# Synergizing Diverse Information Proxies for Reliable Uncertainty Quantification in Large Language Models

**Anonymous ACL submission**

## Abstract

Uncertainty Quantification (UQ) is pivotal for ensuring the reliability of Large Language Models (LLMs), particularly in knowledge-intensive domains. In grey-box and black-box settings, existing UQ methods often yield sub-optimal results due to the limitations of individual information proxies, which restricts their practical utility. This paper investigates these UQ methods and reveals a significant diversity and complementarity in their estimation behaviors. Leveraging these insights, we propose HYPE-UQ, a supervised meta-uncertainty estimation framework that integrates diverse information proxies. Extensive experiments across multiple benchmarks and base models demonstrate that our approach consistently outperforms baseline methods across AUROC, AUPRC, and Brier scores. Furthermore, our ensemble strategy shows only limited degradation in cross-task and cross-domain generalization in Out-of-Distribution evaluations.

## 1 Introduction

The rapid evolution of Large Language Models (LLMs), such as GPT-5 (OpenAI, 2025a), alongside breakthroughs in reasoning models like OpenAI-o1 (OpenAI et al., 2024) and DeepSeek-R1 (Guo et al., 2025), has unlocked transformative potential across diverse domains, ranging from software engineering (Qian et al., 2024) to the life sciences (Gottweis et al., 2025). However, the increasing sophistication of these systems is accompanied by persistent reliability challenges, most notably hallucinations (Ji et al., 2023), inconsistent output quality (Zhou et al., 2024), and the generation of unsafe content (Nolfi, 2024). While these vulnerabilities are often intrinsic to current training paradigms (Wang and Sennrich, 2020) or autoregressive generation mechanisms (Zhang et al., 2024), the inability to accurately quantify the associated uncertainty prevents systems from proactively flagging potential failures, thereby hindering their deployment in mission-critical domains.

A promising way to address the challenges of uncertainty quantification (UQ) is to estimate and communicate the confidence associated with LLM outputs, thereby facilitating the detection of potential errors (Geng et al., 2024; Ren et al., 2023; Yoffe et al., 2025). When UQ methods are used to estimate response confidence, they can be categorized into white-box, grey-box, and black-box approaches based on the level of information accessibility (Liu et al., 2024). White-box methods are applicable in scenarios where the model's internal states are fully accessible (Azaria and Mitchell, 2023; Burns et al., 2023); grey-box methods typically leverage token-level probabilities to estimate uncertainty (Kuhn et al., 2023; Malinin and Gales, 2021; Xu et al., 2020; Yoffe et al., 2025); and black-box methods rely solely on the generated content, accessed through either a single generation (Zhou et al., 2023) or multiple sampling iterations (Kadavath et al., 2022; Xiong et al., 2024). In this paper, we focus exclusively on non-white-box methods (i.e., grey-box and black-box), because the internal states of many proprietary LLMs are inaccessible, rendering white-box techniques impractical.

The information proxies available in non-white-box settings are inherently constrained. For instance, verbalized confidence—a prominent black-box UQ method—has been shown to exhibit significant overconfidence due to its ignorance of underlying token-level probability distributions (Jiang et al.; Si et al., 2023; Singh et al., 2024), while token-level entropy often fails to account for semantic equivalence, struggling to measure uncertainty accurately when synonyms are represented by distinct tokens (Nikitin et al., 2024). In this paper, we demonstrate that although this informational incompleteness is a primary driver of the performance gap in UQ, it can be effectively mitigated. Specifically, we show that by ensembling these complementary and diverse uncertainty proxies, it

is possible to train a robust and generalized model that achieves superior performance in confidence estimation for LLM responses.

The main contributions of this work are summarized as follows:

1. We provide a systematic evaluation of existing non-white-box UQ methods for LLMs and empirically demonstrate the significant diversity and complementarity among them.

2. We propose Heterogeneous Proxy Ensemble for Uncertainty Quantification (HYPE-UQ), a novel ensemble framework designed to aggregate various non-white-box proxies into a robust UQ model, significantly enhancing the reliability of uncertainty estimation for LLM responses.

3. We conduct comprehensive experiments across multiple benchmarks. The results verify that our framework consistently outperforms standalone information proxies and similar ensemble baselines across various metrics, including AUROC, AUPRC, and Brier scores.

## 2 Related Work

**Uncertainty quantification for LLMs.** Uncertainty quantification methods can be categorized into white-box, grey-box, and black-box methods based on the accessibility of information proxies (Liu et al., 2024). White-box methods mainly focus on uncertainty metrics probed from internal network activations, such as hidden layer outputs (Azaria and Mitchell, 2023; Burns et al., 2023). Singh et al. (2024) found that semantic relevance also affects the calibration of uncertainty scores, and thus incorporated it into uncertainty quantification, demonstrating superior performance. Due to the unavailability of internal states in closed-source LLMs, a variety of grey-box and black-box uncertainty estimation methods have been developed recently. For "off-the-shelf" LLMs, black-box techniques such as "sample and aggregation" (Xiong et al., 2024; Kadavath et al., 2022), "prompt elicitation" (Zhou et al., 2023), and multi-agent discussion (Li et al., 2024; Cohen et al., 2023) have been developed. However, lossy information proxies induced by decoding strategies can make black-box methods unreliable (Pham et al., 2024); thus, grey-box UQ methods based on token-level logit entropy have been proposed (Yoffe et al., 2025; Malinin and Gales, 2021; Xu et al., 2020). Kuhn et al. (2023) found that logit probabilities are unable to handle semantic equivalence in practice, and thus proposed to incorporate semantic information into the final metric. The overall performance gap between grey-box and black-box methods remains unclear. Some studies suggest that grey-box methods are slightly better than black-box methods, but the difference is trivial (Xiong et al., 2024).

**Ensemble for uncertainty quantification.** Ensemble methods have been proven effective for enhancing the accuracy of a set of predictors, particularly when the optimal predictor is not contained within the initial hypothesis space (Breiman, 1996), and have also been used in the domain of uncertainty quantification (Wang and Ji, 2023). For instance, Deep Ensembles capture predictive variance by aggregating multiple models trained with different initializations (Lakshminarayanan et al., 2017; Fathullah et al., 2023). As demonstrated in computer vision tasks, increasing member diversity—such as by exploring the parameter space via Laplacian approximation—can significantly improve the overall quality of UQ (Wang and Ji, 2023). In the realm of LLMs, ensemble has recently been applied to UQ through several specialized approaches. For example, Xiao et al. (2025) aggregates multiple consistency-based metrics to achieve robust performance gains, while Chen and Mueller (2024) combines two forms of verbalized confidence to mitigate the biases inherent in single-response elicitation. Other frameworks have explored the integration of heterogeneous information proxies, such as combining consistency with verbalized confidence (Xiong et al., 2024), or token entropy with semantic equivalence (Nikitin et al., 2024). Despite these advancements, a systematic investigation into the full spectrum of available information proxies remains lacking, and a unified framework for their optimal integration has yet to be established. This paper aims to address these gaps.

## 3 Method

This section establishes the formal framework for our uncertainty quantification approach and describes how we harness the diversity of multiple information proxies through ensemble learning.

### 3.1 Problem Formulation

We consider an i.i.d. dataset

$$\mathcal{D} = \{(x_i, y_i, \hat{y}_i)\}_{i=1}^{N}, \quad (1)$$

where $x_i$ is the user query, $y_i$ is the ground truth answer, and $\hat{y}_i = f_\theta(x_i)$ is the answer generated by LLM $f_\theta$. We denote the correctness of $\hat{y}_i$ by

$$c_i = \mathbf{1}[\hat{y}_i = y_i]. \quad (2)$$

During the inference of an LLM, a variety of information proxies $\mathbf{P}(i) = \{P_k(i)\}_{k=1}^{K} \in \mathbb{R}^K$ can be obtained to estimate its confidence in $\hat{y}_i$. These proxies include verbalized confidence $P_{vc}(i)$, answer consistency $P_{cons}(i)$, and token-level entropy $P_{entropy}(i)$; in Section 4.1, we introduce representative information proxies. With each information proxy, we can derive quantified confidence $C_k(i) \in [0, 1]$ with an estimation function $g$:

$$C_k(i) = g(P_k(i)). \quad (3)$$

In non-white-box LLM settings, we observe that different information proxies exhibit significantly divergent predictive strengths. For example, P(True) or verbalized confidence excel at capturing uncertainty via semantic reasoning, whereas token entropy is more effective at the probabilistic level. Even when these proxies yield similar overall performance, the underlying distributions of their successful predictions can be markedly different. In this paper, we reveal that an ensemble of heterogeneous information proxies $\{P_k\}_{k=1}^{K}$ yields significant improvements through their inherent complementarity. Our framework aggregates confidence derived from different information proxies:

$$C_{agg}(i) = \sum_{k=1}^{K} \alpha_k C_k(i) \quad (4)$$

Here $\mathcal{L}$ denotes the loss function that measures the discrepancy between the aggregated confidence and response correctness. Denoting $\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_K\}$, the optimization target of our ensemble framework is to adjust $\alpha$ so as to minimize discrepancy between aggregated confidence and correctness:

$$\alpha^* = \arg\min_{\alpha} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(C_{agg}(i), c_i) \quad (5)$$

### 3.2 Ensemble Framework of Information Proxies

**Proposition 1.** *The performance of an uncertainty quantification ensemble is maximized by increasing the individual predictive accuracy of its members and minimizing their inter-member correlation. See proof in Appendix A.1.*

Since various information proxies leverage distinct information sources to quantify uncertainty, it is reasonable to hypothesize that the inter-family prediction correlation is significantly lower than the intra-family correlation. Based on this premise, we satisfy the diversity requirement of the ensemble by selecting at least one representative method from each distinct information proxy family.

