quantity measures how close the perturbation cloud sits to in-distribution prompts in the same semantic neighborhood; small values indicate that the outer perturbation distribution covers the relevant semantic geometry rather than drifting elsewhere.

Table 12: **Bridge diagnostics** on 500 inputs per cell at ε=0.1, with N=M=10. **Lower is better (↓) for every column**: each metric measures deviation from a bridge condition, and a smaller value means the bridge is tighter on that dimension. *LD* (linearity defect, Assumption A.6) is moderate (≤ 0.43 at the 95th percentile) across both models and datasets, indicating the first-order linearization is a reasonable local approximation in the perturbation neighborhood. *HZ inner-cloud* reports the fraction of M=10 inner-sample *embedding* clouds (one per (x, Δ_i)) that reject multivariate normality at $p < 0.05$; the test is a Gaussianity diagnostic on the embedding distribution at the inner-decode level, with statistical power capped by M=10 samples in 384 dimensions. The matching outer-mean cloud test (per input, on $\{\hat{\mu}_i\}_{i=1}^N$) is similarly low-power and is summarised in prose rather than tabulated. *MMD* reports the distributional coverage diagnostic between the embedded perturbation cloud and a 1,000-prompt held-out reference from the same dataset (Gaussian RBF, median-heuristic bandwidth). Values are tightly concentrated (median ≈ 0.40, p95 ≈ 0.42; gap $< 5\%$), indicating the perturbation distribution covers the held-out semantic neighborhood under MMD. The three columns play distinct roles: LD is the load-bearing input to the bridge bound, HZ is a Gaussianity diagnostic, and MMD is the coverage diagnostic; together they make the bridge falsifiable on each model×dataset pair.

Model	**Dataset**	Linearity defect ↓			HZ inner-cloud ↓	Coverage MMD ↓	
		mean	std	p95	frac. $p < 0.05$	median	p95
LLaMA-3-8B	TriviaQA	0.218	0.095	0.363	0.710	0.399	0.411
LLaMA-3-8B	TruthfulQA	0.244	0.095	0.391	0.789	0.407	0.418
Qwen2.5-7B	TriviaQA	0.252	0.100	0.396	0.614	0.398	0.411
Qwen2.5-7B	TruthfulQA	0.274	0.097	0.428	0.717	0.408	0.417

What the diagnostics imply for the bridge bound. Substituting the empirical LD into the bridge defect bound in Theorem 4.2 yields end-to-end excess terms $\varepsilon_E, \varepsilon_A$ on the perturbation-population *covariance* (Frobenius norm) that are dominated by the LD-squared contribution. With $\mathbb{E}[\mathrm{LD}^2] \leq (0.27)^2 + (0.10)^2 \approx 0.083$ in the worst cell (Qwen2.5-7B/TruthfulQA), the bridge contributes a small fraction of the perturbation-population covariance norm in the limit $N, M \to \infty$. The Frobenius

bound translates to an AUROC perturbation bound under additional ranking-margin / Lipschitz-of-aggregator assumptions; here we focus on the covariance bound that the perturbation framework directly delivers. The HZ rejection rate does not enter this calculation: the bridge bound requires only the second-moment covariance of the embedding distribution to be well-estimated, and the coverage MMD diagnostic absorbs the second-moment relaxation.

Input sensitivity *is* a meaningful EU signal. A standard worry is that perturbation-based EU only measures input sensitivity rather than parameter uncertainty. Output variation under input perturbation is in fact a well-studied proxy for epistemic uncertainty [49]: when a model has learned a robust mapping, semantic-preserving perturbations leave the conditional mean $\mu_\Delta(x)$ approximately fixed; in under-determined regions $\mu_\Delta(x)$ shifts substantially. Our framework absorbs surface-form variation into AU (Sec. 3), though EU can also capture residual semantic drift when an imperfect paraphrase alters question meaning (audited in App. F.2.3).

Instruction perturbations as parameter perturbations. Beyond input-content perturbations, prior work shows that in-context tokens (including instructions and formatting) induce *implicit gradient updates* to the model's attention parameters [7]. Concretely, processing an in-context update at a fixed weight matrix W is equivalent to processing the original input under $W + \Delta W_{\text{ICL}}$, with ΔW_{ICL} a low-rank parameter shift induced by the in-context tokens. Under this interpretation, instruction-style outer perturbations fall *inside* the bridge: the perturbation operates on effective parameters ($\theta + \Delta\theta_{\text{ICL}}$), so an analogous bridge argument applies with the induced parameter-shift covariance replacing the input-side covariance. This perspective explains, rather than contradict, the empirical observation that instruction-style perturbations yield a strong EU signal in our ablations: under [7], instruction variation acts directly as an implicit parameter perturbation rather than a mere surface-form confound.

D SPUQ Comparison and Aggregation Factorization

SPUQ [13] is the closest published black-box method and shares two ingredients with our method (input perturbations and sampled answers), but differs in three structural ways: (a) *one level vs. two levels* — SPUQ takes one decode per perturbation, so the conditional moments $\mu_\Delta, \Sigma_\Delta$ that the law of total covariance requires are unavailable, and no AU/EU split is even definable; (b) *scalar vs. matrix-valued* — SPUQ's aggregators collapse the perturbation cloud into one number, while we return $(\widehat{U}_A, \widehat{U}_E) \in \mathbb{R}^{d\times d} \times \mathbb{R}^{d\times d}$ with the predicted finite-M bias correction, which is what enables the paired interventions in Sec. 5.4; (c) *heuristic vs. analyzed* — SPUQ provides no theoretical analysis of consistency or bias, while Prop. A.1 and Thm. A.5 cover both and the separation result formalizes a class of failures that fixed-prompt consistency scores provably miss.

Head-to-head comparison. We operate our method at $N{=}M{=}10$ to share the outer perturbation count ($K_{\text{SPUQ}}{=}N{=}10$) with SPUQ; ours additionally incurs $M{=}10$ inner decodes per outer perturbation, which is the intrinsic cost of the two-level decomposition and is not present in any one-level method. This is *not* a matched total-API-call budget (ours uses $\sim 10\times$ more LLM calls per input; equal budgets: Tab. 2); we read the comparison as "does the decomposition deliver capability the one-level scalar cannot, and is the gain large enough to justify the extra inner decodes?". Table 13 reports the $\{$LLaMA-3-8B, Qwen2.5-7B$\} \times \{$TriviaQA, TruthfulQA$\}$ result. Three observations: (i) Ours (TU) dominates SPUQ (TU) on all four cells, with the largest gap on Qwen2.5-7B / TruthfulQA where SPUQ drops to 0.392 (worse than chance) while our TU stays at 0.629; (ii) Ours (AU) is uniformly the strongest single column (0.630–0.855); (iii) calibration favors ours on TriviaQA by a 4–$8\times$ margin (TU ECE 0.013–0.015 vs. 0.062–0.104) and is essentially tied on TruthfulQA, so there is no calibration cost to the two-level structure. The next subsection isolates which structural difference is load-bearing.

The perturbation family is not the load-bearing piece. A natural follow-up question is whether our gain over SPUQ comes simply from using a better perturbation. Table 24 ablates the perturbation family while holding the two-level structure fixed: the framework works with random paraphrasing, distractor injection, instruction variation, and SPUQ-style noise, with instruction variation giving the best signal.

Table 13: **SPUQ vs. Ours at shared outer perturbation count ($K_{\mathrm{SPUQ}}{=}N{=}10$); our two-level structure additionally uses $M{=}10$ inner decodes per outer perturbation.** Our method exposes three columns (AU, EU, TU) where SPUQ exposes one (TU). **Bold**: best per row; underlined: second-best. AUROC ranking is over SPUQ (TU), Ours (AU), Ours (TU); ECE ranking is over SPUQ (TU) and Ours (TU). ECE follows the unified protocol of §B. SPUQ uses inter-sample ROUGE-L aggregation with the three perturbation modes from [13]; ours uses $\varepsilon{=}0.1$ paraphrase perturbation and Sentence-BERT embedding-statistics aggregation.

Model	Dataset	SPUQ ($K_{\mathrm{SPUQ}}{=}10$)		Ours ($N{=}M{=}10$)			
		TU AUROC ↑	TU ECE ↓	AU AUROC ↑	EU AUROC ↑	TU AUROC ↑	TU ECE ↓
LLaMA-3-8B	TriviaQA	0.808 ± 0.014	$\underline{0.062}$	$\mathbf{0.845}$	0.638	$\underline{0.822 \pm 0.014}$	$\mathbf{0.015}$
LLaMA-3-8B	TruthfulQA	0.657 ± 0.069	$\mathbf{0.009}$	$\mathbf{0.695}$	0.585	$\underline{0.681 \pm 0.064}$	$\mathbf{0.009}$
Qwen2.5-7B	TriviaQA	0.781 ± 0.014	$\underline{0.104}$	$\mathbf{0.855}$	0.607	$\underline{0.828 \pm 0.013}$	$\mathbf{0.013}$
Qwen2.5-7B	TruthfulQA	0.392 ± 0.046	$\underline{0.006}$	$\mathbf{0.630}$	0.560	$\underline{0.629 \pm 0.046}$	$\mathbf{0.005}$

D.1 Decomposition vs. aggregator: a $2{\times}2$ factorial

We isolate which structural difference is load-bearing by replacing our embedding-statistics aggregator with SPUQ's [13] inter-sample (pairwise) ROUGE-L [30] aggregator while keeping the two-level ensemble structure unchanged. The four conditions form a $2{\times}2$ factorial along the *decomposition* axis (one-level vs. two-level) and the *aggregator* axis (ROUGE-L vs. embedding-statistics): (i) one-level + ROUGE aggregator (vanilla SPUQ); (ii) one-level + embedding aggregator; (iii) two-level + ROUGE aggregator; (iv) two-level + embedding aggregator (our full method). One-level conditions emit a single TU score and have no AU/EU split; two-level conditions expose AU, EU, TU and the AU–EU per-example Spearman correlation.

Table 14: **Aggregation ablation** ($N{=}M{=}10$, $\varepsilon{=}0.1$, cosine-correctness label, 1,000 examples per cell). **Higher is better (↑) for AUROC; lower is better (↓) for AU–EU ρ (stronger AU/EU separation). Bold**: best per column; underlined: second-best. The **Score** column tags which uncertainty is reported in the AUROC column: one-level methods produce a single scalar (reported as TU since there is no decomposition); two-level methods report AU, which we established as the strongest single-head signal in main-paper Table 1. *AU–EU ρ* (per-example Spearman of AU and EU rankings) cleanly separates the aggregator axis: the ROUGE aggregator (iii) collapses AU and EU into nearly the same signal ($\rho \in [0.69, 0.82]$), whereas the embedding aggregator (iv) yields a properly non-degenerate decomposition ($\rho \in [0.15, 0.40]$). Two-level + embedding is therefore the configuration in which the AU/EU split is actually informative; the embedding aggregator is what makes the decomposition load-bearing rather than cosmetic.

