NEURIPS 2026 · MAIN CONFERENCE

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 [*]Work done at IBM Research

RPI RENSSELAER POLYTECHNIC INSTITUTE 1824

IBM

NEURAL INFORMATION PROCESSING SYSTEMS

2026 · Sydney

01 | Why Decompose Uncertainty?

Model is uncertain → **Identify the source** → **Choose the action**

The right action depends on **why** the model is uncertain.

SOURCE: ALEATORIC · AU

Ambiguous question

“What is the biggest city in the U.S. with only one word in its name?”

Action → ask the user to clarify

SOURCE: EPISTEMIC · EU

Missing knowledge

“Who is a famous person from the Galapagos Islands?”

Action → retrieve evidence or abstain

Existing black-box methods

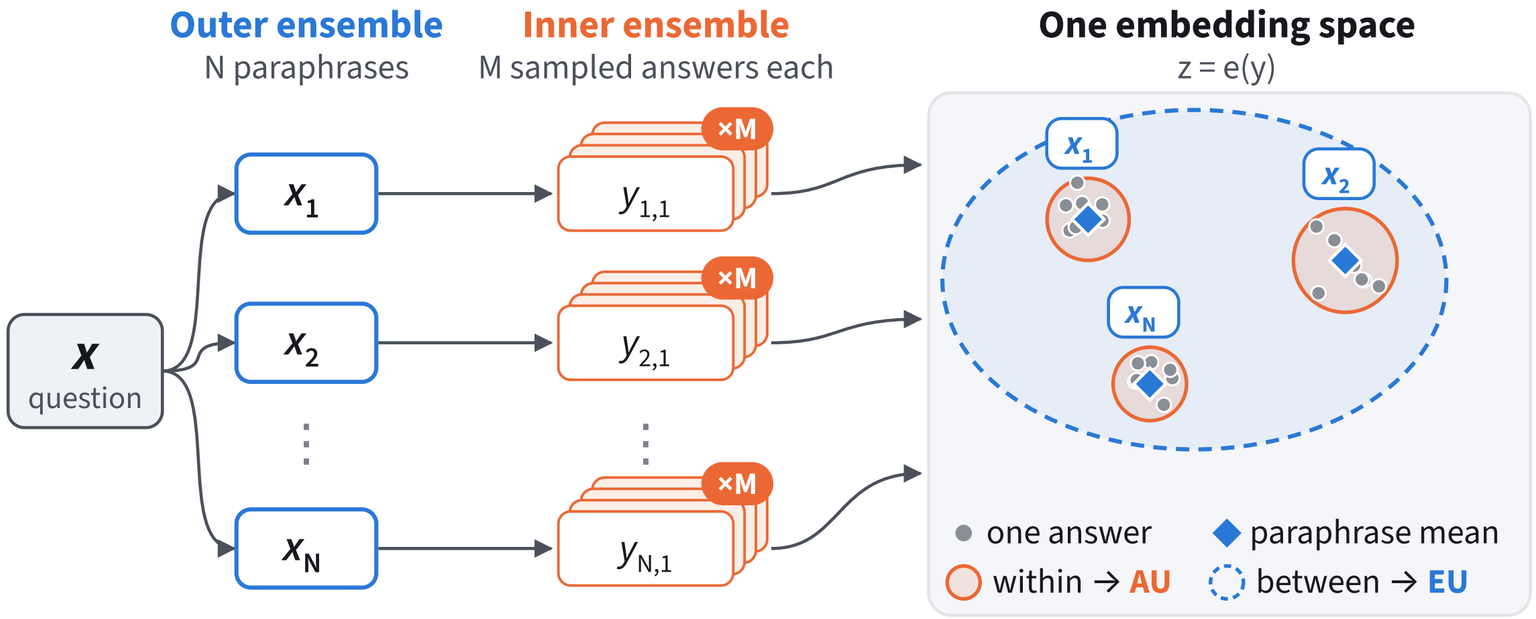
Semantic entropy, self-consistency, SPUQ: output a single score. They flag unreliable answers, but cannot tell why the model is uncertain.

Bayesian & deep ensembles

Separate AU and EU in principle, but require white-box access (weights or logits) and are too expensive for LLMs.

Goal: separate AU and EU using only black-box text generations—with no retraining or token logits.

