Improving Semantic Uncertainty Quantification in Language Model Question-Answering via Token-Level Temperature Scaling

Tom A. Lamb University of Oxford
Desi R. Ivanova University of Oxford
Philip H. S. Torr University of Oxford
Tim G. J. Rudner University of Toronto & Vijil

Website Code

Abstract

Calibration is central to reliable semantic uncertainty quantification, yet prior work has largely focused on discrimination, neglecting calibration. As calibration and discrimination capture distinct aspects of uncertainty, focusing on discrimination alone yields an incomplete picture. We address this gap by systematically evaluating both aspects across a broad set of confidence measures. We show that current approaches, particularly fixed-temperature heuristics, produce systematically miscalibrated and poorly discriminative semantic confidence distributions. We demonstrate that optimising a single scalar temperature, which, we argue, provides a suitable inductive bias, is a surprisingly simple yet effective solution. Our exhaustive evaluation confirms that temperature scaling consistently improves semantic calibration, discrimination, and downstream entropy, outperforming both heuristic baselines and more expressive token-level recalibration methods on question-answering tasks.

1 INTRODUCTION

Calibration, how well predicted confidences match observed frequencies, is fundamental to reliable uncertainty quantification (UQ). However, prior work on semantic UQ for language models (LMs) has focused largely on *discrimination*, evaluating measures such as semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024; Santilli et al., 2025) without assessing calibration. This is a critical omission: perfect discrimination does not imply calibration, nor does perfect calibration necessarily imply accurate discrimination (Huang et al., 2020). Because discrimination and calibration capture distinct aspects of uncertainty, evaluating semantic UQ methods by discrimination alone yields an incomplete, and potentially misleading, assessment of reliability.

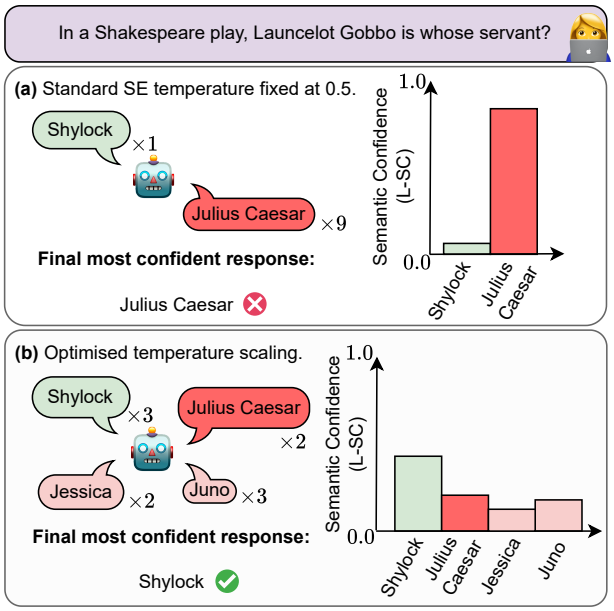


Figure (1): **Temperature Scaling Improves Semantic Uncertainty Quantification**. Same model using different temperatures generates 10 responses for the same input, which are clustered into semantic groups. We compute the L-SC semantic confidence measure of Kuhn et al. (2023) defined in Equation (3). Panel **(a)** uses the standard temperature of 0.5; panel **(b)** uses a temperature optimised on a calibration set. We find that optimised temperature scaling improves both semantic calibration and discrimination.

For instance, given the question *"In a Shakespeare play, Launcelot Gobbo is whose servant?"* in Figure 1, a semantically well-calibrated model should assign similarly high confidence to the semantically equivalent answers *"Shylock"* and *"Shylock in Merchant of Venice"*

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

while assigning low confidence to semantically incorrect answers such as *"Jessica"*. Accurate alignment between confidence and semantic correctness is essential for trustworthy natural language generation; yet, as Figure 1 shows, standard approaches can produce systematically overconfident semantic distributions.

As semantic calibration in LMs remains underexplored, how best to recalibrate models semantically is an open question. While token-level recalibration is well-established, its translation to semantic prediction space remains undetermined (Kuhn et al., 2023; Xie et al., 2024). This transition is fundamental: as semantic confidence measures are derived from token-level probabilities, any token-level miscalibration may translate into semantic miscalibration (Farquhar et al., 2024; Murray and Chiang, 2018). Bridging this gap is essential for grounding semantic uncertainty quantification and may reveal surprisingly simple approaches to semantic recalibration.

Temperature scaling (TS) is a well-established method for calibrating token-level probabilities (Guo et al., 2017) and controlling diversity in generative LMs. While typically treated as a fixed heuristic in semantic UQ (Kuhn et al., 2023), we argue that optimising TS as a single global scalar provides a superior inductive bias compared to more expressive, and expensive, per-token recalibration methods like ATS. This global constraint acts as a natural regulariser, preventing the model from overfitting to filler tokens. Instead, by capturing the uncertainty of the sequence as a whole, TS more effectively reflects the overall meaning being conveyed. TS is thus a surprisingly simple and efficient tool for improving LM reliability, one that practitioners can adopt without altering existing workflows. Since there is no generally agreed-upon way to define distributions in semantic prediction space, a rigorous study of semantic calibration and discrimination requires evaluating a broad range of semantic confidence measures, rather than relying on a small, fixed set as done in prior work (Kuhn et al., 2023; Farquhar et al., 2024). To address this, we consider several different approaches to extracting semantic uncertainty from LMs.

Our contributions are as follows:

1. We provide the **first systematic evaluation of both semantic calibration and discrimination** across a wide range of semantic confidence measures, revealing that **base models and fixed-temperature heuristics** are **poorly calibrated and weakly discriminative**, thereby exposing fundamental limitations of prior work.

2. We show that **optimised token-level temperature scaling**, a single scalar, is a surprisingly simple yet highly effective approach for semantic UQ, **outperforming** fixed-temperature baselines and **complex post-hoc calibration methods** such as ATS (Xie et al., 2024) and Platt scaling (Platt et al., 1999).

3. We demonstrate that optimisation-based temperature scaling **improves downstream semantic entropy**, exposing the suboptimality of heuristics prevalent in the literature. Moreover, we show that a more **principled selection of the final response**, based directly on semantic confidence distributions rather than through an ad-hoc separate greedy-decoding procedure as done in prior work, yields **superior discrimination**.

4. We validate the **robustness** of these findings through an **exhaustive evaluation** across multiple LMs, QA datasets, calibration metrics, and generation sample sizes.

Code to reproduce our experiments is available at:
`github.com/tomalamb/semantic-calibration`

2 RELATED WORK

Confidence and Calibration for LMs. Confidence estimation in language models (LMs) typically relies on token-level likelihoods (Kadavath et al., 2022), post-processing techniques (Malinin and Gales, 2021), or verbalised confidence scores (Kadavath et al., 2022). Calibration, the alignment between confidence and predictive correctness (Flach, 2016), is fundamental to reliability, yet often degrades following post-training procedures like RLHF (Achiam et al., 2023; Kadavath et al., 2022). This miscalibration necessitates post-hoc recalibration techniques such as temperature scaling (TS) (Guo et al., 2017) or Platt scaling (Platt et al., 1999). More recently, Xie et al. (2024) proposed Adaptive Temperature Scaling (ATS) to generate token-specific temperatures; however, this requires a computationally expensive transformer-based head compared to standard scalar methods such as TS.

Semantic Uncertainty Quantification. Traditional UQ techniques, including Bayesian (Blundell et al., 2015; Yang et al., 2023), latent-space (Mukhoti et al., 2023; Liu et al., 2020), and ensemble-based (Lakshminarayanan et al., 2017) approaches, were primarily designed for classification tasks. More specific methods for LMs often focus on token-level instantiations of generations (Malinin and Gales, 2021), neglecting the underlying semantics of generations (Kuhn et al., 2023). With the rise of open-ended generative tasks, research has shifted towards quantifying uncertainty over conveyed *meanings* rather than literal token sequences. This has motivated multi-sampling methods that cluster responses by semantic equivalence, typically via Nat-

ural Language Inference (NLI) (Williams et al., 2018), to define *semantic confidence measures* as distributions over meanings (Kuhn et al., 2023; Lin et al., 2023; Nikitin et al., 2024). The entropy of these semantic measures is subsequently used to discriminate between correct and incorrect responses (Kuhn et al., 2023).

The Transition to Semantic Calibration. While semantic UQ has improved discrimination, the calibration of semantic confidence measures remains underexplored. Consequently, it remains unknown how well existing methods align semantic confidence with actual correctness, or whether established token-level techniques can effectively translate to the more complex setting of meanings. We argue that temperature scaling (TS) is uniquely well-suited for this transition. Unlike expressive methods such as ATS which optimise for local, per-token calibration goals, TS imposes a single global constraint over the entire sequence. We posit that this constraint provides a superior inductive bias for semantics: it regularises against overfitting to semantically hollow filler tokens, forcing the calibration process to reflect the uncertainty of the sequence as a whole. Thus, TS emerges not merely as a heuristic, but as a principled, computationally cheap, and robust tool for reliable semantic UQ.

3 SEMANTIC CONFIDENCE

We consider an autoregressive LM, p, over a vocabulary $\mathcal{V}$ and denote the set of possible token sequences as $\mathcal{V}^*$. As established in Section 1, the absence of a unique way to define distributions over meanings motivates a broader and more holistic investigation into semantic UQ than the limited scope considered in prior work (Kuhn et al., 2023). Consequently, we define and evaluate seven semantic confidence measures; while two originate from existing work (E-SC and L-SC), we introduce five novel measures (ML-SC, IC-SC, B-SC, T-SC, and G-SC). Evaluating across this diverse set of semantic measures strengthens our conclusions, as we show in Figure 3 that improvements from TS persist consistently across all measures.

For an input prompt and ground truth label sampled from a data distribution $(\boldsymbol{x}, \boldsymbol{y}) \sim p_{\mathcal{D}}$, with $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{V}^*$, we generate m responses from a LM: $\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(m)} \sim p(\cdot \mid \boldsymbol{x})$, where each $\boldsymbol{y}^{(i)} \in \mathcal{V}^*$. We then follow Kuhn et al. (2023) and cluster responses based on which responses are semantically equivalent using natural language inference (NLI). This produces $k \in \mathbb{N}$ semantic clusters, $C_1, \ldots, C_k$, where k depends on the input $\boldsymbol{x}$ and generations $\{\boldsymbol{y}^{(i)}\}_{i=1}^{m}$. For each of our proposed measures, we define a score for each cluster C_i. The confidence measures are then found by normalising these scores over all generated clusters.

Empirical Semantic Confidence (E-SC). Following Farquhar et al. (2024), we define the *empirical semantic confidence* (E-SC) as the empirical counting measure over clusters:

$$p^{\text{E-SC}}(C_i \mid x) \propto |C_i|, \quad \forall i \in [k], \tag{1}$$

where $[k] \coloneqq \{1, 2, \ldots, k\}$. We note that this is the same distribution used by Farquhar et al. (2024) to compute the semantic entropy of black-box models.

Likelihood-Based Semantic Confidence (L-SC). Within this work, we use length-normalised sequence likelihoods following the common practice of correcting for the exponential decay of raw likelihoods with sequence length (Murray and Chiang, 2018; Malinin and Gales, 2021).

For each semantic cluster C_i, we compute its score, denoted $s(C_i \mid \boldsymbol{x})$, by summing the length-normalised likelihoods of the samples within it:

$$s(C_i \mid \boldsymbol{x}) \coloneqq \sum_{\boldsymbol{y} \in C_i} p(\boldsymbol{y} \mid \boldsymbol{x})^{\frac{1}{|\boldsymbol{y}|}}, \quad \forall i \in [k]. \tag{2}$$

Normalising these scores yields the *Likelihood-based semantic confidence* (L-SC) measure:

$$p^{\text{L-SC}}(C_i \mid \boldsymbol{x}) \propto s(C_i \mid \boldsymbol{x}), \quad \forall i \in [k]. \tag{3}$$

We note that this measure was originally alluded to in Kuhn et al. (2023) and explicitly formulated in this form in Farquhar et al. (2024).

Mean Likelihood-Based Semantic Confidence (ML-SC). Summing length-normalised likelihoods may bias scores toward larger clusters, so we compute the mean score, $\bar{s}(C_i \mid \boldsymbol{x})$, for each cluster as:

$$\bar{s}(C_i \mid \boldsymbol{x}) \coloneqq \frac{s(C_i \mid \boldsymbol{x})}{|C_i|}, \quad \forall i \in [k].$$

Normalising these scores yields the *mean likelihood-based semantic confidence* (ML-SC) measure:

$$p^{\text{ML-SC}}(C_i \mid \boldsymbol{x}) \propto \bar{s}(C_i \mid \boldsymbol{x}), \quad \forall i \in [k].$$

Bayesian Semantic Confidence (B-SC). We introduce a Bayesian-inspired semantic confidence measure that combines the E-SC and L-SC approaches. Specifically, we adopt the empirical distribution from Equation (1) as a prior over clusters:

$$\pi(C_i \mid \boldsymbol{x}) \coloneqq p^{\text{E-SC}}(C_i \mid \boldsymbol{x}), \quad \forall i \in [k].$$

We then define the cluster-level likelihood as the product of the length-normalised likelihoods for all responses $\boldsymbol{y}^{(1):(m)} \coloneqq (\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(m)})$ assigned to C_i:

$$\bar{p}\left(\boldsymbol{y}^{(1):(m)} \mid C_i, \boldsymbol{x}\right) = \prod_{\boldsymbol{y} \in C_i} p(\boldsymbol{y} \mid \boldsymbol{x})^{\frac{1}{|\boldsymbol{y}|}} \quad \forall i \in [k].$$

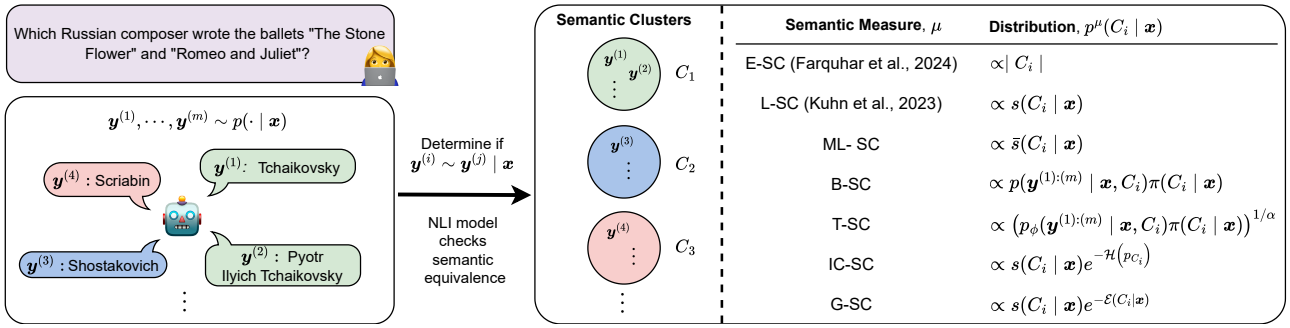


Figure (2): **Semantic Confidence Measures.** Figure showing how existing (E-SC; Farquhar et al., 2024) and (L-SC; Kuhn et al., 2023)) as well as a broad set of novel baseline semantic measures (ML-SC, B-SC, T-SC, IC-SC, and G-SC) are computed. For an input $\boldsymbol{x}$, we sample multiple responses from the model $p(\cdot \mid \boldsymbol{x})$, and use an NLI model to assess bidirectional entailment, determining whether responses $\boldsymbol{y}^i$ and $\boldsymbol{y}^j$ are semantically equivalent ($\boldsymbol{y}^i \sim \boldsymbol{y}^j \mid \boldsymbol{x}$). Here, $s(C \mid \boldsymbol{x})$ denotes the sum, and $\bar{s}(C \mid \boldsymbol{x})$ the average, $\mathcal{H}(p_{C_i})$ the entropy, and $\mathcal{E}(C_i \mid \boldsymbol{x})$ the entropy of the length-normalised log-likelihoods of generations within cluster C. See Section 3 for details.

