# Towards Reliable Truth-Aligned Uncertainty Estimation in Large Language Models

**Ponhvoan Srey**[1] **Quang Minh Nguyen**[2] **Xiaobao Wu**[3*] **Anh Tuan Luu**[1,4*]
[1]Nanyang Technological University [2]KAIST
[3]Shanghai Jiao Tong University [4]VinUniversity
{ponhvoan002, anhtuan.luu}@ntu.edu.sg
qm.nguyen@kaist.ac.kr xiaobaowu@sjtu.edu.cn

## Abstract

Uncertainty estimation (UE) aims to quantify uncertainty in large language model (LLM) outputs to flag potential hallucinations and support reliable deployment. However, UE metrics often exhibit unstable performance across models and datasets, limiting their practical utility. We formalise this phenomenon as *proxy failure*, arising because UE metrics are derived from model behaviour rather than being explicitly grounded in factual correctness. We show that when mutual information between a UE score and correctness is low, the score's discriminability must be close to chance. To address this, we propose *Truth AnChoring* (TAC), a lightweight post-hoc calibration method that maps raw UE scores to truth-aligned correctness probabilities. Even with noisy and few-shot supervision, TAC continues to improve calibration, providing a practical calibration protocol. Our findings highlight the limitations of treating heuristic UE metrics as direct indicators of uncertainty about factual correctness, and position our TAC as a principled step toward more reliable uncertainty estimation for LLMs.[†]

## 1 Introduction

Large Language Models (LLMs) have advanced rapidly in capability (Achiam et al., 2023; Guo et al., 2025; Yang et al., 2025) and are being deployed across a broad range of applications, including in high-risk domains, such as the medical (Nori et al., 2023; Thirunavukarasu et al., 2023; Goh et al., 2024) and financial fields (Wu et al., 2023; Lopez-Lira and Tang, 2023). Yet, they still generate falsehoods or hallucinate with high fluency and apparent confidence, making errors difficult to detect from the generations alone (Zhang et al., 2025). This motivates uncertainty estimation (UE). UE aims to assign an uncertainty score to each response, and flag potential hallucinations. This enables appropriate and timely interventions on the response, like abstention or verification (Liu et al., 2025).

However, recent studies show that UE metrics do not behave stably across configurations, with scores that appear effective in one setting degrading sharply in another (Vashurin et al., 2025; Wang et al., 2025a; Shorinwa et al., 2025; Tan et al., 2025; Tomov et al., 2025). We argue that this is because these UE metrics are not directly measuring correctness. For example, predictive entropy quantifies uncertainty of the next-token distribution, and therefore, reflects local continuation stability rather than semantic truth. In the same manner, many other UE metrics function as heuristics or *proxies* for correctness as they may correlate with truth in some contexts, but that relationship is fragile.

In this work, we present a formal study of this phenomenon, which we call ***proxy failure***. Our contributions are three-fold: (i) We demonstrate explicitly the non-discriminability of UE scores under low information regimes, and highlight how this can manifest in entropy as a concrete example. (ii) We can alleviate this issue through *Truth AnChoring (TAC)*. By connecting uncertainty scores to correctness labels, they can be post-hoc calibrated to align with truth, yielding significant reductions in the expected calibration error (ECE) (Figure 1). (iii) Finally, we provide a practical route for when gold annotations are scarce. Even weak truth signals, such as few-shot and noisy supervision, are still useful for learning truth-aligned uncertainty estimates. More broadly, our results caution that reliable UE cannot rest on heuristics alone. We position TAC as a crucial calibration layer over existing UE scores to inject an explicit notion of correctness to align them with factuality. With these contributions, we hope to steer the field toward more principled post-hoc calibration.

*Corresponding Authors.
†The code repository is available at https://github.com/ponhvoan/TruthAnchor/.



*Findings of the Association for Computational Linguistics: EMNLP 2026*, pages 31306–31321
October 24-29, 2026 ©2026 Association for Computational Linguistics

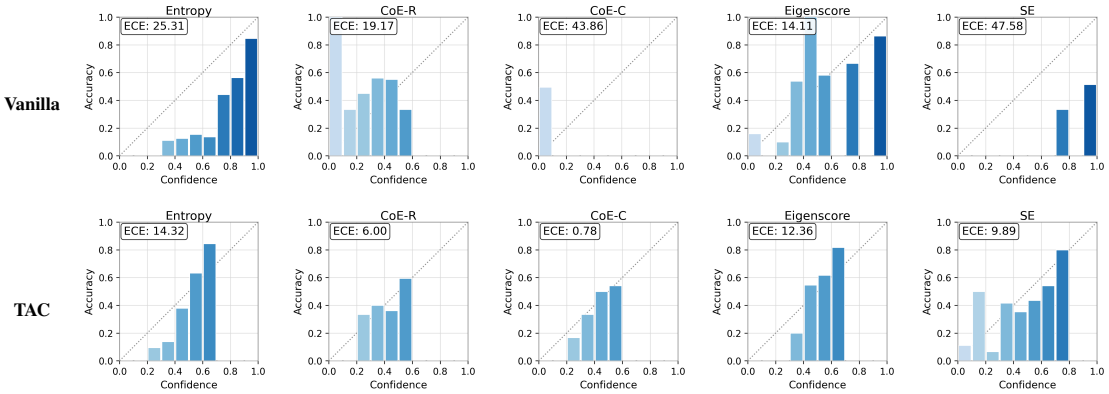


Figure 1: Reliability diagrams of widely used and recent uncertainty methods, and our Truth-AnChored (TAC) scores.

## 2 Related Work

Existing UE methods typically derive uncertainty scores from the LLM's own behaviour. Popular approaches examine the next-token probability distribution, and compute the uncertainty of this underlying distribution via measures such as entropy (Huang et al., 2023), perplexity (Si et al., 2023), or their variants (Malinin and Gales, 2021; Shih et al., 2023). Others rely on response consistency across multiple samples, both at the output token level (Manakul et al., 2023; Kuhn et al., 2023; Lin et al., 2024) and representation level (Chen et al., 2024; Li et al., 2026), with the assumption that if the model emits equivalent responses, then it has sufficient parametric knowledge and has generated a correct answer. Others avoid sampling altogether due to its computational requirements, and extract signals from internal representations and formulate a measure to track uncertainty (Wang et al., 2025b; Sriramanan et al., 2024; Yin et al., 2024). Evidently, the prevailing paradigm is to formulate a score that should, in principle, track factuality. Unfortunately, as noted in recent works (Vashurin et al., 2025; Shorinwa et al., 2025), performance of such metrics varies across datasets. Our work reveals these metrics can only serve as imperfect proxies for truth, and can fail to meaningfully represent uncertainty in general.

Closely related to our work are probe-based UE methods that train a lightweight classifier on LLM internal states to predict hallucination risk, resulting in strong uncertainty estimates (Azaria and Mitchell, 2023; Burns et al., 2023; Ji et al., 2024; Srey et al., 2025). Most efforts are focused on curating more informative features, such as Mahalanobis distance of embeddings (Vazhentsev et al., 2025), lookback ratios over context (Chuang et al., 2024), and attention activation maps (He et al., 2024). Our work contributes to and differs from this suite of methods in two ways. First, we establish that probe-based UE is principled and well-grounded, thus explaining their superior empirical performance. Second, we emphasise a minimal and previously overlooked instantiation: post-hoc calibration of the raw uncertainty score itself. Most pertinent to our work is CUE (Li et al., 2025), which learns a corrector to adjust uncertainty scores using correctness supervision with additional BERT-based embedding features. Our work is complementary but distinct. We provide a formal account of *why* such correction is needed, and empirically demonstrate even the scalar proxy alone can be turned to a truth-aligned estimator.

## 3 Proxy Failure of Uncertainty Scores

We now formulate the proxy failure phenomenon studied in this work. Our goal is to answer the following question: *When does a proxy uncertainty score fail to meaningfully discriminate between correct and incorrect generations?* With Proposition 1, we establish that this arises whenever the uncertainty score carries little information about correctness.