It should be noted that the UQ methods examined in this study are established techniques curated from existing literature. We therefore assume their predictive reliability is sufficiently high such that they do not introduce detrimental noise into the ensemble framework.

Furthermore, as individual accuracy remains a critical determinant of collective performance, we implement a selective integration strategy. Within each information proxy family, we evaluate all candidate methods and incorporate the optimal performer into the ensemble. Rather than relying on standard model-reported confidence scores, which may be poorly calibrated, we employ the *TrustScore*, a more reliable proxy for individual accuracy by measuring the agreement between the model and the underlying data density, as the selection criterion (Jiang et al., 2018). This dual-track approach ensures that the framework maximizes predictive accuracy while maintaining the requisite structural diversity.

**Proposition 2.** *The integration of ensemble members characterized by high mutual redundancy or suboptimal individual accuracy yields diminishing marginal returns and may potentially degrade the overall performance of the ensemble framework. See proof in Appendix A.2.*

Once the optimal representative from each information proxy family is integrated, the marginal utility of adding secondary methods from the same family diminishes. Due to their informational redundancy and relatively lower individual accuracy, such additions offer negligible information gain and may instead introduce noise that compromises the ensemble's calibration. Consequently, in our framework we incorporate only the top-performing

**Algorithm 1** Ensemble framework for HYPE-UQ

**Input:** Dataset $\mathcal{D} = \{(x_i, y_i, \hat{y}_i)\}_{i=1}^{N}$ and labels $c_i = \mathbf{1}[\hat{y}_i = y_i] \in \{0, 1\}$
**Input:** Candidate UQ methods $\{u_j\}_{j=1}^{M}$, proxy families $\{P_k\}_{k=1}^{K}$, and confidence scores $\{C_j\}_{j=1}^{M}$, with $K \leq M$
**Output:** Trained GBDT meta-estimator $f_{\alpha^*}$ as the final uncertainty estimator

Partition methods according to their proxy families
  $S_k = \{j : u_j \text{ uses proxy } P_k\}, \quad k = 1, \ldots, K$
  $\bigcup_{k=1}^{K} S_k = \{1, \ldots, M\}$
Initialize feature set $F \leftarrow \emptyset$
**for** $k = 1, \ldots, K$ **do**
  Select the best method within proxy family $k$
  $j_k^* \leftarrow \arg\max_{j \in S_k} \text{TrustScore}(C_j, c; \mathcal{D})$
  $F \leftarrow F \cup \{u_{j_k^*}\}$
**end for**
**for** each instance $i$ in $\mathcal{D}$ **do**
  Construct the meta-estimator input
  $z(i) \leftarrow [C_{j_1^*}(i), C_{j_2^*}(i), \ldots, C_{j_K^*}(i)]$
**end for**
Fit the GBDT meta-estimator by solving
  $\alpha^* \leftarrow \arg\min_{\alpha} E[f_\alpha(z), c]$
**return** $f_{\alpha^*}$

method from each family as the core ensemble member. While broader combinations within a single proxy family remain theoretically possible, they are excluded from the current scope to ensure maximum structural diversity and computational efficiency, leaving more complex intra-family aggregations for future investigation.

Finally, we address the selection of the meta-estimator employed to aggregate the ensemble outputs into a unified uncertainty estimate. We adopt Gradient Boosted Decision Trees (GBDT) (Friedman, 2001) as our meta-estimator because some LLM uncertainty feature distributions are typically highly concentrated within narrow numerical intervals (Xiong et al., 2024). GBDT's recursive partitioning mechanism allows the model to perform "localized refinement"—resolving complex decision boundaries within high-density regions to provide the granular sensitivity necessary for distinguishing subtle variations in predictive confidence. The overall procedure of our framework is summarized in Algorithm 1.

## 4 Empirical Selection of Information Proxies

Building upon the theoretical framework, this section provides an empirical analysis of the diversity and complementarity of information proxies to justify the selection of input features for HYPE-UQ.

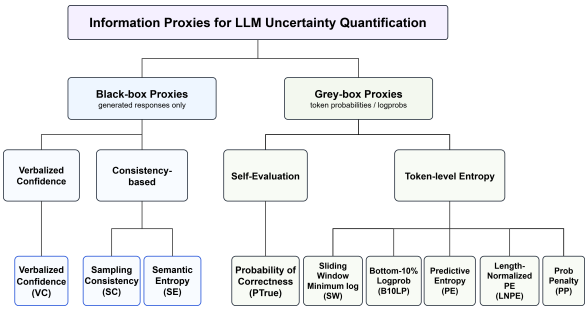



Figure 1: Taxonomy of Information Proxy Methods.

### 4.1 The Diversity of Information Proxies

In non-white-box settings, the information sources for UQ are primarily LLM responses (black-box) and token-level logit distributions (grey-box). These signals provide different perspectives on aleatoric uncertainty (stochastic noise in generation) and epistemic uncertainty (lack of factual knowledge). As illustrated in Figure 1, we categorize UQ methods discussed in this paper into four types, and more details are demonstrated in Appendix B.1:

**Verbalized Confidence (Xiong et al., 2024; Zhang et al., 2024)**: LLM $f_\theta$ is prompted to directly generate a confidence score $P_{vc}$ alongside its answer. The process is defined as

$$(\hat{y}, P_{vc}) \sim f_\theta(\cdot|x).$$

Verbalized confidence elicits the model's perception of its own epistemic uncertainty.

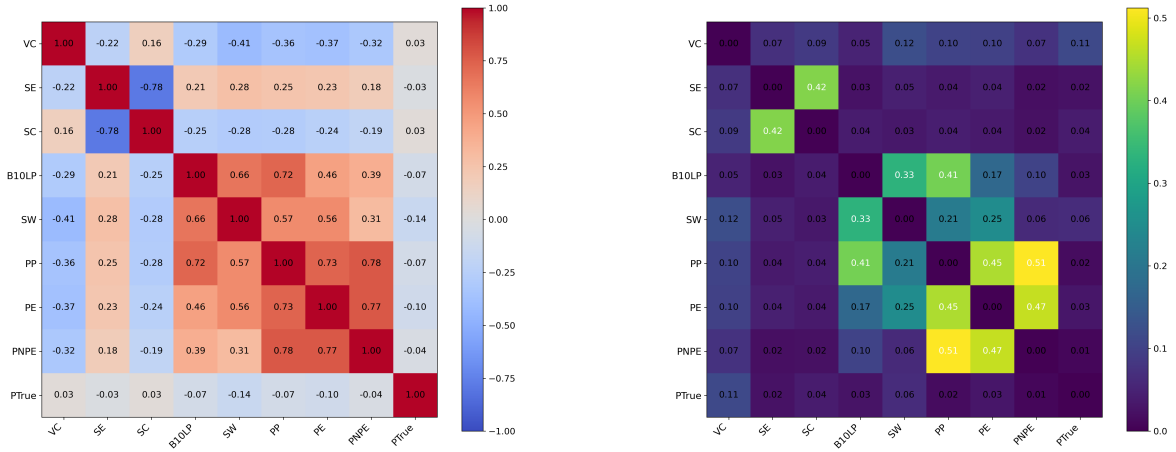
**Self-Evaluation (Kadavath et al., 2022)**: A separate evaluation step, such as "P(True)", is employed where a specialized evaluator model assesses the correctness of a generated answer $\hat{y}$. This decouples the generation process from the confidence estimation:

$$P_{PTrue} \sim f_\phi(\text{True} \mid \hat{y}, x).$$

Self-evaluation offers an independent assessment of epistemic uncertainty from the perspective of another prompted evaluator LLM $f_\phi$.

**Consistency-based Confidence (Gao et al., 2024; Kuhn et al., 2023)**: These methods, including Sampling Consistency (SC) and Semantic Entropy (SE), generate $N$ responses and quantify confidence via their semantic (dis)agreement. A common metric is the frequency of the modal answer:

$$P_{cons} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[y_i = \hat{y}] \tag{6}$$




(a) Spearman

(b) Mutual information

Figure 2: Pairwise dependency analysis of information proxy methods.

**Token-level Entropy (Fu et al., 2025; Malinin and Gales, 2021)**: These methods leverage the logit distribution to calculate statistical metrics such as "sliding window logprobs" or Predictive Entropy (PE), which primarily quantify aleatoric uncertainty in the decoding process. They aggregate token probabilities to indicate uncertainty:

$$P_{entropy} = -\frac{1}{T}\sum_{t=1}^{T} P(y_t \mid y_{<t}, q) \log P(y_t \mid y_{<t}, q) \tag{7}$$

### 4.2 Selection of Information Proxies

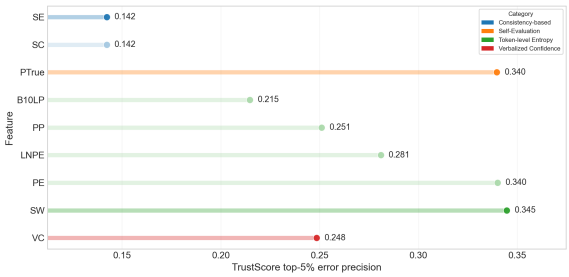



Figure 3: TrustScore-based proxy selection. Proxy features are shown on the vertical axis and TrustScore on the horizontal axis; colors indicate proxy families.

Building on the pre-defined taxonomy introduced in Section 4.1, we validate the relationship between information proxies by utilizing Spearman's rank correlation and Mutual Information (MI) to capture both monotonic and arbitrary statistical dependencies between UQ methods on TriviaQA (Joshi et al., 2017), thereby ensuring a multi-dimensional assessment of feature orthogonality.