Combining the likelihood and the prior yields the posterior distribution over clusters:

$$p^{\text{B-SC}}(C_i \mid \boldsymbol{x}) \propto \bar{p}\left(\boldsymbol{y}^{(1):(m)} \mid C_i, \boldsymbol{x}\right) \pi(C_i \mid \boldsymbol{x}), \quad (4)$$

for all $i \in [k]$. We refer to this as the *Bayesian semantic confidence* (B-SC) measure.

Tempered-Bayesian Posterior (T-SC). We consider a tempered version of the Bayesian posterior in eq. (4) by introducing a scaling parameter $\alpha \in \mathbb{R}_{>0}$:

$$p_\alpha^{\text{T-SC}}(C_i \mid \boldsymbol{x}) \propto \left(\bar{p}\left(\boldsymbol{y}^{(1):(m)} \mid C_i, \boldsymbol{x}\right) \pi(C_i \mid \boldsymbol{x})\right)^{1/\alpha}, \quad (5)$$

for all $i \in [k]$. When $\alpha < 1$, the posterior becomes sharper, concentrating more heavily on the clusters with the highest likelihood, whilst $\alpha > 1$ produces flatter distributions. The case $\alpha < 1$ is referred to as producing a cold posterior (Wenzel et al., 2020).

Internal Consistency Semantic Confidence (IC-SC). We introduce an unnormalised entropy-penalised semantic confidence measure that accounts for the internal consistency of clusters. Given an input $\boldsymbol{x}$ and a cluster C_i, we define the internal agreement between the likelihoods of the responses within this cluster via:

$$p_{C_i}(\boldsymbol{y}_j) := \frac{p(\boldsymbol{y}_j \mid \boldsymbol{x})^{\frac{1}{|\boldsymbol{y}|}}}{\sum_{\boldsymbol{y}_k \in C_i} p(\boldsymbol{y}_k \mid \boldsymbol{x})^{\frac{1}{|\boldsymbol{y}|}}}$$

and compute its internal entropy:

$$\mathcal{H}(p_{C_i}) = -\sum_{\boldsymbol{y}_k \in C_i} p_{C_i}(\boldsymbol{y}_k) \log p_{C_i}(\boldsymbol{y}_k).$$

We then penalise the scores in eq. (2) with this entropy:

$$p^{\text{IC-SC}}(C_i \mid \boldsymbol{x}) \propto s(C_i \mid \boldsymbol{x}) e^{-\mathcal{H}(p_{C_i})},$$

for all $i \in [k]$. This downweights clusters containing responses with different likelihoods.

Gibbs Semantic Confidence (G-SC). We consider the E-SC measure again as a prior, π, over clusters as we did for eq. (4). The energy function over clusters, $\mathcal{E}(\cdot \mid \boldsymbol{x})$, is the negative (length-normalised) log-likelihood of the samples within this cluster:

$$\mathcal{E}(C_i \mid \boldsymbol{x}) = -\log \bar{p}\left(\boldsymbol{y}^{(1):(m)} \mid C_i, \boldsymbol{x}\right), \quad \forall i \in [k].$$

Using this energy, we then form the Gibbs distribution: (Cantoni and Picard, 2004) as

$$p^{\text{G-SC}}(C_i \mid \boldsymbol{x}; \alpha) \propto \pi(C_i \mid \boldsymbol{x}) e^{-\alpha \mathcal{E}(C_i \mid \boldsymbol{x})} \quad \forall i \in [k].$$

Here, $\alpha \in \mathbb{R}_{>0}$ is a scaling parameter that controls the importance of the likelihood in forming the Gibbs posterior. This is similar to the T-SC measure defined in Equation (5), but differs in that we only scale the likelihood term by α, whilst the prior remains unchanged.

Diversity of Behaviour. In Section 4, we show that SC measures yield distinct uncertainty profiles, confirming there is no single uniformly best way to define semantic confidence distributions. This motivates the introduction of new measures to complement existing ones, ensuring a more robust and comprehensive evaluation of semantic UQ.

3.1 Token-Level Recalibration

We present several calibration techniques, often applied for token-level calibration, that we will compare in the context of semantic calibration.

Temperature Scaling (TS) Given an input prompt $\boldsymbol{x} \in \mathcal{V}^*$ and logits $\boldsymbol{z}_t \in \mathbb{R}^{|\mathcal{V}|}$ at decoding step t from an LM $p(\cdot \mid \boldsymbol{x})$, the output probabilities are computed as $p(y_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t}; \tau) = \sigma\left(\boldsymbol{z}_t / \tau\right)$, where $\sigma : \mathbb{R}^{|\mathcal{V}|} \to \Delta^{|\mathcal{V}|-1}$ is the softmax function and $\tau > 0$ is a scalar temperature.

Adaptive Temperature Scaling (ATS) ATS (Xie et al., 2024) replaces the global scalar $\tau \in \mathbb{R}_{>0}$ with token position-specific temperatures via a learned prediction head. Given input $\boldsymbol{x} \in \mathcal{V}^*$ and final hidden representations $\boldsymbol{h} \in \mathbb{R}^{d_{\text{model}} \times n}$, ATS applies a transformation $\psi_{\boldsymbol{\theta}} : \mathbb{R}^{d_{\text{model}} \times n} \to \mathbb{R}^n$, implemented as a single-layer transformer block (Vaswani et al., 2017), to produce a scalar temperature for each token position:

$$p(y_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t}; \boldsymbol{\theta}) = \sigma\left(\boldsymbol{z}_t / \boldsymbol{\tau}_t\right).$$

for $\boldsymbol{\tau}^{-1} = \exp(\psi_{\boldsymbol{\theta}}(\boldsymbol{h}))$. All operations using $\boldsymbol{\tau} \in \mathbb{R}^n$ are performed element-wise.

Platt Scaling Platt scaling is a classic technique used to recalibrate deep learning models (Platt et al., 1999; Niculescu-Mizil and Caruana, 2005; Guo et al., 2017). However, it is computationally expensive to directly apply it to LMs given the size of their vocabularies. Therefore, following Xie et al. (2024), we restrict the affine Platt transformation on the logits of a model to be diagonal:

$$p(y_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t}; \boldsymbol{\theta}) = \sigma(\text{diag}(\boldsymbol{w})\boldsymbol{z}_t + \boldsymbol{b}),$$

where $\boldsymbol{\theta} = (\boldsymbol{w}, \boldsymbol{b}) \in \mathbb{R}^{2|\mathcal{V}|}$ are the learnable parameters.

Temperature Scaling's Promise for Semantic Recalibration. Despite its lower expressivity, we argue that TS provides a superior inductive bias for semantic calibration. Unlike Platt scaling, which can distort token-level likelihood rankings, TS strictly preserves the likelihood rankings of tokens, ensuring that the relative ordering of semantically important tokens remains intact. Moreover, compared to ATS which optimises for local per-token goals, TS enforces a single global constraint across the entire generation. This global focus acts as a regulariser against overfitting to meaningless filler tokens, preventing the method from minimising loss by merely fitting frequent, non-semantic words. Instead, TS forces the calibration to reflect the uncertainty of the sequence as a whole, thereby better capturing the overall meaning conveyed. We provide an extended analysis of these mechanisms, including empirical evidence of ATS overfitting to filler tokens, in Section G. In this context, TS represents an Occam's razor-style methodology (MacKay, 2003).

Hypothesis: Temperature scaling yields better semantic calibration and overall semantic UQ than more expressive methods like Platt scaling and ATS.

3.2 Calibration Loss Functions

We consider two alternative losses for learning the calibration parameters $\boldsymbol{\theta}$ (the temperature τ for TS, ATS head, or Platt scaling parameters).

Negative Log-Likelihood (NLL). As a strictly proper scoring rule (Gneiting and Raftery, 2007), NLL, equivalent to standard cross-entropy with one-hot targets, is often a natural choice for achieving calibration:

$$\ell_{\text{NLL}}(p(\cdot \mid \boldsymbol{x}, \boldsymbol{y}_{<t}; \boldsymbol{\theta}), y_t) = -\log p(y_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t}; \boldsymbol{\theta}). \quad (6)$$

Selective Smoothing (SS). Introduced by Xie et al. (2024), the selective smoothing loss minimises the NLL for correct token predictions while maximising the entropy for incorrect token predictions:

$$\begin{aligned} &\ell_{\text{SS}}(p(\cdot \mid \boldsymbol{x}, \boldsymbol{y}_{<t}; \boldsymbol{\theta}), y_t) \\ &\quad = -(1-\alpha)\log p(y_t \mid \boldsymbol{x}, \boldsymbol{y}_{<t}; \boldsymbol{\theta})\mathbf{1}(\hat{y}_t = y_t) \\ &\qquad - \frac{\alpha}{|\mathcal{V}|}\sum_{y \in \mathcal{V}} \log p(y \mid \boldsymbol{x}, \boldsymbol{y}_{<t}; \boldsymbol{\theta})\mathbf{1}(\hat{y}_t \neq y_t), \end{aligned} \quad (7)$$

where $\hat{y}_t = \arg\max_{y \in \mathcal{V}} p(y \mid \boldsymbol{x}, \boldsymbol{y}_{<t}; \boldsymbol{\theta})$ is the model's top token prediction, $\mathbf{1}(\cdot)$ is the indicator function, and $\alpha \in [0, 1]$ controls the balance between the two terms.

Optimisation Objective. The parameters $\boldsymbol{\theta}$ are optimised by minimising the expected loss over the distribution, $p_{\mathcal{D}}$:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{x}, \boldsymbol{y}) \sim p_{\mathcal{D}}}\left[\ell\left(p(\cdot \mid \boldsymbol{x}; \boldsymbol{\theta}), \boldsymbol{y}\right)\right], \quad (8)$$

where ℓ denotes the loss function and can be the sequence-level aggregation of the token-level NLL of Equation (6) or SS of Equation (7). We optimise this objective using Stochastic Gradient Descent (SGD) over a calibration set of samples.

4 EMPIRICAL EVALUATION

Models and Datasets. We evaluate Llama-3.1-8B-Instruct (Dubey et al., 2024), Ministral-8B-Instruct-2410 (MistralAI, 2024), and Qwen-2.5-7B-Instruct (Team, 2024) on generative short-form question answering (QA). We focus on this task to enable direct comparison with prior work on semantic uncertainty quantification (Kuhn et al., 2023; Farquhar et al., 2024) and because semantic correctness is well-defined, unlike for more nuanced tasks such as summarisation or creative writing (see Section 5). We use TriviaQA (Joshi et al., 2017) and Natural Questions (NQ; Kwiatkowski et al. 2019) for closed-book QA, and SQuAD (Rajpurkar, 2016) for open-book QA. Each dataset is split into calibration and test sets (Section A.1), with calibration further divided into training and validation splits for hyperparameter selection (Section A.2). We use few-shot prompting throughout (Kuhn et al., 2023; Aichberger et al., 2025), with 10 in-context examples for TriviaQA and NQ, and 4 for SQuAD due to longer contexts.

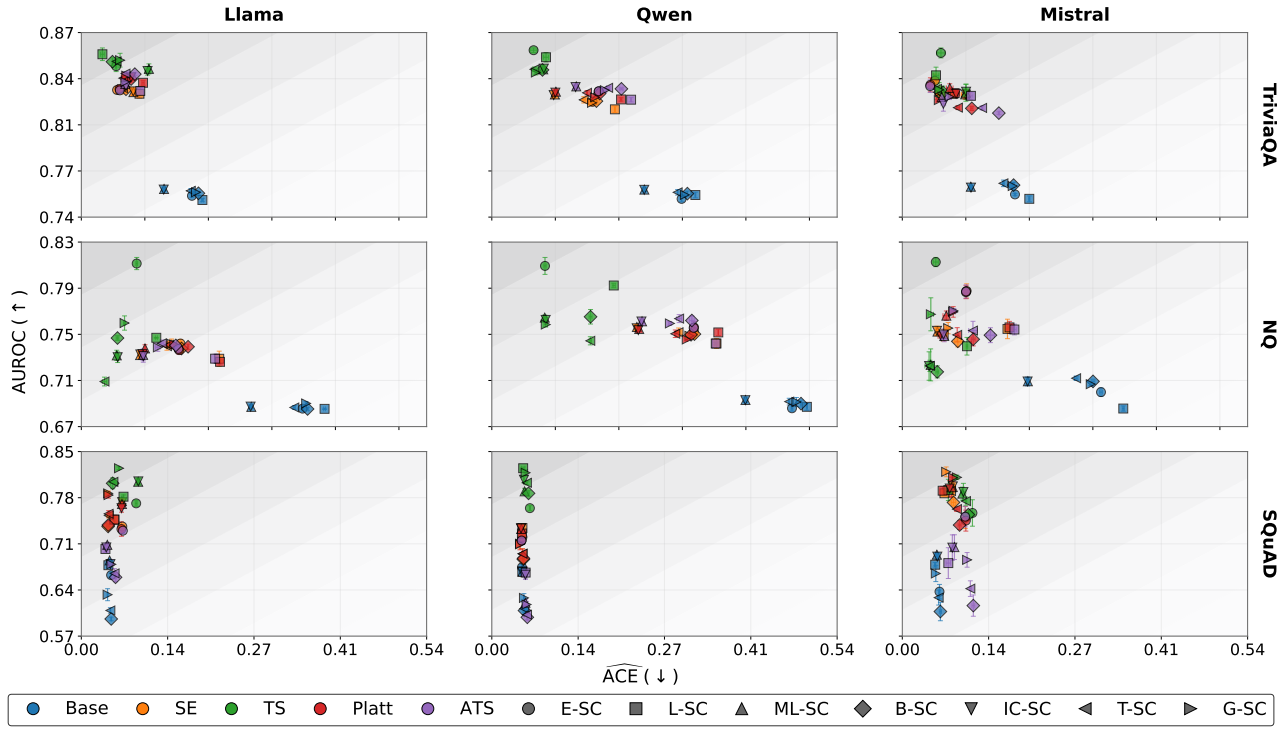


Figure (3): **Uncertainty Metrics of SC Measures Across Methods.** Mean and standard error of $\widehat{\text{ACE}}$ (↓) and AUROC (↑) scores for SC measures across baseline, calibration methods, and datasets. *Closer to top left of plots indicates better discrimination and calibration, and hence better overall semantic uncertainty quantification.*

Calibration Methods and Baselines. We compare a range of calibration methods and baselines:

- *Uncalibrated base model* (Base): the original instruction-tuned LM with temperature fixed at $\tau = 1.0$ (used by Farquhar et al. (2024)).
- *Base model with SE temperature setting* (SE): the original instruction-tuned LM, with temperature fixed at $\tau = 0.5$, the setting used in semantic entropy (Kuhn et al., 2023).
- *Temperature scaling* (TS): Optimised single scalar temperature parameter.
- *Adaptive temperature scaling* (ATS): Adaptive temperature scaling methods, optimising a temperature prediction head as detailed in Section 3.1.
- *Platt scaling* (Platt): Platt scaling optimising a diagonal affine transformation as in Section 3.1.