**Notation.** Given a query $X = x$ drawn from $\mathcal{X}$, a language model induces a conditional distribution $P(\cdot \mid x)$ over responses $Y \in \mathcal{Y}$. We assume a fixed truth function $g^\star : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$, and define correctness label $C = g^\star(X, Y)$, with $C = 1$ indicating correct response. We write $p_c = P(C = 1)$ for the correctness prior. Here, we consider the confidence score $S(X, Y)$, which we simply take as the inverse of uncertainty. For any real-valued score

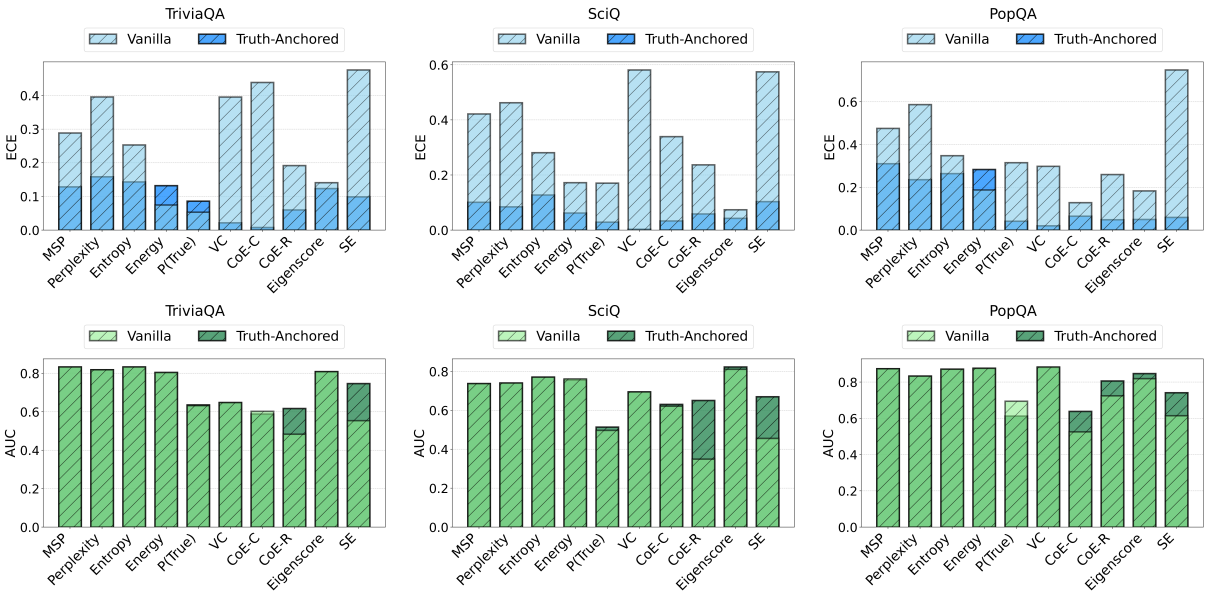


Figure 2: Our proposed TAC significantly reduces ECE (↓) and frequently increases AUC (↑), improving calibration and discrimination.

$S$, define the population Area Under the receiver-operating Characteristic curve (AUC) (Davis and Goadrich, 2006) as

$$\mathrm{AUC}(S) = P(S_+ > S_-) \tag{1}$$

where $S_+ \sim P_+ = P(S \mid C = 1)$ and $S_- \sim P_- = P(S \mid C = 0)$. AUC is our primary measure of discriminability, where $AUC(S) = 1/2$ corresponds to chance ranking.

### 3.1 Mutual Information Controls Discriminability.

Proposition 1 formalises the intuition that *if an uncertainty score carries little information about correctness, then its discriminative ability must be close to chance.*

**Proposition 1.** *Let $S$ be any real-valued score with continuous class-conditional distributions, and $p_c = P(C = 1) \in (0, 1)$. Then,*

$$\left|\mathrm{AUC(S)} - \frac{1}{2}\right| \leq \sqrt{\frac{I(C, S)}{2p_c(1 - p_c)}}, \tag{2}$$

*where $I$ denotes mutual information. In particular, if $I(C, S)$ is small, then the discriminability of $S$ must be close to chance.*

We defer proofs to Appendix A. Next, we show how this can manifest in entropy as a concrete example.

### 3.2 Non-Discriminability of Entropy

Predictive entropy illustrates proxy failure clearly. A model may commit early to an incorrect concept, yet continue to answer fluently with low local uncertainty. A correct continuation can exhibit an almost identical entropy pattern. Entropy, thus, is sensitive to local continuation structure rather than semantic truth, a global characteristic of the entire sequence. The consequence is that entropy can carry vanishing information about correctness and can therefore, by Proposition 1, fail to discriminate. Proposition 2 formalises this occurrence.

**Proposition 2.** *Fix a truth function $g^\star$ and any predictive entropy proxy $S_{pe}$. For any $\varepsilon > 0$, there exists a conditional distribution $P(\cdot \mid x)$ such that*

$$I(C, S_{pe}) \leq \varepsilon \tag{3}$$

*Consequently, by Proposition 1,*

$$\left|\mathrm{AUC}(S_{pe}) - \frac{1}{2}\right| \leq \sqrt{\frac{\varepsilon}{2p_c(1 - p_c)}} \tag{4}$$

*In particular, the AUC of predictive entropy can be made arbitrarily close to chance.*

*Proof Idea.* We provide a proof by construction. For each query, we consider a setting in which the model selects between one correct and one incorrect response that diverge only at a key semantic commitment, while the remaining tokens are routine continuations. This makes the entropy profiles indistinguishable despite different correctness labels. We construct our required distribution $P$ to be a mixture of this setting and an arbitrary LLM. Under $P$, we show that mutual information between predictive entropy and correctness vanishes to zero.

## 4 Truth Anchoring as a Remedy

We have demonstrated that proxy metrics can fail as they may only reflect model behaviour but carry little information about global semantic correctness. This suggests that reliable uncertainty estimation cannot rely on proxies alone. Instead, we argue that uncertainty estimation should explicitly target, and therefore, be *anchored* to, a notion of truth or correctness.

Given an information variable $\mathcal{I}$ available at inference time, the natural target is the posterior probability of correctness $p^\star(\mathcal{I}) = P(C = 1 \mid \mathcal{I})$. Under a strictly proper scoring rule $\ell$, the *truth-anchored* posterior $p^\star$ is exactly the Bayes optimal predictor. That is, among all predictors $q(\mathcal{I}) \in [0, 1]$, it uniquely minimises the population risk $\mathcal{R}(q) = \mathbb{E}\left[\ell(q(\mathcal{I}), C)\right]$ (Gneiting and Raftery, 2007). Thus, rather than treating a raw proxy score itself as uncertainty, we aim to learn this posterior probability that is optimally aligned with truth under a chosen scoring rule. We call the learning of this mapping from raw scalar to truth-aligned score ***Truth AnChoring (TAC)***. This is consistent with converging evidence that probe-based methods showcase superior performance.

### 4.1 Proxy Score Truth Anchoring

A particularly simple instantiation arises when the available information is just the scalar uncertainty score $S$, such as predictive entropy or response consistency. In this case, the Bayes-optimal target reduces to $p^\star(S) = P(C = 1|S)$, namely, the posterior correctness conditioned on the score itself. This suggests a minimal post-hoc calibration strategy. Rather than using $S$ directly as uncertainty with respect to the truth, we learn a mapping

$$m : \mathbb{R} \to [0, 1] \tag{5}$$

from the raw score to its truth-aligned estimate.

### 4.2 Implementation

**Mapper architecture.** We instantiate $m_\theta$ as a lightweight multilayer perceptron (MLP) operating on the one-dimensional input $S$. Concretely, the mapper consists of an input layer, three hidden layers with ReLU activations, and a scalar output head followed by a sigmoid function. From a raw score $S_i$, the mapper produces $\hat{p}_i = m_\theta(S_i) \in (0, 1)$. In practice, a small network is sufficiently capable as the input is one-dimensional. In our experiments, we use a hidden dimension of 32 with three hidden layers, and optimise with Adam using a learning rate of 0.01, with early stopping on validation AUC.

**Objective.** Given score–label pairs $\{S_i, C_i\}_{i=1}^n$, we optimise the binary cross-entropy (BCE) loss as the main objective,

$$\begin{aligned}\mathcal{L}_{\text{BCE}} &= \frac{1}{n}\sum_{i=1}^{n} \ell\big(m_\theta(S_i),\, C_i\big) \\ &= -\frac{1}{n}\sum_{i=1}^{n}\Big[C_i \log m_\theta(S_i) \\ &\quad + (1 - C_i)\log\big(1 - m_\theta(S_i)\big)\Big]\end{aligned} \tag{6}$$

$\mathcal{L}_{\text{BCE}}$ is a strictly proper scoring loss, so its population optimum is the correctness posterior $p^\star(S)$.

Since a primary evaluation metric is AUC, we also consider augmenting the objective with an auxiliary pairwise ranking loss, which is widely used to improve ordering quality and is closely connected to AUC-oriented learning (Cao et al., 2007). Specifically, for positive–negative pairs $(i, j)$ with $C_i = 1$ and $C_j = 0$, we use a logistic pairwise loss

$$\mathcal{L}_{\text{rank}}(\theta) = \frac{1}{|\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}} \log\Big(1 + \exp\big(-(z_i - z_j)\big)\Big), \tag{7}$$

where $z_i$ and $z_j$ are pre-sigmoid logits of the mapper and $\mathcal{P} = \{(i, j) : C_i = 1,\, C_j = 0\}$ is the set of positive–negative pairs. Intuitively, this term encourages correct responses to receive higher scores than incorrect ones.

Our final objective is

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{BCE}}(\theta) + \phi_{\text{rank}}\,\mathcal{L}_{\text{rank}}(\theta), \tag{8}$$

where $\phi_{\text{rank}}$ controls the strength of the ranking regulariser. By default, we use $\phi_{\text{rank}} = 1$. We perform sensitivity analysis of $\phi_{\text{rank}}$ in Section 6.7.