The resulting correlation matrices in Figure 2a and Figure 2b exhibit a clear clustered structure. In the Spearman's rank correlation heatmap, PTrue exhibits the lowest correlations with most other proxies, with a maximum correlation of only 0.14. For the other three proxy families, the inter-group correlation ranges from 0.16 to 0.41. In contrast to high within-group correlation (e.g., an absolute value of 0.78 for the consistency-based family, 0.31–0.78 for the token-level entropy family), these results confirm that proxies within the same family form highly correlated sub-clusters. Given that Mutual Information reveals similar trends, we conclude that correlations across families are generally significantly weaker than correlations within the same family. Accordingly, we retain only one representative proxy from each family as a component of the final ensemble.

For within-family representative selection, we rank candidate proxies within each family by TrustScore-style top-5% suspicious-sample error precision and select the top-performing proxy as that family's representative in the final ensemble. Performance results for all candidates are summarized in Figure 3.

Based on proxy performance measured by this TrustScore-style filtering criterion, we select the candidate with the highest within-family score as the family representative; for the consistency-based

| Base Model | Type | UQ method | TriviaQA | | | Natural Questions | | | GPQA | | | HotpotQA | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | AUROC ↑ | AUPRC ↑ | Brier ↓ | AUROC ↑ | AUPRC ↑ | Brier ↓ | AUROC ↑ | AUPRC ↑ | Brier ↓ | AUROC ↑ | AUPRC ↑ | Brier ↓ |
| GPT-4.1 | Single Proxy | SW | 0.7457 | 0.9544 | 0.0763 | 0.6645 | 0.6704 | 0.2311 | 0.6788 | 0.6361 | 0.2301 | 0.7128 | 0.8674 | 0.1056 |
| | Single Proxy | VC | 0.6446 | 0.9382 | 0.0781 | 0.6510 | 0.6186 | 0.2276 | 0.6253 | 0.6095 | 0.2394 | 0.6657 | 0.9108 | 0.1785 |
| | Single Proxy | SE | 0.7119 | 0.9443 | 0.0769 | 0.6069 | 0.5980 | 0.2394 | 0.5888 | 0.5586 | 0.2436 | 0.6651 | 0.9203 | 0.1213 |
| | Single Proxy | PTrue | 0.5240 | 0.8897 | 0.0867 | 0.5373 | 0.5340 | 0.2474 | 0.4891 | 0.4835 | 0.2503 | 0.5319 | 0.8653 | 0.1654 |
| | Ensemble | BSDetector (black-box) | 0.6385 | 0.9243 | 0.1062 | 0.6407 | 0.6323 | 0.2834 | 0.5959 | 0.5698 | 0.3705 | 0.6179 | 0.9007 | 0.1652 |
| | Ensemble | SAR (grey-box) | 0.7706 | 0.9621 | 0.0776 | 0.6763 | 0.6778 | 0.2279 | 0.7047 | 0.6386 | 0.2255 | 0.7154 | 0.8954 | 0.1748 |
| | Ensemble | SIMBA UQ* (hybrid) | 0.8009 | 0.9654 | 0.0687 | 0.7024 | 0.6930 | 0.2184 | **0.6915** | 0.6979 | 0.2261 | 0.8013 | 0.9513 | **0.0973** |
| | Ensemble | HYPE-UQ* (ours) | **0.8210** | **0.9696** | **0.0681** | **0.7071** | **0.6978** | **0.2178** | 0.6892 | **0.7158** | **0.2255** | **0.8137** | **0.9643** | 0.0998 |
| DeepSeek-V3 | Single Proxy | SW | 0.6513 | 0.9078 | 0.1132 | 0.5706 | 0.5447 | 0.2462 | 0.5580 | 0.6501 | 0.2362 | 0.6719 | 0.8847 | 0.2437 |
| | Single Proxy | VC | 0.6063 | 0.8898 | 0.1142 | 0.5740 | 0.5380 | 0.2432 | 0.6681 | 0.7394 | 0.2208 | 0.6145 | 0.8321 | 0.2741 |
| | Single Proxy | SE | 0.7141 | 0.9207 | 0.1045 | 0.6218 | 0.5886 | 0.2384 | 0.5915 | 0.6779 | 0.2324 | 0.7188 | 0.8906 | 0.1996 |
| | Single Proxy | PTrue | 0.7311 | 0.9385 | 0.1151 | 0.6535 | 0.6070 | 0.2402 | 0.4876 | 0.6326 | 0.2380 | 0.6953 | 0.8811 | 0.2014 |
| | Ensemble | BSDetector (black-box) | 0.7917 | 0.9568 | 0.1069 | 0.6508 | 0.6447 | 0.2991 | 0.5668 | 0.6613 | 0.2655 | 0.7356 | 0.9062 | 0.2689 |
| | Ensemble | SAR (grey-box) | 0.7963 | 0.9572 | 0.1008 | 0.6581 | 0.6566 | 0.2328 | 0.6290 | 0.7411 | 0.2267 | 0.8013 | 0.9415 | 0.1945 |
| | Ensemble | SIMBA UQ* (hybrid) | 0.8087 | 0.9508 | **0.0933** | 0.6804 | 0.6434 | 0.2248 | 0.7195 | 0.8069 | 0.2091 | **0.8157** | 0.9601 | 0.0959 |
| | Ensemble | HYPE-UQ* (ours) | **0.8197** | **0.9629** | 0.0941 | **0.6967** | **0.6701** | **0.2224** | **0.7400** | **0.8307** | **0.1990** | 0.8099 | **0.9653** | **0.0751** |
| LLaMA3-70B | Single Proxy | SW | 0.6095 | 0.9028 | 0.1086 | 0.5637 | 0.5307 | 0.2459 | 0.5104 | 0.4781 | 0.2500 | 0.6675 | 0.5191 | 0.2170 |
| | Single Proxy | VC | 0.7391 | 0.9321 | 0.0955 | 0.6623 | 0.5848 | 0.2260 | 0.6090 | 0.5616 | 0.2350 | 0.7481 | 0.6452 | 0.1796 |
| | Single Proxy | SE | 0.6265 | 0.9023 | 0.1013 | 0.6147 | 0.5452 | 0.2360 | 0.5791 | 0.5420 | 0.2440 | 0.5600 | 0.4053 | 0.2375 |
| | Single Proxy | PTrue | 0.6728 | 0.9158 | 0.1053 | 0.5766 | 0.5176 | 0.2471 | 0.6475 | 0.6210 | 0.2423 | 0.6669 | 0.4657 | 0.2118 |
| | Ensemble | BSDetector (black-box) | 0.7335 | 0.9319 | 0.1059 | 0.6506 | 0.5915 | 0.3716 | 0.6718 | 0.6390 | 0.3396 | 0.6959 | 0.4948 | 0.4538 |
| | Ensemble | SAR (grey-box) | 0.7408 | 0.9392 | 0.1036 | 0.6728 | 0.6367 | 0.2333 | 0.6644 | 0.6464 | 0.2402 | 0.7259 | 0.5762 | 0.2080 |
| | Ensemble | SIMBA UQ* (hybrid) | 0.7732 | 0.9387 | 0.0924 | 0.6891 | 0.6298 | **0.2223** | **0.6832** | 0.6393 | 0.2296 | 0.7823 | **0.6899** | 0.1821 |
| | Ensemble | HYPE-UQ* (ours) | **0.7841** | **0.9467** | **0.0921** | **0.6929** | **0.6370** | 0.2247 | 0.6769 | **0.6467** | **0.2293** | **0.7945** | 0.6706 | **0.1784** |

Table 1: Performance comparison across datasets. Best results within each base model are bolded. (* means supervised ensemble framework.)

family, SE and SC tie under $\alpha = 0.1$, and we retain SE as the representative, yielding the pre-screened set:

$$\mathbf{VC} + \mathbf{PTrue} + \mathbf{SE} + \mathbf{SW}.$$

## 5 Experiments

### 5.1 Experimental Setup

**Datasets and Statistics**: To ensure a comprehensive evaluation across diverse knowledge domains and varying difficulty levels, we selected four QA datasets: TriviaQA (Joshi et al., 2017), Natural Questions (NQ) (Kwiatkowski et al., 2019), GPQA (Rein et al., 2023), and HotpotQA (Yang et al., 2018). These datasets form a progressive complexity spectrum, allowing us to assess uncertainty estimation performance under increasingly challenging reasoning levels. Positive (correct) and negative (incorrect) samples were constructed based on the accuracy of the generated answers. Detailed answer accuracy statistics for each base model are provided in Appendix C.3.

**Model and Prompt Configuration**: We evaluate uncertainty estimation robustness across three representative LLMs—GPT-4.1 (OpenAI, 2025b), DeepSeek-V3 (DeepSeek-AI et al., 2025), and LLaMA3-70B (Grattafiori et al., 2024)—covering both closed-source and open-source models for a more comprehensive comparative evaluation. Multi-dimensional features are extracted via structured prompts designed for joint answer-confidence generation and PTrue verification; the specific prompt templates and model configurations are detailed in Appendix C.

**Baseline Features**: To validate our proposed scheme, we compare it against established uncertainty estimation baselines, including both ensemble members and other ensemble UQ methods (Chen and Mueller, 2024; Luo et al., 2026; Bhattacharjya et al., 2025). A comprehensive list and the formal mathematical definitions of these baseline features are provided in Appendix B.2.

**Evaluation Metrics**: To evaluate the performance of uncertainty estimation in distinguishing correct from incorrect answers, we employ three standard metrics: AUROC to measure overall discriminative capability, AUPRC to assess error detection on imbalanced datasets, and the Brier score to measure overall predictive performance.