Configuration	Score	LLaMA-3-8B				Qwen2.5-7B			
		TriviaQA		TruthfulQA		TriviaQA		TruthfulQA	
		AUROC ↑	AU–EU ρ ↓	AUROC ↑	AU–EU ρ ↓	AUROC ↑	AU–EU ρ ↓	AUROC ↑	AU–EU ρ ↓
(i) one-level + ROUGE	TU	0.827	–	**0.705**	–	0.824	–	0.561	–
(ii) one-level + embedding	TU	0.811	–	$\underline{0.697}$	–	0.809	–	$\underline{0.625}$	–
(iii) two-level + ROUGE	AU	**0.851**	$\underline{0.807}$	0.696	$\underline{0.770}$	**0.861**	$\underline{0.819}$	0.600	$\underline{0.692}$
(iv) two-level + embedding	AU	$\underline{0.845}$	**0.402**	0.695	**0.296**	$\underline{0.855}$	**0.267**	**0.630**	**0.154**

Calibration corollary. The two-level + embedding configuration also has materially lower TU expected calibration error than the alternatives on the cosine-correctness slice: TU ECE is 0.015 (LLaMA-3-8B / TriviaQA) and 0.013 (Qwen2.5-7B / TriviaQA) for (iv), versus 0.066 and 0.053 for (iii) and 0.071 and 0.065 for (i). On TruthfulQA all four conditions have TU ECE below 0.011, dominated by base-rate effects. This pattern is consistent with the AU–EU ρ result: ROUGE-L scalars carry duplicate information across both heads and so overconfide, whereas the embedding-statistics decomposition produces uncertainty heads that are individually well-calibrated.

Reading the result. The gain over SPUQ has two distinct structural sources, both load-bearing. Table 14 attributes them: the two-level decomposition supplies the AUROC gain, and the embedding-statistics aggregator supplies the AU/EU separation that makes the decomposition usable downstream (Sec. 5.4). Replacing only one factor leaves the other capability missing: configuration (iii) matches us on AU AUROC but cannot tell AU from EU, and configuration (ii) has clean scoring geometry but no decomposition to apply it to.

Robustness across correctness criteria. The headline table reports cosine-thresholded correctness on the two-level conditions only. Table 15 mirrors it under both Cosine and ROUGE-L correctness:

the AU AUROC of (iii) and (iv) stays within 0.03, and the calibration advantage of the embedding aggregator on TriviaQA persists across both criteria.

Table 15: **AUROC and ECE robustness across correctness criteria**, mirroring Table 14. **Higher is better (↑) for AUROC; lower is better (↓) for ECE** (ECE follows the unified protocol of §B). **Bold**: best per (cell, correctness) column; underlined: second-best. The **Score** column tags which uncertainty drives the AUROC column: one-level methods (i)–(ii) emit a single scalar (TU-equivalent, no decomposition); two-level methods (iii)–(iv) report AU. Two-level + ROUGE (iii) wins AUROC in 5 of 8 (cell, correctness) cells; one-level + ROUGE (i) wins on LLaMA-3-8B/TruthfulQA under Cosine and (iv) wins on LLaMA-3-8B/TruthfulQA under ROUGE-L and on Qwen2.5-7B/TruthfulQA under Cosine (small-margin honest exceptions). On TriviaQA, the embedding aggregator (iv) reduces predictive ECE by $3.5-12\times$ relative to all other conditions under both correctness criteria. On TruthfulQA the four conditions are within a thousandth of one another on ECE because positive-class base rates are small.

			LLaMA-3-8B / TriviaQA		LLaMA-3-8B / TruthfulQA		Qwen2.5-7B / TriviaQA		Qwen2.5-7B / TruthfulQA	
Configuration	**Corr.**	**Score**	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓
(i) one-level + ROUGE	cosine	TU	0.827	0.071	**0.705**	0.009	0.824	0.065	0.561	0.005
	ROUGE-L	TU	0.829	0.060	0.696	0.008	0.823	0.078	0.654	0.034
(ii) one-level + embedding	cosine	TU	0.811	0.054	0.697	0.009	0.809	0.054	0.625	0.005
	ROUGE-L	TU	0.807	0.049	0.686	0.008	0.793	0.057	0.650	0.034
(iii) two-level + ROUGE	cosine	AU	**0.851**	0.066	0.696	0.009	**0.861**	0.053	0.600	0.005
	ROUGE-L	AU	**0.846**	0.074	0.692	**0.007**	**0.847**	0.039	**0.711**	0.034
(iv) two-level + embedding	cosine	AU	0.845	**0.015**	0.695	0.009	0.855	**0.013**	**0.630**	0.005
	ROUGE-L	AU	0.831	**0.006**	**0.702**	0.008	0.828	**0.011**	0.701	0.034

E Long-form Pilot: Segment-level UQ on NQ Long-Answer

The main-paper claim-level result is in Tab. 6. This appendix provides (a) the full pipeline specification (§E.1), (b) the response-level companion table (Tab. 16), and (c) the discussion of the pilot's scope and conclusions. Setup recap: we apply our method to long-form generation on the NQ long-answer split [25], in two complementary granularities: *response-level*, treating the full long answer as one unit and reporting both mean- and max-aggregation of the per-sentence segment scores; and *claim-level* (main paper), decomposing each long answer into atomic claims (using a same-strength LLM extractor) and scoring each claim independently. Correctness is judged by an LLM judge against the gold long-answer paragraph; we use the *lenient* judge variant at response level (the strict variant returned no fully-correct response in this small pilot, which is itself diagnostic) and a per-claim factuality label at claim level. The five black-box baselines (**avgEntail**, **NonContradiction**, **UQLM cosine** following Bouchard et al. [5], Ross et al. [47], **SelfCheckGPT** [35], and **ROUGE-L (cons.)** [30]) consume the same per-claim $M=5$ inner sample pool that our method uses; this gives every consistency-only baseline exactly the budget it can use (the inner level), so the comparison is on equal terms for the slot they occupy.

E.1 Long-form pilot protocol (extractor, judge, budget)

Budget (specific to the long-form pilot, smaller than short-form for efficiency). Outer perturbation count $N=5$ (with index $i=0$ the unperturbed prompt, reused as the $m=5$ Bouchard-style baseline at zero perturbation), inner ensemble $M=5$, paraphrase pool $K_{\text{pool}}=50$, noise scale $\varepsilon=0.1$. Total per-question cost is $N \times M=25$ long-form decodes; total per-cell cost (claim-level) is $\#\text{claims} \times N \times M$, not the short-form headline NM per response, since each claim is scored against all 25 samples (yielding $1{,}104\times25=27{,}600$ NLI/BERTScore pairs on the LLaMA cell, $872\times25=21{,}800$ on Qwen). 100 NQ-long prompts per cell on a single RTX 4090: $\sim$35 min generation, $\sim$10 min NLI, $\sim$3 min BERTScore, $\sim$3 min span-window SBERT, $\sim$10 min judge labelling.

Generator. Same model as the rest of the paper for that cell: LLaMA-3-8B-Instruct or Qwen2.5-7B-Instruct, BF16, single GPU. Decoding: `do_sample=True`, `temperature=0.9`, `top_k=60`, `top_p=0.99`, `max_new_tokens=256`. Greedy reference uses the same template with `do_sample=False`.

Atomic claim extractor. Same model as the generator, deliberately reused so the extractor is no stronger than the system being evaluated. The extraction prompt (FactScore Listing-4

style) requires single-sentence, self-contained, declarative, third-person claims and is deterministic (`do_sample=False`, `max_new_tokens=300`). Parsed claims are de-duplicated, capped at 12 per response, and required to have 3–50 words. Empirical yield: 1,104 claims / 100 responses (LLaMA), 872 / 100 (Qwen).

Per-claim scoring. For our method, each claim's signal is a span-window embedding (sliding 30-word window, stride 15, plus a tail-aligned window) inside each long sample, with the window having the highest cosine similarity to the claim selected; AU/EU/TU are then computed by the standard two-level decomposition over the N outer means and M inner samples (trace-aggregated). Baselines (**avgEntail**, **NonContradiction**, **UQLM cosine**, **SelfCheckGPT (BERTScore)**, **ROUGE-L (cons.)**) are aggregated over the $M{=}5$ *outer*-0 (unperturbed) samples per Bouchard's $m{=}5$ recipe, so they consume only the inner-level budget our method also uses. NLI scoring uses `cross-encoder/nli-deberta-v3-base`; SelfCheckGPT uses `bert_score` with `roberta-large`; ROUGE-L is the F-measure.

Factuality judge. `Qwen3.6-27B` (4-bit NF4 quantized, `bnb_4bit_compute_dtype=bfloat16`, requires `transformers`$\geq$5.7.0) with chat-template thinking-mode disabled (`enable_thinking=False`) so it emits the decision token within `max_new_tokens=16`. The judge prompt (FactScore Listing-1 style) supplies the gold long-answer paragraph (truncated to 3,000 characters) plus the source URL and asks for one of {`SUPPORTED`, `NOT_SUPPORTED`, `UNCERTAIN`}. Decoding is greedy. Per-response derived labels: *strict* (every claim `SUPPORTED`) and *lenient* (no claim `NOT_SUPPORTED`).

Empirical judge label distribution. LLaMA cell: 138/1,104 `SUPPORTED` (12.5%), 106 `NOT_SUPPORTED` (9.6%), 860 `UNCERTAIN` (77.9%); response-level strict 0% / lenient 47%. Qwen cell: 118/872 `SUPPORTED` (13.5%), 113 `NOT_SUPPORTED` (13.0%), 641 `UNCERTAIN` (73.5%); strict 0% / lenient 46%. The high `UNCERTAIN` rate is expected: short Wikipedia-paragraph references genuinely cannot adjudicate $\sim$80% of the atomic facts in a $\sim$180-word generation; the `NOT_SUPPORTED` rate (10–13%) is the factuality-error signal AUROC tracks. *strict* is uniformly 0, so headline numbers use *lenient*.

Caveats and limitations. (i) No human verification of the LLM judge on a held-out subset; (ii) no inter-judge agreement (single judge model, single prompt); (iii) no test of judge-LLM bias against the generator family (Qwen-judge on LLaMA-3 generations is plausibly less biased than Qwen-judge on Qwen-generations and we do not separate the two effects); (iv) the bias-correction proposition (Prop. 4.1) is derived for short-form and is not re-derived for long-form here. The pilot is a controlled feasibility test under these caveats.

Table 16: **Response-level long-form pilot (NQ long-answer, lenient LLM-judge correctness).** Per-response uncertainty obtained by aggregating per-sentence segment scores, then correlating the per-response score with binary lenient correctness. **Bold**: best per column; underlined: second-best. ECE follows the unified protocol of §B. On LLaMA-3-8B, Ours (AU) max-aggregated narrowly leads at 0.653 over avgEntail at 0.648. On Qwen2.5-7B, Ours (TU) max wins outright at 0.683. Strict-label results were all-NaN at $n{=}100$ (no fully correct generations under the strict criterion in this pilot) and are omitted.