Producing semantic clusters. We form semantic clusters using the DeBERTa-V2-XXLarge NLI model (He et al., 2021), following the same methodology as prior work (Kuhn et al., 2023). This approach has been shown to produce consistent semantic clusterings for short-form generative question answering tasks, which are the focus of both previous studies and our evaluation for direct comparison (Kuhn et al., 2023; Farquhar et al., 2024). We further corroborate these findings in an experiment detailed in section F.8.

Selecting a Final Response. For each semantic measure in Section 3, we identify the most confident cluster. Since responses in clusters are semantically equivalent, any member could serve as the model's final answer. For robustness against minor variations, we randomly sample a subset of up to four responses from this top cluster and compare each against the ground truth. If at least one sampled response is correct, we mark the model's response as correct. This makes correctness evaluation less brittle to superficial phrasing differences and better reflects whether the model has identified the correct underlying meaning.

Measuring Correctness. In our QA experiments, we view correctness as binary, defined as $c = \mathbf{1}(\hat{\boldsymbol{y}} \sim \boldsymbol{y} \mid \boldsymbol{x})$, where $\sim$ denotes semantic equivalence given $\boldsymbol{x}$, and $\hat{\boldsymbol{y}} \sim p(\cdot \mid \boldsymbol{x}; \boldsymbol{\theta})$ is a model's final response. To precisely determine the equivalence of a response to a ground truth reference, we use a combination of criteria that includes soft matching, SQuAD-F1 and Rouge-L metrics (Kuhn et al., 2023; Farquhar et al., 2024; Aichberger et al., 2025). We discuss and justify this evaluation setup in more detail in Section B.2. In addition, in Section B.3, we note some common edge cases within this evaluation setup that we carefully mitigate in this work and which are not explicitly addressed in prior work using similar setups (Kuhn et al., 2023; Farquhar et al., 2024; Aichberger et al., 2025).

Table (1): **Discrimination Comparison of Entropy for Qwen.** Mean and standard error of AUROC (↑) values. (a) reports SE_{conf}, where correctness is determined by the most confident semantic cluster under a given SC measure. (b) reports $\text{SE}_{\text{vanilla}}$ from Kuhn et al. (2023), where correctness is determined via greedy decoding. Bold entries denote the best result within each SC measure per dataset, and underlined entries indicate the best overall per dataset.

		E-SC	L-SC	IC-SC	G-SC
TriviaQA	Base	$0.766_{\pm 0.002}$	$0.761_{\pm 0.002}$	$0.765_{\pm 0.002}$	$0.767_{\pm 0.002}$
	SE	$0.834_{\pm 0.002}$	$0.833_{\pm 0.001}$	$0.830_{\pm 0.002}$	$0.834_{\pm 0.002}$
	Platt	$0.839_{\pm 0.002}$	$0.840_{\pm 0.001}$	$0.836_{\pm 0.001}$	$0.839_{\pm 0.002}$
	ATS	$0.843_{\pm 0.001}$	$0.840_{\pm 0.001}$	$0.838_{\pm 0.002}$	$0.843_{\pm 0.001}$
	TS	$\mathbf{0.853_{\pm 0.002}}$	$\underline{\mathbf{0.865}}_{\pm 0.002}$	$\underline{\mathbf{0.864}}_{\pm 0.004}$	$\mathbf{0.851_{\pm 0.002}}$
NQ	Base	$0.693_{\pm 0.003}$	$0.691_{\pm 0.002}$	$0.692_{\pm 0.003}$	$0.693_{\pm 0.003}$
	SE	$0.749_{\pm 0.003}$	$0.752_{\pm 0.002}$	$0.746_{\pm 0.003}$	$0.750_{\pm 0.003}$
	Platt	$0.746_{\pm 0.003}$	$0.755_{\pm 0.005}$	$0.752_{\pm 0.002}$	$0.746_{\pm 0.003}$
	ATS	$0.759_{\pm 0.003}$	$0.754_{\pm 0.005}$	$0.743_{\pm 0.001}$	$\mathbf{0.759_{\pm 0.002}}$
	TS	$\mathbf{0.769_{\pm 0.004}}$	$\underline{\mathbf{0.795}}_{\pm 0.007}$	$\underline{\mathbf{0.789}}_{\pm 0.003}$	$0.747_{\pm 0.002}$
SQuAD	Base	$0.569_{\pm 0.007}$	$0.605_{\pm 0.004}$	$0.598_{\pm 0.003}$	$0.571_{\pm 0.008}$
	SE	$0.666_{\pm 0.007}$	$0.655_{\pm 0.006}$	$0.669_{\pm 0.005}$	$0.665_{\pm 0.007}$
	Platt	$0.665_{\pm 0.007}$	$0.649_{\pm 0.002}$	$0.660_{\pm 0.002}$	$0.666_{\pm 0.007}$
	ATS	$0.579_{\pm 0.009}$	$0.649_{\pm 0.002}$	$0.649_{\pm 0.006}$	$0.591_{\pm 0.009}$
	TS	$\underline{\mathbf{0.780}}_{\pm 0.003}$	$\mathbf{0.704_{\pm 0.004}}$	$\mathbf{0.772_{\pm 0.002}}$	$\underline{\mathbf{0.780}}_{\pm 0.003}$

a SE_{conf} (correctness via most confident cluster).

		E-SC	L-SC	IC-SC	G-SC
TriviaQA	Base	$0.755_{\pm 0.002}$	$0.755_{\pm 0.002}$	$0.754_{\pm 0.002}$	$0.755_{\pm 0.002}$
	SE	$0.833_{\pm 0.001}$	$0.833_{\pm 0.001}$	$0.833_{\pm 0.001}$	$0.833_{\pm 0.001}$
	Platt	$0.832_{\pm 0.002}$	$0.832_{\pm 0.002}$	$0.832_{\pm 0.002}$	$0.832_{\pm 0.002}$
	ATS	$0.834_{\pm 0.001}$	$0.832_{\pm 0.002}$	$0.832_{\pm 0.002}$	$0.834_{\pm 0.001}$
	TS	$\mathbf{0.849_{\pm 0.001}}$	$\mathbf{0.844_{\pm 0.001}}$	$\underline{\mathbf{0.857}}_{\pm 0.002}$	$\mathbf{0.848_{\pm 0.002}}$
NQ	Base	$0.690_{\pm 0.002}$	$0.689_{\pm 0.002}$	$0.689_{\pm 0.002}$	$0.691_{\pm 0.002}$
	SE	$0.749_{\pm 0.001}$	$\mathbf{0.745_{\pm 0.001}}$	$0.745_{\pm 0.001}$	$\mathbf{0.750_{\pm 0.001}}$
	Platt	$0.745_{\pm 0.001}$	$0.743_{\pm 0.003}$	$0.744_{\pm 0.003}$	$0.744_{\pm 0.002}$
	ATS	$0.740_{\pm 0.002}$	$0.743_{\pm 0.003}$	$0.741_{\pm 0.002}$	$0.740_{\pm 0.002}$
	TS	$\underline{\mathbf{0.758}}_{\pm 0.001}$	$0.737_{\pm 0.002}$	$\mathbf{0.751_{\pm 0.004}}$	$0.745_{\pm 0.001}$
SQuAD	Base	$0.590_{\pm 0.001}$	$0.595_{\pm 0.001}$	$0.595_{\pm 0.002}$	$0.595_{\pm 0.001}$
	SE	$0.653_{\pm 0.002}$	$0.653_{\pm 0.001}$	$0.653_{\pm 0.002}$	$0.653_{\pm 0.002}$
	Platt	$0.653_{\pm 0.002}$	$0.652_{\pm 0.004}$	$0.649_{\pm 0.004}$	$0.653_{\pm 0.002}$
	ATS	$0.575_{\pm 0.007}$	$0.652_{\pm 0.004}$	$0.624_{\pm 0.002}$	$0.578_{\pm 0.009}$
	TS	$\mathbf{0.707_{\pm 0.007}}$	$\mathbf{0.717_{\pm 0.005}}$	$\underline{\mathbf{0.748}}_{\pm 0.003}$	$\mathbf{0.707_{\pm 0.007}}$

b $\text{SE}_{\text{vanilla}}$ (correctness via greedy-decoding).

Calibration and Discrimination Metrics. We measure semantic calibration via Adaptive Calibration Error ($\widehat{\text{ACE}}$) (Nixon et al., 2019); Section F.10 shows that our calibration results are robust to the choice of calibration metric including ECE (Naeini et al., 2015) and CORP-MCB (Dimitriadis et al., 2021) metrics). We assess the discrimination of semantic measures across methods by reporting AUROC scores (Bradley, 1997). See Section B.1 for more details.

Reporting Results. We perform four independent runs, which include both calibration and evaluation stages. We compute the mean and standard error of each evaluation metric over the four inference runs.

4.1 Optimised Temperatures Improve Semantic Uncertainty Quantification

Figure 3 shows $\widehat{\text{ACE}}$ and AUROC evaluated across test sets, semantic confidence (SC) measures, and methods. Overall, optimised Temperature Scaling (TS) proves to be a simple and robust method for improving semantic uncertainty quantification, echoing the key findings of (Guo et al., 2017) non-trivially in a more complex domain. In particular, TS outperforms complex methods such as Platt scaling and ATS, whose performance is less robust in a semantic setting. It consistently produces results toward the top-left (lower $\widehat{\text{ACE}}$, higher AUROC) on both closed-book (TriviaQA, NQ) and open-book (SQuAD) datasets. On the open-book SQuAD dataset, where base models are already well-calibrated, improvements are primarily in discrimination, and G-SC consistently yields the strongest AUROC. Conversely, on the closed-book TriviaQA and NQ datasets, base models exhibit poor calibration. In this setting, E-SC and G-SC provide the best balance, delivering consistent improvements in both $\widehat{\text{ACE}}$ and AUROC. Across all experiments, E-SC, L-SC, and G-SC emerge as the most effective semantic measures. Finally, and crucially, optimised TS outperforms the fixed, ad-hoc temperature settings from prior work, including $\tau = 1.0$ from Farquhar et al. (2024) (Base) and $\tau = 0.5$ from Kuhn et al. (2023) (SE).

Key Takeaway: Optimised temperature scaling improves semantic calibration and discrimination.

4.2 Optimised Token-Level Temperature Improves Semantic Entropy

We evaluate our methods on the downstream task of discriminating correct from incorrect instances using semantic entropy (SE) (Kuhn et al., 2023). We compare our principled variant, SE_{conf}, which selects the final answer from the most confident semantic class (see Section 4), against the ad-hoc baseline from prior work, $\text{SE}_{\text{vanilla}}$. The latter determines correctness using a greedily decoded final answer, while estimating uncertainty by sampling from a temperature-smoothed distribution (Kuhn et al., 2023). As a result, the reported performance of the Base method can differ between panels (a) and (b) of Section 4, despite using the same model and temperature ($\tau = 1.0$), due to the differences in how the final response is selected for semantic correctness evaluation.

Section 4 reports the results for the Qwen model across datasets. TS consistently improves the discriminative power of each semantic measure's entropy under both formulations, crucially outperforming the fixed-temperature heuristics (e.g., $\tau \in \{0.5, 1.0\}$) from prior

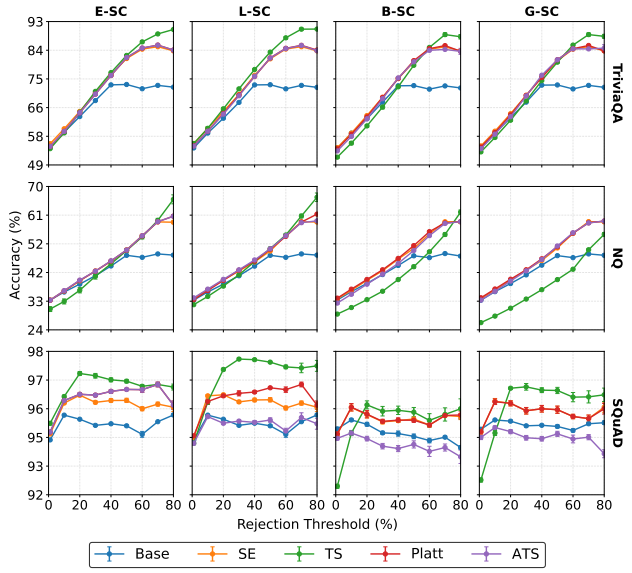


Figure (4): **Selective Accuracy for Qwen Across Rejection Rates and Datasets.** Mean selective accuracy (↑) curves, comparing E-SC, L-SC, B-SC and G-SC measures across baseline and calibration methods. Error bars show one standard error across runs.

work once again (Kuhn et al., 2023; Farquhar et al., 2024). Gains are particularly notable on well-calibrated datasets like SQuAD, highlighting the favourable trade-off TS provides between calibration and discrimination. Moreover, a direct comparison reveals that our SE_{conf} shown in panel (a) approach generally achieves higher discrimination (AUROC) than $\text{SE}_{\text{vanilla}}$ shown in panel (b). This validates deriving the final answer according to the semantic measures used for uncertainty quantification provides a more principled and effective method for downstream semantic UQ.

Key Takeaway: Temperature scaling enhances semantic entropy discrimination. Selecting final responses via semantic confidence yields superior discrimination to adhoc prior approaches.