We note that the ranking term is auxiliary. Unlike BCE, it is not a strictly proper scoring rule, and does not by itself target the correctness posterior. We include it to better align training with AUC. Collectively, the BCE term provides probabilistic grounding, while the pairwise term can help sharpen ranking when raw proxy is informative but poorly ordered.

| Method | TriviaQA | | | | | | SciQ | | | | | | PopQA | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ECE (↓) | | | AUC (↑) | | | ECE (↓) | | | AUC (↑) | | | ECE (↓) | | | AUC (↑) | | |
| | Vanilla | TAC | Δ | Vanilla | TAC | Δ | Vanilla | TAC | Δ | Vanilla | TAC | Δ | Vanilla | TAC | Δ | Vanilla | TAC | Δ |
| **Qwen-3-4B** | | | | | | | | | | | | | | | | | | |
| MSP | 63.68 | 5.52 | -58.16 | 73.82 | 73.58 | -0.24 | 55.97 | 3.89 | -52.08 | 70.07 | 70.02 | -0.05 | 80.21 | 6.07 | -74.13 | 33.02 | 69.14 | 36.12 |
| Perplexity | 62.71 | 6.12 | -56.59 | 73.60 | 73.48 | -0.12 | 51.53 | 9.95 | -41.58 | 69.98 | 69.98 | 0.00 | 80.03 | 4.65 | -75.38 | 30.43 | 61.09 | 30.66 |
| Entropy | 53.45 | 5.97 | -47.47 | 73.51 | 73.87 | 0.37 | 42.20 | 2.05 | -40.15 | 69.04 | 67.19 | -1.85 | 72.57 | 3.70 | -68.87 | 33.93 | 63.13 | 29.20 |
| Energy | 16.97 | 4.05 | -12.91 | 45.89 | 49.65 | 3.75 | 11.61 | 3.31 | -8.30 | 50.25 | 54.63 | 4.39 | 41.70 | 4.20 | -37.51 | 55.28 | 59.13 | 3.85 |
| P(True) | 41.20 | 5.92 | -35.28 | 63.13 | 63.33 | 0.20 | 41.85 | 1.83 | -40.03 | 51.40 | 55.46 | 4.06 | 56.48 | 2.91 | -53.57 | 83.91 | 83.33 | -0.57 |
| VC | 35.48 | 14.58 | -20.90 | 83.27 | 82.79 | -0.48 | 57.45 | 1.34 | -56.11 | 63.15 | 63.71 | 0.56 | 20.64 | 27.54 | 6.90 | 82.18 | 81.87 | -0.31 |
| CoE-C | 8.24 | 8.60 | 0.36 | 70.14 | 70.39 | 0.25 | 3.35 | 7.64 | 4.29 | 53.71 | 54.83 | 1.12 | 20.65 | 3.95 | -16.70 | 74.62 | 76.55 | 1.92 |
| CoE-R | 19.76 | 6.69 | -13.07 | 52.95 | 65.96 | 13.01 | 29.21 | 5.97 | -23.25 | 37.19 | 62.35 | 25.16 | 4.89 | 3.40 | -1.48 | 73.81 | 81.41 | 7.60 |
| Eigenscore | 43.99 | 2.40 | -41.59 | 74.18 | 74.17 | -0.01 | 42.04 | 6.77 | -35.27 | 68.53 | 68.58 | 0.05 | 66.49 | 2.86 | -63.63 | 35.30 | 57.91 | 22.61 |
| SE | 66.82 | 5.27 | -61.55 | 49.77 | 65.13 | 15.37 | 50.65 | 5.00 | -45.65 | 44.31 | 69.38 | 25.07 | 80.22 | 4.20 | -76.03 | 36.77 | 42.18 | 5.40 |
| **Ministral-3-8B** | | | | | | | | | | | | | | | | | | |
| MSP | 28.86 | 12.85 | -16.02 | 83.32 | 83.32 | 0.00 | 42.17 | 10.10 | -32.07 | 73.76 | 73.76 | 0.00 | 47.57 | 31.05 | -16.52 | 87.37 | 87.37 | 0.00 |
| Perplexity | 39.57 | 15.84 | -23.73 | 81.88 | 81.88 | 0.00 | 46.23 | 8.44 | -37.79 | 74.05 | 74.05 | 0.00 | 58.70 | 23.57 | -35.13 | 83.32 | 83.32 | 0.00 |
| Entropy | 25.31 | 14.32 | -10.99 | 83.31 | 83.31 | 0.00 | 28.09 | 12.75 | -15.34 | 77.08 | 77.08 | 0.00 | 34.80 | 26.41 | -8.39 | 87.10 | 87.10 | 0.00 |
| Energy | 7.47 | 13.21 | 5.74 | 80.45 | 80.45 | 0.00 | 17.22 | 6.20 | -11.02 | 76.23 | 75.59 | -0.64 | 18.78 | 28.35 | 9.57 | 87.61 | 87.61 | 0.00 |
| P(True) | 5.33 | 8.59 | 3.27 | 63.21 | 63.58 | 0.37 | 17.06 | 2.90 | -14.16 | 49.73 | 51.38 | 1.65 | 31.52 | 4.15 | -27.37 | 69.46 | 61.27 | -8.19 |
| VC | 39.54 | 2.17 | -37.37 | 64.84 | 64.84 | 0.00 | 58.13 | 0.32 | -57.81 | 69.51 | 69.51 | 0.00 | 29.83 | 2.00 | -27.83 | 88.35 | 88.17 | -0.18 |
| CoE-C | 43.86 | 0.78 | -43.08 | 60.26 | 58.85 | -1.41 | 33.94 | 3.32 | -30.62 | 62.13 | 63.06 | 0.93 | 12.84 | 6.49 | -6.34 | 52.53 | 63.85 | 11.32 |
| CoE-R | 19.17 | 6.00 | -13.17 | 48.34 | 61.70 | 13.35 | 23.68 | 5.85 | -17.84 | 34.92 | 65.08 | 30.16 | 25.98 | 4.86 | -21.12 | 72.44 | 80.63 | 8.19 |
| Eigenscore | 14.11 | 12.36 | -1.75 | 80.90 | 80.90 | 0.00 | 7.39 | 4.27 | -3.12 | 81.08 | 82.29 | 1.21 | 18.35 | 5.00 | -13.35 | 81.92 | 84.70 | 2.79 |
| SE | 47.58 | 9.89 | -37.69 | 55.39 | 74.68 | 19.30 | 57.45 | 10.34 | -47.11 | 45.57 | 67.02 | 21.45 | 74.90 | 5.93 | -68.97 | 61.48 | 74.15 | 12.67 |
| **Gemma-2-9B** | | | | | | | | | | | | | | | | | | |
| MSP | 31.96 | 10.67 | -21.29 | 71.46 | 71.46 | 0.00 | 38.69 | 7.99 | -30.70 | 72.95 | 72.91 | -0.04 | 65.10 | 2.64 | -62.46 | 67.70 | 67.31 | -0.39 |
| Perplexity | 29.55 | 6.22 | -23.33 | 70.71 | 70.14 | -0.57 | 29.03 | 3.37 | -25.65 | 70.40 | 69.45 | -0.95 | 60.10 | 6.65 | -53.44 | 67.97 | 67.97 | 0.00 |
| Entropy | 22.97 | 7.52 | -15.45 | 72.84 | 72.91 | 0.07 | 21.02 | 9.47 | -11.55 | 73.99 | 73.77 | -0.22 | 46.75 | 17.07 | -29.67 | 79.14 | 79.14 | 0.00 |
| Energy | 13.43 | 4.38 | -9.04 | 67.11 | 67.28 | 0.17 | 24.46 | 7.23 | -17.23 | 54.84 | 51.42 | -3.42 | 64.72 | 2.94 | -61.77 | 39.41 | 63.13 | 23.71 |
| P(True) | 31.91 | 2.16 | -29.75 | 60.19 | 60.05 | -0.13 | 38.10 | 2.61 | -35.49 | 48.23 | 48.08 | -0.15 | 66.03 | 18.19 | -47.84 | 81.88 | 81.88 | 0.00 |
| VC | 34.10 | 3.31 | -30.79 | 58.60 | 58.60 | 0.00 | 44.26 | 2.02 | -42.24 | 63.56 | 63.33 | -0.23 | 63.41 | 3.98 | -59.43 | 47.78 | 68.99 | 21.20 |
| CoE-C | 25.20 | 6.39 | -18.82 | 48.72 | 63.24 | 14.52 | 18.16 | 1.22 | -16.94 | 49.03 | 53.22 | 4.19 | 49.59 | 1.91 | -47.68 | 17.52 | 82.21 | 64.69 |
| CoE-R | 9.77 | 4.49 | -5.28 | 49.71 | 60.61 | 10.90 | 7.28 | 2.43 | -4.85 | 49.18 | 53.22 | 4.04 | 44.37 | 5.95 | -38.43 | 18.54 | 81.46 | 62.92 |
| Eigenscore | 12.88 | 5.17 | -7.71 | 70.64 | 70.37 | -0.27 | 14.32 | 5.85 | -8.47 | 70.41 | 70.60 | 0.19 | 38.94 | 1.91 | -37.03 | 59.24 | 61.71 | 2.47 |
| Semantic Entropy | 35.34 | 5.71 | -29.63 | 41.17 | 61.24 | 20.07 | 39.09 | 3.54 | -35.56 | 48.73 | 60.41 | 11.68 | 69.70 | 4.14 | -65.57 | 53.41 | 60.96 | 7.56 |

Table 1: Main results. Δ denotes the change of ECE and AUC after truth anchoring (TAC), with improvements in green.