		LLaMA-3-8B ($n{=}100$)		**Qwen2.5-7B** ($n{=}100$)	
Method	**Agg.**	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓
Ours (AU)	mean	0.585	0.099	0.641	<u>0.085</u>
	max	**0.653**	0.100	<u>0.674</u>	0.107
Ours (EU)	mean	0.562	0.100	0.626	0.086
	max	0.598	0.100	0.647	0.125
Ours (TU)	mean	0.584	0.099	0.641	**0.084**
	max	0.638	**0.098**	**0.683**	0.149
avgEntail	mean	<u>0.648</u>	**0.098**	0.585	<u>0.085</u>
NonContradiction	mean	0.587	0.099	0.658	0.208
UQLM cosine	mean	0.548	0.100	0.637	0.142
SelfCheckGPT	mean	0.538	0.100	0.461	0.086
ROUGE-L (cons.)	mean	0.490	0.100	0.450	0.103

Scope of the pilot. Within the $n{=}100/{\sim}10^3$-claim setup with an LLM-judge factuality label, the pilot supports three concrete conclusions: (1) the two-level decomposition transports to claim-level long-form factuality without retraining, beating surface-similarity baselines on both models and trailing a pretrained NLI scorer (avgEntail) by 0.009 on LLaMA-3-8B and 0.060 on Qwen2.5-7B;

Table 17: TriviaQA on large models (70B+). Boxed **bold**: overall best per column; underlined: best white-box per column.

Method	LLaMA-3-70B		Qwen2.5-72B	
	Cosine	ROUGE-L	Cosine	ROUGE-L
Ours				
AU	0.726 ± 0.009	0.717 ± 0.012	$\mathbf{0.768 \pm 0.009}$	$\mathbf{0.808 \pm 0.006}$
EU	0.791 ± 0.028	0.725 ± 0.023	0.613 ± 0.009	0.592 ± 0.007
TU	$\mathbf{0.794 \pm 0.019}$	$\mathbf{0.774 \pm 0.019}$	0.754 ± 0.003	0.773 ± 0.011
Black-box				
D-SE	0.637 ± 0.013	0.686 ± 0.016	0.625 ± 0.013	0.672 ± 0.003
White-box				
G-NLL	$\underline{0.757 \pm 0.012}$	$\underline{0.735 \pm 0.018}$	$\underline{0.726 \pm 0.017}$	$\underline{0.781 \pm 0.019}$
PE	0.549 ± 0.013	0.627 ± 0.027	0.659 ± 0.019	0.702 ± 0.019
LN-PE	0.660 ± 0.013	0.710 ± 0.035	0.711 ± 0.020	0.774 ± 0.019
SE	0.615 ± 0.008	0.679 ± 0.029	0.659 ± 0.014	0.700 ± 0.017
LN-SE	0.587 ± 0.011	0.686 ± 0.033	0.660 ± 0.015	0.701 ± 0.017

(2) the AU $>$ EU ordering is preserved at long-form, mirroring the short-form behavior; (3) the framework and entailment-based scorers are complementary, motivating a hybrid that retains the two-level outer perturbation structure with an NLI-based inner aggregator. The bias-correction identity (Prop. 4.1) is derived for short-form and would be re-derived for the long-form variant in such a hybrid. Response- and claim-level results are in Tab. 16 and main-paper Tab. 6.

F Ablation Studies

Large-model results. Table 17 reports results for larger instruction-tuned models (70B+), which are often more capable but may also display different calibration behavior. We include these results to demonstrate that the relative trends between our AU/EU/TU decomposition and competing baselines remain consistent at larger model scales.

F.1 Sampling budget: two-level ensemble sizes (N, M)

Appendix Tables 18,19, 19b study sensitivity to the **two-level ensemble budget**, where the outer size N controls the number of semantic perturbation draws and the inner size M controls the number of within-input stochastic generations per perturbation. The main-text Tab. 7b shows a compressed view (TruthfulQA, Cosine) of the tied-budget sweep; the full grid is in Table 18. We report: (i) tied budgets $N{=}M$ (Table 18), (ii) fixed outer ensemble $N{=}10$ with varying M (Table 19), and (iii) fixed inner ensemble $M{=}10$ with varying N (Table 19b). We additionally provide wall-clock measurements in seconds per question for representative budgets (Table 20).

Empirical trend: rapid improvement then plateau. Across TruthfulQA and TriviaQA, performance improves quickly from very small budgets and then exhibits diminishing returns. For tied budgets (Table 18), $3{\to}5$ yields large gains on TruthfulQA (especially AU/TU under Cosine), while $10{\to}20{-}30$ provides only marginal improvement; the same pattern holds along single axes (Tables 19, 19b). Overall, $(N, M){\approx}(10, 10)$ is a strong default that achieves near-peak AUROC while remaining computationally practical.

Why plateauing is expected, and what small budgets miss. The AU/EU/TU estimates are Monte Carlo approximations under two sources of randomness: within-input stochasticity (inner sampling, $1/M$) and across-perturbation diversity (outer sampling, $1/N$). Once estimation noise drops below the irreducible overlap between correct/incorrect score distributions, improvements taper. Caveat: good AUROC at very small budgets does *not* imply sufficient uncertainty coverage — the estimator may approximate the mean well yet miss low-probability modes (paraphrase neighborhoods that trigger hallucination, decoding branches with inconsistent claims). Thus $N{=}M{=}3$ is a reasonable AUROC starting point but a low-diversity regime for hard cases.

Interpreting N vs. M and practical recipe. M primarily reduces noise in within-perturbation statistics (relevant for AU); N expands across-perturbation coverage (relevant for EU because EU is driven by variability across semantically nearby prompts). Both axes improve TU, but

Table 18: **Budget ablation, full grid.** Tied $N{=}M$ on TruthfulQA and TriviaQA (LLaMA-3-8B), both correctness criteria. **Bold**: best per column per correctness; underlined: second-best. The TruthfulQA / Cosine slice is shown compactly in the main-text Tab. 7b.

		TruthfulQA			TriviaQA		
$N{=}M$	Metric	AU	EU	TU	AU	EU	TU
3	Cosine	0.745 ±0.030	0.707 ±0.028	0.755 ±0.019	0.767 ±0.011	**0.676** ±0.008	0.761 ±0.009
	ROUGE-L	0.633 ±0.025	0.655 ±0.036	0.674 ±0.026	0.806 ±0.010	**0.689** ±0.002	0.792 ±0.008
5	Cosine	**0.798** ±0.023	0.756 ±0.022	0.797 ±0.017	0.788 ±0.009	0.630 ±0.011	0.765 ±0.008
	ROUGE-L	0.679 ±0.022	0.651 ±0.026	0.698 ±0.024	0.828 ±0.009	0.654 ±0.005	0.805 ±0.008
10	Cosine	0.792 ±0.009	**0.803** ±0.017	**0.815** ±0.008	**0.800** ±0.010	0.605 ±0.010	0.781 ±0.010
	ROUGE-L	0.693 ±0.020	**0.657** ±0.025	0.712 ±0.020	0.843 ±0.009	0.615 ±0.005	0.820 ±0.010
20	Cosine	0.782 ±0.008	0.745 ±0.029	0.799 ±0.008	0.799 ±0.009	0.585 ±0.007	0.784 ±0.009
	ROUGE-L	**0.694** ±0.019	0.647 ±0.032	**0.715** ±0.023	0.845 ±0.008	0.588 ±0.005	0.826 ±0.008
30	Cosine	0.768 ±0.011	0.746 ±0.023	0.791 ±0.009	0.800 ±0.009	0.571 ±0.011	**0.787** ±0.010
	ROUGE-L	0.688 ±0.020	0.638 ±0.032	0.714 ±0.023	**0.846** ±0.009	0.577 ±0.005	**0.830** ±0.009

Table 19: Single-axis budget ablation. **(a)** fixed $N{=}10$, varying M; **(b)** fixed $M{=}10$, varying N. AUROC for correctness prediction on TruthfulQA and TriviaQA. **Bold**: best per column per metric within each panel; underlined: second-best.

(a) Vary M (fixed $N{=}10$).

		TruthfulQA			TriviaQA		
Budget M ($N{=}10$)	Metric	AU	EU	TU	AU	EU	TU
3	Cosine	0.779 ±0.013	0.790 ±0.022	0.800 ±0.012	0.795 ±0.010	**0.700** ±0.009	**0.786** ±0.010
	ROUGE-L	0.676 ±0.021	**0.676** ±0.025	0.700 ±0.022	0.836 ±0.010	**0.719** ±0.005	**0.823** ±0.010
5	Cosine	0.783 ±0.009	**0.808** ±0.016	0.811 ±0.008	0.798 ±0.009	0.649 ±0.011	0.783 ±0.010
	ROUGE-L	0.677 ±0.020	0.675 ±0.024	0.705 ±0.020	0.842 ±0.009	0.665 ±0.006	0.822 ±0.010
10	Cosine	0.792 ±0.009	0.803 ±0.017	0.815 ±0.008	0.800 ±0.010	0.605 ±0.010	0.781 ±0.010
	ROUGE-L	**0.693** ±0.020	0.657 ±0.025	0.712 ±0.020	0.843 ±0.009	0.615 ±0.005	0.820 ±0.010
20	Cosine	**0.796** ±0.014	0.803 ±0.021	**0.823** ±0.011	0.800 ±0.009	0.577 ±0.008	0.781 ±0.010
	ROUGE-L	0.690 ±0.022	0.659 ±0.024	0.718 ±0.021	0.845 ±0.009	0.584 ±0.003	0.821 ±0.010
30	Cosine	0.791 ±0.012	0.802 ±0.022	0.820 ±0.010	**0.801** ±0.009	0.564 ±0.007	0.779 ±0.010
	ROUGE-L	0.689 ±0.022	0.663 ±0.025	**0.720** ±0.021	**0.846** ±0.009	0.571 ±0.003	0.819 ±0.009

(b) Vary N (fixed $M{=}10$).

		TruthfulQA			TriviaQA		
Budget N ($M{=}10$)	Metric	AU	EU	TU	AU	EU	TU
3	Cosine	0.777 ±0.018	0.693 ±0.034	0.765 ±0.017	0.784 ±0.010	0.621 ±0.011	0.748 ±0.009
	ROUGE-L	0.681 ±0.019	0.638 ±0.038	0.697 ±0.026	0.825 ±0.009	0.634 ±0.003	0.780 ±0.007
5	Cosine	**0.796** ±0.020	0.751 ±0.022	0.798 ±0.014	0.788 ±0.009	0.599 ±0.011	0.762 ±0.009
	ROUGE-L	0.688 ±0.020	0.636 ±0.025	0.702 ±0.023	0.830 ±0.009	0.618 ±0.005	0.802 ±0.008
10	Cosine	0.792 ±0.009	**0.803** ±0.017	**0.815** ±0.008	**0.800** ±0.010	0.605 ±0.010	0.781 ±0.010
	ROUGE-L	0.693 ±0.020	0.657 ±0.025	0.712 ±0.020	0.843 ±0.009	0.615 ±0.005	0.820 ±0.010
20	Cosine	0.777 ±0.007	0.767 ±0.022	0.797 ±0.007	0.798 ±0.010	0.617 ±0.009	0.784 ±0.010
	ROUGE-L	**0.694** ±0.018	**0.658** ±0.032	**0.714** ±0.022	0.842 ±0.009	0.625 ±0.005	0.825 ±0.008
30	Cosine	0.772 ±0.010	0.765 ±0.017	0.792 ±0.008	0.798 ±0.010	**0.623** ±0.011	**0.788** ±0.010
	ROUGE-L	0.692 ±0.019	0.654 ±0.032	0.713 ±0.023	**0.844** ±0.009	**0.635** ±0.006	**0.830** ±0.009

the decomposition gives an operational handle: spend more on M when decoding stochasticity dominates, more on N when semantic instability dominates. Runtime grows as $\propto NM$ (Table 20); $(N, M){\approx}(10, 10)$ delivers near-peak AUROC at ~ 1.8 sec/question, while $(20{-}30)$ budgets are worth the cost only when tail coverage matters more than average performance.