4.3 Selective Prediction

Figure 4 presents selective accuracy results. We observe distinct behaviours across datasets: on SQuAD, TS yields pronounced gains even at low rejection rates, whereas on TriviaQA, benefits appear primarily at high rejection rates by mitigating the most overconfident errors. On NQ, we note a drop in base performance for B-SC and G-SC; this occurs due to lower base accuracy when using B-SC and G-SC for selecting the most probable meaning under the model. Crucially, however, TS consistently drives a sharper rate of improvement (steeper slope) across all settings once rejection begins. This demonstrates that, regardless of the starting point, TS aligns confidence scores more effectively with correctness, ensuring that rejecting low-confidence predictions rapidly and reliably filters incorrect responses.

Key Takeaway: Temperature scaling improves selective accuracy by strictly aligning confidence with correctness, driving a generally steeper rise in accuracy as low confidence examples are filtered out.

4.4 Number of Generations Ablation

Figure 5 shows AUROC and $\widehat{\text{ACE}}$ results for the Llama model on NQ as the number of sampled generations per example varies from 5 to 25. AUROC (bottom row) remains largely stable across sample sizes, indicating that varying the number of sampled generations leads to minor variation in discrimination across measures. Calibration ($\widehat{\text{ACE}}$, top row) improves as the number of samples increases, with gains beginning to saturate beyond roughly 10 samples. Across all sample sizes and measures, we find that TS enhances semantic UQ, primarily by improving calibration. These results justify our use of 10 generations throughout the paper, consistent with prior work (Kuhn et al., 2023).

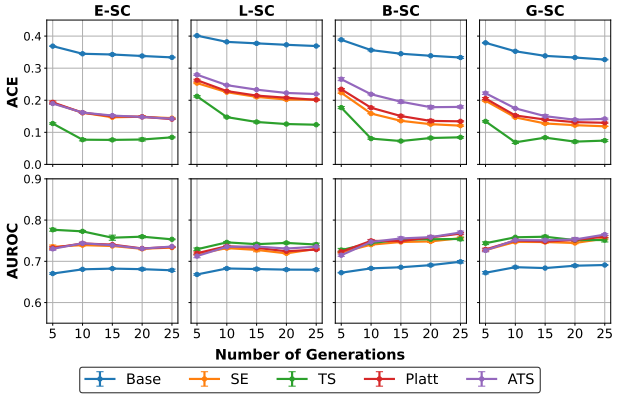


Figure (5): **SC Measures over Varying Numbers of Samples for Llama.** Mean $\widehat{\text{ACE}}$ (↓) and AUROC (↑) on the NQ dataset over varying number of sample generations per example. Error bars show one standard error across runs.

Key Takeaway: Semantic calibration improves with sample size. Temperature scaling consistently boosts performance across all sample counts.

5 DISCUSSION

We demonstrated that existing semantic confidence measures are systematically miscalibrated. Optimising a single scalar temperature parameter (TS), whilst

being surprisingly simple, computationally cheap and easy to implement, consistently improved both semantic calibration and discrimination. Crucially, this enhancement extends to downstream semantic entropy. Simple token-level temperature scaling outperformed both heuristic fixed temperatures and expressive token-level methods like ATS and Platt scaling. Our findings represent a meaningful update of the seminal work by Guo et al. (2017), linking token-level classification to semantic prediction space in modern autoregressive transformer models in the generative QA domain.

The effectiveness of TS is driven by its global inductive bias. Unlike expressive, computationally expensive methods such as ATS (Xie et al., 2024) which optimise for localised, token-specific calibration goals, TS is constrained to a single global scalar for entire sequences. This global constraint regularises against overly specific token-level adjustments that can overfit semantically irrelevant or filler words to reduce calibration loss. We find that this sequence-level bias transfers more reliably to semantic calibration than token-local methods. Overall, the choice of calibration method has a larger impact on performance than the specific form of the semantic measure itself.

Our analysis focused on short-form generative QA where correctness admits a clear binary definition. Extending semantic calibration to tasks with partially correct outputs, such as summarisation, remains challenging as the notion of calibration is less well-defined in such settings (Wei et al., 2024). Developing principled evaluation frameworks for semantic calibration beyond binary correctness is a direction for future research

6 CONCLUSION

We systematically evaluated semantic calibration and discrimination, showing that base models and fixed-temperature heuristics produce miscalibrated and poorly discriminative semantic confidence estimates. We demonstrated that optimising a single scalar temperature parameter consistently improves calibration, discrimination, and downstream semantic entropy, outperforming both heuristic baselines and more sophisticated token-level methods. Consequently, temperature scaling offers a surprisingly simple, cheap, plug-and-play solution for practitioners to enhance semantic reliability without altering existing workflows.

Acknowledgements

TGJR acknowledges support from the Foundational Research Grants program at Georgetown University's Center for Security and Emerging Technology.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*. 2, 16

Aichberger, L., Schweighofer, K., Ielanskyi, M., and Hochreiter, S. (2025). Improving uncertainty estimation through semantically diverse language generation. In *The Thirteenth International Conference on Learning Representations*. 5, 6

Anthropic (2025). Claude 4.5 model family announcement. `https://platform.claude.com/docs/en/about-claude/models/whats-new-claude-4-5`. Accessed: 2026-01. 24

Bachmann, M. (2021). Rapidfuzz: A fast fuzzy string matching library in c++ and python. 16

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR. 2

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159. 7, 14, 15

Cantoni, O. and Picard, J. (2004). *Statistical Learning Theory and Stochastic Optimization: Ecole D'Ete de Probabilites de Saint-Flour XXXI-2001*. Springer. 4

De Leeuw, J., Hornik, K., and Mair, P. (2010). Isotone optimization in r: pool-adjacent-violators algorithm (pava) and active set methods. *Journal of statistical software*, 32:1–24. 15

Desai, S. and Durrett, G. (2020). Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302. 19

Dimitriadis, T., Gneiting, T., and Jordan, A. I. (2021). Stable reliability diagrams for probabilistic classifiers. *Proceedings of the National Academy of Sciences*, 118(8):e2016191118. 7, 15, 16, 25

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*. 5, 16

Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630. 1, 2, 3, 4, 5, 6, 7, 8, 16, 25

Flach, P. A. (2016). Classifier calibration. In *Encyclopedia of machine learning and data mining*. Springer US. 2

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378. 5

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR. 2, 5, 7, 9

He, P., Liu, X., Gao, J., and Chen, W. (2021). Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*. 6

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*. 19

Huang, Y., Li, W., Macheret, F., Gabriel, R. A., and Ohno-Machado, L. (2020). A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association*, 27(4):621–633. 1

Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. (2017). Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*. 5

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al. (2022). Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*. 2

Kuhn, L., Gal, Y., and Farquhar, S. (2023). Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *International Conference on Learning Representations*. 1, 2, 3, 4, 5, 6, 7, 8, 16, 17, 21, 22, 24, 25

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al. (2019). Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466. 5

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30. 2

Lin, Z., Trivedi, S., and Sun, J. (2023). Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*. 3

Liu, J., Lin, Z., Padhy, S., Tran, D., Bedrax Weiss, T., and Lakshminarayanan, B. (2020). Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Advances in neural information processing systems*, 33:7498–7512. 2

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*. 12

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press. 5, 27

Malinin, A. and Gales, M. (2021). Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*. 2, 3

MistralAI (2024). Introducing ministrel: Our new lightweight model. Accessed: 2025-01-17. 5

Mukhoti, J., Kirsch, A., van Amersfoort, J., Torr, P. H., and Gal, Y. (2023). Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24384–24394. 2

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology and Climatology*, 12(4):595–600. 15

Murray, K. and Chiang, D. (2018). Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 212–223. 2, 3

Naeini, M. P., Cooper, G., and Hauskrecht, M. (2015). Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29 (1). 7, 14

Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632. 5

Nikitin, A., Kossen, J., Gal, Y., and Marttinen, P. (2024). Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *arXiv preprint arXiv:2405.20003*. 3

Nixon, J., Dusenberry, M. W., Zhang, L., Jerfel, G., and Tran, D. (2019). Measuring calibration in deep learning. In *CVPR workshops*, volume 2. 7, 14

Platt, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74. 2, 5

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training. 12

Rajpurkar, P. (2016). Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*. 5

Santilli, A., Golinski, A., Kirchhof, M., Danieli, F., Blaas, A., Xiong, M., Zappella, L., and Williamson, S. (2025). Revisiting uncertainty quantification evaluation in language models: Spurious interactions with response length bias results. *arXiv preprint arXiv:2504.13677*. 1

Team, Q. (2024). Qwen2.5: Advancing open-source language models. Accessed: 2025-01-17. 5

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*. 5

Wei, J., Yang, C., Song, X., Lu, Y., Hu, N., Huang, J., Tran, D., Peng, D., Liu, R., Huang, D., et al. (2024). Long-form factuality in large language models. *Advances in Neural Information Processing Systems*, 37:80756–80827. 9

Wenzel, F., Roth, K., Veeling, B., Swiatkowski, J., Tran, L., Mandt, S., Snoek, J., Salimans, T., Jenatton, R., and Nowozin, S. (2020). How good is the bayes posterior in deep neural networks really? In *International Conference on Machine Learning*, pages 10248–10259. PMLR. 4

Williams, A., Nangia, N., and Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics. 3

Xie, J., Chen, A. S., Lee, Y., Mitchell, E., and Finn, C. (2024). Calibrating language models with adaptive temperature scaling. *arXiv preprint arXiv:2409.19817*. 2, 5, 9, 13

Yang, A. X., Robeyns, M., Wang, X., and Aitchison, L. (2023). Bayesian low-rank adaptation for large language models. *arXiv preprint arXiv:2308.13111*. 2

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Not Applicable]
   (b) Complete proofs of all theoretical results. [Not Applicable]
   (c) Clear explanations of any assumptions. [Not Applicable]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

A Dataset Splits, Hyperparameters, and Computational Resources

A.1 Dataset Splits

Our calibration pipeline is organised into three stages: calibration training, calibration validation, and final test evaluation. Each stage relies on dedicated dataset splits to ensure that training, model selection, and evaluation remain strictly separated.

- **Calibration Training.** The data is first divided into calibration-training and calibration-validation subsets. Using the pre-trained base LMs, we fit calibration parameters on the calibration-training split. Specifically, we optimise either a scalar temperature parameter or Platt scaling or temperature head parameters for 2 epochs.
- **Calibration Validation.** The held-out calibration-validation split is used to select the best-performing hyperparameters from the calibration training stage.
- **Final Test Evaluation.** Finally, we evaluate the selected models on a held-out test set. Results are reported on this dataset across methods, SC measures, and evaluation metrics (e.g., $\widehat{\text{ACE}}$ and AUROC). For each setting, we include error bars to reflect performance variability and enable fair comparisons across both methods and measures.

To ensure fair cross-dataset comparison, we constrain the size of each split to be approximately matched across datasets. The exact split sizes for each dataset are provided in Table 2.

Table (2): **Dataset Split Sizes.** For each dataset, we report the number of examples used at different stages of the calibration pipeline: calibration training, calibration validation, and final test evaluation. Split sizes are matched across datasets to ensure fair comparison.

Stage	Split	TriviaQA	Natural Questions	SQuAD
Calibration	Training	59374	61600	59577
	Validation	2000	2000	2000
Final Evaluation	Test	2000	2000	2000

A.2 Hyperparameter settings

Below we list the hyperparameter settings swept over for calibration optimisation. All models are trained using the AdamW optimiser (Loshchilov and Hutter, 2017) with a cosine-annealing learning rate scheduler and a linear warm-up over the first 10% of training examples within the first epoch of training (Radford et al., 2018). Unless otherwise specified, all methods use the following shared hyperparameters: number of epochs $\{2\}$, calibration loss $\{\ell_{\text{NLL}}, \ell_{\text{SS}}\}$, and sweeps over the SS-loss weight $\alpha \in \{0.1, 0.25, 0.5, 0.75\}$ as well as G-SC and T-SC parameter $\alpha \in \{0.5, 0.75, 1.25\}$.

Optimised Temperature Scaling (TS). For scalar temperature scaling, we sweep over:

- Learning rate: $\{10^{-4}\}$.
- Initial temperature τ: $\{1.0\}$.

Adaptive Temperature Scaling (ATS). For adaptive temperature head optimisation, we sweep over:

- Learning rate: $\{10^{-5}\}$ (smaller for stability).
- Weight decay: $\{0.0, 0.01\}$.

- Gradient clipping (max norm): {1.0} (again for stability).
- Temperature head architecture: a single transformer block from the LLaMA-2 family, following Xie et al. (2024).

Platt Scaling (PS). For Platt scaling calibration, we sweep over:

- Learning rate: $\{10^{-5}\}$.
- Weight decay: $\{0.0, 0.01\}$.
- Gradient clipping (max norm): {1.0}.
- Transformation: affine mapping constrained to be diagonal, as in Xie et al. (2024), to reduce vocabulary-scale cost.

Hyperparameter Selection. Hyperparameters are selected independently for each SC measure based on Brier score on the calibration-validation set. We selected based on Brier score to balance calibration and discrimination whilst avoiding degenerate selection that can arise from optimising solely for calibration.

A.3 Computational Resources

All models were trained and evaluated on our internal cluster equipped with NVIDIA A40 GPUs. Calibration training for each method was performed on a single GPU. During evaluation, we executed model inference and likelihood computations on one GPU, while semantic clustering using an NLI model was performed concurrently on a separate GPU. This setup ensures efficient parallelisation of the evaluation pipeline while maintaining manageable memory and runtime requirements.

B Evaluation

B.1 Evaluation Metrics

Below we detail the evaluation metrics that we report for our results presented in both the main paper and within the supplementary material. We frame these metrics in the context of representing an NLG task, specifically QA within this paper, as a binary task, with model's predictions being correct $c = 1$ or incorrect $c = 0$.

Let $(\boldsymbol{x}, \boldsymbol{y}) \sim p_{\mathcal{D}}$ be a dataset sample, where $\boldsymbol{x} \in \mathcal{V}^*$ is the input and $\boldsymbol{y} \in \mathcal{V}^*$ is the ground-truth label. Let p be a language model, and let $\hat{\boldsymbol{y}} \sim p(\cdot \mid \boldsymbol{x})$ be a model's prediction conditioned on the input $\boldsymbol{x}$. The correctness of a model prediction $\hat{\boldsymbol{y}}$ is a function of the input, the ground-truth label, and the prediction itself, denoted as $c(\boldsymbol{x}, \boldsymbol{y}, \hat{\boldsymbol{y}}) \in \{0, 1\}$.