02 | Method: Ensemble of Ensembles



Paraphrase the prompt *N* times, sample *M* answers each.

Spread within a paraphrase
→ **AU** · prompt ambiguity

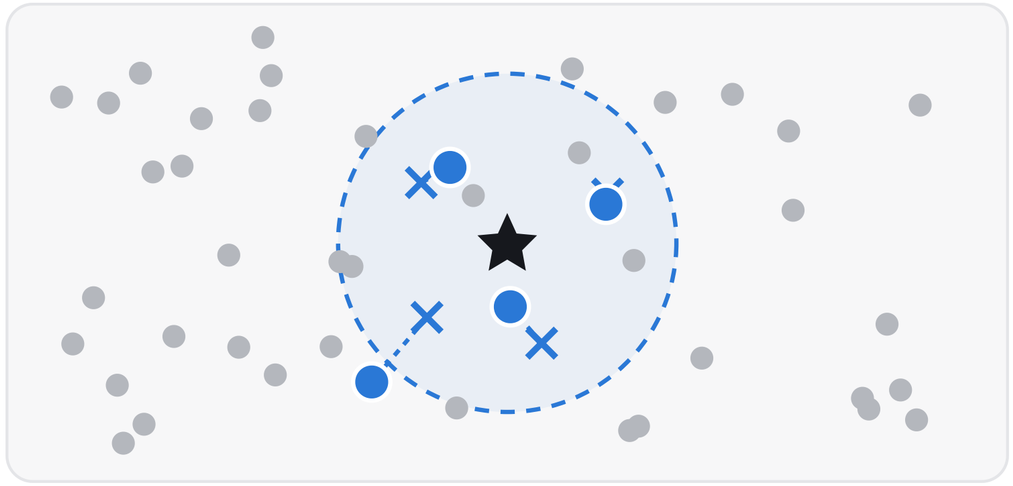
Spread across paraphrases
→ **EU** · knowledge gap

$$\underbrace{\mathrm{Cov}(Z)}_{\text{total (TU)}} = \underbrace{\mathbb{E}_{\Delta}\left[\mathrm{Cov}(Z \mid \Delta)\right]}_{\text{AU: within a paraphrase}} + \underbrace{\mathrm{Cov}_{\Delta}\left(\mathbb{E}[Z \mid \Delta]\right)}_{\text{EU: between paraphrases}}$$

Law of Total Covariance on answer embeddings $Z = e(Y)$.
AU measures answer variation within the same phrasing.
EU measures how the mean answer shifts across equivalent phrasings Δ.

02 | Method: Sampling and Estimators

Outer ensemble: localized paraphrase sampling



1. **Pool** ●: generate candidate rewrites offline & cache.
2. **Perturb** ×: add Gaussian noise around the question ★.
3. **Match** ●: snap to the nearest pool candidate as x^i.

Snapping to cached candidates keeps each prompt natural and meaning-preserving while varying surface phrasing.

Inner ensemble and two-level estimators

For each paraphrase x^i, sample M answers, embed them as $z^{i,j} = e(y^{i,j})$, and compute per-prompt statistics:

$$\hat{\mu}_i = \frac{1}{M}\sum_j z_{i,j}, \qquad \hat{\Sigma}_i = \frac{1}{M-1}\sum_j (z_{i,j} - \hat{\mu}_i)(z_{i,j} - \hat{\mu}_i)^\top$$

AU: average within-paraphrase covariance

$$\hat{U}_A = \frac{1}{N}\sum_i \hat{\Sigma}_i$$

EU (raw): sample covariance of the paraphrase means

$$\hat{U}_E^{\text{raw}} = \frac{1}{N-1}\sum_i (\hat{\mu}_i - \bar{\mu})(\hat{\mu}_i - \bar{\mu})^\top$$

Scalar score for ranking: matrix trace $\mathrm{tr}(\hat{U})$. Total cost: $N \cdot M$ generations per query.

03 | Theory: Exact Finite-Sample Debiasing

1 | Ideally, EU is the covariance of the **true** paraphrase means $\mu^i = \mathrm{E}[Z \mid \Delta^i]$.