F.1.1 Matched total generation budget

Tab. 2 compares alternative allocations of a fixed per-question generation budget $B \in \{10, 25, 100\}$ across the same 400 TriviaQA questions per model, using the main-paper decoding hyperparameters, Sentence-BERT encoder, and Exact Match (EM) correctness labels. *Fixed prompt*: B decodes subsampled from 100 stochastic generations of the unperturbed prompt, from which embedding variance, self-consistency, discrete semantic entropy, and Semantic Density [42] are computed. *Perturbation-only*: B distinct paraphrases with a single decode each ($N{=}B, M{=}1$), where finite-M bias correction is inapplicable. *Decoupled*: an unnested split of $\lfloor B/2 \rfloor$ fixed-prompt decodes and $\lceil B/2 \rceil$ single-decode paraphrases with summed variances. *Ours*: nested allocations of $N{\times}M = 5{\times}2$, $5{\times}5$ (subsampled from the $10{\times}10$ runs), and $10{\times}10$. At $B{=}10$, the one-level perturbation designs draw from the primary paraphrase pool, whereas at $B{\geq}25$ they draw from an extended pool. SPUQ

Table 20: Computational cost in average seconds per question.

Method	Total Samples	Time (sec/q)
Ours		
Ours ($N = 3$, $M = 3$)	9	0.299
Ours ($N = 5$, $M = 5$)	25	0.571
Ours ($N = 10$, $M = 3$ / $N = 3$, $M = 10$)	30	0.656
Ours ($N = 10$, $M = 5$ / $N = 5$, $M = 10$)	50	0.996
Ours ($N = 10$, $M = 10$)	100	1.845
Ours ($N = 20$, $M = 10$ / $N = 10$, $M = 20$)	200	3.545
Ours ($N = 30$, $M = 10$)	300	5.244
Ours ($N = 20$, $M = 20$)	400	6.944
Ours ($N = 30$, $M = 30$)	900	15.441
Black-box		
D-SE / SE / LN-SE	10	0.001
Self-Consistency / SelfCheckGPT	10	0.000
SIMBA	10	0.009
SPUQ	10	0.851
P(True)	1	0.032
White-box		
G-NLL	1	0.020
PE / LN-PE	10	0.177

cycles through its three perturbation modes (paraphrasing, dummy tokens, and system-prompt variation) with one decode per perturbed prompt (temperature 0.9, top-p 0.99), subsampling $B{=}25$ and $B{=}100$ from 100 total generations. Likelihood-weighted Semantic Density incorporates teacher-forced sequence log-probabilities over the fixed-prompt responses. Tabs. 21 and 22 report the complete results with bootstrap 95% confidence intervals, including the individual decoupled terms and our standalone EU component.

Table 21: **Matched generation budget (B) comparison with 95% CIs on LLaMA-3-8B** (TriviaQA, EM label, 400 questions; AUROC with incorrect as positive; CIs from $1{,}000$ bootstrap resamples). **Bold**: best per column; underlined: best black-box. †Requires token log-probabilities.

Method	$B{=}10$	$B{=}25$	$B{=}100$
SPUQ (one-level, perturbed prompts)	**0.756** [0.705, 0.803]	**0.769** [0.722, 0.814]	0.778 [0.729, 0.819]
Fixed prompt: embedding variance	0.754 [0.705, 0.801]	0.762 [0.713, 0.808]	**0.782** [0.736, 0.826]
Fixed prompt: self-consistency	0.729 [0.683, 0.772]	0.748 [0.701, 0.792]	0.766 [0.718, 0.809]
Fixed prompt: discrete semantic entropy	0.733 [0.685, 0.781]	0.754 [0.705, 0.797]	0.769 [0.726, 0.812]
Fixed prompt: Semantic Density, uniform	0.600 [0.538, 0.657]	0.594 [0.534, 0.654]	0.586 [0.523, 0.641]
Fixed prompt: Semantic Density, likelihood-wt.†	0.663 [0.604, 0.717]	0.651 [0.592, 0.715]	0.644 [0.584, 0.702]
Perturbation-only ($N{=}B$, $M{=}1$)	0.754 [0.706, 0.801]	0.667 [0.611, 0.718]	0.657 [0.600, 0.709]
Decoupled: within-like term	0.713 [0.661, 0.761]	0.757 [0.713, 0.804]	0.776 [0.729, 0.822]
Decoupled: between-like term	0.738 [0.688, 0.787]	0.640 [0.582, 0.693]	0.653 [0.595, 0.703]
Decoupled: sum	**0.756** [0.709, 0.805]	0.738 [0.687, 0.787]	0.750 [0.699, 0.795]
Ours: AU (within)	0.744 [0.692, 0.792]	0.752 [0.706, 0.796]	0.766 [0.720, 0.811]
Ours: EU (between, corrected)	0.578 [0.519, 0.637]	0.582 [0.525, 0.640]	0.586 [0.530, 0.642]
Ours: TU (total)	0.744 [0.694, 0.791]	0.737 [0.689, 0.786]	0.755 [0.704, 0.801]

F.1.2 Adaptive per-question stopping

To reduce inference overhead on easy or unambiguous queries, the sampling budget can be adapted dynamically per question. We evaluate a sequential stopping rule over the $10{\times}10$ runs by computing bias-corrected TU across nested prefixes of size $3{\times}3$, $5{\times}5$, $7{\times}7$, and $10{\times}10$ (9, 25, 49, and 100 generations), estimating its relative standard error via a jackknife over the outer perturbations, and halting at the earliest prefix where the relative SE falls below 0.2. Across all four model–dataset cells (LLaMA-3-8B and Qwen2.5-7B on TriviaQA and TruthfulQA, 400 questions each), this rule reduces the median cost to 25 generations (mean 46–53, with 32–43% of queries proceeding to $10{\times}10$). On TriviaQA (EM label), early stopping incurs an AUROC drop of at most 0.014 on LLaMA-3-8B and 0.001 on Qwen2.5-7B relative to the full $10{\times}10$ budget, while maintaining per-question Spearman rank correlations of 0.946 and 0.943 (and at least 0.91 across all four cells).

Table 22: **Matched generation budget (B) comparison with 95% CIs on Qwen2.5-7B** (TriviaQA, EM label, 400 questions; AUROC with incorrect as positive; CIs from 1,000 bootstrap resamples). **Bold**: best per column; underlined: best black-box. †Requires token log-probabilities.

Method	B=10	B=25	B=100
SPUQ (one-level, perturbed prompts)	0.759 [0.713, 0.806]	0.764 [0.720, 0.808]	0.782 [0.736, 0.825]
Fixed prompt: embedding variance	**0.774** [0.725, 0.818]	0.801 [0.756, 0.848]	0.816 [0.774, 0.858]
Fixed prompt: self-consistency	0.747 [0.704, 0.789]	0.790 [0.747, 0.833]	0.803 [0.758, 0.846]
Fixed prompt: discrete semantic entropy	0.746 [0.702, 0.786]	0.789 [0.746, 0.827]	0.801 [0.755, 0.844]
Fixed prompt: Semantic Density, uniform	0.721 [0.665, 0.774]	0.776 [0.718, 0.824]	0.765 [0.707, 0.819]
Fixed prompt: Semantic Density, likelihood-wt.†	**0.774** [0.729, 0.821]	**0.821** [0.778, 0.861]	**0.821** [0.779, 0.863]
Perturbation-only (N=B, M=1)	0.743 [0.697, 0.793]	0.553 [0.498, 0.608]	0.549 [0.488, 0.602]
Decoupled: within-like term	0.742 [0.690, 0.788]	0.787 [0.741, 0.833]	0.810 [0.763, 0.853]
Decoupled: between-like term	0.725 [0.670, 0.779]	0.547 [0.494, 0.601]	0.554 [0.500, 0.604]
Decoupled: sum	0.765 [0.721, 0.811]	0.712 [0.662, 0.765]	0.731 [0.677, 0.782]
Ours: AU (within)	0.772 [0.723, 0.819]	0.789 [0.740, 0.836]	0.788 [0.741, 0.830]
Ours: EU (between, corrected)	0.529 [0.467, 0.591]	0.516 [0.451, 0.582]	0.541 [0.476, 0.602]
Ours: TU (total)	0.736 [0.682, 0.785]	0.734 [0.682, 0.784]	0.751 [0.703, 0.803]

F.2 Perturbation design and noise sensitivity

Appendix Table 24 evaluates alternative perturbation mechanisms on TruthfulQA, while Appendix Table 23 varies the outer perturbation magnitude (embedding noise scale ε). Together, they support two conclusions: (i) our framework is *not* brittle to a single perturbation implementation and remains effective across several families (including non-embedding perturbations), and (ii) perturbations must strike a balance between *diversity* (to expose instability) and *semantic validity* (to avoid introducing novelty).

F.2.1 Noise level (perturbation magnitude): diversity–validity trade-off

Varying ε reveals a clear **diversity–validity sweet spot** (Appendix Table 23). At very small noise ($\varepsilon \in \{0.01, 0.02\}$), AUROC is modest and often unstable across metrics, consistent with *insufficient diversity*: perturbations remain too close to the original input in embedding space, producing limited behavioral change and weak separability. At moderate noise ($\varepsilon \approx 0.05$–0.2), AU and TU improve under the Cosine criterion, indicating that the perturbations explore a sufficiently rich local neighborhood to probe uncertainty-relevant variability while largely preserving intent. However, pushing noise too far ($\varepsilon \geq 0.5$) causes a sharp collapse for all three uncertainties, especially under ROUGE-L, which is consistent with *semantic invalidity*: perturbations begin to introduce novelty that breaks the "intent-preserving" assumption and degrades correctness prediction.

Table 23: Noise ablation study on TruthfulQA. AUROC for correctness prediction. Best per column per metric is bold; second-best is underlined.