### 5.2 Main Results

We compare our framework with the UQ methods selected in Section 4.2 and other ensemble UQ baselines. The main results on three metrics are listed in Table 1, and the statistical significance results are provided in Appendix D.

**Ensembling information proxies yields robust performance gains.** As illustrated in Table 1, ensemble methods outperform all single-proxy UQ methods on AUROC and AUPRC. Notably, HYPE-UQ achieves consistent performance gains over individual UQ methods across all base models and benchmarks. Specifically, compared with the best-performing single-proxy method, our approach

yields an improvement in AUROC ranging from 0.015 to 0.088, and in AUPRC from 0.015 to 0.091, underscoring its superior capability in both predictive discrimination and error detection. Besides, HYPE-UQ outperforms other ensemble baselines (Ye et al., 2023; Chen and Mueller, 2024; Bhattacharjya et al., 2025). We attribute the success of our framework to two main factors: the meta-estimator and diverse information proxies, which we delve into in Sections 5.3 and 5.4, respectively.

**Our framework mitigates the calibration degradation typically associated with ensembling while boosting discrimination.** While prior literature suggests that ensembling can potentially lead to overconfidence and a decline in calibration—particularly in boosted tree algorithms (Ashukha et al., 2020; Niculescu-Mizil and Caruana, 2005)—our results exhibit a distinct pattern: HYPE-UQ achieves substantial discrimination improvements without suffering the expected calibration penalties, as evidenced by a consistent reduction in the overall Brier score. Concretely, the Brier score decreases in every reported setting, with reductions ranging from small but consistent gains (e.g., GPT-4.1 on GPQA: 0.2297→0.2255) to pronounced improvements under harder conditions (e.g., DeepSeek-V3 on GPQA: 0.2208→0.1990). In contrast, other ensemble baselines often suffer from elevated Brier scores compared to single-proxy UQ methods, despite yielding higher AUROC and AUPRC scores. This supports our hypothesis that the diverse proxies integrated by our meta-estimator contribute complementary uncertainty information rather than redundant signals, thereby enhancing the overall quality of probabilistic predictions without compromising model calibration.

### 5.3 Why GBDT as Meta-Estimator

Recent supervised UQ and hallucination-detection methods often use lightweight linear probes or LR-style classifiers as meta-estimators (Orgad et al., 2025; Luo et al., 2026). We therefore compare GBDT with LR and a stronger LR variant using *rank normalization*, which maps each proxy through the empirical CDF estimated on the training fold. This transformation is not test leakage, but it introduces a source-distribution-specific preprocessing step whose mapping must be re-estimated when the dataset, base model, or deployment distribution changes.

Table 2 shows that rank normalization substantially strengthens LR, confirming that skewed proxy marginals are a major limitation of plain LR. However, rank normalization remains a feature-wise monotonic transform: it reshapes marginal distributions but is sensitive to the extreme values in the training set. We therefore use GBDT as the default meta-estimator, since it is inherently invariant to any monotonic transformation and is thus more robust to outliers in the training set. Additional distribution diagnostics, significance analysis, and full meta-estimator comparisons are provided in Appendix C.11.

| Method | AUROC ↑ | AUPRC ↑ | Brier ↓ |
|---|---|---|---|
| LR | $0.789 \pm 0.035$ | $0.947 \pm 0.013$ | 0.099 |
| LR + rank normalization | $\mathbf{0.822} \pm \mathbf{0.021}$ | $\mathbf{0.958} \pm \mathbf{0.006}$ | **0.095** |
| GBDT (ours) | $0.815 \pm 0.029$ | $0.956 \pm 0.010$ | **0.095** |

Table 2: Meta-estimator comparison on TriviaQA / GPT-4.1. Results are averaged over 5 rounds.

### 5.4 Ablation Study

To assess the contribution of each selected proxy, we conduct leave-one-out ablation on TriviaQA / GPT-4.1.

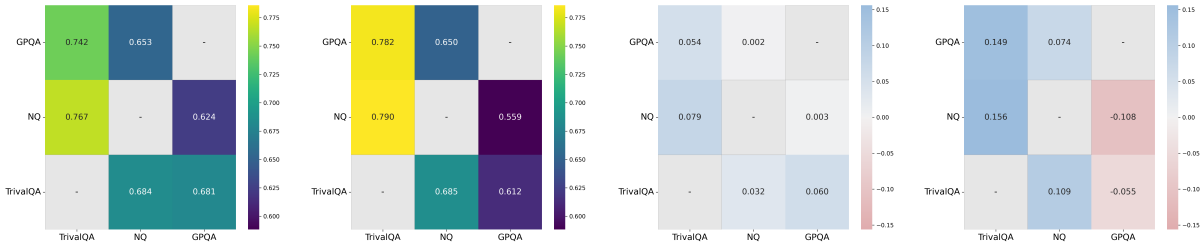
| Category | Feature | AUROC | ΔAUROC |
|---|---|---|---|
| Baseline | full | 0.8490 | – |
| Ablation | SW | 0.7982 | -0.0508 |
| Ablation | VC | 0.8044 | -0.0446 |
| Ablation | SE | 0.7740 | -0.0750 |
| Ablation | PTrue | 0.6919 | -0.1571 |

Table 3: Leave-one-out ablation on TriviaQA / GPT-4.1.

As shown in Table 3, removing any proxy degrades AUROC, confirming that all four selected features contribute non-redundant information. PTrue is the most influential component, whose removal causes the largest drop, consistent with its low correlation with other proxy families in Section 4.1. Additional iterative feature-elimination results in Appendix C.10 show the same trend.

### 5.5 Out-of-Distribution Detection

Orgad et al. (2025) notes that trained detectors/error classifiers exhibit limited generalization across datasets. Following their Train-on-Source, Test-on-Target OOD setup, we evaluate cross-dataset transfer among TriviaQA, NQ, and GPQA; proto-




(a) AUROC, GPT-4.1 (b) AUROC, DeepSeek-V3 (c) ΔAUROC, GPT-4.1 (d) ΔAUROC, DeepSeek-V3

Figure 4: OOD transfer performance and gains over VC.

col details are provided in Appendix C.8.

Figure 4 compares HYPE-UQ with the best-performing single proxy (VC) on GPT-4.1 and DeepSeek-V3. Rows denote source datasets, columns denote target datasets, and cells report AUROC or ΔAUROC over VC. While prior supervised UQ methods often underperform single-proxy entropy in OOD transfer, HYPE-UQ shows similar drops in only 17% of evaluated scenarios, yielding an average AUROC improvement of 0.045 across OOD benchmarks. Notably, in cases where performance improves, we observe an average AUROC gain of 0.060, which substantially outweighs the average decrease of 0.014 in the remaining cases. This yields an overall average AUROC improvement of 0.045 across all OOD benchmarks. These findings suggest that the fusion of multi-dimensional uncertainty signals generally enhances robustness, while also acknowledging the challenges posed by inherent heterogeneity across diverse domain shifts.

## 6 Conclusion

In this study, we investigated the diversity of information proxies across a range of grey-box and black-box LLM UQ methods. Our findings reveal that these methods are highly diverse and complementary, suggesting that their integration can yield a more robust and effective UQ framework. Experimental results across four benchmarks and three base models demonstrate that our ensemble approach consistently enhances both calibration and discrimination performance. Furthermore, our model exhibits remarkable generalization resilience in OOD scenarios, effectively overcoming the generalization bottlenecks commonly associated with supervised UQ methods in prior literature.

Despite these promising results, several directions remain for future investigation. First, while white-box UQ methods were beyond the current scope, we look forward to exploring whether integrating internal model states can further enrich the diversity of our ensemble members. Second, the efficacy of our framework should be validated across a broader spectrum of model architectures and a more diverse range of application domains to further ensure its universality. Finally, while our results demonstrate the generalization resilience of the ensemble model, a more rigorous analysis is warranted to uncover the underlying mechanisms of this phenomenon.

## 7 Limitations

**Scope of evaluation.** Our experiments focus on uncertainty quantification for English, knowledge-intensive question answering benchmarks. Although the selected datasets cover different difficulty levels and domains, they do not exhaust the range of settings in which LLM uncertainty estimates may be used, such as multilingual interaction, long-form generation, dialogue, tool use, safety-critical decision support, or highly specialized professional domains. The empirical results are also limited to three base models and to the availability of non-white-box information proxies. In particular, several grey-box features require access to token-level log-probabilities, which may be unavailable, truncated, or implemented differently across commercial APIs. Therefore, the observed gains should be interpreted as evidence for the evaluated settings rather than a guarantee of universal reliability.

**Assumptions and measurement limitations.** Our framework assumes that response correctness labels provide an adequate target for learning confidence estimates. In practice, correctness can be ambiguous for open-ended questions, and exact-match style evaluation may undercount semantically correct answers. The selected proxies also depend on

prompting, decoding parameters, semantic clustering, and verification prompts, which can introduce dataset- and model-specific behavior. Moreover, our TrustScore-style proxy selection is intended as a practical diagnostic for selecting representative proxies within families, but it is not a definitive measure of causal usefulness. As shown in our diagnostics, the best within-family representative can vary across datasets, suggesting that proxy selection may need to be revisited when the deployment distribution changes.

**Generalization and deployment risks.** HYPE-UQ is a supervised meta-estimator and can inherit distributional biases from its training data, base models, prompts, and benchmark labels. Although our OOD experiments evaluate cross-dataset transfer, they cover only a subset of possible distribution shifts. Incorrect confidence estimates may create false assurance: overconfident wrong predictions could discourage human verification, while underconfident correct predictions could reduce system utility. For this reason, our method should not be treated as a standalone safety mechanism in high-stakes applications. It is better viewed as an auxiliary signal to support selective prediction, auditing, or human review, especially when errors may affect vulnerable users or consequential decisions.