**Remark.** We emphasise that this is intentionally minimal, with the mapper operating only on the raw uncertainty, without introducing additional model features. As such, it isolates the effect of truth anchoring itself. If performance improves after calibration, this directly indicates that the original proxy was not intrinsically truth-aligned but could be made more useful once connected to correctness. Furthermore, TAC mainly helps with calibration and does not overturn the non-discriminability result of Section 3. If a score carries no correctness information, strong discrimination cannot be recovered by post-hoc mapping of that score alone but may be achieved with more information, *e.g.* through inter-score anchoring in Figure 4.

## 5 Experimental Setup

### 5.1 Language Models

We select five models from four popular model families: **Qwen-3-4B** (Yang et al., 2025), **Llama-3.2-3B**, **Llama-3.1-8B** (Grattafiori et al., 2024), **Ministral-3-8B** (Liu et al., 2026), and **Gemma-2-9B** (Riviere et al., 2024) for our main experiments. We report the performance of the Llama models in Appendix B for better readability. For more detailed experiments and analysis, we primarily utilise Ministral-3-8B. We supplement the experiments with more specialised tasks in Section 6.5 with Gemma-2-9B evaluation. For all models, we work with their instruction-tuned versions.

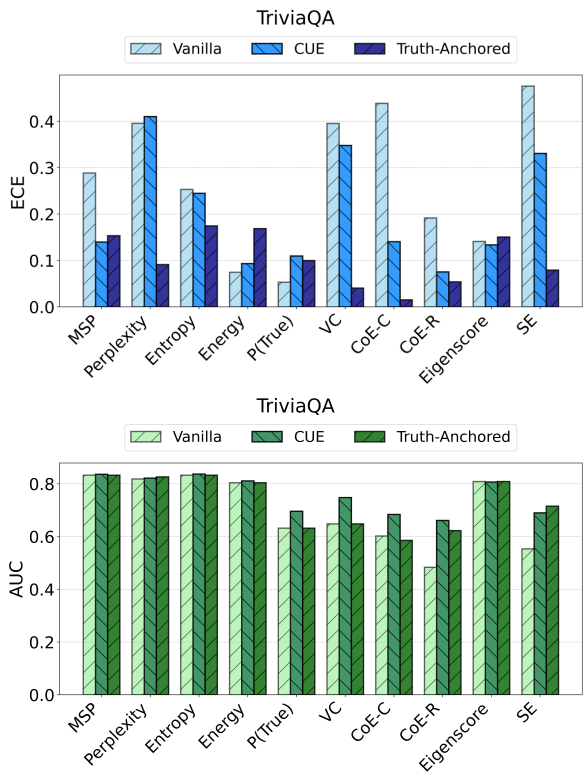


Figure 3: Performance of vanilla, CUE, and truth-anchored scores.

### 5.2 Datasets.

We utilise three representative open-domain question-answering (QA) datasets, covering a wide spectrum of topics: (i) **TriviaQA** (Joshi et al., 2017): broad, general-knowledge factual recall; (ii) **SciQ** (Welbl et al., 2017): domain-specific scientific queries; and (iii) **PopQA** (Mallen et al., 2023): long-tail entity questions. Collectively, these datasets constitute fact-heavy queries across many topics, making them suitable to test factual recall of language models.

### 5.3 Metrics

We evaluate our results using two standard metrics: (i) **AUC** (Davis and Goadrich, 2006): with a value of 1 signifying perfect discriminative performance to rank samples reliably, while a value of 0.5 indicates that estimation is no better than chance; and (ii) **ECE** (Expected Calibration Error; Guo et al., 2017), which assesses the calibration by calculating the expected difference between predicted confidence and empirical accuracy, penalizing over- and under-confidence. In practice, this is computed by partitioning all $n$ confidence scores into $M$ bins and comparing the average bin confidence (conf) of each bin $B_m$ against the ratio of correct predictions (acc). We follow the standard formulation of ECE (Guo et al., 2017) with 10 equal-width bins.

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_M|}{n} |\mathrm{acc}(B_m) - \mathrm{conf}(B_m)| \quad (9)$$

An ECE of 0 means perfect calibration, where the confidence scores exactly match actual empirical accuracy. We note a degenerate case where ECE is zero but AUC is 0.5 (random chance). This happens when a model makes uninformative but perfectly calibrated predictions. Taken together, these two metrics provide a more complete and fairer evaluation of confidence or uncertainty scores.

### 5.4 Baselines

We evaluate 10 UE scores across four families of approaches: (i) **Logit-based**: Maximum Sequence Probability (MSP), Perplexity (Si et al., 2023), Entropy (Huang et al., 2023), Energy (Liu et al., 2020); (ii) **Elicitation**: P(True) (Kadavath et al., 2022), Verbalised Confidence (VC; Tian et al., 2023); (iii) **Internal States**: CoE-C, CoE-R (Wang et al., 2025b); and (iv) **Consistency**: Eigenscore (ES; Chen et al., 2024) and Semantic Entropy (SE; Kuhn et al., 2023). Furthermore, we compare with the correction framework CUE (Li et al., 2025). For CUE, we make use of DeBERTa V3 (He et al., 2021, 2023) as the encoder for the corrector model and train for three epochs.

## 6 Results and Analysis

### 6.1 Main Results

Table 1 and Figure 2 present the main results. First, we observe the proxy failure phenomenon. For example, with Ministral-3-8B, CoE-R and SE fail to discriminate meaningfully on SciQ, although they perform significantly better on TriviaQA and PopQA, highlighting how proxy scores *can*, but often *fail to*, track factuality. We observe a similar occurrence with Gemma-2-9B, with CoE-C and CoE-R performing worse than chance on PopQA, although their AUCs with other models are competitive. These observations highlight the prevalence of proxy failure, where performance varies considerably across models and datasets.

Second, most vanilla scores exhibit poor ECE, undermining their utility as uncertainty indicators. Most importantly, TAC consistently leads to significant reductions in ECE, and frequently improves AUC. For instance,

| # Labels | 8 | 16 | 32 | 64 | 128 | All |
|---|---|---|---|---|---|---|
| Δ(ECE) (↓) | -12.31 | -12.46 | -12.75 | -12.74 | -16.09 | -17.48 |
| Δ(AUC) (↑) | -7.39 | -2.67 | -0.76 | -0.37 | -0.96 | 3.16 |
| **% Corrupt** | **0.5** | **0.4** | **0.3** | **0.2** | **0.1** | **Clean** |
| Δ(ECE) (↓) | -22.81 | -18.95 | -17.96 | -18.54 | -18.78 | -17.48 |
| Δ(AUC) (↑) | -31.15 | -3.32 | -0.62 | 1.49 | 2.76 | 3.16 |

(a) TriviaQA

| # Labels | 8 | 16 | 32 | 64 | 128 | All |
|---|---|---|---|---|---|---|
| Δ(ECE) (↓) | -18.74 | -19.30 | -22.08 | -22.86 | -18.59 | -26.69 |
| Δ(AUC) (↑) | 2.03 | -0.38 | -0.79 | 4.66 | 4.86 | 5.48 |
| **% Corrupt** | **0.5** | **0.4** | **0.3** | **0.2** | **0.1** | **Clean** |
| Δ(ECE) (↓) | -22.27 | -21.08 | -23.62 | -25.45 | -25.72 | -26.69 |
| Δ(AUC) (↑) | -14.71 | -0.56 | 3.89 | 4.15 | 4.64 | 5.48 |

(b) SciQ

| # Labels | 8 | 16 | 32 | 64 | 128 | All |
|---|---|---|---|---|---|---|
| Δ(ECE) (↓) | -16.83 | -20.02 | -11.58 | -6.18 | -12.97 | -21.55 |
| Δ(AUC) (↑) | -51.30 | -0.09 | -6.13 | -2.58 | 0.25 | 2.66 |
| **% Corrupt** | **0.5** | **0.4** | **0.3** | **0.2** | **0.1** | **Clean** |
| Δ(ECE) (↓) | -5.60 | -7.06 | -12.63 | -15.64 | -24.23 | -21.55 |
| Δ(AUC) (↑) | -46.52 | -4.62 | 1.30 | 2.75 | 2.83 | 2.66 |

(c) PopQA

Table 2: Performance (average Δ) under few-shot and noisy label constraints.

on average, with Ministral-3-8B, TAC leads to a change of $-17.48, -26.69, -21.55$ in ECE, $+3.16, +5.48, +2.66$ in AUC, on TriviaQA, SciQ, and PopQA, respectively. We verify the statistical significance of these gains in Table 5. Table 7 illustrates qualitative examples of proxy failure with entropy-derived confidence scores on LLM-generated answers and how we apply TAC as a calibration layer. These results validate TAC as an important post-hoc calibration step.

### 6.2 Comparison with CUE

In Figure 3, we compare TAC against CUE. Figures 7 and 8 show results for SciQ and PopQA, respectively. On average, with Ministral-3-8B, TAC reports a reduction in ECE of $-21.91$ and an increase in AUC of $+3.77$, compared to CUE's $-14.03$ and $+7.26$. Overall, TAC produces considerably better calibrated scores than CUE, without access to BERT-based embeddings, while CUE has a greater advantage in recovering AUC. This trade-off arises because TAC operates only on a scalar proxy and is designed mainly for calibration, whereas CUE utilises the additional information in BERT embeddings that can improve discriminability. On balance, this highlights TAC's strengths in aligning raw scores to achieve better calibrated, and thus, more meaningful uncertainty estimates.