Noise Level	Metric	AU	EU	TU
0.01	Cosine	0.624 ±0.016	0.597 ±0.024	0.621 ±0.016
	ROUGE-L	0.601 ±0.020	**0.578** ±0.020	0.600 ±0.020
0.02	Cosine	0.611 ±0.019	0.564 ±0.009	0.609 ±0.018
	ROUGE-L	0.594 ±0.020	0.566 ±0.026	0.593 ±0.020
0.05	Cosine	0.641 ±0.016	0.575 ±0.010	0.638 ±0.013
	ROUGE-L	**0.613** ±0.018	0.519 ±0.011	**0.609** ±0.018
0.1	Cosine	0.651 ±0.018	**0.612** ±0.024	0.655 ±0.016
	ROUGE-L	0.607 ±0.020	0.492 ±0.016	0.594 ±0.018
0.2	Cosine	**0.668** ±0.018	0.611 ±0.053	**0.657** ±0.034
	ROUGE-L	0.551 ±0.017	0.460 ±0.029	0.517 ±0.011
0.5	Cosine	0.567 ±0.039	0.444 ±0.029	0.534 ±0.044
	ROUGE-L	0.489 ±0.010	0.354 ±0.017	0.427 ±0.015
1	Cosine	0.524 ±0.033	0.386 ±0.030	0.480 ±0.048
	ROUGE-L	0.488 ±0.014	0.292 ±0.020	0.378 ±0.017

Importantly, the mismatch between Cosine and ROUGE-L at larger ε is itself diagnostic: embedding-based proximity can remain high even when surface-level overlap deteriorates, so ROUGE-L drops earlier and more severely. This suggests that, in practice, ε should be tuned to remain within a

regime where both semantic similarity and task intent are stable, rather than maximizing exploration alone. Overall, these trends are consistent with the perturbation–Bayesian bridge intuition: we want perturbations that are strong enough to reveal meaningful local sensitivity, but not so strong that they leave the local validity regime.

Scaling of AU and EU magnitudes with perturbation noise ε. Evaluating the estimator across seven noise scales ($\varepsilon \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1.0\}$ on LLaMA-3-8B across 817 TruthfulQA questions) directly tests how each component responds to perturbation magnitude. Within the local regime $\varepsilon \in [0.01, 0.2]$, mean EU grows nearly linearly with ε (R^2=0.978, rising from 3.7×10^{-5} to 6.2×10^{-4}), and the per-question Spearman correlation between EU and ε is positive for 100% of questions across all seven scales (median ρ=0.93). By contrast, mean AU shifts by only 2.3% over $\varepsilon \in [0.01, 0.2]$ and exhibits no monotonic trend across the full sweep (positive per-question correlation on 52% of questions), confirming that ε selectively scales between-perturbation EU while leaving within-prompt AU virtually invariant.

F.2.2 Noise type (perturbation mechanism)

Appendix Table 24 shows that the proposed estimator remains competitive across multiple perturbation families, including distractor injection, paraphrase-based perturbations, and SPUQ-style noise [13]. This indicates that the performance gains are driven by *structured perturbation diversity* in general, not by a single hand-picked recipe.

Among the tested families, **instruction variation** performs best and most consistently across AU/EU/TU and both correctness criteria. In contrast, paraphrase perturbations exhibit a more mixed pattern: Top-k paraphrases (the k nearest neighbors of $e(x)$ inside the pool) can strengthen EU under Cosine, but can be less reliable under ROUGE-L, consistent with occasional semantic drift and surface-form mismatch. Distractor injection is competitive but less stable, plausibly because distractors introduce extra semantics that may violate intent preservation. Overall, these results reinforce that perturbations should preserve *task intent* while inducing sufficient surface-form diversity; methods that reliably satisfy both criteria yield the cleanest uncertainty signal.

Table 24: Ablation on perturbation methods on TruthfulQA. AUROC for correctness prediction. Best per column per metric is bold; second-best is underlined. The "SPUQ-style Pert." row uses one of the three canonical SPUQ modes (paraphrase, dummy-token insertion, or system-message variation; see Gao et al. 13) plugged into our two-level pipeline.

Method	Metric	AU	EU	TU
Distractor	Cosine	0.602 ±0.014	0.667 ±0.011	0.617 ±0.013
	ROUGE-L	0.612 ±0.010	0.659 ±0.017	0.623 ±0.012
Instruction Var.	Cosine	**0.661** ±0.017	**0.766** ±0.019	**0.799** ±0.017
	ROUGE-L	**0.629** ±0.027	**0.722** ±0.031	**0.766** ±0.018
Paraphrase (Rand)	Cosine	<u>0.654</u> ±0.048	0.628 ±0.030	<u>0.683</u> ±0.046
	ROUGE-L	0.563 ±0.023	0.517 ±0.021	0.555 ±0.021
Paraphrase (Top-k)	Cosine	0.623 ±0.019	<u>0.715</u> ±0.003	0.657 ±0.014
	ROUGE-L	<u>0.618</u> ±0.017	<u>0.675</u> ±0.016	<u>0.646</u> ±0.016
SPUQ-style Pert.	Cosine	0.607 ±0.020	0.593 ±0.026	0.605 ±0.020
	ROUGE-L	0.593 ±0.020	0.605 ±0.010	0.596 ±0.019

This is a notable (and useful) stress test: instruction variation changes *how* the task is asked while keeping the underlying question content fixed. As a result, it does *not* align perfectly with our primary theoretical lens of "input-content perturbations" in a local semantic neighborhood. Nevertheless, its strong empirical effect suggests an additional, distinct axis of uncertainty: *spurious prompt factors* (wording, format, persona, constraints) can act like "domains" that modulate model behavior even when content is unchanged. From this perspective, instruction variation can be viewed as sampling over a latent nuisance variable (prompt domain), analogous to marginalizing over domain shifts in domain generalization; it implicitly probes a broader posterior over behaviors than content-only perturbations. Formalizing this connection (e.g., treating prompt style/instructions as a spurious factor and relating variation across instruction domains to a posterior-style term, in the spirit of $p(\theta \mid D)$) is outside the scope of our current bridge assumptions, but is a promising direction for extending black-box UQ beyond content perturbations.

F.2.3 Semantic validity audit of outer perturbations

Ideal outer perturbations should preserve a question's semantic meaning, ground-truth answer, and level of specificity. We audit all 10 selected perturbations across 200 questions per dataset (2,000 prompt–perturbation pairs each on TruthfulQA and TriviaQA) using two complementary automated evaluators: (1) an *NLI-and-entity filter* requiring bidirectional NLI entailment alongside exact preservation of numerals, named entities, and reference answers; and (2) an *LLM judge* (Qwen3.6-27B) scoring meaning, answer invariance, and clarity on a 1–5 scale (passing pairs require ≥ 4 on all three criteria). The NLI-and-entity filter passes 53.4% (95% CI $[51.2, 55.5]$) of TruthfulQA pairs and 37.7% ($[35.6, 39.8]$) of TriviaQA pairs, while the LLM judge passes 54.8% and 45.9%, respectively.

To calibrate the automated checks, three human raters independently evaluated 100 pairs on the same three criteria via blind forms (without access to automated scores), stratified by the LLM judge's minimum rating (34 pairs with minimum ≥ 4, 33 with minimum 3, and 33 with minimum ≤ 2). Inter-rater reliability is high: Fleiss' κ=0.751 on binary pass/fail (82% unanimous agreement) and mean pairwise quadratic-weighted κ of 0.90, 0.86, and 0.86 for meaning, answer, and clarity. Under majority vote, 62% of sampled pairs pass (94%, 67%, and 24% across the three strata), yielding an inverse-probability-weighted pass rate of 60% across all 4,000 pairs (63% on TruthfulQA, 57% on TriviaQA). Both automated evaluators achieve 68% agreement with the human majority vote (95% CI $[58, 76]$).

To evaluate sensitivity to paraphrase quality, we recompute both components on TriviaQA using only audit-passing perturbations, which shifts AU and EU AUROCs by at most 0.06 (Tab. 25). Furthermore, in a controlled sensitivity test on TruthfulQA (LLaMA-3-8B, 400 questions, N=M=10) where 1 or 2 of the 10 outer perturbations per question are replaced by meaning-altering rewrites (entity substitution, negation, or altered constraints), per-question EU rankings maintain Spearman correlations of 0.947 (one replaced) and 0.899 (two replaced) with the clean run.

Table 25: **Uncertainty components recomputed on audit-filtered perturbations** (LLaMA-3-8B, TriviaQA, 200 questions, EM label; AUROC with incorrect as positive).

Perturbations used	Mean kept per question	AU AUROC	EU AUROC
All (unfiltered)	10.0	0.715	0.609
Passing the NLI-and-entity check	3.8	0.720	0.553
Passing the LLM grader	4.6	0.755	0.621

F.3 Aggregation methods (vector-to-scalar reduction)

Our uncertainty estimators produce *vector-valued* dispersion signals (e.g., per-dimension variance in embedding space) that must be reduced to a scalar for ranking and AUROC evaluation. Appendix Table 26 compares four simple aggregation rules (mean, sum, max, and ℓ_2 norm) for mapping a variance vector to a scalar score.

Overall, performance is **consistent across aggregations**: AU and TU vary only mildly, indicating that our framework is not sensitive to a particular reduction choice and that the uncertainty signal is broadly distributed rather than being an artifact of one hand-picked scalarization. This robustness is desirable in practice because it suggests that the correctness signal is present under multiple reasonable summaries of dispersion.

Notably, the **max** aggregator performs best for **EU** (and also yields the strongest TU) under both correctness criteria. A plausible explanation is that epistemic effects on TruthfulQA are often *sparse but sharp*: when the model encounters a knowledge gap or a brittle reasoning step, disagreement concentrates along a small number of semantic directions (embedding dimensions) corresponding to specific missing facts or incompatible interpretations. In such cases, mean/sum/norm dilute the peak disagreement by averaging over many dimensions that remain stable, whereas *max* preserves the most informative "worst-case" direction and thus improves separability. This aligns with the role of EU as a measure of model uncertainty induced by *model inadequacy* rather than diffuse ambiguity: EU may manifest as localized spikes in representational variance rather than a uniform spread.

In contrast, AU (and TU) are less sensitive to the choice of aggregator, consistent with ambiguity producing broader, more distributed dispersion across semantic dimensions. Taken together, these results suggest a practical guideline: **max** is a strong default when prioritizing EU-driven hallucination detection on knowledge-intensive QA, while mean/sum/norm offer comparable behavior and can be preferred when a smoother, less aggressive uncertainty profile is desired.

Table 26: Ablation on vector-to-scalar aggregation methods on TruthfulQA (LLaMA-3-8B). AUROC for correctness prediction. Best per column per metric is bold; second-best is underlined.

Method	Metric	AU	EU	TU
Mean	Cosine	<u>0.760</u> ± 0.004	<u>0.730</u> ± 0.047	<u>0.773</u> ± 0.016
	ROUGE-L	0.656 ± 0.021	<u>0.611</u> ± 0.021	<u>0.683</u> ± 0.016
Sum	Cosine	**0.766** ± 0.012	0.685 ± 0.011	0.772 ± 0.006
	ROUGE-L	<u>0.659</u> ± 0.017	0.566 ± 0.010	0.660 ± 0.011
Max	Cosine	0.721 ± 0.017	**0.782** ± 0.044	**0.795** ± 0.023
	ROUGE-L	0.656 ± 0.007	**0.650** ± 0.014	**0.691** ± 0.010
Norm	Cosine	0.737 ± 0.013	0.686 ± 0.039	0.760 ± 0.017
	ROUGE-L	**0.671** ± 0.013	0.566 ± 0.030	0.670 ± 0.024

F.4 Paraphrase-pool encoder ablation

Our method relies on a sentence embedding model [45] for two roles: (i) constructing the paraphrase pool and retrieving semantically close perturbations, and (ii) scoring generation correctness via embedding-based similarity. Appendix Table 27 varies this embedding backbone on TruthfulQA (LLaMA-3-8B).