**Artifacts, privacy, and computational cost.** This work uses existing public benchmarks and pretrained or API-served LLMs, and does not introduce new human-subject data collection. We use these artifacts only for research evaluation and follow the access conditions and terms associated with the original datasets, model APIs, and released model checkpoints. Nevertheless, public QA datasets and generated model outputs may contain names, demographic references, or other sensitive content inherited from their original sources. We do not study differential performance across demographic groups, languages, or dialects, so fairness-related claims are outside the scope of this paper. The method also increases computational and monetary cost because several proxies require repeated sampling, verification calls, or access to token-level probabilities. These costs should be considered when reproducing the experiments or deploying the framework at scale.

## References


Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry P. Vetrov. 2020. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Amos Azaria and Tom M. Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Findings of ACL, pages 967–976. Association for Computational Linguistics.

Debarun Bhattacharjya, Balaji Ganesan, Junkyu Lee, Radu Marinescu, Katya Mirylenka, Michael Glass, and Xiao Shou. 2025. 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, Suzhou, China. Association for Computational Linguistics.

Leo Breiman. 1996. Stacked regressions. *Mach. Learn.*, 24(1):49–64.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jiuhai Chen and Jonas Mueller. 2024. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 5186–5200. Association for Computational Linguistics.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. LM vs LM: detecting factual errors via cross examination. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 12621–12640. Association for Computational Linguistics.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, and etc. 2025. Deepseek-v3 technical report.

Yassir Fathullah, Guoxuan Xia, and Mark J. F. Gales. 2023. Logit-based ensemble distribution distillation for robust autoregressive sequence uncertainties. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 582–591. PMLR.

Jerome H. Friedman. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232.

Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. 2025. Deep think with confidence.

Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. 2024. SPUQ: perturbation-based uncertainty quantification for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - Volume 1: Long Papers, St. Julian's, Malta, March 17-22, 2024*, pages 2336–2346. Association for Computational Linguistics.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 6577–6595. Association for Computational Linguistics.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, and etc. 2025. Towards an ai co-scientist.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and etc. 2024. The llama 3 herd of models.

Daya Guo, Dejian Yang, Haowei Zhang, Song, and etc. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Heinrich Jiang, Been Kim, Melody Y. Guan, and Maya R. Gupta. 2018. To trust or not to trust A classifier. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 5546–5557.

Mingjian Jiang, Yangjun Ruan, Sicong Huang, Saifei Liao, Silviu Pitis, Roger Baker Grosse, and Jimmy Ba. Calibrating language models via augmented prompt ensembles.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

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. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Moxin Li, Wenjie Wang, Fuli Feng, Fengbin Zhu, Qifan Wang, and Tat-Seng Chua. 2024. Think twice before trusting: Self-detection for large language models through comprehensive answer reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, Findings of ACL, pages 11858–11875. Association for Computational Linguistics.

Linyu Liu, Yu Pan, Xiaocheng Li, and Guanting Chen. 2024. Uncertainty estimation and quantification for llms: A simple supervised approach.

Wen Luo, Guangyue Peng, Wei Li, Shaohang Wei, Feifan Song, Liang Wang, Nan Yang, Xingxing Zhang, Jing Jin, Furu Wei, and Houfeng Wang. 2026. Two pathways to truthfulness: On the intrinsic encoding of llm hallucinations.

Andrey Malinin and Mark J. F. Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. *Proceedings of the 22nd international conference on Machine learning*.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. 2024. Kernel language entropy: Fine-

grained uncertainty quantification for llms from semantic similarities. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Stefano Nolfi. 2024. On the unexpected abilities of large language models. *Adapt. Behav.*, 32(6):493–502.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, and etc. 2024. Openai o1 system card.

OpenAI. 2025a. GPT-5 System Card. *arXiv preprint arXiv:2601.03267v1*. Version posted 2025-12-19.

OpenAI. 2025b. Introducing gpt-4.1 in the api. OpenAI Product Research Publication. Published April 14, 2025.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2025. Llms know more than they show: On the intrinsic representation of LLM hallucinations. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Chau Pham, Boyi Liu, Yingxiang Yang, Zhengyu Chen, Tianyi Liu, Jianbo Yuan, Bryan A. Plummer, Zhaoran Wang, and Hongxia Yang. 2024. Let models speak ciphers: Multiagent debate through embeddings. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15174–15186. Association for Computational Linguistics.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof qa benchmark.

Jie Ren, Yao Zhao, Tu Vu, Peter J. Liu, and Balaji Lakshminarayanan. 2023. Self-evaluation improves selective generation in large language models. In *Proceedings on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models" at NeurIPS 2023 Workshops, 16 December 2023, New Orleans, Louisiana, USA*, Proceedings of Machine Learning Research, pages 49–64. PMLR.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2023. Prompting gpt-3 to be reliable.

Aniket Kumar Singh, Bishal Lamichhane, Suman Devkota, Uttam Dhakal, and Chandra Dhakal. 2024. Do large language models show human-like biases? exploring confidence—competence gap in ai. *Information*, 15(2).

Chaojun Wang and Rico Sennrich. 2020. On exposure bias, hallucination and domain shift in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3544–3552. Association for Computational Linguistics.

Hanjing Wang and Qiang Ji. 2023. Diversity-enhanced probabilistic ensemble for uncertainty estimation. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 216 of *Proceedings of Machine Learning Research (PMLR)*, pages 2214–2225.

Quan Xiao, Debarun Bhattacharjya, Balaji Ganesan, Radu Marinescu, Katsiaryna Mirylenka, Nhan H. Pham, Michael R. Glass, and Junkyu Lee. 2025. The consistency hypothesis in uncertainty quantification for large language models. In *Conference on Uncertainty in Artificial Intelligence, Rio Othon Palace, Rio de Janeiro, Brazil, 21-25 July 2025*, Proceedings of Machine Learning Research, pages 4636–4651. PMLR.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Jiacheng Xu, Shrey Desai, and Greg Durrett. 2020. Understanding neural abstractive summarization models via uncertainty. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6275–6281, Online. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Wenqian Ye, Yunsheng Ma, Xu Cao, and Kun Tang. 2023. Mitigating transformer overconfidence via lipschitz regularization. In *Uncertainty in Artificial Intelligence, UAI 2023, July 31 - 4 August 2023, Pittsburgh, PA, USA*, Proceedings of Machine Learning Research, pages 2422–2432. PMLR.

Luke Yoffe, Alfonso Amayuelas, and William Yang Wang. 2025. Debunc: Improving large language model agent communication with uncertainty metrics. In *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November*

*4-9, 2025*, pages 23299–23315. Association for Computational Linguistics.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2024. How language model hallucinations can snowball. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pages 59670–59684. PMLR / OpenReview.net.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 5506–5524. Association for Computational Linguistics.

Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José Hernández-Orallo. 2024. Larger and more instructable language models become less reliable. *Nature*, 634(8032):61–68.

# A Proofs of Propositions

## A.1 Proof of Proposition 1

Consider an ensemble that linearly aggregates $K$ member estimators $\{u_k\}_{k=1}^{K}$ of the correctness indicator $c \in \{0, 1\}$ with weights $\{\alpha_k\}$ summing to one, yielding $u_P = \sum_k \alpha_k u_k$. Let $e_k = u_k - c$ denote each member's prediction error. The ensemble mean squared error can be decomposed as

$$\mathbb{E}[(u_P - c)^2] = \sum_k \alpha_k^2 \, \mathbb{E}[e_k^2] + \sum_{k \neq l} \alpha_k \alpha_l \, \mathbb{E}[e_k e_l].$$

The first term is a weighted sum of individual error magnitudes, while the second is governed by pairwise error correlations. Hence, minimizing the ensemble error simultaneously requires (i) reducing each $\mathbb{E}[e_k^2]$, i.e. increasing individual accuracy, and (ii) reducing $\mathbb{E}[e_k e_l]$ for $k \neq l$, i.e. decreasing inter-member correlation. This establishes the claim. □

## A.2 Proof of Proposition 2

Suppose a new member $u_{K+1}$ is added to an existing ensemble of size $K$. Let $\rho$ denote the average error correlation between $u_{K+1}$ and the existing members, and let $\sigma_{K+1}^2 = \mathbb{E}[e_{K+1}^2]$ denote its individual error variance. By the bias–variance decomposition of the augmented ensemble,

$$\Delta \, \mathbb{E}[(u_P - c)^2] \;\propto\; \sigma_{K+1}^2 + (K-1)\, \rho \, \sigma_{K+1} \bar{\sigma} - \bar{\sigma}^2,$$

where $\bar{\sigma}$ is the average error magnitude of the existing members. When either $\rho \to 1$ (high mutual redundancy) or $\sigma_{K+1}$ is large relative to $\bar{\sigma}$ (suboptimal individual accuracy), the marginal contribution $\Delta$ is non-negative and can become strictly positive, meaning the additional member yields diminishing returns or actively degrades the ensemble. □

# B UQ Methods Discussed in This Paper

## B.1 Single Proxy Features

### B.1.1 Verbalized Confidence

**Verbal Confidence (VC)**: Prompts the model to self-report its confidence (e.g., on a 1–10 scale) for its output, and only averages these confidence scores for outputs belonging to the "majority answer" cluster (if multiple samples are obtained).

### B.1.2 Self-Evaluation

**PTrue**: Measures the model's self-assessed confidence in the correctness of its prediction by estimating the probability assigned to the predicted answer being true, thereby reflecting the model's calibrated belief in its own output.