Expected Calibration Error (ECE) Calibration errors aim to quantify the misalignment between a model's predicted confidence and its actual correctness. The (L1) expected calibration error is defined as:

$$\mathrm{ECE} = \mathbb{E}_{\boldsymbol{x} \sim p_{\mathcal{D}}} \left[\left| \mathbb{P}(c = 1 \mid p(\hat{\boldsymbol{y}} \mid \boldsymbol{x}) = p) - p \right| \right]. \tag{9}$$

Following Naeini et al. (2015), ECE is empirically estimated by binning predictions from a dataset sample into M intervals, denoted $\{B_m\}_{i=1}^{M}$. The weighted average of the absolute accuracy-confidence difference is then computed over the bins to estimate the expectation in Equation (9):

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

where:

$$\mathrm{conf}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} p(\hat{\boldsymbol{y}}_i \mid \boldsymbol{x}_i), \quad \mathrm{acc}(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} c\big(\boldsymbol{x}_i, \boldsymbol{y}_i, \hat{\boldsymbol{y}}_i\big).$$

We list two separate estimates of this metric, which arise from different binning schemes for confidence values within the $[0, 1]$ interval:

- $\widehat{\mathbf{ECE}}$**:** An even partition where each bin has an equal range provides the standard empirical approximation of the ECE.
- $\widehat{\mathbf{ACE}}$**:** The empirical *adaptive calibration error* (Nixon et al., 2019) is obtained using a binning scheme where each of the M bins contains an equal number of examples, with the bin ranges varying to accommodate this constraint.

$\widehat{\mathrm{ACE}}$ is generally considered a more robust metric than $\widehat{\mathrm{ECE}}$ because it mitigates issues with uneven sample sizes across bins, which can lead to unreliable and noisy bin estimates in sparse regions of the confidence spectrum (Nixon et al., 2019). Therefore, throughout this work, we report the $\widehat{\mathrm{ACE}}$.

AUROC The Area Under the Receiver Operating Characteristic Curve (AUROC) measures how well confidence scores distinguish between correct and incorrect responses. The ROC curve is constructed by varying a confidence threshold λ and plotting the true positive rate (TPR) against the false positive rate (FPR) at each threshold (Bradley, 1997). At its core, the AUROC value represents the probability that a randomly chosen correct response will receive a higher confidence score than a randomly chosen incorrect response (Bradley, 1997).

Given a dataset of N examples with inputs $\boldsymbol{x}_i \in \mathcal{V}^*$, model-generated responses $\hat{\boldsymbol{y}}_i \in \mathcal{V}^*$, correctness indicators $c_i = \mathbf{1}(\hat{\boldsymbol{y}}_i \sim \boldsymbol{y}_i \mid \boldsymbol{x})$, and model confidence scores $p(\hat{\boldsymbol{y}}_i \mid \boldsymbol{x}_i)$ in response $\hat{\boldsymbol{y}}_i$, we define:

$$\mathrm{TPR}(\lambda) = \frac{\sum_{i:c_i=1} \mathbf{1}(p(\hat{\boldsymbol{y}}_i \mid \boldsymbol{x}_i) \geq \lambda)}{\sum_{i:c_i=1} 1}, \quad \mathrm{FPR}(\lambda) = \frac{\sum_{i:c_i=0} \mathbf{1}(p(\hat{\boldsymbol{y}}_i \mid \boldsymbol{x}_i) \geq \lambda)}{\sum_{i:c_i=0} 1}.$$

AUROC is then computed as:

$$\mathrm{AUROC} = \int_0^1 \mathrm{TPR}(\lambda)\, d\mathrm{FPR}(\lambda),$$

A higher AUROC indicates better uncertainty quantification through the lens of ranking, where a model's confidence scores can better discriminate between correct and incorrect predictions (Bradley, 1997). An AUROC of 0.5 corresponds to a confidence metric that is no better than random guessing. An AUROC below 0.5 indicates that the metric is inverted, in the sense that it systematically assigns higher scores to incorrect predictions than to correct ones.

Brier Score (BS) The Brier Score (BS) is a proper scoring rule that evaluates both calibration and correctness (Murphy, 1973). It is defined and decomposes into interpretable terms as follows:

$$\begin{aligned}
\mathrm{BS} &= \mathbb{E}_{(\mathbf{x},\boldsymbol{y})\sim p_{\mathcal{D}}}\left[(p(\hat{\mathbf{y}} \mid \mathbf{x}) - c)^2\right] \\
&= \underbrace{\mathbb{E}_{\mathbf{x}\sim p_{\mathcal{D}}}\left[(p(\hat{\mathbf{y}} \mid \mathbf{x}) - \mathbb{P}(c = 1 \mid \mathbf{x}))^2\right]}_{\text{Calibration}} + \underbrace{\mathbb{E}_{\mathbf{x}\sim p_{\mathcal{D}}}\mathbb{V}[c \mid \mathbf{x}]}_{\text{Refinement}} \\
&= \underbrace{\mathbb{E}_{\mathbf{x}\sim p_{\mathcal{D}}}\left[(p(\hat{\mathbf{y}} \mid \mathbf{x}) - \mathbb{P}(c = 1 \mid \mathbf{x}))^2\right]}_{\text{Calibration}} + \underbrace{\mathbb{V}[c]}_{\text{Uncertainty}} - \underbrace{\mathbb{V}_{\mathbf{x}\sim p_{\mathcal{D}}}[\mathbb{E}[c \mid \mathbf{x}]]}_{\text{Resolution}}.
\end{aligned}$$

The last line uses the law of total variance to split the refinement term into uncertainty and resolution components. Here, c is a Bernoulli random variable due to being either 0 or 1, which allows simple interchange of probability and expectation.

We can interpret the three terms as follows: *calibration* measures the agreement between the model's predicted probabilities and the true frequencies of outcomes; *resolution* quantifies how much the true outcome probabilities vary across inputs, rewarding models that assign distinct probabilities to different cases (and is closely related to discrimination); and finally, *uncertainty* represents the inherent, irreducible variance in the data, which cannot be reduced by the model.

CORP-MCB. The CORP methodology (Dimitriadis et al., 2021) provides an alternative framework that addresses limitations of standard binning algorithms like $\widehat{\mathrm{ECE}}$ and $\widehat{\mathrm{ACE}}$ through a non-parametric, isotonic regression approach.

Given a calibration set $\mathcal{D} = \{(p_1, c_1), \ldots, (p_N, c_N)\}$, where p_i is the model's predicted probability for input x_i and $c_i \in \{0, 1\}$ denotes the correctness of the response $\hat{y}_i$, we estimate the conditional event probability (CEP), $\mathbb{P}(c = 1 \mid p)$, using isotonic least-squares regression:

$$f_* = \operatorname*{arg\,min}_{f\in\mathcal{M}} \sum_{i=1}^{N} (f(p_i) - c_i)^2, \tag{10}$$

where $\mathcal{M} := \{f : [0, 1] \to [0, 1] \mid f(p_1) \le f(p_2) \le \cdots \le f(p_N)\}$ denotes the set of monotonic functions mapping forecasts in $[0, 1]$ to recalibrated forecasts in $[0, 1]$. For each input forecast p_i, the recalibrated forecast is $\tilde{p}_i = f_*(p_i)$. The optimisation problem in eq. (10) is solved non-parametrically using the pool-adjacent-violators (PAV) algorithm (De Leeuw et al., 2010), which runs in $\mathcal{O}(N)$.

Given a score function $S : [0, 1] \times \{0, 1\} \to \mathbb{R}_{\ge 0}$, we define the following empirical scores over the dataset $\mathcal{D}$:

$$S_p = \frac{1}{N}\sum_{i=1}^{N} S(p_i, c_i), \quad S_{\tilde{p}} = \frac{1}{N}\sum_{i=1}^{N} S(\tilde{p}_i, c_i), \quad S_r = \frac{1}{N}\sum_{i=1}^{N} S(r, c_i),$$

where $S_p, S_{\tilde{p}}, S_r$ denote the empirical scores of the original forecasts, recalibrated forecasts, and a constant reference forecast r, respectively. The decomposition of S_p is given by:

$$S_p = \underbrace{(S_p - S_{\tilde{p}})}_{\text{MCB}} - \underbrace{(S_r - S_{\tilde{p}})}_{\text{DSC}} + \underbrace{S_r}_{\text{UNC}}.$$

Here, the terms correspond to: *MCB*, a miscalibration term that is non-negative and zero when predictions are perfectly calibrated; *DSC*, a discrimination term measuring the ability of the model to separate correct from incorrect outcomes; and *UNC*, the inherent uncertainty in the data. The MCB term provides an additional quantitative measure of calibration.

We use the Brier score as our score function S, as it is a proper scoring rule. Following Dimitriadis et al. (2021), the constant reference forecast is taken as the empirical correctness $r = \frac{1}{N}\sum_{i=1}^{N} c_i$. This choice, together with using PAV-transformed recalibrated predictions, satisfies the calibration condition specified in their work (c.f. Equation 4).

Selective Prediction via Selective Accuracy Fix a threshold $\eta \in [0, 1]$ and define the *coverage set*:

$$S_\eta = \left\{ \mathbf{x} \in \mathcal{V}^* \;\middle|\; p(\hat{\mathbf{y}} \mid \mathbf{x}) \geq \eta, \text{ for } \hat{\boldsymbol{y}} \sim p(\cdot \mid \boldsymbol{x}) \right\}.$$

This is the set of examples for which the model's confidence exceeds the threshold η. The *selective accuracy* is then the average correctness over this set:

$$\mathcal{A}_{\text{sel}}(\eta) = \frac{1}{|S_\eta|} \sum_{\mathbf{x} \in S_\eta} c(\mathbf{x}, \mathbf{y}, \hat{\mathbf{y}}).$$

Varying η traces out a selective accuracy curve, showing how model accuracy changes as we focus on increasingly confident predictions. For a well-calibrated confidence score, one expects $\mathcal{A}_{\text{sel}}(\eta)$ to generally increase as lower-confidence examples are filtered out.

B.2 Evaluation of Accuracy

To evaluate accuracy in generative QA tasks, we apply a multi-step procedure that balances correctness assessment with computational efficiency and overall cost:

- **Initial Text Cleaning:** The model's response is preprocessed by discarding any extraneous text beyond the first direct answer to the question.
- **Direct Answer Matching:** If any reference answer is present verbatim in the response, the response is considered correct.
- **Fuzzy Matching:** When a direct match is absent, we apply fuzzy string matching (Bachmann, 2021) using string-distance metrics. Responses exceeding a similarity threshold of 90 are classified as correct.
- **SQuAD F1 Evaluation:** For remaining unmatched responses, we compute the SQuAD-F1 score. Responses with F1 above 50.0 are considered correct as done by Farquhar et al. (2024).

An alternative evaluation approach could involve using a model such as Llama-3.1 (Dubey et al., 2024) or GPT-4 (Achiam et al., 2023) as a judge to assess equivalence with reference answers. However, this introduces additional cost and latency. Our current methodology follows prior work (Kuhn et al., 2023; Farquhar et al., 2024), which relies on token-level matching and has been shown to be effective in practice for short-form response QA tasks such as those that we work with in this paper. Unlike (Farquhar et al., 2024), we find that combining multiple accuracy checks beyond SQuAD-F1 is necessary to reduce both false positives and negatives in correctness assessment.

B.3 Practical Considerations for Semantic Clustering and Evaluation Accuracy in a Semantic Uncertainty Pipeline

We note that our clustering procedure is identical to that of prior work (Kuhn et al., 2023; Farquhar et al., 2024), and our complete evaluation of correctness pipeline is detailed in Section B.2. As discussed in the aforementioned section, while prior work often adopts a single correctness criterion, we find that relying on only one measure is insufficient to reliably judge answer correctness. In addition, we now identify several practical pitfalls that are not adequately addressed as far as we can tell in the existing literature that affect both accuracy metrics and semantic clustering via NLI models. Below, we highlight these issues and our remedies.

- **Clustering Numeric Responses.** Off-the-shelf NLI models frequently fail to recognise equivalence between numeric and verbal forms of answers (e.g. "20" vs. "twenty"). Without intervention, this leads to over-clustering. *Remedy:* we normalise all numeric responses to their digit form before clustering (e.g. "twenty" $\rightarrow$ "20").

- **Clustering and Evaluating Dates.** Standard NLI-based criteria struggle with varied date formats and often produce false positives for nearby dates (e.g. "20th December 1988" vs. "19/12/1988"). Moreover, the standard components of the evaluation criteria discussed in Section B.2 also penalise model outputs that consist of full dates when the ground truth answers provide only a year. *Remedy:* we first parse each date into an ISO-8601 string (YYYY-MM-DD). Then, for cases where only a year is required, we accept any normalised date within that year. Finally, we apply a hierarchical correctness check: exact match on year, month, and day only if the ground truth specifies those levels of granularity.

If unaddressed, these issues can confound both semantic clustering and evaluation scores. By applying normalisation and hierarchical checking for numbers and dates, we ensure that our reported metrics reflect true semantic correctness rather than artifacts of formatting.

B.4 Evaluation Computational Complexity

Let $M \in \mathbb{N}$ denote the number of generations sampled per input, and let $L \in \mathbb{Z}$ denote the average sequence length. The overall time complexity of our evaluation pipeline decomposes into the following steps:

- **Sample generation.** Generate M samples via multinomial sampling from the autoregressive LM. Using KV caching, each forward step requires retrieving prior keys and values, yielding a total cost of $\mathcal{O}(ML)$. This step is the dominant cost as it is inherently sequential, requiring L separate model invocations.
- **Text normalisation.** Normalise each sample to extract the model's final answer, discarding extraneous generated content. This step is negligible in cost.
- **Likelihood computation.** Compute the length-normalised log-likelihood of each normalised sample under the model, requiring an additional forward pass. This also scales as $\mathcal{O}(ML)$, but is substantially cheaper than generation because the entire sequence is processed in a single parallel pass.
- **Clustering.** Perform semantic clustering by running $\binom{M}{2} = \mathcal{O}(M^2)$ pairwise NLI comparisons to determine entailment relations between responses. Although quadratic in M, this step is far cheaper than LM generation or likelihood scoring, since it uses a much smaller NLI model.
- **Semantic confidence calculation.** Compute semantic confidence measures using the cluster structure and log-likelihoods. This is negligible compared to the other steps.

In practice, the computational cost is dominated by autoregressive sample generation, which scales as $\mathcal{O}(ML)$ with KV caching. Semantic clustering is formally quadratic in the number of samples ($\mathcal{O}(M^2)$), but is comparatively inexpensive because it relies on a much smaller NLI model and can be efficiently batched. Likelihood computation contributes modestly at $\mathcal{O}(ML)$, while text normalisation and semantic-confidence calculations are negligible. Overall, a modest sample budget ($M \approx 10$) provides representative coverage while keeping runtime tractable; performance gains saturate around 10–15 samples (Section 4.4), consistent with prior observations (Kuhn et al., 2023).

C On the Necessity of Jointly Evaluating Calibration and Discrimination in UQ

As discussed in the introduction, a central contribution of this work is the joint evaluation of *discrimination* and *calibration* for semantic uncertainty quantification (UQ). Prior work on semantic UQ has largely focused on discrimination, typically evaluating confidence-derived quantities such as semantic entropy using ranking-based metrics (e.g., AUROC), while neglecting calibration. This is a fundamental limitation: calibration and discrimination capture distinct aspects of predictive uncertainty, and strong performance on one does not imply strong performance on the other. Consequently, discrimination, only evaluation provides an incomplete—and potentially misleading, assessment of UQ reliability.