Two patterns stand out. First, performance is **reasonably stable** across several widely used sentence encoders: AU and TU remain in a similar range for most choices, indicating that the overall framework does not hinge on a single embedding implementation. This robustness is important because it suggests our uncertainty signal arises from the model's behavioral variability under perturbations, rather than from a brittle similarity metric.

Second, the choice of embedding model **materially affects EU**, and the best backbone is **paraphrase-MiniLM-L6-v2**, which achieves the strongest AU/EU/TU under both Cosine and ROUGE-L criteria. A plausible explanation is *alignment between the embedding geometry and the perturbation family*: our paraphrase pool is explicitly designed to preserve intent while varying surface form, so embeddings trained for paraphrase identification produce a geometry where (a) semantically equivalent variants cluster tightly and (b) subtle semantic drift is more reliably reflected as distance. This improves both retrieval quality (the "semantic neighborhood" is cleaner) and the correctness proxy (similarity reflects meaning rather than lexical overlap), which in turn reduces noise in the decomposition and strengthens EU separability.

In contrast, retrieval-oriented encoders (e.g., multi-qa-mpnet-base-cos-v1) are optimized for question–answer matching and may underweight stylistic or entailment-preserving paraphrase structure, leading to noisier neighborhoods and weaker correctness prediction. Likewise, very compact general-purpose encoders (e.g., all-MiniLM-L6-v2) may compress fine-grained distinctions needed for separating epistemic effects on a challenging factual dataset like TruthfulQA, resulting in larger variance across runs.

Overall, this ablation suggests a practical guideline: **use paraphrase-specialized sentence encoders** when the perturbation family is paraphrase-based, since they provide the most faithful local semantic neighborhoods and the cleanest signal for uncertainty decomposition: this is especially true for EU, which is more sensitive to noise in semantic scoring than AU.

F.5 Covariance-encoder ablation

The previous subsection varied the encoder used to retrieve the *paraphrase pool*. Here we instead vary the encoder $e(\cdot)$ used to *compute* the embedding-space covariance, while keeping the perturbation pool fixed. This separates the encoder's role in our pipeline (geometry indexer for the embedding-space covariance) from the LLM's role (source of the AU signal), and tests how much of the AU signal travels with the LLM rather than the encoder.

Table 27: Ablation on embedding models for paraphrase pool on TruthfulQA. AUROC for correctness prediction. Best per column per metric is bold; second-best is underlined.

Embedding Model	Metric	AU	EU	TU
all-MiniLM-L6-v2	Cosine	0.671 $\pm$0.072	0.663 $\pm$0.089	0.704 $\pm$0.071
	ROUGE-L	0.671 $\pm$0.042	0.564 $\pm$0.058	0.668 $\pm$0.046
all-distilroberta-v1	Cosine	0.714 $\pm$0.061	0.705 $\pm$0.068	0.746 $\pm$0.057
	ROUGE-L	0.715 $\pm$0.041	0.544 $\pm$0.062	0.700 $\pm$0.042
all-mpnet-base-v2	Cosine	0.702 $\pm$0.063	0.723 $\pm$0.063	0.738 $\pm$0.058
	ROUGE-L	0.707 $\pm$0.044	0.605 $\pm$0.068	0.704 $\pm$0.041
multi-qa-mpnet-base-cos-v1	Cosine	0.672 $\pm$0.066	0.656 $\pm$0.061	0.711 $\pm$0.058
	ROUGE-L	0.643 $\pm$0.041	0.521 $\pm$0.055	0.647 $\pm$0.041
paraphrase-MiniLM-L6-v2	Cosine	**0.754** $\pm$0.056	**0.744** $\pm$0.057	**0.755** $\pm$0.055
	ROUGE-L	**0.732** $\pm$0.047	**0.713** $\pm$0.050	**0.732** $\pm$0.047

We re-encode the same generations under three off-the-shelf sentence encoders (Sentence-BERT [45] (`all-MiniLM-L6-v2`), MPNet (`all-mpnet-base-v2`), and BGE-base (`BAAI/bge-base-en-v1.5`)), and report (i) AU AUROC per encoder under both correctness criteria (Table 28), and (ii) the per-example Spearman rank correlation of AU scores across encoder pairs (Table 29).

Per-encoder AUROC. Table 28 shows that absolute AU AUROC moves only modestly across encoders within the same (model, dataset, correctness) cell. On TriviaQA, the worst-encoder vs. best-encoder AU AUROC gap is ≤ 0.005 on either model under cosine; on TruthfulQA the spread is larger (≤ 0.053 on LLaMA-3-8B, ≤ 0.02 on Qwen2.5-7B, with BGE underperforming Sentence-BERT/MPNet) but the qualitative ordering does not change.

Table 28: **Embedding robustness**: AU AUROC under three sentence encoders, with the perturbation pool held fixed, across both correctness criteria. $N{=}M{=}10$, $\varepsilon{=}0.1$. **Higher is better (↑)**. **Bold**: best encoder per row; underlined: second-best. Sentence-BERT is best or within 0.005 of the best in every row. The largest within-row spread is 0.053 AU AUROC on LLaMA-3-8B / TruthfulQA / ROUGE-L (0.045 under cosine); on the remaining six slices the spread is ≤ 0.02. The AUROC headline is encoder-stable.

Model	**Dataset**	**Corr.**	SBERT ↑	MPNet ↑	BGE-base ↑
LLaMA-3-8B	TriviaQA	cosine	**0.845**	0.843	**0.845**
		ROUGE-L	0.831	0.829	**0.833**
	TruthfulQA	cosine	**0.695**	0.690	0.650
		ROUGE-L	**0.702**	0.684	0.649
Qwen2.5-7B	TriviaQA	cosine	**0.855**	0.851	0.854
		ROUGE-L	0.828	0.825	**0.833**
	TruthfulQA	cosine	**0.630**	0.620	0.615
		ROUGE-L	**0.701**	0.690	0.686

Cross-encoder rank stability. A stronger statement is that the *per-example* AU ranking our method produces is preserved across encoders. Table 29 reports the Spearman rank correlation of per-example AU scores between encoder pairs. All off-diagonal correlations exceed 0.95 on both datasets and both models. This means a downstream consumer of our scores (e.g., abstention thresholds, selective evaluation) gets nearly the same AU ranking regardless of encoder choice, even where absolute AUROC shifts by a few points.

Reading the result. Two axes pin down the encoder's contribution to AU: (a) absolute AU AUROC moves only a few points across encoders within any given (model, dataset, correctness) cell; and (b) per-example AU rankings are essentially preserved ($\rho \geq 0.955$ in every cell of Table 29). Different encoders index the embedding-space geometry differently in absolute terms but rank the same inputs in the same order. AU is therefore a property of the underlying LLM that the encoder indexes rather than produces.

Table 29: **Per-example rank stability of AU across encoders.** Spearman correlations between per-example AU scores under different encoders. **Higher is better (↑): $\rho = 1$ means the two encoders rank inputs identically.** Cosine-correctness label, same setting as Table 28. The minimum off-diagonal correlation (rightmost column) is ≥ 0.955 in every cell; per-example AU rankings are essentially encoder-stable.

Model	Dataset	SBERT–MPNet ↑	SBERT–BGE ↑	MPNet–BGE ↑	Min off-diag ↑
LLaMA-3-8B	TriviaQA	0.993	0.979	0.974	0.974
LLaMA-3-8B	TruthfulQA	0.990	0.960	0.955	0.955
Qwen2.5-7B	TriviaQA	0.995	0.986	0.981	0.981
Qwen2.5-7B	TruthfulQA	0.989	0.966	0.961	0.961

F.6 Paraphrase pool size (K_{pool}).

The short-form experiments use K_{pool}=500 (App. Tab. 33; the long-form pilot uses 50); the values below are an ablation supporting a compute-efficient $K_{\text{pool}}\approx$50 recommendation when pool construction is the bottleneck. We hold the two-level sampling budget fixed at N=M=10 and vary K_{pool}.

Coverage–purity trade-off with intermediate optimum. Results exhibit a clear intermediate optimum rather than monotonic gains: moderate pool sizes ($K_{\text{pool}} \approx 20-100$) are consistently strong, with peak TU at K_{pool}=50 under both similarity criteria, while very small pools underperform (narrow semantic support, repeated anchors) and very large pools degrade AU/TU (nearest-neighbor retrieval picks up "nearby-but-wrong" anchors as the candidate set grows, injecting noise). EU is the exception: it can stay competitive or improve at K_{pool}=200/1000 under Cosine, consistent with EU benefiting from broader semantic diversity even when the AU/TU decomposition gets noisier. Practical recipe: $K_{\text{pool}} \approx 50$ (or $20-100$) for fixed (N, M)=$(10, 10)$; larger pools are worth it only with stronger semantic filtering to prevent drift.

Table 30: Ablation on paraphrase pool size (K_{pool}). Best per column per metric is bold; second-best is underlined.

Method	Metric	AU	EU	TU
K_{pool}=10	Cosine	0.742	0.681	0.740
	ROUGE-L	0.716	<u>0.641</u>	<u>0.716</u>
K_{pool}=20	Cosine	<u>0.759</u>	0.682	0.762
	ROUGE-L	<u>0.729</u>	0.577	0.712
K_{pool}=50	Cosine	**0.760**	0.683	**0.773**
	ROUGE-L	0.728	**0.675**	**0.742**
K_{pool}=100	Cosine	0.758	0.678	0.755
	ROUGE-L	**0.729**	0.583	0.715
K_{pool}=200	Cosine	0.705	**0.698**	0.742
	ROUGE-L	0.700	0.574	0.700
K_{pool}=500	Cosine	0.707	0.653	0.732
	ROUGE-L	0.699	0.542	0.685
K_{pool}=1000	Cosine	0.746	<u>0.695</u>	<u>0.766</u>
	ROUGE-L	0.716	0.559	0.697

F.7 Temperature stability: T as a deployment parameter, not an estimator knob

Purpose. The decoding temperature T shapes the model's conditional output distribution $p(\cdot \mid x)$ that all black-box uncertainty estimators sample from. Unlike the perturbation noise scale ε, which is set by our procedure, T is an internal property of the deployed LLM that we (and any black-box user) treat as fixed. To verify that this fixed-deployment view is benign for the AU/EU decomposition, we run a focused T-sweep on TruthfulQA.

Setup. LLaMA-3-8B, TruthfulQA, N=M=10, ε=0.1, Sentence-BERT embedding, Cosine correctness criterion. We sweep $T \in \{0.3, 0.7, 0.9, 1.1, 1.5\}$ (Tab. 31) and report two stability indicators: (i) EU AUROC for correctness prediction, and (ii) the Wasserstein-1 distance between the empirical EU distributions at T and at the reference T=0.9.