### 6.3 Noisy and Few-shot Supervision

Table 2 summarises the average change in ECE and AUC after TAC with few-shot or noisy labels. Individual evaluation metrics are deferred to Figures 9 and 10. Calibration improvements persist with limited and noisy supervision, although AUC gains can weaken in the few-shot regime. This hints at TAC as a realistic calibration protocol that does not demand perfect supervision.

| Train on | Test on | | |
|---|---|---|---|
| | TriviaQA | SciQ | PopQA |
| TriviaQA | 9.60 / 73.35<br>(-17.48) / (+3.16) | 13.30 / 69.40<br>(-19.84) / (+5.00) | 25.98 / 77.35<br>(-9.35) / (+0.19) |
| SciQ | 12.58 / 71.40<br>(-14.50) / (+1.21) | 6.45 / 69.88<br>(-26.69) / (+5.48) | 16.73 / 71.12<br>(-18.60) / (-6.04) |
| PopQA | 20.56 / 70.35<br>(-6.52) / (+0.16) | 18.76 / 64.12<br>(-14.44) / (-0.28) | 13.78 / 79.82<br>(-21.55) / (+2.66) |

(a) TAC

| Train on | Test on | | |
|---|---|---|---|
| | TriviaQA | SciQ | PopQA |
| TriviaQA | 15.79 / 76.10<br>(-11.29) / (+5.91) | 31.57 / 65.66<br>(-1.56) / (+1.26) | 31.68 / 83.32<br>(-3.65) / (+6.16) |
| SciQ | 4.65 / 66.45<br>(-22.43) / (-3.74) | 20.03 / 73.37<br>(-13.11) / (+8.97) | 23.39 / 73.06<br>(-11.93) / (-4.10) |
| PopQA | 18.61 / 70.62<br>(-8.47) / (+0.43) | 19.51 / 62.96<br>(-13.63) / (-1.44) | 17.65 / 84.06<br>(-17.68) / (+6.90) |

(b) CUE

Table 3: Average performance and Δ compared to vanilla scores under transfer settings.

### 6.4 Cross-Task Transfer

Complementing the noisy and few-shot constraints, we further consider another realistic out-of-distribution (OOD) setting, where the scores and correctness labels are obtained from a different query set. Under OOD constraints, TAC reports average changes in ECE / AUC of $-13.88/+0.24$, compared to CUE's $-10.28/-0.23$, over the vanilla scores. We break down the average changes by each individual train–test configuration in Table 3. TAC remains useful under train-test mis-

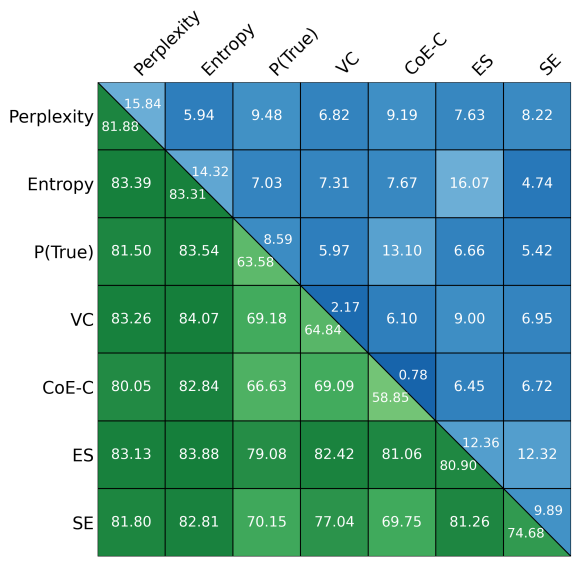


Figure 4: Pairwise TAC. Upper triangle shows ECE, and lower triangle shows AUC. Darker colours indicate more favourable results.

match, although gains are weaker than in-domain calibration. The transfer results are not uniform: transfer between TriviaQA and SciQ is relatively stable, especially when trained on TriviaQA and tested on SciQ. In contrast, transfer involving PopQA is weaker. For example, training on SciQ and testing on PopQA reduces AUC by 6.04 points. Similarly, using PopQA as the source gives weaker transfer to TriviaQA and SciQ than in-domain training. This is likely due to PopQA's long-tail entity distribution, which represents a larger shift from TriviaQA and SciQ, than the shift between the latter two datasets.

Overall, this suggests that the learned calibration is partially transferable across related QA tasks, especially for calibration, but weakens as the shift between train and test distributions widens. Therefore, TAC can be regarded as a task-sensitive calibration. When possible, in-domain calibration is preferable, while cross-task TAC remains a practical option when target labels are unavailable.

| Dataset | ECE (↓) | | | AUC (↑) | | |
|---|---|---|---|---|---|---|
| | Vanilla | TAC | Δ | Vanilla | TAC | Δ |
| CSQA | 14.16 | 4.03 | -10.13 | 56.85 | 63.10 | +6.25 |
| MedMCQA | 23.12 | 5.59 | -17.53 | 56.16 | 64.28 | +8.12 |

Table 4: Average performance and change on CSQA and MedMCQA with Gemma-2-9B.

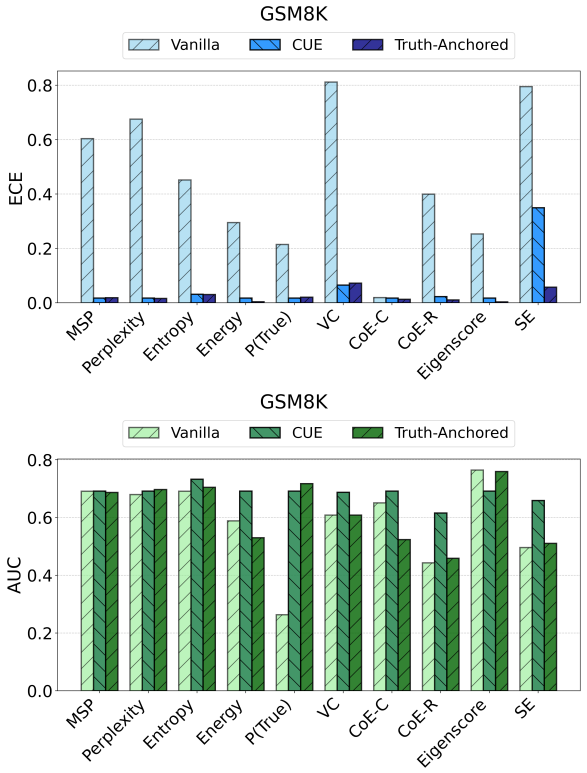


Figure 5: Performance on GSM8K with Ministral-3-8B.

## 6.5 Specialised Datasets

We have mainly focused on fact-heavy question-answering datasets. Here, we provide a preliminary analysis of applying our TAC to more specialised datasets, namely the mathematical dataset **GSM8K** (Cobbe et al., 2021) with Ministral-3-8B (Figure 5), commonsense reasoning task **CSQA** (Talmor et al., 2019) and medical-domain dataset **MedMCQA** (Pal et al., 2022) with Gemma-2-9B (Table 4). TAC yields consistent improvements, similar to the short-form factual datasets. Evaluation with tasks beyond QA, such as open-ended generation, remains future work.

## 6.6 Inter-score Anchoring

TAC explicitly aims to improve calibration, but it cannot recover high discriminability if the initial score is uninformative. To achieve this, additional information is necessary. Here, we explore this setting, where we assume access to two sources of uncertainty scores, and combine the two scores to inject external information not available within each score alone. Specifically, we consider distinct raw scores $S_i$ and $S_j$, $i \neq j$, and learn the mapping $m_\theta(S_{\text{concat}})$ from the concatenated input $S_{\text{concat}} = [S_i, S_j]$, our pairwise variant of TAC. We provide the results in Figure 4.

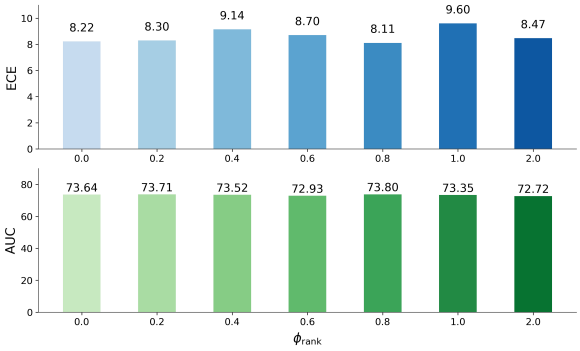


Figure 6: Sensitivity analysis of $\phi_{\text{rank}}$.

With a combination of two uncertainty scores, stronger discrimination can be recovered. We hypothesise this is because each score captures a different facet of truth uncertainty, especially for those scores belonging to different families, or formulated using distinct sources of information. For example, CoE-C and VC scores are derived with different information. CoE-C utilises internal state information, by tracking the changes between successive layers in their embeddings, whereas VC operates at the output token level. With pairwise TAC, AUC noticeably improves. However, it remains to be seen how we can optimally combine scores to achieve even better discriminability.

### 6.7 Sensitivity Analysis

Figure 6 presents the sensitivity analysis of $\phi_{\text{rank}}$. While the ranking loss can lead to gains when $\phi_{\text{rank}} > 0$, it is not consistently helpful as performance is unstable, especially for ECE. We believe this happens because the ranking loss optimises relative ordering, while the main BCE objective optimises calibrated correctness probabilities. These goals are related but not identical: pushing pairs apart may help AUC in some cases but can also distort the probability mapping learned by BCE, hurting calibration. In addition, because TAC operates on a scalar score, the auxiliary term may not be able to effectively rank pairs given this weak signal.