In this section, we present a simple illustrative example demonstrating the logical independence of calibration and discrimination, motivating the need to evaluate both jointly. To the best of our knowledge, this work is the first, in the context of semantic uncertainty quantification for LLMs, to systematically assess both properties within a unified framework.

We consider a binary prediction setting consistent with the evaluation framework used throughout this work. Let $\mathcal{D} = \{(x_i, y_i, c_i)\}_{i=1}^{n}$ denote a dataset of triples, where $x_i \in \mathcal{V}^*$ is an input, $y_i \in \mathcal{V}^*$ is a model-generated output, and $c_i \in \{0, 1\}$ indicates whether y_i is correct given x_i. The model assigns a confidence score to each prediction via $p(y_i \mid x_i)$.

For simplicity, consider two examples:

$$\mathcal{D} = \{(x_1, y_1, 0), (x_2, y_2, 1)\},$$

corresponding to one incorrect and one correct generated response.

Perfect Calibration $\not\Rightarrow$ Discrimination. Suppose the model assigns identical confidence to both predictions:

$$p(y_1 \mid x_1) = p(y_2 \mid x_2) = 0.5.$$

This model is perfectly calibrated, as the average predicted confidence matches the empirical correctness rate. However, the confidence scores provide no discriminative signal, preventing the model from ranking correct predictions above incorrect ones (e.g., AUROC = 0.5), despite perfect calibration.

Perfect Discrimination $\not\Rightarrow$ Calibration. Now suppose the model assigns higher confidence to the correct prediction:

$$p(y_1 \mid x_1) = 0.9 \quad \text{and} \quad p(y_2 \mid x_2) = 0.99.$$

This yields perfect discrimination, as the correct prediction is always ranked above the incorrect one. However, the model is poorly calibrated: the predicted confidence values do not reflect the empirical correctness frequencies and are substantially overconfident.

These examples illustrate that calibration and discrimination are complementary but logically independent. Consequently, evaluating uncertainty quantification methods solely via discrimination is insufficient; both properties must be assessed jointly to obtain a reliable evaluation.

D Prompts

We present the exact prompts used for both calibration training and evaluation in Figure 6. For closed-book datasets, we use $m = 10$ in-context examples, while for open-book datasets, we use $m = 4$. For each dataset, examples from the training set (which is not used within any part of our pipeline) are uniformly sampled to serve as in-context examples, and these examples are fixed across both training and evaluation for all methods. The use of in-context examples encourages the model to produce concise outputs, containing only the final response.

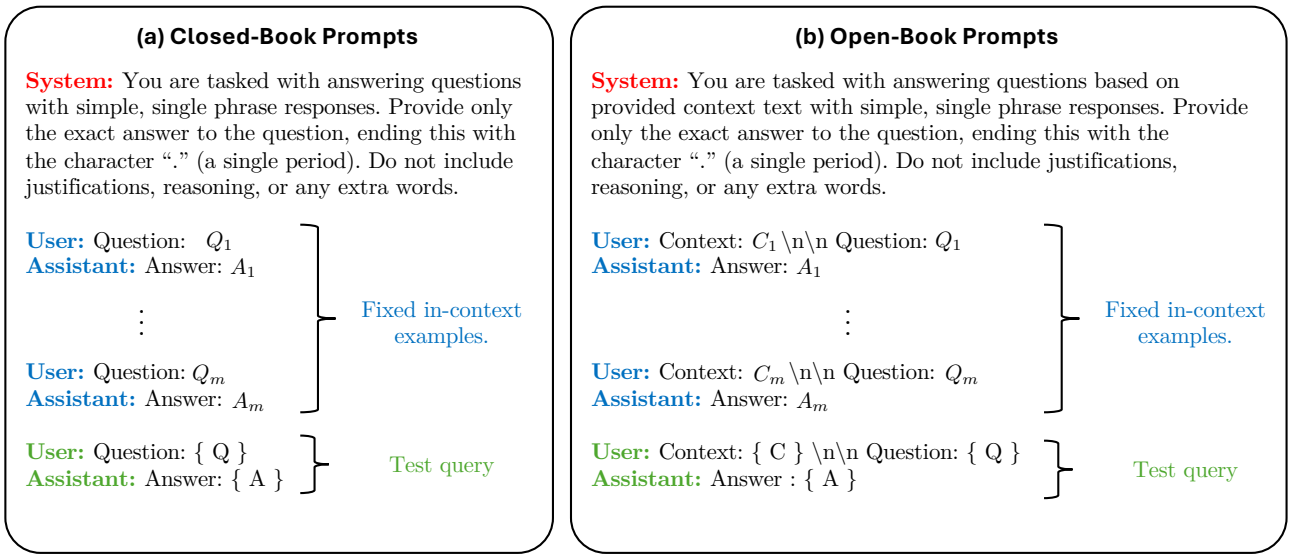


Figure (6): **Training and Evaluation Prompts.** Prompts applied for both calibration and evaluation across all methods, shown separately for (a) closed-book datasets (TriviaQA and Natural Questions) and (b) open-book datasets (SQuAD).

E On the Standard Practice of Task-Specific Calibration

In our evaluation, we adopt a protocol where calibration parameters such as scalar temperature parameters or transformer calibration heads are optimised separately for each model on every new dataset or domain. Regarding the practical application of this task-based, held-out calibration, we argue that this workflow aligns seamlessly with the standard paradigm of transfer learning, where pre-trained models are adapted to specific domains via learnable parameters. A prominent example of this paradigm is Low-Rank Adaptation (LoRA) (Hu et al., 2021), where small datasets are used to adapt general models to specific domains.

Task-specific temperature optimisation can be viewed as a lightweight, stripped-down version of this adaptation: instead of injecting new knowledge or capabilities, it adjusts the model's beliefs to align with the uncertainty of the new domain. Therefore, our approach fits naturally within modern deployment frameworks. Furthermore, prior work by Desai and Durrett (2020) demonstrates that pre-trained transformers require task-specific recalibration to remain reliable when shifting to new domains. This evidence suggests that post-hoc recalibration is not merely an optional step, but a necessity for reliably adapting pre-trained language models to downstream tasks. Consequently, we argue that optimising a scalar temperature on a small held-out set is a highly practical, computationally efficient, and methodologically sound procedure for current community practices.

F Supplementary Results

F.1 Model Accuracies

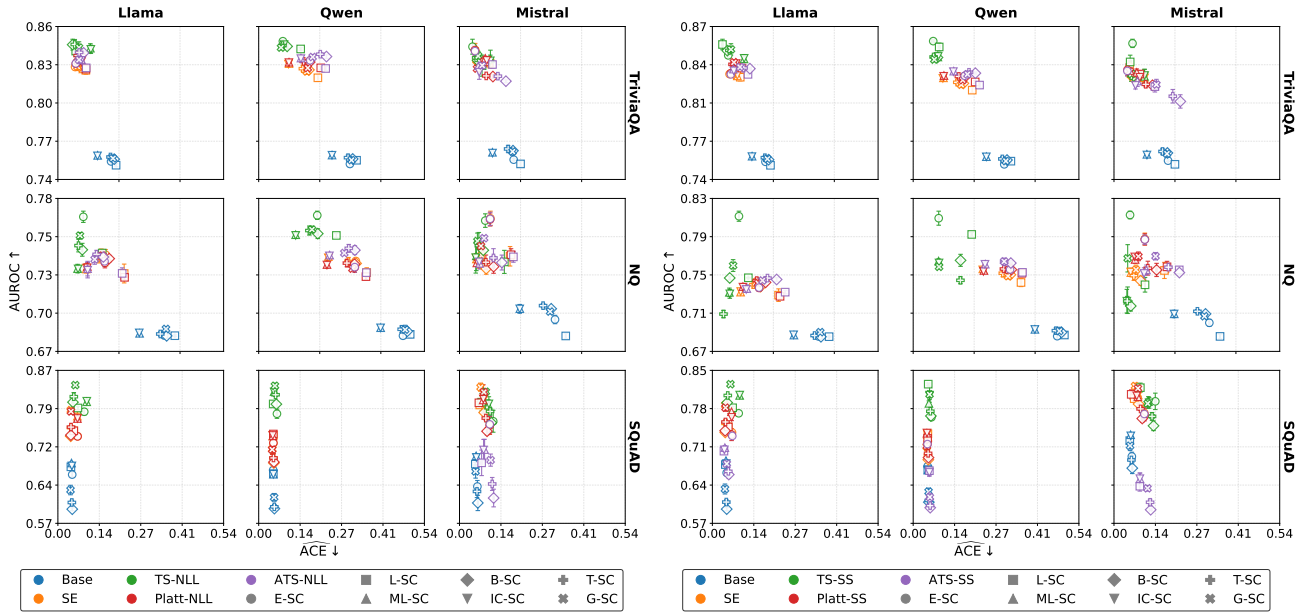
We report the accuracies of the base models on the held-out test sets for reference. These are the beam search results giving a indication of general model capabilities on the datasets before further calibrating model semantic confidence.

Table (3): **Base Model Accuracies (%).** Test-set accuracies (↑) of the base models through beam search.

Model	TriviaQA	Natural Questions	SQuAD
Llama	71.9	42.5	95.5
Qwen	53.0	33.2	94.4
Mistral	64.8	33.2	94.2

F.2 Comparison of Recalibration Loss Functions

We compare the calibration and discrimination of models across semantic confidence measures, recalibration methods, and baselines, focusing on the calibration loss function used for recalibration. The results are shown in Figure 7.



(a) Calibration methods with NLL loss.

(b) Calibration methods with SS loss.

Figure (7): **Comparison of Uncertainty Metrics of Semantic Measures Across Calibration Methods Using SS and NLL Losses.** Mean and standard deviation across four inference runs of $\widehat{\text{ACE}}$ (↓) and AUROC (↑) scores for semantic confidence measures for each model across baselines and calibration methods. We compare using a (a) NLL and (b) SS calibration loss.

F.3 Main Figure Results in Tabular Form

To supplement the results shown in Figure 3, we include the same results in tabular form that allow for more fine grained and precise comparison between methods. We present these in Table 4, Table 5, and Table 6.

Table (4): **Uncertainty Metrics for SC Measures Across Methods for Llama.** Mean and standard error of $\widehat{\text{ACE}}$ (↓) and AUROC (↑) scores for SC measures across measures baseline and recalibration methods for Llama.