2 | In practice, each mean averages M answers, carrying $1/M$ inner noise:

$$\hat{\mu}_i = \mu_i + \epsilon_i, \qquad \mathrm{Cov}(\epsilon_i \mid \Delta_i) = \tfrac{1}{M}\,\Sigma_i$$

3 | As a result, the naive EU estimator absorbs $1/M$ of the **within-paraphrase** covariance:

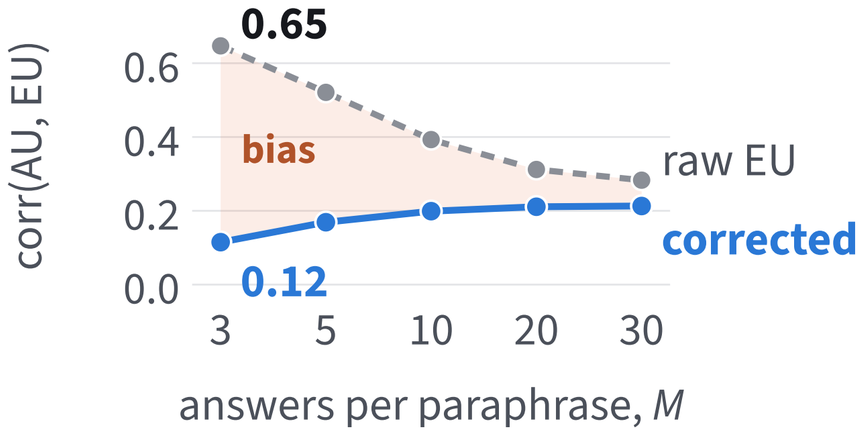
$$\mathbb{E}\big[\hat{U}_E^{\mathrm{raw}}\big] = U_E + \tfrac{1}{M}\,U_A$$

4 | **Closed-form bias correction (unbiased for any $M \geq 2$):**

$$\hat{U}_E = \hat{U}_E^{\mathrm{raw}} - \tfrac{1}{M}\,\hat{U}_A, \qquad \hat{U}_T = \hat{U}_A + \hat{U}_E$$

Empirical check: removing AU leakage

Spearman correlation ρ(AU, EU) on TriviaQA (LLaMA-3-8B, $N = 10$).



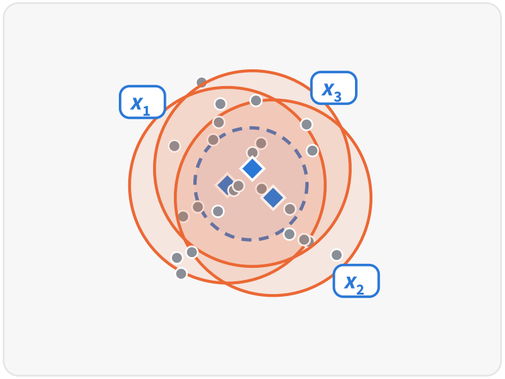
Without correction, AU leaks into EU at small M (ρ = **0.65** at $M = 3$). Subtracting $\hat{U}^A / M$ removes this inflation immediately (ρ = **0.12**), matching the predicted $1/M$ rate.

04 | Qualitative Behavior on Real Questions

High AU · Low EU

Ambiguous prompt. "Biggest" can mean population or area. Each wording yields diverse cities, with a similar mix across rewrites.

What is the biggest city in the U.S. with only one word in its name?



x_1 *What is the biggest city in the U.S. with only 1 word in its name?*

Austin ×2 | Chicago ×2 | Phoenix ×2 | Tucson | Omaha | Kansas | Dallas

x_2 *Which is the biggest city in the United States with only one word in its name?*

Phoenix ×2 | Chicago ×2 | Austin ×2 | Oklahoma | Jackson | Atlanta | Detroit

x_3 *What would be the biggest city in the U.S. with only one word in its name?*

Phoenix ×4 | Omaha ×2 | Austin ×2 | Oklahoma | Chicago

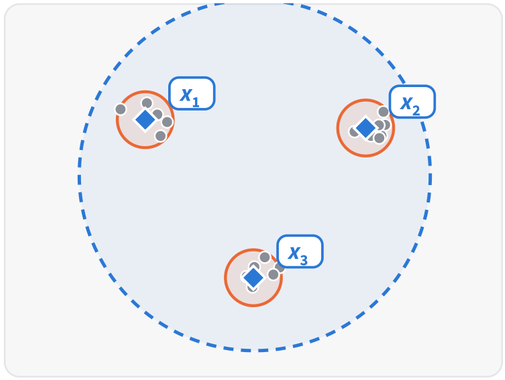
Within each wording: diverse answers. **Across** wordings: similar mean.