## 7 Conclusion

In this work, we formalise proxy failure in LLM uncertainty estimation, explaining the phenomenon where performance of widely used heuristic scores degrades substantially across configurations. We argue that truth anchoring (TAC), our proposed post-hoc calibration method on uncertainty scores, even with imperfect supervision, offers a principled and practical remedy, paving the way for more reliable and meaningful uncertainty estimates.

## Limitations

Truth anchoring successfully targets calibration, as supported by the significant reduction in ECE. On the other hand, discriminability is not guaranteed to improve, and can sometimes slightly worsen. Inter-score anchoring can be useful, but the optimal combination remains to be analysed. Furthermore, the scope of our study is limited to fact-heavy QA datasets. Future work may consider other specialised or long-form reasoning-intensive tasks. Finally, we examined moderately sized LLMs, ranging from 3 to 9 billion parameters. Experiments with larger and more capable models could strengthen our claims.

## Ethical Considerations

Even though our method is shown to improve confidence calibration, by design, it can only provide meaningful uncertainty measures in applications with clear ground truth. Users should avoid deploying Truth Anchoring in open-ended tasks, due to the risk of biased or subjective outputs being represented as confident.

## Acknowledgements

This research is supported by the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A*STAR, as well as supported by Alibaba Group and NTU Singapore through Alibaba-NTU Global e-Sustainability CorpLab (ANGEL). We would like to thank the anonymous reviewers for their valuable feedback.

## References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

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*.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, page 129–136, New York, NY, USA. Association for Computing Machinery.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. INSIDE: LLMs' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. 2024. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1419–1436, Miami, Florida, USA. Association for Computational Linguistics.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Jesse Davis and Mark Goadrich. 2006. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 233–240, New York, NY, USA. Association for Computing Machinery.

Tilmann Gneiting and Adrian E Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378.

Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A Cool, Zahir Kanjee, Andrew S Parsons, Neera Ahuja, et al. 2024. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA network open*, 7(10):e2440969.

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

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1321–1330. JMLR.org.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Jinwen He, Yujia Gong, Zijin Lin, Cheng'an Wei, Yue Zhao, and Kai Chen. 2024. LLM factoscope: Uncovering LLMs' factual discernment through measuring inner states. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10218–10230, Bangkok, Thailand. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2023. Look before you leap: An exploratory study of uncertainty measurement for large language models. *arXiv preprint arXiv:2307.10236*.

Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. 2024. LLM internal states reveal hallucination risk faced with a query. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 88–104, Miami, Florida, US. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel 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 (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. 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, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

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*.

Rui Li, Jing Long, Muge Qi, Heming Xia, Lei Sha, Peiyi Wang, and Zhifang Sui. 2025. Towards harmonized uncertainty estimation for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22938–22953, Vienna, Austria. Association for Computational Linguistics.

Xiaomin Li, Zhou Yu, Ziji Zhang, Yingying Zhuang, Swair Shah, Narayanan Sadagopan, and Anurag Beniwal. 2026. Semantic volume: Quantifying and detecting both external and internal uncertainty in LLMs.

In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 31751–31759.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*.

Alexander H Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, et al. 2026. Ministral 3. *arXiv preprint arXiv:2601.08584*.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. 2025. 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.

Alejandro Lopez-Lira and Yuehua Tang. 2023. Can ChatGPT forecast stock price movements? Return predictability and large language models. *arXiv preprint arXiv:2304.07619*.

Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. 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, Singapore. Association for Computational Linguistics.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Andy Shih, Dorsa Sadigh, and Stefano Ermon. 2023. Long horizon temperature scaling. In *International Conference on Machine Learning*, pages 31422–31434. PMLR.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. 2025. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*, 58(3):1–38.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Lee Boyd-Graber, and Lijuan Wang. 2023. Prompting GPT-3 to be reliable. In *The Eleventh International Conference on Learning Representations*.

Ponhvoan Srey, Xiaobao Wu, and Anh Tuan Luu. 2025. Unsupervised hallucination detection by inspecting reasoning processes. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22117–22129, Suzhou, China. Association for Computational Linguistics.

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

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Hexiang Tan, Fei Sun, Sha Liu, Du Su, Qi Cao, Xin Chen, Jingang Wang, Xunliang Cai, Yuanzhuo Wang, Huawei Shen, and Xueqi Cheng. 2025. Too consistent to detect: A study of self-consistent errors in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4755–4765, Suzhou, China. Association for Computational Linguistics.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature Medicine*, 29(8):1930–1940.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn,

and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Tim Tomov, Dominik Fuchsgruber, Tom Wollschläger, and Stephan Günnemann. 2025. The illusion of certainty: Uncertainty quantification for LLMs fails under ambiguity. *arXiv preprint arXiv:2511.04418*.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. 2025. Benchmarking uncertainty quantification methods for large language models with LM-polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248.

Artem Vazhentsev, Lyudmila Rvanova, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. 2025. Token-level density-based uncertainty quantification methods for eliciting truthfulness of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2246–2262, Albuquerque, New Mexico. Association for Computational Linguistics.

Kevin Wang, Subre Abdoul Moktar, Jia Li, Kangshuo Li, and Feng Chen. 2025a. Measuring aleatoric and epistemic uncertainty in LLMs: Empirical evaluation on ID and OOD QA tasks. *arXiv preprint arXiv:2511.03166*.

Yiming Wang, Pei Zhang, Baosong Yang, Derek F. Wong, and Rui Wang. 2025b. Latent space chain-of-embedding enables output-free LLM self-evaluation. In *The Thirteenth International Conference on Learning Representations*.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. BloombergGPT: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Fan Yin, Jayanth Srinivasa, and Kai-Wei Chang. 2024. Characterizing truthfulness in large language model generations with local intrinsic dimension. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2025. Siren's song in the AI ocean: A survey on hallucination in large language models. *Computational Linguistics*, 51(4):1373–1418.

## A Proofs

In this section, we provide the proofs to Proposition 1 and Proposition 2.

**Lemma 1.** *For any score $S$, with $P_+ = P(S \mid C = 1)$ and $P_- = P(S \mid C = 0)$,*

$$\left|\mathrm{AUC}(S) - \frac{1}{2}\right| \leq \delta(P_+, P_-), \tag{10}$$

*where $\delta$ is the total variation distance.*

*Proof.* Let $S_+ \sim P_+$ and $S_- \sim P_-$ be continuous random variables with PDFs $p_+(s)$ and $p_-(s)$, and CDFs $F_+(s)$ and $F_-(s)$, respectively. By law of total probability, we can condition AUC on the positive class $S_+ = s$

$$\begin{aligned}\mathrm{AUC} &= P(S_+ > S_-) \\ &= \int_{-\infty}^{\infty} P(S_- < S_+ \mid S_+ = s)\, p_+(s)\, ds \\ &= \int_{-\infty}^{\infty} F_-(s)\, p_+(s)\, ds\end{aligned} \tag{11}$$

Next, draw $S'_+$ such that $S_+, S'_+ \overset{i.i.d}{\sim} P_+$. Similarly,

$$\frac{1}{2} = P(S_+ > S'_+) = \int_{-\infty}^{\infty} F_+(s)\, p_+(s)\, ds \tag{12}$$

Then, combining Equation (11) and 12, and applying absolute value to both sides and the triangle inequality yields

$$\left|\mathrm{AUC} - \frac{1}{2}\right| \leq \int_{-\infty}^{\infty} |F_-(s) - F_+(s)|\, p_+(s)\, ds \tag{13}$$

By definition, the total variation $\delta$ between $P_+$ and $P_-$ is

$$\delta\left(P_+, P_-\right) = \sup_A |P_+(A) - P_-(A)| \tag{14}$$

Taking $A = (-\infty, s]$, $F_+(s) = P_+(A)$ and $F_-(s) = P_-(A)$, for all $s$, we have

$$|F_+(s) - F_-(s)| \leq \delta\left(P_+, P_-\right) \tag{15}$$

Thus,

$$\begin{aligned}\left|\mathrm{AUC}-\frac{1}{2}\right| &\le \int_{-\infty}^{\infty} \delta\left(P_+, P_-\right) p_+(s)\, ds \\ &= \delta\left(P_+, P_-\right) \int_{-\infty}^{\infty} p_+(s)\, ds \\ &= \delta\left(P_+, P_-\right)\end{aligned} \tag{16}$$

□

**Proposition 1.** *Let $S$ be any real-valued score with continuous class-conditional distributions, and $p_c = P(C=1) \in (0,1)$. Then,*

$$\left|\mathrm{AUC(S)}-\frac{1}{2}\right| \le \sqrt{\frac{I(C,S)}{2p_c(1-p_c)}}, \tag{2}$$

*where $I$ denotes mutual information. In particular, if $I(C,S)$ is small, then the discriminability of $S$ must be close to chance.*