		$\widehat{\text{ACE}}$							AUROC						
		E-SC	L-SC	ML-SC	B-SC	IC-SC	T-SC	G-SC	E-SC	L-SC	ML-SC	B-SC	IC-SC	T-SC	G-SC
TriviaQA	Base	$0.184_{\pm 0.002}$	$0.174_{\pm 0.002}$	$0.180_{\pm 0.001}$	$0.130_{\pm 0.001}$	$0.191_{\pm 0.002}$	$0.130_{\pm 0.002}$	$0.172_{\pm 0.002}$	$0.757_{\pm 0.001}$	$0.755_{\pm 0.001}$	$0.757_{\pm 0.001}$	$0.759_{\pm 0.001}$	$0.752_{\pm 0.001}$	$0.760_{\pm 0.001}$	$0.758_{\pm 0.002}$
	SE	$0.070_{\pm 0.003}$	$0.056_{\pm 0.001}$	$0.064_{\pm 0.003}$	$0.082_{\pm 0.002}$	$0.092_{\pm 0.002}$	$0.082_{\pm 0.002}$	$0.066_{\pm 0.003}$	$0.832_{\pm 0.003}$	$0.832_{\pm 0.002}$	$0.832_{\pm 0.003}$	$0.831_{\pm 0.002}$	$0.829_{\pm 0.003}$	$0.830_{\pm 0.002}$	$0.833_{\pm 0.003}$
	TS	$\mathbf{0.049_{\pm 0.001}}$	$\mathbf{0.055_{\pm 0.001}}$	$\mathbf{0.061_{\pm 0.001}}$	$0.104_{\pm 0.004}$	$\mathbf{0.033_{\pm 0.003}}$	$0.106_{\pm 0.002}$	$\mathbf{0.056_{\pm 0.004}}$	$\mathbf{0.852_{\pm 0.003}}$	$\mathbf{0.849_{\pm 0.004}}$	$\mathbf{0.853_{\pm 0.005}}$	$\mathbf{0.846_{\pm 0.003}}$	$\mathbf{0.858_{\pm 0.005}}$	$\mathbf{0.847_{\pm 0.004}}$	$\mathbf{0.853_{\pm 0.003}}$
	Platt	$0.078_{\pm 0.001}$	$0.060_{\pm 0.002}$	$0.071_{\pm 0.002}$	$0.078_{\pm 0.002}$	$0.097_{\pm 0.002}$	$0.077_{\pm 0.001}$	$0.066_{\pm 0.003}$	$0.840_{\pm 0.002}$	$0.833_{\pm 0.002}$	$0.842_{\pm 0.001}$	$0.841_{\pm 0.002}$	$0.837_{\pm 0.002}$	$0.842_{\pm 0.002}$	$0.840_{\pm 0.002}$
	ATS	$0.084_{\pm 0.001}$	$0.061_{\pm 0.002}$	$0.069_{\pm 0.002}$	$\mathbf{0.070_{\pm 0.004}}$	$0.093_{\pm 0.003}$	$\mathbf{0.069_{\pm 0.005}}$	$0.070_{\pm 0.002}$	$0.843_{\pm 0.001}$	$0.832_{\pm 0.002}$	$0.838_{\pm 0.002}$	$0.835_{\pm 0.001}$	$0.831_{\pm 0.003}$	$0.836_{\pm 0.001}$	$0.844_{\pm 0.001}$
NQ	Base	$0.356_{\pm 0.002}$	$0.348_{\pm 0.002}$	$0.354_{\pm 0.002}$	$0.267_{\pm 0.002}$	$0.383_{\pm 0.001}$	$0.268_{\pm 0.002}$	$0.335_{\pm 0.002}$	$0.685_{\pm 0.002}$	$0.685_{\pm 0.002}$	$0.690_{\pm 0.001}$	$0.687_{\pm 0.001}$	$0.685_{\pm 0.001}$	$0.687_{\pm 0.002}$	$0.686_{\pm 0.002}$
	SE	$0.158_{\pm 0.002}$	$0.156_{\pm 0.001}$	$0.142_{\pm 0.003}$	$0.093_{\pm 0.003}$	$0.217_{\pm 0.002}$	$0.093_{\pm 0.003}$	$0.135_{\pm 0.002}$	$0.737_{\pm 0.004}$	$0.741_{\pm 0.004}$	$0.741_{\pm 0.004}$	$0.732_{\pm 0.003}$	$0.728_{\pm 0.006}$	$0.731_{\pm 0.003}$	$0.739_{\pm 0.004}$
	TS	$\mathbf{0.057_{\pm 0.002}}$	$\mathbf{0.087_{\pm 0.001}}$	$\mathbf{0.068_{\pm 0.002}}$	$\mathbf{0.057_{\pm 0.002}}$	$\mathbf{0.118_{\pm 0.002}}$	$\mathbf{0.056_{\pm 0.002}}$	$\mathbf{0.036_{\pm 0.003}}$	$\mathbf{0.746_{\pm 0.002}}$	$\mathbf{0.810_{\pm 0.005}}$	$\mathbf{0.759_{\pm 0.006}}$	$0.730_{\pm 0.005}$	$\mathbf{0.746_{\pm 0.000}}$	$0.731_{\pm 0.004}$	$0.709_{\pm 0.004}$
	Platt	$0.168_{\pm 0.001}$	$0.154_{\pm 0.002}$	$0.146_{\pm 0.001}$	$0.096_{\pm 0.005}$	$0.218_{\pm 0.000}$	$0.100_{\pm 0.002}$	$0.138_{\pm 0.001}$	$0.739_{\pm 0.002}$	$0.736_{\pm 0.004}$	$0.741_{\pm 0.002}$	$\mathbf{0.732_{\pm 0.002}}$	$0.726_{\pm 0.002}$	$\mathbf{0.737_{\pm 0.003}}$	$0.740_{\pm 0.002}$
	ATS	$0.149_{\pm 0.003}$	$0.153_{\pm 0.002}$	$0.120_{\pm 0.003}$	$0.097_{\pm 0.004}$	$0.211_{\pm 0.000}$	$0.097_{\pm 0.002}$	$0.127_{\pm 0.003}$	$0.740_{\pm 0.003}$	$0.737_{\pm 0.004}$	$0.738_{\pm 0.003}$	$0.731_{\pm 0.005}$	$0.728_{\pm 0.004}$	$0.732_{\pm 0.005}$	$\mathbf{0.741_{\pm 0.003}}$
SQuAD	Base	$0.047_{\pm 0.001}$	$\mathbf{0.047_{\pm 0.001}}$	$\mathbf{0.041_{\pm 0.001}}$	$0.046_{\pm 0.001}$	$0.043_{\pm 0.001}$	$0.045_{\pm 0.001}$	$0.046_{\pm 0.001}$	$0.594_{\pm 0.003}$	$0.662_{\pm 0.005}$	$0.632_{\pm 0.009}$	$0.679_{\pm 0.002}$	$0.677_{\pm 0.003}$	$0.685_{\pm 0.002}$	$0.607_{\pm 0.004}$
	SE	$\mathbf{0.042_{\pm 0.002}}$	$0.065_{\pm 0.001}$	$0.042_{\pm 0.002}$	$0.064_{\pm 0.001}$	$0.052_{\pm 0.001}$	$0.064_{\pm 0.001}$	$0.043_{\pm 0.003}$	$0.737_{\pm 0.004}$	$0.737_{\pm 0.003}$	$0.788_{\pm 0.004}$	$0.775_{\pm 0.005}$	$0.748_{\pm 0.006}$	$0.776_{\pm 0.002}$	$0.754_{\pm 0.004}$
	TS	$0.049_{\pm 0.003}$	$0.086_{\pm 0.002}$	$0.059_{\pm 0.001}$	$0.090_{\pm 0.002}$	$0.066_{\pm 0.001}$	$0.090_{\pm 0.002}$	$0.051_{\pm 0.001}$	$\mathbf{0.804_{\pm 0.008}}$	$\mathbf{0.773_{\pm 0.005}}$	$\mathbf{0.827_{\pm 0.002}}$	$\mathbf{0.807_{\pm 0.005}}$	$\mathbf{0.783_{\pm 0.003}}$	$\mathbf{0.806_{\pm 0.005}}$	$\mathbf{0.806_{\pm 0.002}}$
	Platt	$0.043_{\pm 0.002}$	$0.063_{\pm 0.003}$	$0.043_{\pm 0.001}$	$0.063_{\pm 0.001}$	$0.052_{\pm 0.001}$	$0.064_{\pm 0.002}$	$\mathbf{0.043_{\pm 0.002}}$	$0.739_{\pm 0.003}$	$0.73_{\pm 0.01}$	$0.786_{\pm 0.005}$	$0.765_{\pm 0.006}$	$0.748_{\pm 0.007}$	$0.772_{\pm 0.003}$	$0.757_{\pm 0.005}$
	ATS	$0.054_{\pm 0.001}$	$0.065_{\pm 0.001}$	$0.047_{\pm 0.001}$	$\mathbf{0.039_{\pm 0.001}}$	$\mathbf{0.038_{\pm 0.001}}$	$\mathbf{0.041_{\pm 0.001}}$	$0.052_{\pm 0.001}$	$0.659_{\pm 0.007}$	$0.731_{\pm 0.004}$	$0.679_{\pm 0.007}$	$0.705_{\pm 0.005}$	$0.702_{\pm 0.005}$	$0.709_{\pm 0.004}$	$0.665_{\pm 0.007}$

Table (5): **Uncertainty Metrics for SC Measures Across Methods for Qwen.** Mean and standard error of $\widehat{\text{ACE}}$ (↓) and AUROC (↑) scores for SC measures across measures baseline and recalibration methods for Qwen.

		$\widehat{\text{ACE}}$							AUROC						
		E-SC	L-SC	ML-SC	B-SC	IC-SC	T-SC	G-SC	E-SC	L-SC	ML-SC	B-SC	IC-SC	T-SC	G-SC
TriviaQA	Base	$0.308_{\pm 0.001}$	$0.299_{\pm 0.001}$	$0.304_{\pm 0.000}$	$0.240_{\pm 0.000}$	$0.320_{\pm 0.001}$	$0.240_{\pm 0.000}$	$0.293_{\pm 0.001}$	$0.763_{\pm 0.002}$	$0.759_{\pm 0.002}$	$0.762_{\pm 0.002}$	$0.765_{\pm 0.001}$	$0.761_{\pm 0.002}$	$0.766_{\pm 0.001}$	$0.764_{\pm 0.002}$
	SE	$0.165_{\pm 0.002}$	$0.159_{\pm 0.001}$	$0.158_{\pm 0.001}$	$0.097_{\pm 0.002}$	$0.194_{\pm 0.001}$	$0.099_{\pm 0.002}$	$0.145_{\pm 0.002}$	$0.832_{\pm 0.002}$	$0.831_{\pm 0.001}$	$0.830_{\pm 0.002}$	$0.836_{\pm 0.002}$	$0.825_{\pm 0.002}$	$0.837_{\pm 0.002}$	$0.833_{\pm 0.002}$
	TS	$\mathbf{0.080_{\pm 0.001}}$	$\mathbf{0.066_{\pm 0.005}}$	$\mathbf{0.069_{\pm 0.002}}$	$\mathbf{0.082_{\pm 0.003}}$	$\mathbf{0.085_{\pm 0.004}}$	$\mathbf{0.081_{\pm 0.003}}$	$\mathbf{0.067_{\pm 0.002}}$	$\mathbf{0.855_{\pm 0.002}}$	$\mathbf{0.870_{\pm 0.002}}$	$\mathbf{0.853_{\pm 0.002}}$	$\mathbf{0.856_{\pm 0.000}}$	$\mathbf{0.864_{\pm 0.004}}$	$\mathbf{0.855_{\pm 0.001}}$	$\mathbf{0.855_{\pm 0.002}}$
	Platt	$0.170_{\pm 0.001}$	$0.168_{\pm 0.002}$	$0.163_{\pm 0.002}$	$0.100_{\pm 0.002}$	$0.203_{\pm 0.001}$	$0.101_{\pm 0.003}$	$0.150_{\pm 0.001}$	$0.837_{\pm 0.002}$	$0.839_{\pm 0.001}$	$0.834_{\pm 0.001}$	$0.838_{\pm 0.001}$	$0.833_{\pm 0.001}$	$0.839_{\pm 0.001}$	$0.838_{\pm 0.002}$
	ATS	$0.204_{\pm 0.003}$	$0.168_{\pm 0.002}$	$0.178_{\pm 0.002}$	$0.132_{\pm 0.003}$	$0.219_{\pm 0.001}$	$0.133_{\pm 0.002}$	$0.183_{\pm 0.003}$	$0.841_{\pm 0.001}$	$0.839_{\pm 0.001}$	$0.840_{\pm 0.002}$	$0.842_{\pm 0.001}$	$0.833_{\pm 0.002}$	$0.843_{\pm 0.001}$	$0.842_{\pm 0.001}$
NQ	Base	$0.487_{\pm 0.001}$	$0.473_{\pm 0.001}$	$0.480_{\pm 0.001}$	$0.399_{\pm 0.001}$	$0.496_{\pm 0.001}$	$0.399_{\pm 0.001}$	$0.468_{\pm 0.001}$	$0.690_{\pm 0.003}$	$0.687_{\pm 0.002}$	$0.692_{\pm 0.004}$	$0.693_{\pm 0.002}$	$0.688_{\pm 0.003}$	$0.693_{\pm 0.002}$	$0.692_{\pm 0.003}$
	SE	$0.319_{\pm 0.002}$	$0.318_{\pm 0.002}$	$0.311_{\pm 0.002}$	$0.228_{\pm 0.002}$	$0.354_{\pm 0.002}$	$0.228_{\pm 0.002}$	$0.294_{\pm 0.002}$	$0.745_{\pm 0.003}$	$0.747_{\pm 0.002}$	$0.744_{\pm 0.003}$	$0.750_{\pm 0.003}$	$0.738_{\pm 0.004}$	$0.751_{\pm 0.002}$	$0.747_{\pm 0.003}$
	TS	$\mathbf{0.155_{\pm 0.002}}$	$\mathbf{0.084_{\pm 0.002}}$	$\mathbf{0.085_{\pm 0.003}}$	$\mathbf{0.085_{\pm 0.001}}$	$\mathbf{0.192_{\pm 0.002}}$	$\mathbf{0.083_{\pm 0.002}}$	$\mathbf{0.155_{\pm 0.002}}$	$\mathbf{0.759_{\pm 0.006}}$	$\mathbf{0.799_{\pm 0.007}}$	$0.753_{\pm 0.002}$	$\mathbf{0.756_{\pm 0.002}}$	$\mathbf{0.784_{\pm 0.002}}$	$\mathbf{0.758_{\pm 0.002}}$	$0.740_{\pm 0.003}$
	Platt	$0.314_{\pm 0.001}$	$0.318_{\pm 0.001}$	$0.307_{\pm 0.001}$	$0.231_{\pm 0.002}$	$0.357_{\pm 0.000}$	$0.232_{\pm 0.002}$	$0.289_{\pm 0.001}$	$0.745_{\pm 0.003}$	$0.751_{\pm 0.005}$	$0.741_{\pm 0.001}$	$0.748_{\pm 0.001}$	$0.747_{\pm 0.003}$	$0.749_{\pm 0.001}$	$0.746_{\pm 0.003}$
	ATS	$0.315_{\pm 0.004}$	$0.318_{\pm 0.001}$	$0.280_{\pm 0.002}$	$0.236_{\pm 0.004}$	$0.353_{\pm 0.002}$	$0.237_{\pm 0.003}$	$0.295_{\pm 0.004}$	$0.756_{\pm 0.004}$	$0.750_{\pm 0.005}$	$\mathbf{0.754_{\pm 0.001}}$	$0.755_{\pm 0.002}$	$0.738_{\pm 0.001}$	$0.755_{\pm 0.001}$	$\mathbf{0.757_{\pm 0.003}}$
SQuAD	Base	$0.051_{\pm 0.001}$	$0.047_{\pm 0.000}$	$0.051_{\pm 0.001}$	$0.048_{\pm 0.000}$	$0.048_{\pm 0.000}$	$0.048_{\pm 0.000}$	$0.051_{\pm 0.001}$	$0.534_{\pm 0.004}$	$0.605_{\pm 0.004}$	$0.554_{\pm 0.007}$	$0.597_{\pm 0.002}$	$0.598_{\pm 0.003}$	$0.597_{\pm 0.002}$	$0.537_{\pm 0.004}$
	SE	$\mathbf{0.050_{\pm 0.001}}$	$0.048_{\pm 0.001}$	$0.044_{\pm 0.001}$	$0.046_{\pm 0.000}$	$0.048_{\pm 0.002}$	$\mathbf{0.046_{\pm 0.000}}$	$0.049_{\pm 0.001}$	$0.618_{\pm 0.007}$	$0.654_{\pm 0.006}$	$0.643_{\pm 0.009}$	$0.669_{\pm 0.005}$	$0.669_{\pm 0.005}$	$0.667_{\pm 0.004}$	$0.627_{\pm 0.008}$
	TS	$0.058_{\pm 0.001}$	$0.060_{\pm 0.001}$	$0.053_{\pm 0.002}$	$0.051_{\pm 0.001}$	$0.049_{\pm 0.003}$	$0.052_{\pm 0.002}$	$0.055_{\pm 0.001}$	$\mathbf{0.726_{\pm 0.005}}$	$\mathbf{0.702_{\pm 0.004}}$	$\mathbf{0.760_{\pm 0.005}}$	$\mathbf{0.748_{\pm 0.005}}$	$\mathbf{0.767_{\pm 0.002}}$	$\mathbf{0.729_{\pm 0.005}}$	$\mathbf{0.743_{\pm 0.003}}$
	Platt	$0.050_{\pm 0.001}$	$\mathbf{0.047_{\pm 0.000}}$	$\mathbf{0.044_{\pm 0.001}}$	$\mathbf{0.046_{\pm 0.000}}$	$\mathbf{0.047_{\pm 0.002}}$	$0.046_{\pm 0.000}$	$\mathbf{0.048_{\pm 0.001}}$	$0.619_{\pm 0.007}$	$0.649_{\pm 0.002}$	$0.642_{\pm 0.009}$	$0.668_{\pm 0.005}$	$0.660_{\pm 0.002}$	$0.667_{\pm 0.004}$	$0.627_{\pm 0.008}$
	ATS	$0.056_{\pm 0.001}$	$\mathbf{0.047_{\pm 0.000}}$	$0.055_{\pm 0.001}$	$0.053_{\pm 0.001}$	$0.053_{\pm 0.001}$	$0.053_{\pm 0.001}$	$0.056_{\pm 0.001}$	$0.523_{\pm 0.006}$	$0.649_{\pm 0.002}$	$0.544_{\pm 0.005}$	$0.593_{\pm 0.008}$	$0.596_{\pm 0.009}$	$0.547_{\pm 0.005}$	$0.527_{\pm 0.008}$

Table (6): **Uncertainty Metrics for SC Measures Across Methods for Mistral.** Mean and standard error of $\widehat{\text{ACE}}$ (↓) and AUROC (↑) scores for SC measures across measures baseline and recalibration methods for Mistral.