AU: 86th pct | **EU: 12th pct** **Ambiguity → clarify "biggest"**

Low AU · High EU

Knowledge gap. Repeats one answer per wording, but switches answers completely when the question is rephrased.

Who is a famous person from the Galapagos Islands?



x_1 *What is a famous person from Galapagos Islands?*

Charles Darwin ×10

x_2 *How is a famous person from the Galapagos Islands?*

Lonesome George ×10

x_3 *Who are famous people who live in the Galapagos Islands?*

David Attenborough ×10

Within each wording: one repeated answer. **Across** wordings: conflicting answers.

AU: 17th pct | **EU: 96th pct** **Knowledge gap → retrieve or abstain**

LLaMA-3-8B on TruthfulQA (3 of $N = 10$ paraphrases shown, $M = 10$ answers each). Percentiles are across 400 questions.

05 | Error Detection Across QA Benchmarks

STEP 1
Generate answers & compute AU, EU, TU.

→

STEP 2
Label each answer right or wrong vs. reference.

→

STEP 3
Evaluate AUROC: do higher scores rank errors first?

20 settings = 5 QA benchmarks (TriviaQA, TruthfulQA, SciQA, GPQA, CoQA) × 2 LLMs (LLaMA-3-8B, Qwen2.5-7B) × 2 metrics (Cosine, ROUGE-L); vs. 11 baselines.

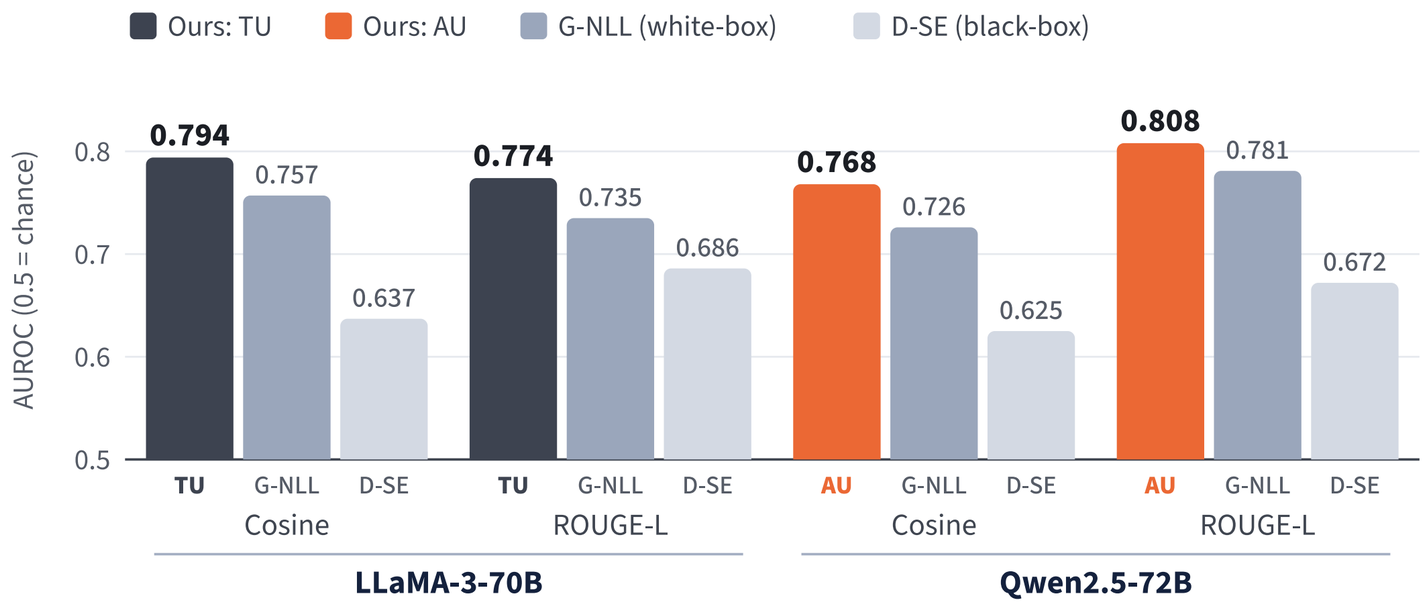
12/20 settings where our AU or TU is the #1 black-box method

11/20 settings where our method also beats all white-box baselines

4/4 settings where our method is #1 overall on 70B+ models (right)

Scaling to 70B+ models: TriviaQA on LLaMA-3-70B and Qwen2.5-72B

Error-detection AUROC (higher is better) vs. best white-box (G-NLL) and black-box (D-SE) baselines.