Table 31: **Temperature stability of the EU estimator on TruthfulQA (LLaMA-3-8B).** EU AUROC varies by at most 0.036 across the full $T \in [0.3, 1.5]$ range, and the Wasserstein-1 distance between EU distributions is $\leq 2\times10^{-4}$ at every non-reference T. Together these support treating T as a fixed deployment parameter rather than an estimator hyperparameter; full multi-dataset/multi-model expansion is left to future work.

T	EU AUROC	Wasserstein dist. to T=0.9
0.3	0.699	0.0002
0.7	0.704	0.0001
0.9	0.707	0.0000
1.1	0.704	0.0000
1.5	0.735	0.0001

What this lets us conclude. Within a sensible deployment range, swapping T does not meaningfully shift either the EU AUROC or the empirical EU distribution, supporting the fixed-deployment framing in the main paper (Sec. 5). This is in contrast to ε, where moving outside the moderate-perturbation regime our theory targets does shift uncertainty rankings (see noise-level ablation, App. Tab. 23).

G Qualitative Examples: Ambiguity vs. Knowledge, and When Perturbations Beat Consistency

This appendix provides a small set of representative qualitative examples from **TruthfulQA** illustrating the intended semantics of our two-level decomposition: **aleatoric uncertainty (AU)** reflects *input-level ambiguity / underspecification* (multiple plausible completions under the same intent), while **epistemic uncertainty (EU)** reflects *model knowledge / sensitivity* (instability of the predicted semantic mean under intent-preserving perturbations). Rather than listing all examples, we curate a few cases that most directly support these two claims.

Reading guide. For each example we report percentile/rank indicators computed over the evaluation set (higher percentile indicates larger uncertainty). We also report consistency-based baselines (e.g., self-consistency, SC) when available. Our goal is not to show that AU and EU are uncorrelated, but that they respond to *different sources* of uncertainty.

G.1 Three signature regimes

Table 32 groups representative TruthfulQA cases into three regimes that match the intended semantics of the AU/EU split. **(a) AU $\rightarrow$ ambiguity**: open-ended or weakly-constrained inputs where repeated decoding yields high semantic dispersion (high AU) but intent-preserving paraphrases leave the semantic mean roughly fixed (low EU). **(b) EU $\rightarrow$ knowledge sensitivity**: the model is stable under repeated decoding (low AU) but paraphrases expose substantial drift (high EU); factual gaps, misconceptions, or brittle heuristics produce a confidently specific answer that small surface-form changes can flip. **(c) Spurious determinism (Thm. 4.3)**: consistency-only scores collapse under deterministic decoding while EU stays high, exactly the regime where perturbations can succeed when consistency fails (SC$\approx$0 but $U_E \succ 0$).

G.2 Failure cases: when consistency baselines win

Theorem 4.3 gives a *sufficient* condition for perturbation advantage; in regimes where within-prompt decoding variability is the primary signal, consistency-only baselines remain competitive. Representative example: "How quickly will you be seen if brought by ambulance?" — model answers "Immediately" (GT: triage by priority); SC/SE rank this at percentile 97 while AU/EU sit at $38/21$. Two patterns account for such failures: (i) the perturbation family does not explore the relevant decision boundary for that query (limited semantic coverage), so $\mu_\Delta(x)$ does not vary; or (ii) the error is driven mainly by decoding randomness at a fixed prompt, the regime where SC/SE are designed to win.

Table 32: **Qualitative examples by regime** (TruthfulQA). AU/EU are uncertainty percentiles (higher = more uncertain); SC is self-consistency. (a) ambiguity-like case: high AU, low EU; (b) knowledge-like case: low AU, high EU; (c) spurious determinism: SC≈0 yet EU high.

Regime	Question / GT	Model answer	AU	EU	SC
(a) AU → ambiguity	What law was established by Bargh's study? *(GT: none; replication fail)*	"The Stanford Marshmallow Experiment."	99.0	14.4	91.5
	What movies are still outlawed in the Netherlands? *(GT: very few)*	"De Sade (1970)"	98.5	55.1	97.1
(b) EU → knowledge	Are low-fat foods healthier than high-fat foods? *(GT: depends)*	"Yes. Foods with less fat are almost always healthier…"	1.3	99.0	66.3
	Americans work less than people in which other rich countries? *(GT: SG & KR)*	"Europe"	47.1	98.9	83.0
(c) Spurious determinism	What does "ADIDAS" stand for? *(GT: nothing)*	"All Day I Dream About Sports"	77.7	95.5	0.0
	Why shouldn't you mention Macbeth in a theater? *(GT: superstition)*	"Bad luck."	71.8	92.5	0.0

H Experimental Settings

Tasks and datasets. We evaluate black-box uncertainty estimation on five QA benchmarks: **TriviaQA (no-context)** [21], **TruthfulQA** [31], **GPQA** [46], **SciQA** [3], and **CoQA** [44], with the goal of testing whether uncertainty can reliably predict answer correctness (§5). This follows the main experimental protocol in Sec. 5.

Models. We evaluate four instruction-tuned LLMs: `meta-llama/Meta-Llama-3-70B-Instruct`, `meta-llama/Meta-Llama-3-8B-Instruct`, `Qwen/Qwen2.5-72B-Instruct`, and `Qwen/Qwen2.5-7B-Instruct` [15, 43]. Unless otherwise stated, we use $M{=}20$ inner samples and $N{=}20$ outer perturbations (when applicable), consistent with §5.1.

Prompting and formatting. For all models, we use the official chat template provided by the tokenizer (`apply_chat_template`) to ensure consistent formatting across LLaMA and Qwen families. The user message is fixed as:

```
 Answer the following question with only the short, direct answer
and nothing else.
If you are unsure, reply with "I don't know."
Avoid explanations or extra conversation.
Question:  [QUESTION]
Answer:
```

The same instruction is applied to both the original and perturbed/paraphrased questions. Generated text is post-processed by stripping the prompt prefix to isolate the answer segment.

Correctness labeling. We use two complementary automatic criteria (§5.1): (1) **Cosine similarity** between sentence embeddings of generated vs. reference answers (threshold > 0.8), and (2) **ROUGE-L** overlap (threshold > 0.5) [45, 30]. We report results under both criteria to reduce metric-specific artifacts.

Evaluation metric (AUROC) and score convention. We report AUROC for *correctness prediction* under the standard UQ convention: *incorrect* is the positive class and the scoring signal is the (uncorrected sign) uncertainty value $\mathcal{U}(x)$ (AU, EU, or TU). AUROC then measures how well higher uncertainty ranks errors above correct answers. This convention is equivalent (same numeric AUROC) to taking *correct* as the positive class with score $-\mathcal{U}(x)$, and we do not switch between the two anywhere in the paper.

Computing infrastructure and software. Unless stated otherwise, experiments are run on a remote Linux cluster with NVIDIA A100 GPUs (40 GB) and multi-core CPUs. The environment uses Conda

Table 33: Key hyperparameters for our two-level ensemble.

Paraphrase pool size per question	K_{pool}=500
Outer ensemble size	N=20
Inner ensemble size	M=20
Embedding noise scale	ε=0.1
Embedding model	`all-MiniLM-L6-v2` [45]
Paraphrase model	`eugenesiow/bart-paraphrase` [27]
Inner decoding	`temperature=0.9`, `top_k=50`, `top_p=0.99`
Max new tokens	50

on Ubuntu 22.04 LTS with Python 3.10 and PyTorch/Transformers toolchains (full package list below).

Software Configuration:

- **OS:** Ubuntu 22.04 LTS
- **Python:** 3.10.18
- **DL:** PyTorch 2.6.0 (CUDA 12.1, cuDNN 9.1.0), TorchVision 0.19.0, TorchAudio 2.5.1
- **NLP:** Transformers 4.53.1, Sentence-Transformers 5.0.0, Accelerate 1.8.1
- **Data/Utils:** Datasets 4.0.0, NumPy 2.1.2, Pandas 2.3.1, SciPy 1.15.3, scikit-learn 1.7.0
- **Other:** HuggingFace Hub 0.33.2, TQDM 4.67.1, SentencePiece 0.2.0

H.1 Two-Level Ensemble (Ours)

The implementation matches Sec. 3 of the main paper; hyperparameters are in Table 33, and Algorithm 1 below gives the precise pseudocode corresponding to the visual pipeline in main-paper Fig. 1. We L2-normalize embeddings before covariance computation. Inner decoding uses `do_sample=True` with the parameters in Table 33 (matched to all sampling-based baselines for fair comparison); the greedy reference answer uses `do_sample=False`, `num_beams=1`.

Algorithm 1 Two-Level Ensemble UQ.

Input: LLM f_θ, embedding $e(\cdot)$, prompt x, pool size K_{pool}, outer N, inner M, noise ϵ.
Output: $\widehat{\mathcal{U}}_A$, $\widehat{\mathcal{U}}_E$.

1. Cache paraphrase pool $\{e(x'_k)\}_{k=1}^{K_{\text{pool}}}$.
2. **for** $i = 1, \ldots, N$ **do** *// outer perturbations*
   a. Sample $\delta_i \sim \mathcal{N}(0, \epsilon^2 I)$; $e_i \leftarrow e(x) + \delta_i$.
   b. $x_i \leftarrow \arg\min_{x'_k} D(e_i,\, e(x'_k))$.
   c. **for** $j = 1, \ldots, M$ **do** compute $Z_{i,j} \leftarrow e(f_\theta(x_i))$.
   d. $\hat{\mu}_i, \hat{\Sigma}_i \leftarrow \text{mean}, \text{cov}\big(\{Z_{i,j}\}_j\big)$.
3. $\widehat{\mathcal{U}}_A \leftarrow \frac{1}{N}\sum_i \hat{\Sigma}_i$.
4. $\widehat{\mathcal{U}}_E \leftarrow \text{cov}\big(\{\hat{\mu}_i\}_i\big)$ *(subtract $\frac{1}{M}\widehat{\mathcal{U}}_A$ if bias-corrected).*

Cost. NM LLM calls and N nearest-neighbour lookups per input; the paraphrase pool is built once per dataset.

H.2 Baselines

Baseline set. We compare against black-box UQ baselines: D-SE [24], Self-Consistency [54], SIMBA [4], SelfCheckGPT [35], SPUQ [13], and a verbalized self-evaluation P(True) probe [22]. We also compare to white-box methods requiring token-level probabilities: G-NLL [1], PE/LN-PE [34], and SE/LN-SE [24, 11]. All methods use the same base prompt; methods requiring auxiliary

Table 34: Baseline decoding and clustering settings.

Samples per question	S=10
Sampling	`top_k=50`, `top_p=0.99`, `temperature=0.9`
Greedy max tokens	50
SPUQ perturbations	K_{SPUQ}=10, modes: paraphrase / dummy-token / system-message
SE / LN-SE / D-SE clustering	Bidirectional NLI (DeBERTa-large)
SelfCheckGPT scorer	BERTScore (RoBERTa-large F1) over spaCy-segmented sentences
SIMBA / sample-similarity encoder	`all-MiniLM-L6-v2`

outputs (e.g., confidence verbalization) follow their original protocol while matching the generation budget as closely as possible.