*Proof.* Let $M = P(S) = p_c P_+ + (1-p_c)P_-$. Since $C$ is binary, mutual information can be decomposed as

$$I(C,S) = p_c \mathrm{KL}(P_+ \| M) + (1-p_c)\mathrm{KL}(P_- \| M) \tag{17}$$

We first relate $P_+$ and $P_-$ to $M$. Because

$$\begin{aligned}P_+ - M &= P_+ - (p_c P_+ + (1-p_c)P_-) \\ &= (1-p_c)(P_+ - P_-),\end{aligned} \tag{18}$$

so

$$\delta(P_+, M) = (1-p_c)\, \delta(P_+, P_-) \tag{19}$$

Similarly, $M - P_- = p_c(P_+ - P_-)$, so

$$\delta(P_-, M) = p_c\, \delta(P_+, P_-) \tag{20}$$

Applying Pinsker's inequality to each term in Equation (17),

$$\begin{aligned}\mathrm{KL}(P_+ \| M) &\ge 2(1-p_c)^2\, \delta(P_+, P_-)^2, \\ \mathrm{KL}(P_- \| M) &\ge\ 2p_c^2\, \delta(P_+, P_-)^2\end{aligned} \tag{21}$$

Substituting into Equation (17) yields

$$\begin{aligned}I(C,S) \ge\ & 2p_c(1-p_c)^2\, \delta(P_+, P_-)^2 \\ &+ 2(1-p_c)p_c^2\, \delta(P_+, P_-)^2 \\ =& 2p_c(1-p_c)\delta\left(P_+, P_-\right)^2\end{aligned} \tag{22}$$

Hence,

$$\delta\left(P_+, P_-\right) \le \sqrt{\frac{I(C,S)}{2p_c(1-p_c)}} \tag{23}$$

Finally, by Lemma 1,

$$\left|\mathrm{AUC(S)}-\frac{1}{2}\right| \le \sqrt{\frac{I(C,S)}{2p_c(1-p_c)}}, \tag{24}$$

□

**Proposition 2.** *Fix a truth function $g^\star$ and any predictive entropy proxy $S_{pe}$. For any $\varepsilon > 0$, there exists a conditional distribution $P(\cdot \mid x)$ such that*

$$I(C, S_{pe}) \le \varepsilon \tag{3}$$

*Consequently, by Proposition 1,*

$$\left|\mathrm{AUC}(S_{pe})-\frac{1}{2}\right| \le \sqrt{\frac{\varepsilon}{2p_c(1-p_c)}} \tag{4}$$

*In particular, the AUC of predictive entropy can be made arbitrarily close to chance.*

*Proof.* Let $\varepsilon > 0$. Assume for every query $x$, both correct and wrong responses can be generated. For each $x$, choose one correct $y_+(x)$ and one wrong response $y_-(x)$. Let $\tau(x)$ denote the first token position at which the two sequences differ.

**Step 1: Construct a base setting with zero information.** Define a conditional distribution $P_0(\cdot \mid x)$ by $P_0(y_+ \mid x) = P_0(y_- \mid x) = 1/2$.

Under $P_0$, for any $x$, the next-token distribution is deterministic at every step except the first divergence time $\tau(x)$, where it is uniform over two tokens. Hence, along either support trajectory,

$$H\left(P_0(\cdot \mid x, y_{<t})\right) = \begin{cases}\log 2, & t = \tau(x), \\ 0, & t \ne \tau(x).\end{cases} \tag{25}$$

The entire entropy profile is identical along $y^{(1)}(x)$ and $y^{(0)}(x)$, and so

$$S_{pe}(x, y^{(1)}(x)) = S_{pe}(x, y^{(0)}(x)) = s_0(x). \tag{26}$$

Under $P_0$, $P_+ = P_-$, by construction, $I_{P_0}(C, S_{pe}) = 0$.

**Step 2: Mix in with arbitrary component.** Let $P_1(\cdot \mid x)$ be any conditional distribution, for example, the distribution induced by the actual deployed LLM. Let $B \sim \mathrm{Bern}(\lambda)$, where $\lambda \in (0,1)$, and

$$P(\cdot \mid x) = (1-\lambda)P_0 + \lambda P_1 \tag{27}$$

**Step 3: Bound mutual information.** Using chain rule,

$$\begin{aligned} I(C; S_{pe}) &\leq I(C; S_{pe}, B) \\ &= I(C; B) + I(C; S_{pe}|B) \end{aligned} \tag{28}$$

Since $B \sim \text{Bern}(\lambda)$, $I(C, B) \leq H(B) = h(\lambda)$, where $h(\lambda)$ is the binary entropy. Further,

$$\begin{aligned} I(C; S_{pe} \mid B) = (1-\lambda) I(C; S_{pe} \mid B = 0) \\ + \lambda I(C; S_{pe} \mid B = 1) \end{aligned} \tag{29}$$

Under $B = 0$, the distribution is exactly $P_0$ with zero information. Under $B = 1$, since $C$ is binary,

$$I(C; S_{pe} \mid B = 1) \leq H(C \mid B = 1) \leq \log 2 \tag{30}$$

Therefore,

$$I(C; S_{pe}) \leq h(\lambda) + \lambda \log 2 \tag{31}$$

RHS tends to 0 as $\lambda \to 0$. Hence, for any $\varepsilon$, $\exists \lambda$ such that $h(\lambda) + \lambda \log 2 \leq \varepsilon$. This yields $I(C; S_{pe}) \leq \varepsilon$, and the AUC bound follows directly from Proposition 1.

□

## B Supplementary Experiments

**Qualitative examples.** Table 7 illustrates instances of the proxy failure phenomenon, where entropy fails to track correctness. We provide four successful examples of TAC calibrating confidence toward the correctness label. For overly confident scores on incorrect answers, TAC pushes confidence down; conversely, for initially low-confidence scores on correct answers, TAC pulls confidence up. Additionally, we highlight a false negative case, where TAC wrongly suppresses confidence on a correct generation. These examples demonstrate that on aggregate, TAC improves truth alignment but, as a scalar calibration method, cannot guarantee correction of every individual response.

**Statistical significance.** To evaluate the statistical significance of our main results in Table 1, we additionally run the Ministral-3-8B experiments using five different seeds. We observe that TAC performance is stable, with small variability across runs. Further, we perform two-sided one-sample $t$-tests on the average $\Delta$(AUC) values against zero. On TriviaQA, SciQ, PopQA, the $p$-values are $9.6 \times 10^{-7}$, $6.1 \times 10^{-6}$, and $1.2 \times 10^{-6}$, respectively. These results suggest that the observed AUC improvements, on average, are statistically significant, and are not artifacts of stochastic MLP training.

| Dataset | $\Delta$(AUC) (SD) | $\Delta$(ECE) (SD) |
|---|---|---|
| TriviaQA | +3.24 (0.14) | -19.35 (1.16) |
| SciQ | +5.01 (0.36) | -25.69 (1.37) |
| PopQA | +2.85 (0.14) | -18.76 (2.77) |

Table 5: Five-seed average changes with Ministral-3-8B.

**More comparison with CUE.** We present more comparison with CUE on SciQ Figure 7 and PopQA in Figure 8.

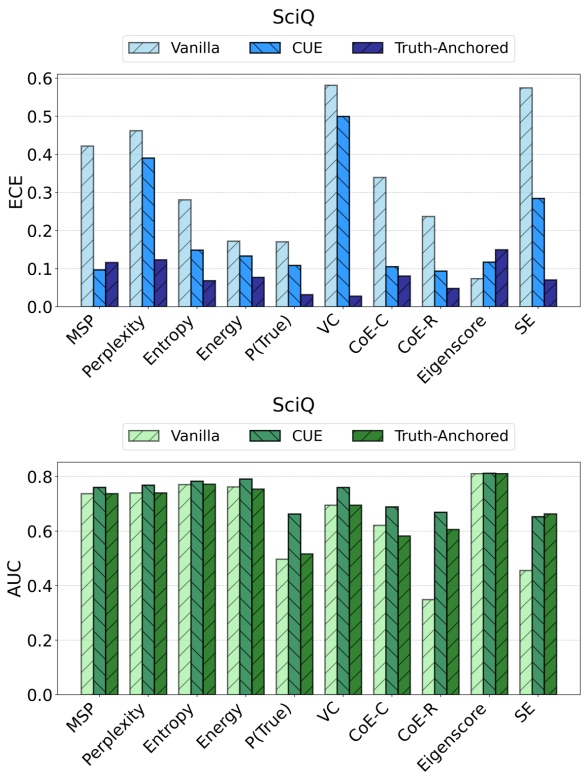


Figure 7: Performance of vanilla, CUE, and truth-anchored scores.