06 | Targeted Intervention Validation

Intervene on one uncertainty source at a time (1,000 paired examples each, LLaMA-3-8B) and measure the gain in error-detection AUROC:

1. Clarify Prompt · AmbigQA

Prediction: **AU** improves more

AMBIGUOUS *Voice of Alvin in Alvin and the Chipmunks?*
Justin Long ×4 · Jason Lee ×3 · 2 others

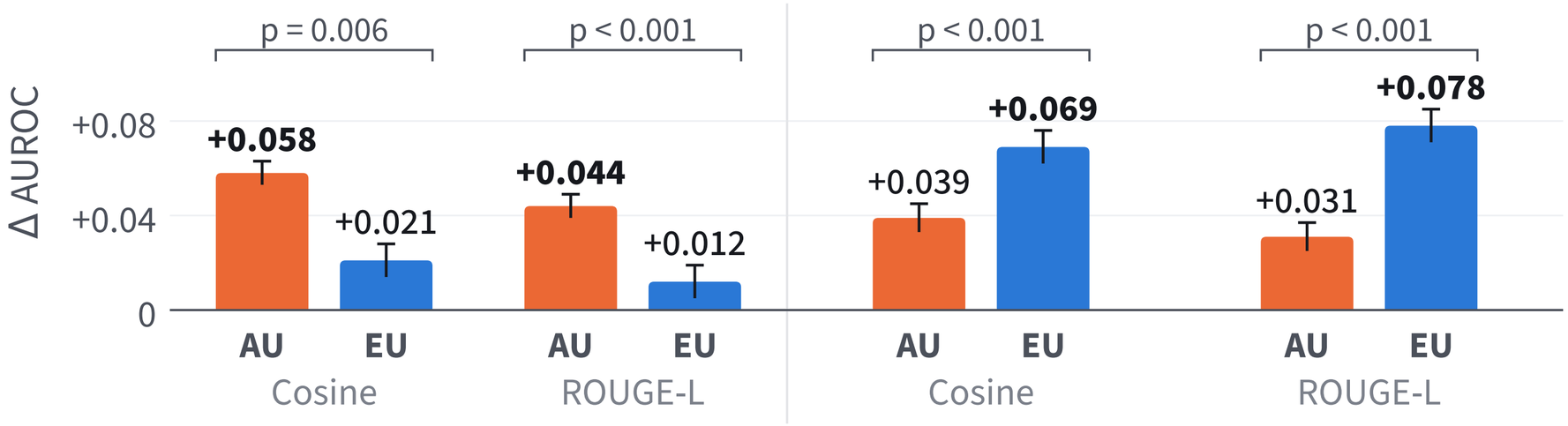
CLARIFIED *Who played Alvin in the 2007 film?*
Justin Long ×7 · Jason Lee ×2 · 1 other

2. Add Evidence · NQ

Prediction: **EU** improves more

NO CONTEXT *What does HP mean in War and Order?*
High Point · High-Pressure · Horsepower · …

+ CONTEXT + gold passage *"… health points, shortened to HP…"*
health / hit points ×8



Clarifying prompts boosts **AU** **2.8–3.7× more** than EU; adding evidence boosts **EU** **1.8–2.5× more** than AU (all four $p \leq 0.006$). AU and EU selectively track prompt ambiguity and knowledge gaps.

07 | Matched-Budget Efficiency & Encoder Stability

Matched generation budget $B = N \cdot M$ (TriviaQA, EM)

Method	AU / EU split	LLaMA-3-8B			Qwen2.5-7B		
		B=10	25	100	B=10	25	100
Ours: AU	✓	**0.744**	**0.752**	**0.766**	**0.772**	**0.789**	**0.788**
Ours: TU	✓	**0.744**	**0.737**	**0.755**	**0.736**	**0.734**	**0.751**
SPUQ (one level)	–	**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 ($M = 1$)	–	0.754	*0.667*	*0.657*	0.743	*0.553*	*0.549*

Across all budgets B, our AU matches leading one-level baselines (overlapping 95% CIs) while decomposing AU and EU. Single-decode perturbation ($M = 1$) drops sharply without inner averaging.