### B.1.3 Consistency-based UQ

**Sampling Consistency (SC)**: Measures the proportion of the most frequent (semantically similar) answer across multiple samples from the model. A higher proportion indicates greater model certainty.

**Semantic Entropy (SE)**: Leverages a vector model to compute semantic similarity between multiple generated responses, clusters them, and then calculates a normalized entropy value based on these clusters.

### B.1.4 Token-level Entropy

**Bottom-10% Logprob (B10LP)**: Calculates the average probability of the 10% least likely tokens within a generated sequence.

**Sliding Window Min (SW)**: Scans the sequence using multi-scale sliding windows (e.g., covering 5% to 30% of the sequence length) to identify the minimum average probability within any window.

**Prob Penalty (PP)**: Quantifies the consistency of internal token confidence by calculating a penalty based on the deviation (gap) of token probabilities below the sequence's average, thus penalizing outputs with significant confidence fluctuations.

**Predictive Entropy (PE)**: Accumulates the entropy values across all token positions in a sequence, quantifying the overall uncertainty throughout the entire generation process.

**Length-Normalized PE (LNPE)**: Normalizes the predictive entropy by the sequence length to represent the average uncertainty per token, making it comparable across sequences of different lengths.

## B.2 Ensemble Proxy Features

### B.2.1 BSDetector

Assigns a confidence score to a model's answer, even when the model is a black box (no access to logits/probabilities). Confidence is a weighted combination of two signals:

**Observed Consistency (O)**: Re-sample the model multiple times for the same prompt (e.g., by prompt variations and higher temperature) to get a set of alternative answers. Measure whether these answers are consistent with the original answer, using an NLI-based contradiction check (and an exact-match fallback for short/closed-form answers). Aggregate consistency across samples to obtain O.

**Self-reflection Certainty (S)**: Ask the model to judge its own answer using a discrete choice (e.g., Correct / Incorrect / Not sure) rather than a raw numeric score (to reduce overconfidence). Convert to a scalar (e.g., 1 / 0 / 0.5) and average across multiple reflection prompts to obtain S.

**Final confidence**:

$$C = \beta O + (1 - \beta) S$$

Typical uses. Detect likely hallucinations/incorrect answers; select the most reliable answer among multiple candidates; improve reliability of LLM-as-a-judge by filtering low-confidence judgments.

### B.2.2 SAR

Improves uncertainty estimates for free-form generation by addressing a bias: not all tokens (or sentences) matter equally for meaning, but standard entropy-based uncertainty treats them similarly. This can make uncertainty dominated by low-importance tokens. SAR reweights uncertainty by semantic relevance at two levels:

**Token-level (TOKENSAR)**: Compute each token's relevance by measuring how much the sentence meaning changes when that token is removed (semantic similarity drop). Normalize relevance weights and compute a relevance-weighted token uncertainty:

$$\mathrm{TOKENSAR}(s) = \sum_i \Big( -\log p(z_i \mid \cdot) \Big) \tilde{R}(z_i)$$

This focuses uncertainty on meaningful tokens, not filler words.

**Sentence-level (SENTSAR)**: Sample multiple candidate sentences/answers. Assign each sentence a relevance/representativeness score based on its semantic agreement with other samples (weighted by generation likelihood). Use this relevance to shift sentence-level uncertainty so that more representative (semantically consistent) samples contribute less uncertainty.

**Combined SAR**: Convert token-level uncertainty into a shifted sentence likelihood and apply the sentence-level shifting to obtain the final uncertainty measure. Practical note: SAR typically requires access to token probabilities/logits (i.e., not purely black-box APIs).

### B.2.3 SIMBA UQ

SIMBA UQ (Bhattacharjya et al., 2025) is a similarity-based black-box uncertainty estimation method defined on multiple sampled responses to the same question. The method assumes that a response is more reliable when it is supported by other independently generated responses, and more uncertain when it is semantically isolated from them.

**Similarity-based support**: Given a question $x$ and a set of sampled responses $S = \{s_1, \ldots, s_K\}$, SIMBA computes pairwise similarities between answers and derives, for each answer $s_j$, a support score from the remaining samples. A simple form is the average similarity

$$U_{\mathrm{SIM}}(s_j) = \frac{1}{K-1} \sum_{k \neq j} g(s_j, s_k),$$

where $g(\cdot, \cdot)$ is a lexical or semantic similarity function. Higher agreement indicates lower uncertainty, whereas semantically isolated answers are assigned higher uncertainty.

**Learned aggregation**: Beyond direct averaging, SIMBA can employ a supervised scorer over similarity-derived statistics. In our experiments, the pairwise similarity matrix is summarized by mean similarity, maximum similarity, minimum similarity, standard deviation, centrality-style support, and similarity to the most central response. These statistics are then mapped to a confidence score by a lightweight classifier.

**Practical note**: SIMBA is well suited to non-white-box settings because it only requires multiple sampled responses together with a similarity func-

tion over generated texts. In this paper, we instantiate this family using the eight sampled answers for each question, compute lexical and embedding-based pairwise similarities, and evaluate both direct similarity aggregation and a supervised similarity-based scorer.

# C Extra Experiment Setup

## C.1 Candidate Feature Pool

Nine candidate UQ features are extracted per response, spanning the four proxy families described in Section 4.1. Table 4 lists each feature together with its proxy family.

| Feature | Abbr. | Family |
|---|---|---|
| Verbal Confidence$^\star$ | VC | Verbalized Confidence |
| PTrue$^\star$ | PTrue | Self-Evaluation |
| Semantic Entropy$^\star$ | SE | Consistency-based |
| Sampling Consistency | SC | Consistency-based |
| Sliding Window Min$^\star$ | SW | Token-level Entropy |
| Bottom-10% Logprob | B10LP | Token-level Entropy |
| Prob Penalty | PP | Token-level Entropy |
| Predictive Entropy | PE | Token-level Entropy |
| Length-Normalized PE | LNPE | Token-level Entropy |

Table 4: Full candidate feature pool. The four features retained after TrustScore-based selection are marked with $\star$.

## C.2 TrustScore-Based Proxy Selection Diagnostics

To further examine whether the proxy selection in Section 4.2 is sensitive to the operationalization of TrustScore, we implement a TrustScore-style diagnostic following the distance-ratio intuition of Jiang et al. (2018). For each scalar proxy, we fit a one-dimensional threshold classifier in a five-fold out-of-fold protocol, construct class-wise high-density sets from the training fold, and compute the TrustScore ratio for held-out samples. We then rank candidate proxies within each family by the error precision among the lowest-confidence top 5% samples. We report results for both $\alpha = 0$ (no density filtering) and $\alpha = 0.1$ (removing the lowest-density 10% of each class).

**Experimental settings.** For each dataset and each proxy feature, we use correctness labels $c \in \{0, 1\}$, where 1 denotes a correct response. The feature is first oriented on the training fold so that larger values indicate higher correctness likelihood. We then fit a one-dimensional threshold classifier by maximizing Youden's $J$ statistic on the training fold. Following the TrustScore procedure, we build separate high-density sets for correct and incorrect samples using the $k$th-nearest-neighbor radius with $k = 10$. For a held-out instance $x$, the TrustScore is computed as the ratio between the distance from $x$ to the nearest high-density set of the non-predicted class and the distance from $x$ to the high-density set of the predicted class. We use five stratified folds, aggregate out-of-fold scores, and flag the lowest-confidence 5% samples as suspicious. The reported precision and coverage are computed against the true incorrect responses in this flagged subset.

The TrustScore-style diagnostic preserves the TriviaQA-selected proxy set (VC, PTrue, SE, SW) under both density-filtering settings. On NQ and GPQA, the selected representative within the consistency-based or token-level family can change with the dataset and the density-filtering parameter, while VC and PTrue remain the selected representatives of their single-proxy families. These results support the family-level selection principle used in our framework and indicate that the precise within-family representative can be dataset-sensitive.

## C.3 Dataset Construction

Table 6 reports the answer accuracy of each base model across the three evaluation datasets for which repeated statistical runs are reported.

For each dataset–model combination, each question is answered once to produce the response used for token-level feature extraction. Correctness is judged by exact-match after lower-casing and punctuation stripping. To control class imbalance without distorting the natural difficulty distribution, we apply stratified subsampling capped at 2,000 instances per dataset–model pair. Instances with missing values in any of the nine features are excluded prior to training.

## C.4 Generation & Confidence Prompt

Each question is submitted with the following prompt to elicit a joint answer–confidence response:

```
{question}

Please provide your answer along with
a confidence level (1-10, where 10 is
highest confidence). Please also explain
the reason why you chose this answer.
Format your response as JSON:
{
  "answer": "your answer",
  "confidence": confidence_level,
  "reason": "full_reason"
}
```

| Dataset | $\alpha$ | Verbalized Confidence | Self-Evaluation | Consistency-based | Token-level Entropy |
|---|---|---|---|---|---|
| TriviaQA | 0 | VC (0.176 / 0.092) | PTrue (0.463 / 0.244) | SE (0.142 / 0.074) | SW (0.297 / 0.156) |
| TriviaQA | 0.1 | VC (0.248 / 0.130) | PTrue (0.340 / 0.179) | SE (0.142 / 0.074) | SW (0.345 / 0.180) |
| NQ | 0 | VC (0.643 / 0.068) | PTrue (0.874 / 0.092) | SE (0.449 / 0.047) | LNPE (0.624 / 0.066) |
| NQ | 0.1 | VC (0.788 / 0.083) | PTrue (0.853 / 0.090) | SE (0.494 / 0.052) | LNPE (0.675 / 0.071) |
| GPQA | 0 | VC (0.661 / 0.057) | PTrue (0.949 / 0.095) | SE (0.598 / 0.051) | SW (0.651 / 0.056) |
| GPQA | 0.1 | VC (0.736 / 0.063) | PTrue (0.819 / 0.082) | SC (0.942 / 0.081) | PE (0.756 / 0.065) |

Table 5: TrustScore-style within-family proxy selection. Each cell reports the selected proxy and its top-5% suspicious-sample error precision / error coverage.

| Dataset | Domain | GPT-4.1 | DeepSeek-V3 | Llama3-70B |
|---|---|---|---|---|
| TriviaQA | Common Sense | 90.4% | 86.3% | 87.1% |
| NQ | Natural Science | 52.5% | 50.0% | 46.4% |
| GPQA | Graduate-Level Science | 61.2% | 41.7% | 47.5% |

Table 6: Answer accuracy of each base model across datasets.