**More LLMs.** Table 6 shows the performance with Llama models (Grattafiori et al., 2024).

**Noisy and few-shot supervision.** Figures 9 and 10 plot the individual TAC scores for each few-shot or noisy setting for TriviaQA.

| Method | TriviaQA | | | | | | SciQ | | | | | | PopQA | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ECE (↓) | | | AUC (↑) | | | ECE (↓) | | | AUC (↑) | | | ECE (↓) | | | AUC (↑) | | |
| | Vanilla | TAC | Δ | Vanilla | TAC | Δ | Vanilla | TAC | Δ | Vanilla | TAC | Δ | Vanilla | TAC | Δ | Vanilla | TAC | Δ |
| **Llama-3.2-3B** | | | | | | | | | | | | | | | | | | |
| MSP | 48.25 | 13.41 | -34.84 | 76.15 | 76.15 | 0.00 | 47.96 | 13.23 | -34.73 | 75.79 | 75.79 | 0.00 | 70.96 | 2.74 | -68.22 | 53.46 | 61.51 | 8.05 |
| Perplexity | 58.46 | 6.51 | -51.95 | 69.19 | 67.11 | -2.09 | 50.47 | 8.70 | -41.77 | 72.88 | 72.25 | -0.63 | 80.05 | 2.28 | -77.76 | 48.25 | 57.41 | 9.16 |
| Entropy | 37.83 | 18.88 | -18.95 | 84.60 | 84.60 | 0.00 | 35.06 | 15.56 | -19.49 | 76.28 | 76.28 | 0.00 | 58.48 | 1.71 | -56.77 | 65.41 | 64.20 | -1.21 |
| Energy | 17.07 | 5.39 | -11.68 | 78.92 | 79.12 | 0.20 | 15.55 | 13.84 | -1.71 | 74.09 | 74.09 | 0.00 | 35.56 | 5.95 | -29.61 | 69.93 | 63.01 | -6.92 |
| P(True) | 57.55 | 6.15 | -51.40 | 50.74 | 50.99 | 0.25 | 61.05 | 2.33 | -58.72 | 61.00 | 61.00 | 0.00 | 71.68 | 25.49 | -46.19 | 84.13 | 84.13 | 0.00 |
| VC | 17.78 | 20.38 | 2.60 | 85.56 | 85.56 | 0.00 | 47.73 | 14.02 | -33.71 | 66.41 | 66.41 | 0.00 | 8.15 | 30.11 | 21.96 | 94.77 | 94.77 | 0.00 |
| CoE-C | 20.90 | 7.63 | -13.27 | 59.18 | 73.38 | 14.20 | 17.62 | 7.28 | -10.33 | 69.79 | 70.36 | 0.57 | 2.50 | 1.61 | -0.89 | 52.84 | 67.07 | 14.23 |
| CoE-R | 22.71 | 4.07 | -18.64 | 74.43 | 75.30 | 0.87 | 40.04 | 4.99 | -35.05 | 59.86 | 60.77 | 0.91 | 31.08 | 3.53 | -27.54 | 65.55 | 80.35 | 14.80 |
| Eigenscore | 27.70 | 12.15 | -15.56 | 68.37 | 78.05 | 9.67 | 20.57 | 7.20 | -13.37 | 69.81 | 66.94 | -2.87 | 65.55 | 3.37 | -62.18 | 40.44 | 58.29 | 17.85 |
| Semantic Entropy | 59.45 | 6.58 | -52.86 | 52.08 | 65.69 | 13.61 | 66.03 | 6.16 | -59.87 | 43.86 | 64.07 | 20.21 | 78.07 | 3.23 | -74.84 | 49.36 | 55.30 | 5.94 |
| **Llama-3.1-8B** | | | | | | | | | | | | | | | | | | |
| MSP | 30.31 | 8.53 | -21.78 | 75.79 | 75.79 | 0.00 | 46.84 | 6.49 | -40.35 | 81.93 | 82.39 | 0.46 | 47.85 | 5.61 | -42.24 | 75.42 | 77.36 | 1.94 |
| Perplexity | 29.34 | 7.96 | -21.37 | 73.87 | 73.87 | 0.00 | 44.29 | 9.91 | -34.38 | 81.88 | 81.57 | -0.31 | 58.16 | 13.63 | -44.53 | 71.12 | 71.12 | 0.00 |
| Entropy | 20.53 | 11.66 | -8.87 | 76.91 | 76.91 | 0.00 | 33.65 | 6.38 | -27.26 | 82.31 | 82.43 | 0.12 | 34.24 | 22.18 | -12.06 | 77.40 | 77.40 | 0.00 |
| Energy | 10.21 | 8.52 | -1.69 | 70.53 | 70.53 | 0.00 | 17.64 | 4.28 | -13.36 | 67.54 | 65.14 | -2.40 | 10.57 | 6.91 | -3.66 | 76.97 | 76.56 | -0.41 |
| P(True) | 38.56 | 4.12 | -34.44 | 62.16 | 62.21 | 0.05 | 53.99 | 4.18 | -49.81 | 62.86 | 65.90 | 3.04 | 57.01 | 17.04 | -39.96 | 80.66 | 80.66 | 0.00 |
| VC | 38.98 | 2.16 | -36.82 | 62.87 | 63.62 | 0.75 | 57.62 | 4.20 | -53.42 | 67.74 | 68.67 | 0.93 | 35.24 | 17.15 | -18.09 | 75.97 | 76.15 | 0.18 |
| CoE-C | 46.47 | 8.89 | -37.59 | 62.43 | 63.08 | 0.64 | 31.04 | 6.20 | -24.84 | 66.22 | 66.79 | 0.58 | 19.90 | 3.53 | -16.37 | 43.10 | 58.30 | 15.20 |
| CoE-R | 16.40 | 3.92 | -12.48 | 54.60 | 65.19 | 10.60 | 20.65 | 4.19 | -16.46 | 38.01 | 63.98 | 25.97 | 19.24 | 6.31 | -12.93 | 68.03 | 75.92 | 7.89 |
| Eigenscore | 9.63 | 4.75 | -4.88 | 76.42 | 76.76 | 0.34 | 12.89 | 9.81 | -3.08 | 83.16 | 82.69 | -0.47 | 21.09 | 3.60 | -17.49 | 69.30 | 75.57 | 6.28 |
| SE | 40.39 | 3.85 | -36.54 | 42.88 | 68.68 | 25.80 | 58.44 | 3.65 | -54.78 | 49.34 | 74.63 | 25.29 | 68.49 | 3.21 | -65.28 | 53.25 | 60.28 | 7.03 |

Table 6: Additional results with Llama models.

| Question | LLM Answer | Label | Vanilla → TAC Confidence |
|---|---|---|---|
| Which businessman on his elevation to the House of Lords, by Gordon Brown, in 2009, took the title Baron of Clapton? | Lord Alan Sugar | Correct | 0.172 → 0.271 |
| Which 19th century British novelist, who worked as a Surveyor for the Post Office, was reputedly instrumental in the introduction of the pillar box in 1852? | Charles Dickens | Incorrect (Anthony Trollope) | 0.877 → 0.591 |
| Approximately 20% of the atmosphere is made of which element? | Nitrogen | Incorrect (Oxygen) | 0.890 → 0.594 |
| What model describing atomic structure is often referred to as a planetary model? | Rutherford model | Incorrect (Bohr model) | 0.692 → 0.267 |
| What is another word for cellular drinking? | Pinocytosis | Correct | 0.632 → 0.416 (false negative) |

Table 7: Qualitative examples of proxy failure and subsequent TAC calibration. Vanilla confidence is derived from entropy. Green indicates movement toward the correctness label, with 1 representing correct generations.

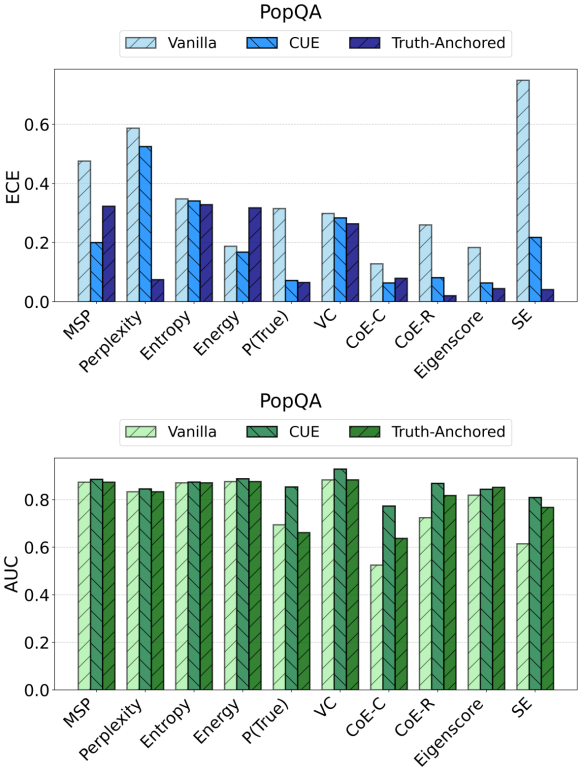


Figure 8: Performance of vanilla, CUE, and truth-anchored scores.

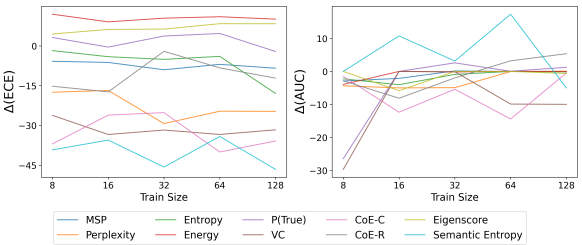


Figure 9: Δ(ECE) and Δ(AUC) on TriviaQA with few-shot supervision.

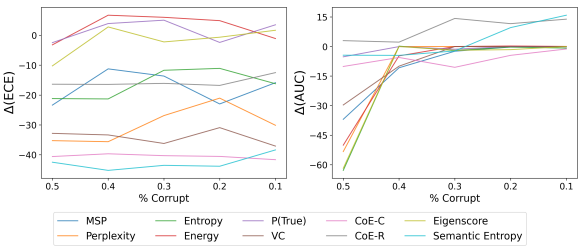


Figure 10: Δ(ECE) and Δ(AUC) on TriviaQA with noisy supervision.