Stability across sentence encoders (AU AUROC)

Model · Dataset	SBERT	MPNet	BGE	min ρ
LLaMA · TriviaQA	0.845	0.843	0.845	0.974
LLaMA · TruthfulQA	0.695	0.690	0.650	0.955
Qwen · TriviaQA	0.855	0.851	0.854	0.981
Qwen · TruthfulQA	0.630	0.620	0.615	0.961

min ρ: lowest Spearman rank correlation of per-question AU between any two encoders. Rankings stay highly consistent ($\rho \geq 0.955$): the signal lives in the LLM's behavior, not the encoder choice.

At a matched generation budget, two-level sampling matches one-level error detection while providing the AU / EU split.

Summary

Two-Level Ensemble. Sample N meaning-preserving paraphrases × M answers each. Spread **within** a paraphrase gives **AU**; spread **across** paraphrases gives **EU**—from text outputs alone.

Theoretical Guarantees. Closed-form correction $\hat{U}^E = \hat{U}^E_{\text{raw}} - \hat{U}^A / M$ makes EU unbiased for any $M \geq 2$, backed by a Bayesian bridge and provable separation from fixed-prompt consistency.

Empirical Validation. Top black-box error detector in 12/20 benchmarks (and 4/4 on 70B+ models); paired interventions confirm AU and EU selectively track ambiguity and knowledge gaps.

Actionable Decisions. High **AU** diagnoses prompt ambiguity → **ask for clarification**; high **EU** diagnoses missing knowledge → **retrieve evidence or abstain**.

Code & Paper

github.com/wma17/Black-box_UQ_for_LLMs_Ensemble_of_Ensemble

maw6@rpi.edu

Supported by the IBM–Rensselaer Future of Computing Research Collaboration

BACKUP

Score-Level Shifts Under Paired Interventions

Score	Clarify (AmbigQA)	Evidence (NQ)
Ours: AU (within)	−0.013	**−0.610**
Ours: EU (between)	**−0.218**	**−0.604**
Ours: TU (total)	**−0.156**	**−0.725**
FIXED-PROMPT BASELINES		
Embedding variance	−0.038	**−0.546**
Self-Consistency	−0.022	**−0.518**
D-SE	−0.031	**−0.576**
Semantic Density, uniform	+0.018	**+0.079**
PERTURBED PROMPTS, ONE LEVEL		
SPUQ	**−0.106**	**−0.836**

- **Clarifying the prompt (AmbigQA):** Significantly lowers perturbation-aware scores (EU, TU, SPUQ), while all fixed-prompt baselines stay unchanged—confirming our separation theorem (Thm. 4.3).
- **Adding evidence (NQ):** Significantly lowers uncertainty across almost all methods as the knowledge gap closes.
- **Why two levels matter:** SPUQ outputs a single scalar that drops under both interventions, whereas our decomposition separates ambiguity from missing knowledge.

Standardized shift = mean of (after − before) / SD of per-question changes (LLaMA-3-8B, 1,000 pairs per intervention). Bold: 95% bootstrap CI excludes zero.

BACKUP

Large Models and Sampling Budget

TriviaQA on 70B-scale models (AUROC)

Method	LLaMA-3-70B		Qwen2.5-72B	
	Cosine	ROUGE-L	Cosine	ROUGE-L
Ours: AU	**0.726**	**0.717**	**0.768**	**0.808**
Ours: EU	**0.791**	**0.725**	**0.613**	**0.592**
Ours: TU	**0.794**	**0.774**	**0.754**	**0.773**
BLACK-BOX				
D-SE	0.637	0.686	0.625	0.672
WHITE-BOX				
G-NLL	0.757	0.735	0.726	0.781
PE	0.549	0.627	0.659	0.702
LN-PE	0.660	0.710	0.711	0.774
SE	0.615	0.679	0.659	0.700
LN-SE	0.587	0.686	0.660	0.701

Bold: best per column. Our method leads all four columns, outperforming white-box baselines.

Tied budget $N = M$ (TruthfulQA, LLaMA-3-8B)

$N = M$	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

AUROC saturates by $N = M = 10$, which serves as our compute-efficient default for ablations and interventions.