Inference is run with `temperature` $= 0.6$, `max_tokens` $= 4{,}000$, `logprobs` $=$ `True`, and `top_logprobs` $= 5$. The returned per-token log-probabilities are used directly to compute all five token-level entropy features (SW, B10LP, PP, PE, LNPE). For sampling-based features (SC and SE), the same prompt is issued $N = 8$ times per question under identical inference settings, yielding eight independent completions.

### C.5 Verification Prompt and PTrue Extraction

PTrue is obtained by a separate verification call in which the model judges the correctness of its own answer:

```
Question: {question}

Proposed Answer: {answer}

Is the proposed answer true or false?
The answer is:
```

The call is made with `temperature` $= 0.0$, `max_tokens` $= 1$, `logprobs` $=$ `True`, and `top_logprobs` $= 20$, so that only the first decoded token's distribution is returned. The PTrue score is then

$$P_{\text{PTrue}} = \frac{\exp(\ell_{\text{True}})}{\exp(\ell_{\text{True}}) + \exp(\ell_{\text{False}})},$$

where $\ell_{\text{True}}$ and $\ell_{\text{False}}$ are the log-probabilities of the tokens "`True`" and "`False`" in the top-20 candidates. Instances for which neither token appears in the top-20 are discarded.

### C.6 Semantic Entropy Computation

Semantic entropy is computed over the $N = 8$ sampled completions per question. Two responses are assigned to the same semantic cluster when a bidirectional NLI check (`deberta-large-mnli`) finds that they mutually entail each other; otherwise they form separate clusters. Entropy over the resulting cluster distribution is

$$\text{SE} = -\sum_{c} p(c) \log p(c),$$

where $p(c) = |\{i : r_i \in c\}|/N$, and the value is normalized by $\log N$ to map to $[0, 1]$.

### C.7 SAR Implementation Note

SAR (Ye et al., 2023) requires a semantic similarity function $g(\cdot, \cdot)$ for both token and sentence relevance weighting. The original paper uses a cross-encoder; we substitute a Sentence-Transformer model (`all-MiniLM-L6-v2`) with $g(s_i, s_j) = (\cos(\mathbf{e}_i, \mathbf{e}_j) + 1)/2$, mapping cosine similarity to $[0, 1]$. Token-level uncertainty is restricted to the answer span parsed from the JSON response; if extraction fails, the full completion is used as fallback. All other aspects of the SAR formulation are unchanged.

### C.8 OOD Transfer Protocol

We follow a Train-on-Source, Test-on-Target paradigm for cross-dataset OOD evaluation. For each source dataset among TriviaQA, NQ, and GPQA, the meta-estimator is trained on the source dataset and evaluated on each remaining target dataset. The source-domain data is split 80/20 into training and validation sets, stratified by correctness label. The meta-estimator is selected on the validation set and then applied directly to the full target dataset without any target-domain supervision.

### C.9 GBDT Training Protocol

The meta-estimator is the scikit-learn gradient boosting classifier. Hyperparameters are chosen by randomized search ($n_{\text{iter}} = 20$, scored by AUROC) over the grid shown in Table 7. Cross-validation uses 3-fold stratified group splitting with question IDs as groups, preventing responses to the same question from appearing in both the training and validation folds simultaneously.

| Hyperparameter | Candidates |
|---|---|
| `n_estimators` | 50, 100, 200 |
| `learning_rate` | 0.01, 0.05, 0.1 |
| `max_depth` | 3, 4, 5 |
| `subsample` | 0.8, 0.9, 1.0 |

Table 7: Hyperparameter search grid for GBDT.

The search-and-evaluation procedure is repeated 20 times with seeds $42+1$ through $42+20$; the run with the highest held-out AUROC is reported. The held-out test set is 20% of the data, split at the question-group level to apply the same no-leakage constraint as cross-validation.

### C.10 Additional Ablation Results

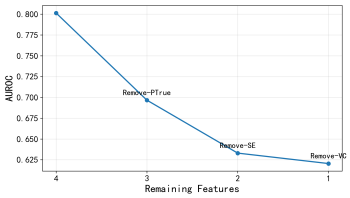



Figure 5: Iterative feature-elimination ablation on TriviaQA / GPT-4.1.

### C.11 Why GBDT as Meta-Estimator

The main text reports the compact comparison among LR, LR with rank normalization, and GBDT. Here we provide the full diagnostic analysis, including marginal feature distributions, bootstrap significance testing, and additional averaging baselines. Rank normalization is fit only on the training fold and then applied to validation/test instances, so it does not use target labels or test-set statistics. All methods share the same four core proxies, group-aware cross-validation, isotonic calibration, and the same randomized search budget; we report mean $\pm$ std over 5 random seeds, with significance assessed by paired bootstrap 95% CI.

Figure 6 shows the standardized marginal distributions of the four core proxies. The proxies are strongly non-Gaussian: *verbal_confidence* exhibits kurtosis $95.86$, *ptrue* exhibits kurtosis $12.20$ with mass piling at the upper boundary and a heavy left tail. This property persists under alternative normalization schemes, indicating it is intrinsic to the proxy signals rather than an artifact of scaling. It directly violates the scale-homogeneous contribution assumption on which logistic regression (LR) relies as a meta-estimator.

To test whether the skewed marginals account for LR's underperformance, we keep the LR classifier unchanged and apply rank normalization to the input features, mapping each proxy's empirical distribution to a Uniform reference via its empirical CDF estimated on the training fold. Relative to LR, LR with rank normalization yields an AUROC improvement of $+0.031$ to $+0.033$ (95% CI excludes $0$, significant), and is statistically indistinguishable from GBDT on AUROC, AUPRC, and Brier (GBDT $-$ LR with rank normalization has an AUROC $\Delta = -0.007$ with CI including $0$; Table 8). The LR–GBDT gap thus traces back to the skewed marginals of the proxy features, not to the absence of useful signal in these proxies.

| Method | AUROC ↑ | AUPRC ↑ | Brier ↓ |
|---|---|---|---|
| UniformAvg | $0.787 \pm 0.038$ | $0.945 \pm 0.013$ | 0.097 |
| WeightedAvg | $0.805 \pm 0.031$ | $0.951 \pm 0.010$ | 0.097 |
| LR | $0.789 \pm 0.035$ | $0.947 \pm 0.013$ | 0.099 |
| LR + rank normalization | $\mathbf{0.822 \pm 0.021}$ | $\mathbf{0.958 \pm 0.006}$ | **0.095** |
| GBDT (ours) | $0.815 \pm 0.029$ | $0.956 \pm 0.010$ | **0.095** |

Table 8: Full meta-estimator comparison on TriviaQA / GPT-4.1, 5 seeds.

**Takeaway.** Taken together, the results indicate that correcting marginal skew is important. We therefore interpret LR with rank normalization as a diagnostic upper-bound control rather than the final estimator: it shows that LR can become competitive once each proxy marginal is adaptively reshaped, but this requires an empirical-CDF mapping tied to the source training distribution. This mapping is not test leakage, yet it is an additional source-distribution-specific transformation that may change when the dataset, base model, or deployment distribution changes. In addition, rank normalization is feature-wise and monotonic; it improves LR by correcting marginal skew but does not explicitly model interactions among proxies. GBDT attains statistically comparable performance without a separate rank-mapping preprocessing layer and directly learns local thresholds and cross-proxy interactions within the estimator

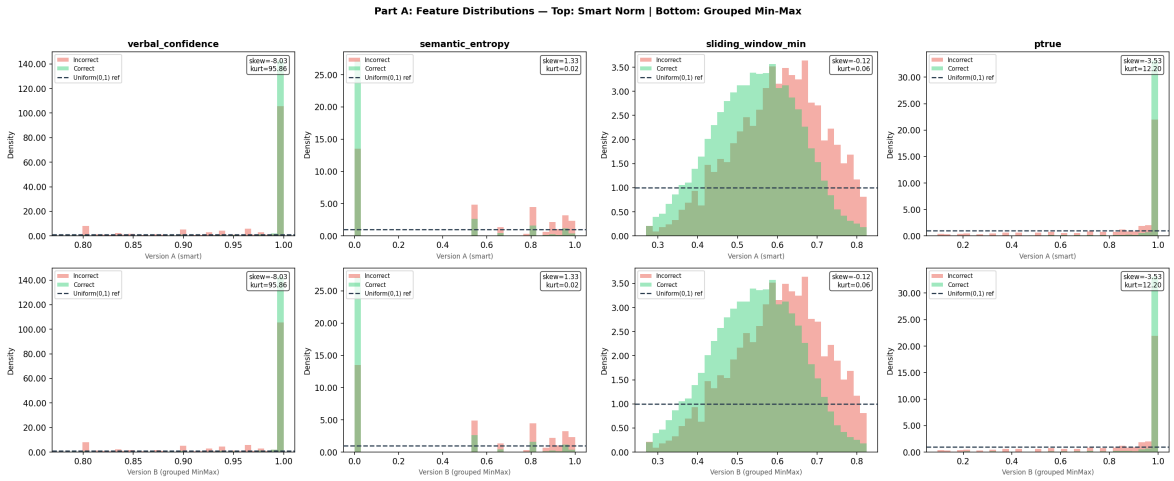



Figure 6: Standardized marginal distributions of the four core proxies on TriviaQA / GPT-4.1 (Correct vs. Incorrect overlaid), annotated with skewness and kurtosis. Within each panel, the horizontal axis shows the standardized proxy value and the vertical axis shows the estimated density.

itself, which is more consistent with our ensemble motivation.