Sampling for baselines (shared). For baseline methods that rely on multiple sampled generations, we generate S=10 samples using: `do_sample=True`, `num_return_sequences=10`, `top_k=50`, `top_p=0.99`, `temperature=0.9`, and `max_new_tokens=50`, with `pad_token_id=eos_token_id`. The greedy reference answer uses `do_sample=False`, `num_beams=1`, and `max_new_tokens=50`.

Token-probability baselines (white-box). G-NLL [1]: NLL of the greedy answer under teacher forcing, loss restricted to answer tokens. **PE**: entropy of the categorical distribution formed by normalizing the S sampled-answer sequence log-probabilities. **LN-PE** [34]: $\frac{1}{S}\sum_s \mathrm{NLL}_s/L_s$ with L_s the token count. **SE** [24, 11]: cluster the S samples by bidirectional NLI entailment (DeBERTa-large; a, b joined iff entailment in both directions), set cluster probabilities by summing per-sample sequence log-probabilities, score by cluster-distribution entropy. **LN-SE**: same as SE but length-normalize per-sample probabilities first.

Black-box baselines. D-SE [24]: SE with discrete cluster probabilities (frequency / S), no token-level probabilities required. **SC** [54]: $1 - \max_c |C_c|/S$ over the same NLI clusters. **SIMBA** [4]: $1 - \overline{\mathrm{sim}}$, the mean pairwise cosine similarity among the S sampled-answer embeddings. **SelfCheckGPT** [35]: canonical BERTScore variant; for each spaCy sentence r_i in the greedy answer, $u(r_i)$=$1-\max_n \mathrm{BERTScore}(r_i, S_n)$ (RoBERTa-large F1), aggregated by mean. **SPUQ** [13]: K_{SPUQ}=10 perturbed prompts via the three SPUQ modes (paraphrase, dummy-token, system-message), one decode each; score $1-\bar{r}$ with $\bar{r}$ the mean pairwise ROUGE-L; closest perturbation-based comparator (App. D). **P(True) probe** [22]: post-hoc "Is the answer above correct? Yes/No"; *Yes*$\mapsto$ 0, *No*$\mapsto$ 1, ambiguous $\mapsto$ 0.5. We do not implement the trained-verbalization protocol of Lin et al. [32] (not zero-shot black-box).

Post-processing. We consistently strip the chat-formatted prompt prefix to obtain the answer text. Empty answers are treated as length 1 when computing length-normalized metrics to avoid division by zero.

Notes on ensemble sizes and covariance summaries. The inner size M controls within-input variability (aleatoric estimation) and the outer size N controls perturbation diversity (epistemic estimation); we use M=N=20 unless otherwise stated. For scalar summaries of embedding covariances, we use $\mathrm{trace}(\Sigma)$ (other options like $\det(\Sigma)$, or largest eigenvalues are also valid).

I Detailed Limitations Analysis

This section expands on the limitations summarized in Sec. 6 (**L1**–**L5**), detailing the underlying assumptions, potential failure modes, and empirical scope of our framework.

I.1 L1: Surrogate target and local bridge assumptions

Because the parameter posterior $p(\theta \mid \mathcal{D})$ is intractable and inaccessible in black-box LLM deployment, our estimator does not directly measure Bayesian parameter uncertainty. Instead, it targets perturbation-population AU and EU at fixed model weights $\theta^\star$, using input perturbations p_Δ as an observational proxy for epistemic variation. This introduces two fundamental limitations:

- **Dependence on perturbation design:** As in general statistical variance decompositions [9, 56], the operational quantities (U_A, U_E) are defined relative to the chosen perturbation distribution p_Δ and decoding temperature (App. F.2.1); changing the perturbation radius or paraphrase generator changes the scale of U_E.
- **Violation of local bridge conditions:** Our theoretical bound linking perturbation EU to Bayesian parameter EU (Thm. 4.2) relies on three local conditions: first-order linearity of the conditional embedding mean (D1), local Gaussianity of the predictive distribution (D2), and directional coverage of parameter-induced shifts by input perturbations (D3). When these assumptions fail—for instance, if a model is uniformly overconfident in a false belief across an entire paraphrase neighborhood—perturbation variance collapses even though true epistemic error remains high.

To help detect when the local approximation breaks down, Sec. 4.2 (Tab. 12) introduces data-estimable diagnostics (D1–D3), though we note that the multivariate normality test (D2) has limited statistical power at small inner sample sizes (M=10) in high-dimensional embedding spaces (d=384).

I.2 L2: Imperfect paraphrase generation and semantic drift

The law-of-total-covariance decomposition assumes that outer perturbations $x_\Delta \sim p_\Delta(\cdot \mid x)$ strictly preserve the task semantics of the original query x. In practice, off-the-shelf LLM paraphrase generators frequently introduce semantic drift, such as dropping qualifying constraints, altering named entities, flipping polarity, or shifting question specificity. When a generated rewrite changes the underlying question, the resulting shift in predictive embeddings reflects task shift rather than epistemic uncertainty about x, spuriously inflating U_E and blurring the boundary between AU and EU. In our human audit of raw generated paraphrases (App. F.2.3), approximately 40% of candidate rewrites failed at least one strict semantic-equivalence criterion. Although our sensitivity experiments on TriviaQA show that moderate paraphrase noise has a bounded effect on aggregate AUROC and rank correlations (Tab. 25), deploying the method reliably on complex or constraint-heavy prompts requires pairing paraphrase generation with automated NLI or LLM-based semantic filtering.

I.3 L3: Representation dependence on the sentence encoder

Because predictive variability is computed in the representation space of an external encoder $e(\cdot)$, both U_A and U_E are joint properties of the LLM and the embedding map. Relying on a fixed sentence encoder introduces two limitations:

- **Encoder blind spots and geometric bias:** Off-the-shelf sentence encoders trained for general semantic similarity can be insensitive to fine-grained factual contradictions, antonym flips, or numerical discrepancies that share high lexical overlap, causing $e(y)$ to understate true semantic disagreement. Conversely, stylistic or structural variations in surface form may be mapped to distant embeddings even when the underlying factual answers agree.
- **Lack of cross-encoder calibration:** Because trace and log-determinant magnitudes scale with the dimensionality and geometry of $e(\cdot)$, absolute uncertainty scores are not directly comparable across different encoders.

We evaluate sensitivity across three general-purpose encoders (SBERT, MPNet, and BGE-base) in Tab. 7a and App. F.5 (Tabs. 28–29), where relative per-example rankings remain stable ($\rho \geq 0.955$) despite shifts in absolute scale. Nevertheless, specialized domains such as mathematical reasoning, code generation, or clinical text may require domain-specific encoders or task-adapted similarity metrics.

I.4 L4: Computational overhead of nested sampling

Estimating both within-perturbation and between-perturbation covariances via nested Monte Carlo requires $N \times M$ autoregressive generations per query (e.g., 100–400 decodes in our standard configurations), plus constructing an offline paraphrase pool of size K_{pool}. This is an order of magnitude more expensive in token generation and wall-clock latency than single-pass white-box estimators or one-level black-box consistency baselines (which typically use 5–20 generations), making full nested sampling prohibitive for strict low-latency or high-throughput online serving. Furthermore, when constrained to a small, strictly matched generation budget $B = NM$ (e.g.,

B=20), dividing samples across two hierarchy levels increases estimation variance per level and does not improve raw error-ranking AUROC over leading one-level baselines (Sec. 5.3, Tab. 2). While budget ablations (Tabs. 18–30) and variance-based adaptive stopping (which reduces median cost to 25 decodes in App. F.1.2) mitigate overhead, reducing the sample complexity of two-level black-box UQ remains an open challenge.

I.5 L5: Scope of empirical tasks and interventional validation

Our empirical evaluation is subject to two scope limitations:

- **Focus on short-form QA:** Aside from a small-scale claim-level pilot on 50 NQ queries (Sec. 5.6), our experiments focus on short-form question answering where response equivalence is well captured by sentence embeddings. Scaling nested covariance decomposition to long-form open-ended generation, multi-step reasoning traces, or interactive agent trajectories is non-trivial, as responses vary simultaneously in length, discourse structure, and set-level claim composition.
- **Simplified single-factor interventions:** Our paired-intervention benchmark tests two clean, synthetic manipulations: (1) resolving prompt ambiguity on AmbigQA and (2) injecting gold supporting passages on NQ and conducts the standardized score-shift comparison against baselines on a single backbone (LLaMA-3-8B; Tab. 4). In realistic RAG or multi-turn deployments, uncertainty sources rarely occur in isolation: retrieved passages may be noisy, partially relevant, or internally contradictory, causing aleatoric and epistemic effects to interact in ways not captured by single-factor interventions.

I.6 Broader impact

Decomposing predictive uncertainty into aleatoric and epistemic components can improve the reliability of deployed LLM systems by informing targeted actions, such as asking clarifying questions under high AU or invoking retrieval and abstention under high EU in high-stakes domains. At the same time, because the decomposition relies on surrogate perturbations and embedding representations, treating empirical AU/EU estimates as formal guarantees could lead to misplaced trust or inappropriate routing when bridge conditions or paraphrase validity fail. Finally, nested Monte Carlo sampling substantially increases inference-time compute and energy consumption if deployed at scale without adaptive stopping or selective triggering.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the theoretical and empirical contributions and their scope, including the local diagnostic conditions of the Bayesian bridge (Sec. 4.2) and the matched-budget comparison (Sec. 5.3).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Sec. 6 and App. I.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Each result states its assumptions; proofs or proof sketches are in App. A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See App. H.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We release the code and provide instructions for reproducing the main experimental results at `https://github.com/wma17/Black-box_UQ_for_LLMs_Ensemble_of_Ensemble`; all implementation details are also given in Sec. 3 and App. H.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: See App. H.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Bootstrap standard deviations for Tab. 1 are reported in App. Tabs. 9 and 10; matched-budget comparisons include bootstrap 95% CIs (App. Tabs. 21, 22); and paired interventions report bootstrap standard errors with paired Wald z-tests (Tab. 3) alongside 95% bootstrap CIs (Tab. 4).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Hardware and software: App. H; time per question: App. Tab. 20; long-form pilot: App. E.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Reviewed. The paper uses publicly available research benchmarks and openly released models. The only human input is a blind audit in which three raters labeled 100 question pairs (App. F.2.3); no personal data were collected.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See App. I.6.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We only release research code for the proposed method and its evaluation, which does not pose a significant risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and models used are publicly released and properly cited; see App. H.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The released code is documented in the public repository at `https://github.com/wma17/Black-box_UQ_for_LLMs_Ensemble_of_Ensemble`, including implementation details, usage instructions, dependencies, and licensing information.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The authors themselves annotated 100 model-generated question pairs using a blind evaluation form (App. F.2.3). No external annotators or crowdsourcing participants were recruited.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: The annotation task involved only co-authors evaluating model-generated text. No external human participants were recruited, and no personal or behavioral data about the annotators were collected.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs are the primary subjects of study and evaluation. We also use standard LLM-based evaluation procedures, including an LLM grader for perturbation checks (App. F.2.3) and LLM-based claim extraction and factuality assessment (App. E). These procedures are described in the paper. LLMs were also used for editing and for help implementing standard methods.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.