# D Main results with statistical significance

| Base Model | Feature | TriviaQA | | | Natural Questions | | | GPQA | | | HotpotQA | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AUROC | AUPRC | Brier | AUROC | AUPRC | Brier | AUROC | AUPRC | Brier | AUROC | AUPRC | Brier |
| GPT-4.1 | SW | 0.7436±0.0018 | 0.9545±0.0051 | 0.0777±0.0006 | 0.6648±0.0274 | 0.6712±0.0255 | 0.2305±0.0072 | 0.6796±0.0314 | 0.6370±0.0394 | 0.2301±0.0076 | 0.7128 | 0.8674 | 0.1056 |
| | VC | 0.6846±0.0049 | 0.9382±0.0017 | 0.0781±0.0021 | 0.6510±0.0262 | 0.6186±0.0211 | 0.2276±0.0068 | 0.6253±0.0344 | 0.6095±0.0378 | 0.2394±0.0051 | 0.6657 | 0.9108 | 0.1785 |
| | SE | 0.7117±0.0291 | 0.9440±0.0058 | 0.0772±0.0016 | 0.6074±0.0250 | 0.5996±0.0192 | 0.2292±0.0065 | 0.5962±0.0364 | 0.5629±0.0273 | 0.2501±0.0077 | 0.6651 | 0.9203 | 0.1213 |
| | PTrue | 0.5236±0.0089 | 0.8901±0.0009 | 0.0869±0.0012 | 0.5376±0.0258 | 0.5340±0.0193 | 0.2472±0.0025 | 0.4898±0.0325 | 0.4835±0.0290 | 0.2510±0.0014 | 0.5319 | 0.8653 | 0.1654 |
| | BSDetector | 0.6385±0.0113 | 0.9243±0.0031 | 0.1062±0.0014 | 0.6407±0.0120 | 0.6323±0.0125 | 0.2834±0.0061 | 0.5959±0.0187 | 0.5698±0.0226 | 0.3705±0.0091 | 0.6179 | 0.9007 | 0.1652 |
| | SAR | 0.7706±0.0074 | 0.9621±0.0032 | 0.0776±0.0019 | 0.6763±0.0169 | 0.6778±0.0191 | 0.2279±0.0065 | 0.7047±0.0253 | 0.6386±0.0338 | 0.2255±0.0088 | 0.7154 | 0.8954 | 0.1748 |
| | SIMBA UQ | 0.8009±0.0002 | 0.9654±0.0002 | 0.0687±0.0001 | 0.7024±0.0020 | 0.6930±0.0020 | 0.2184±0.0006 | 0.6915±0.0041 | 0.6979±0.0073 | 0.2261±0.0012 | 0.8013 | 0.9513 | 0.0973 |
| | GBDT (ours) | 0.8230±0.0036 | 0.9693±0.0036 | 0.0690±0.0034 | 0.7089±0.0232 | 0.6979±0.0262 | 0.2181±0.0074 | 0.6901±0.0322 | 0.7172±0.0451 | 0.2262±0.0091 | 0.8137 | 0.9643 | 0.0998 |
| DeepSeek-V3 | SW | 0.6513±0.0475 | 0.9078±0.0108 | 0.1132±0.0018 | 0.5706±0.0181 | 0.5447±0.0157 | 0.2462±0.0022 | 0.5580±0.0025 | 0.6501±0.0056 | 0.2362±0.0005 | 0.6719 | 0.8847 | 0.2437 |
| | VC | 0.6063±0.0648 | 0.8898±0.0150 | 0.1142±0.0093 | 0.5740±0.0167 | 0.5380±0.0122 | 0.2432±0.0031 | 0.6681±0.0049 | 0.7394±0.0126 | 0.2208±0.0032 | 0.6145 | 0.8321 | 0.2741 |
| | SE | 0.7141±0.0109 | 0.9207±0.0023 | 0.1045±0.0000 | 0.6218±0.0183 | 0.5886±0.0171 | 0.2384±0.0041 | 0.5915±0.0248 | 0.6779±0.0269 | 0.2324±0.0022 | 0.7188 | 0.8906 | 0.1996 |
| | PTrue | 0.7311±0.0211 | 0.9385±0.0063 | 0.1151±0.0041 | 0.6535±0.0199 | 0.6070±0.0197 | 0.2402±0.0020 | 0.4876±0.0050 | 0.6326±0.0092 | 0.2380±0.0004 | 0.6953 | 0.8811 | 0.2014 |
| | BSDetector | 0.7917±0.0126 | 0.9568±0.0036 | 0.1069±0.0028 | 0.6508±0.0120 | 0.6447±0.0142 | 0.2991±0.0055 | 0.5668±0.0164 | 0.6613±0.0177 | 0.2655±0.0067 | 0.7356 | 0.9062 | 0.2689 |
| | SAR | 0.7963±0.0140 | 0.9572±0.0026 | 0.1008±0.0046 | 0.6581±0.0157 | 0.6566±0.0176 | 0.2328±0.0054 | 0.6290±0.0198 | 0.7411±0.0146 | 0.2267±0.0062 | 0.8013 | 0.9415 | 0.1945 |
| | SIMBA UQ | 0.8087±0.0020 | 0.9508±0.0013 | 0.0933±0.0002 | 0.6804±0.0026 | 0.6434±0.0018 | 0.2248±0.0007 | 0.7195±0.0012 | 0.8069±0.0011 | 0.2091±0.0004 | 0.8157 | 0.9601 | 0.0959 |
| | GBDT (ours) | 0.8197±0.0153 | 0.9629±0.0015 | 0.0941±0.0064 | 0.6967±0.0194 | 0.6701±0.0210 | 0.2224±0.0052 | 0.7400±0.0232 | 0.8307±0.0115 | 0.1990±0.0056 | 0.8099 | 0.9653 | 0.0751 |
| LLaMA3-70B | SW | 0.6095±0.0257 | 0.9028±0.0106 | 0.1086±0.0017 | 0.5637±0.0194 | 0.5307±0.0227 | 0.2459±0.0020 | 0.5104±0.0360 | 0.4781±0.0298 | 0.2500±0.0008 | 0.6675 | 0.5191 | 0.2170 |
| | VC | 0.7391±0.0304 | 0.9321±0.0084 | 0.0955±0.0037 | 0.6623±0.0226 | 0.5848±0.0217 | 0.2260±0.0052 | 0.6090±0.0416 | 0.5616±0.0307 | 0.2350±0.0091 | 0.7481 | 0.6452 | 0.1796 |
| | SE | 0.6265±0.0285 | 0.9023±0.0075 | 0.1013±0.0045 | 0.6147±0.0196 | 0.5452±0.0151 | 0.2360±0.0048 | 0.5791±0.0327 | 0.5420±0.0220 | 0.2440±0.0057 | 0.5600 | 0.4053 | 0.2375 |
| | PTrue | 0.6728±0.0365 | 0.9158±0.0096 | 0.1053±0.0030 | 0.5766±0.0264 | 0.5176±0.0167 | 0.2471±0.0021 | 0.6475±0.0401 | 0.6210±0.0418 | 0.2423±0.0053 | 0.6669 | 0.4657 | 0.2118 |
| | BSDetector | 0.7335±0.0168 | 0.9319±0.0047 | 0.1059±0.0026 | 0.6506±0.0121 | 0.5915±0.0116 | 0.3716±0.0058 | 0.6718±0.0159 | 0.6390±0.0180 | 0.3396±0.0097 | 0.6959 | 0.4948 | 0.4538 |
| | SAR | 0.7408±0.0229 | 0.9392±0.0064 | 0.1036±0.0038 | 0.6728±0.0177 | 0.6367±0.0190 | 0.2333±0.0059 | 0.6644±0.0323 | 0.6464±0.0350 | 0.2402±0.0114 | 0.7259 | 0.5762 | 0.2080 |
| | SIMBA UQ | 0.7732±0.0038 | 0.9387±0.0019 | 0.0924±0.0001 | 0.6891±0.0016 | 0.6298±0.0000 | 0.2223±0.0004 | 0.6832±0.0031 | 0.6393±0.0032 | 0.2246±0.0007 | 0.7823 | 0.6899 | 0.1821 |
| | GBDT (ours) | 0.7841±0.0289 | 0.9467±0.0081 | 0.0921±0.0049 | 0.6929±0.0233 | 0.6370±0.0263 | 0.2247±0.0061 | 0.6769±0.0486 | 0.6467±0.0520 | 0.2293±0.0131 | 0.7945 | 0.6706 | 0.1784 |

Table 9: Performance comparison across datasets. Results are reported as mean $\pm$ standard deviation over 5 independent random seeds when available. For HotpotQA, we report the point estimates from the main results table. Each seed corresponds to an independent 80/20 stratified group split; the reported run is the one achieving the highest held-out AUROC across 20 repeated trials.