		$\widehat{\text{ACE}}$							AUROC						
		E-SC	L-SC	ML-SC	B-SC	IC-SC	T-SC	G-SC	E-SC	L-SC	ML-SC	B-SC	IC-SC	T-SC	G-SC
TriviaQA	Base	$0.175_{\pm 0.001}$	$0.177_{\pm 0.001}$	$0.174_{\pm 0.001}$	$0.108_{\pm 0.001}$	$0.200_{\pm 0.001}$	$0.108_{\pm 0.001}$	$0.159_{\pm 0.001}$	$0.792_{\pm 0.002}$	$0.787_{\pm 0.001}$	$0.792_{\pm 0.002}$	$0.791_{\pm 0.001}$	$0.785_{\pm 0.002}$	$0.791_{\pm 0.001}$	$0.793_{\pm 0.002}$
	SE	$\mathbf{0.062_{\pm 0.001}}$	$0.051_{\pm 0.006}$	$0.059_{\pm 0.004}$	$0.100_{\pm 0.007}$	$0.080_{\pm 0.002}$	$0.099_{\pm 0.006}$	$0.056_{\pm 0.003}$	$0.842_{\pm 0.002}$	$0.849_{\pm 0.004}$	$0.843_{\pm 0.003}$	$0.842_{\pm 0.002}$	$0.842_{\pm 0.001}$	$0.842_{\pm 0.003}$	$0.843_{\pm 0.002}$
	TS	$0.064_{\pm 0.002}$	$0.061_{\pm 0.007}$	$0.058_{\pm 0.001}$	$0.100_{\pm 0.004}$	$\mathbf{0.053_{\pm 0.004}}$	$0.100_{\pm 0.004}$	$\mathbf{0.055_{\pm 0.002}}$	$\mathbf{0.843_{\pm 0.003}}$	$\mathbf{0.865_{\pm 0.003}}$	$\mathbf{0.844_{\pm 0.004}}$	$\mathbf{0.843_{\pm 0.004}}$	$\mathbf{0.852_{\pm 0.004}}$	$0.843_{\pm 0.005}$	$\mathbf{0.845_{\pm 0.006}}$
	Platt	$0.109_{\pm 0.002}$	$0.044_{\pm 0.003}$	$\mathbf{0.057_{\pm 0.004}}$	$0.085_{\pm 0.005}$	$0.082_{\pm 0.004}$	$0.075_{\pm 0.005}$	$0.088_{\pm 0.003}$	$0.834_{\pm 0.001}$	$0.847_{\pm 0.004}$	$0.839_{\pm 0.001}$	$0.842_{\pm 0.002}$	$0.842_{\pm 0.002}$	$\mathbf{0.845_{\pm 0.002}}$	$0.835_{\pm 0.001}$
	ATS	$0.152_{\pm 0.001}$	$\mathbf{0.044_{\pm 0.002}}$	$0.074_{\pm 0.003}$	$\mathbf{0.065_{\pm 0.004}}$	$0.108_{\pm 0.003}$	$\mathbf{0.062_{\pm 0.002}}$	$0.126_{\pm 0.001}$	$0.832_{\pm 0.001}$	$0.846_{\pm 0.003}$	$0.840_{\pm 0.002}$	$0.837_{\pm 0.004}$	$0.841_{\pm 0.002}$	$0.840_{\pm 0.002}$	$0.834_{\pm 0.001}$
NQ	Base	$0.300_{\pm 0.003}$	$0.313_{\pm 0.002}$	$0.297_{\pm 0.002}$	$0.197_{\pm 0.001}$	$0.348_{\pm 0.002}$	$0.198_{\pm 0.001}$	$0.274_{\pm 0.003}$	$0.722_{\pm 0.001}$	$0.715_{\pm 0.003}$	$0.720_{\pm 0.001}$	$0.722_{\pm 0.003}$	$0.705_{\pm 0.002}$	$0.722_{\pm 0.002}$	$0.724_{\pm 0.001}$
	SE	$0.087_{\pm 0.005}$	$0.101_{\pm 0.003}$	$0.072_{\pm 0.003}$	$0.055_{\pm 0.001}$	$0.166_{\pm 0.003}$	$0.055_{\pm 0.000}$	$0.065_{\pm 0.005}$	$0.748_{\pm 0.004}$	$0.780_{\pm 0.004}$	$0.756_{\pm 0.004}$	$\mathbf{0.754_{\pm 0.003}}$	$0.755_{\pm 0.006}$	$0.754_{\pm 0.003}$	$0.752_{\pm 0.004}$
	TS	$\mathbf{0.055_{\pm 0.002}}$	$\mathbf{0.053_{\pm 0.004}}$	$\mathbf{0.045_{\pm 0.004}}$	$\mathbf{0.042_{\pm 0.007}}$	$\mathbf{0.102_{\pm 0.001}}$	$\mathbf{0.045_{\pm 0.006}}$	$\mathbf{0.043_{\pm 0.002}}$	$0.728_{\pm 0.004}$	$\mathbf{0.798_{\pm 0.003}}$	$0.76_{\pm 0.01}$	$0.732_{\pm 0.009}$	$0.745_{\pm 0.006}$	$0.73_{\pm 0.01}$	$0.732_{\pm 0.004}$
	Platt	$0.112_{\pm 0.003}$	$0.100_{\pm 0.002}$	$0.080_{\pm 0.003}$	$0.061_{\pm 0.005}$	$0.169_{\pm 0.002}$	$0.069_{\pm 0.005}$	$0.086_{\pm 0.002}$	$0.749_{\pm 0.004}$	$0.779_{\pm 0.004}$	$0.766_{\pm 0.003}$	$0.751_{\pm 0.001}$	$\mathbf{0.756_{\pm 0.004}}$	$\mathbf{0.764_{\pm 0.003}}$	$0.752_{\pm 0.005}$
	ATS	$0.139_{\pm 0.006}$	$0.100_{\pm 0.004}$	$0.081_{\pm 0.004}$	$0.065_{\pm 0.004}$	$0.177_{\pm 0.003}$	$0.066_{\pm 0.002}$	$0.111_{\pm 0.007}$	$\mathbf{0.752_{\pm 0.005}}$	$0.779_{\pm 0.004}$	$\mathbf{0.767_{\pm 0.002}}$	$0.752_{\pm 0.004}$	$0.755_{\pm 0.003}$	$0.751_{\pm 0.004}$	$\mathbf{0.754_{\pm 0.006}}$
SQuAD	Base	$\mathbf{0.060_{\pm 0.002}}$	$\mathbf{0.059_{\pm 0.001}}$	$\mathbf{0.053_{\pm 0.000}}$	$\mathbf{0.055_{\pm 0.001}}$	$\mathbf{0.052_{\pm 0.001}}$	$\mathbf{0.054_{\pm 0.001}}$	$\mathbf{0.057_{\pm 0.002}}$	$0.676_{\pm 0.010}$	$0.697_{\pm 0.007}$	$0.716_{\pm 0.008}$	$0.733_{\pm 0.003}$	$0.725_{\pm 0.002}$	$0.735_{\pm 0.002}$	$0.690_{\pm 0.008}$
	SE	$0.081_{\pm 0.003}$	$0.099_{\pm 0.002}$	$0.069_{\pm 0.001}$	$0.080_{\pm 0.003}$	$0.066_{\pm 0.000}$	$0.079_{\pm 0.004}$	$0.077_{\pm 0.002}$	$\mathbf{0.789_{\pm 0.004}}$	$0.77_{\pm 0.01}$	$\mathbf{0.821_{\pm 0.005}}$	$\mathbf{0.806_{\pm 0.010}}$	$0.799_{\pm 0.003}$	$\mathbf{0.81_{\pm 0.01}}$	$\mathbf{0.801_{\pm 0.003}}$
	TS	$0.104_{\pm 0.001}$	$0.110_{\pm 0.003}$	$0.086_{\pm 0.001}$	$0.096_{\pm 0.004}$	$0.073_{\pm 0.003}$	$0.095_{\pm 0.002}$	$0.100_{\pm 0.001}$	$0.777_{\pm 0.003}$	$\mathbf{0.78_{\pm 0.01}}$	$0.815_{\pm 0.002}$	$0.800_{\pm 0.009}$	$\mathbf{0.80_{\pm 0.01}}$	$0.798_{\pm 0.007}$	$0.791_{\pm 0.004}$
	Platt	$0.090_{\pm 0.002}$	$0.100_{\pm 0.002}$	$0.080_{\pm 0.001}$	$0.075_{\pm 0.004}$	$0.064_{\pm 0.001}$	$0.077_{\pm 0.003}$	$0.086_{\pm 0.004}$	$0.766_{\pm 0.005}$	$0.77_{\pm 0.01}$	$0.815_{\pm 0.004}$	$0.804_{\pm 0.008}$	$0.801_{\pm 0.004}$	$0.80_{\pm 0.01}$	$0.782_{\pm 0.002}$
	ATS	$0.112_{\pm 0.002}$	$0.099_{\pm 0.002}$	$0.102_{\pm 0.001}$	$0.079_{\pm 0.002}$	$0.072_{\pm 0.003}$	$0.082_{\pm 0.001}$	$0.107_{\pm 0.002}$	$0.68_{\pm 0.01}$	$0.77_{\pm 0.01}$	$0.730_{\pm 0.008}$	$0.74_{\pm 0.01}$	$0.73_{\pm 0.02}$	$0.74_{\pm 0.01}$	$0.700_{\pm 0.008}$

F.4 Full Results for SE_{conf} and $\text{SE}_{\text{vanilla}}$

In Section 4.2, we compared two formulations of semantic entropy: (i) SE_{conf}, which defines correctness relative to the most confident semantic class, providing a principled alignment between entropy computation and correctness evaluation; and (ii) $\text{SE}_{\text{vanilla}}$, which follows the original implementation using a separate greedy-decoding procedure to determine correctness (Kuhn et al., 2023).

Importantly, SE_{conf} uses the same distribution that generates the final response to estimate both confidence and uncertainty, providing a principled and self-consistent alignment between prediction and uncertainty. In contrast, SEvanilla relies on separate distributions for correctness and uncertainty, effectively mixing distinct model beliefs

Table (7): **AUROC for SE_{conf} Across Baseline and Calibration Methods for Qwen.** Mean and standard error of AUROC (↑) scores computed using SE_{conf}, where correctness is defined relative to the most confident semantic class. Results are reported for TriviaQA, NQ, and SQuAD across baseline and calibration methods. Bold values indicate the best result within each SC measure for a dataset, while underlined bold values denote the overall best result for that dataset.

		E-SC	L-SC	ML-SC	B-SC	IC-SC	T-SC	G-SC
TriviaQA	Base	$0.766_{\pm 0.002}$	$0.761_{\pm 0.002}$	$0.764_{\pm 0.002}$	$0.765_{\pm 0.002}$	$0.765_{\pm 0.002}$	$0.766_{\pm 0.002}$	$0.767_{\pm 0.002}$
	SE	$0.834_{\pm 0.002}$	$0.833_{\pm 0.001}$	$0.832_{\pm 0.002}$	$0.835_{\pm 0.002}$	$0.830_{\pm 0.002}$	$0.836_{\pm 0.002}$	$0.834_{\pm 0.002}$
	TS	$\mathbf{0.853_{\pm 0.002}}$	$\underline{\mathbf{0.865}}_{\pm 0.002}$	$\mathbf{0.849_{\pm 0.002}}$	$\mathbf{0.854_{\pm 0.000}}$	$\underline{\mathbf{0.864}}_{\pm 0.004}$	$\mathbf{0.853_{\pm 0.001}}$	$\mathbf{0.851_{\pm 0.002}}$
	Platt	$0.839_{\pm 0.002}$	$0.840_{\pm 0.001}$	$0.836_{\pm 0.001}$	$0.837_{\pm 0.001}$	$0.836_{\pm 0.001}$	$0.838_{\pm 0.001}$	$0.839_{\pm 0.002}$
	ATS	$0.843_{\pm 0.001}$	$0.840_{\pm 0.001}$	$0.840_{\pm 0.002}$	$0.843_{\pm 0.001}$	$0.838_{\pm 0.002}$	$0.843_{\pm 0.001}$	$0.843_{\pm 0.001}$
NQ	Base	$0.693_{\pm 0.003}$	$0.691_{\pm 0.002}$	$0.695_{\pm 0.003}$	$0.694_{\pm 0.002}$	$0.692_{\pm 0.003}$	$0.694_{\pm 0.002}$	$0.693_{\pm 0.003}$
	SE	$0.749_{\pm 0.003}$	$0.752_{\pm 0.002}$	$0.749_{\pm 0.003}$	$0.751_{\pm 0.002}$	$0.746_{\pm 0.003}$	$0.752_{\pm 0.002}$	$0.750_{\pm 0.003}$
	TS	$\mathbf{0.769_{\pm 0.004}}$	$\underline{\mathbf{0.795}}_{\pm 0.007}$	$\mathbf{0.754_{\pm 0.002}}$	$\mathbf{0.783_{\pm 0.003}}$	$\underline{\mathbf{0.789}}_{\pm 0.003}$	$\underline{\mathbf{0.784}}_{\pm 0.004}$	$0.747_{\pm 0.002}$
	Platt	$0.746_{\pm 0.003}$	$0.755_{\pm 0.005}$	$0.743_{\pm 0.001}$	$0.749_{\pm 0.001}$	$0.752_{\pm 0.002}$	$0.749_{\pm 0.001}$	$0.746_{\pm 0.003}$
	ATS	$0.759_{\pm 0.003}$	$0.754_{\pm 0.005}$	$0.754_{\pm 0.001}$	$0.755_{\pm 0.001}$	$0.743_{\pm 0.001}$	$0.756_{\pm 0.001}$	$\mathbf{0.759_{\pm 0.002}}$
SQuAD	Base	$0.569_{\pm 0.007}$	$0.605_{\pm 0.004}$	$0.574_{\pm 0.006}$	$0.597_{\pm 0.002}$	$0.598_{\pm 0.003}$	$0.597_{\pm 0.002}$	$0.571_{\pm 0.008}$
	SE	$0.666_{\pm 0.007}$	$0.655_{\pm 0.006}$	$0.666_{\pm 0.006}$	$0.670_{\pm 0.005}$	$0.669_{\pm 0.005}$	$0.668_{\pm 0.004}$	$0.665_{\pm 0.007}$
	TS	$\underline{\mathbf{0.780}}_{\pm 0.003}$	$\mathbf{0.704_{\pm 0.004}}$	$\underline{\mathbf{0.774}}_{\pm 0.004}$	$\underline{\mathbf{0.752}}_{\pm 0.005}$	$\mathbf{0.772_{\pm 0.002}}$	$\underline{\mathbf{0.732}}_{\pm 0.005}$	$\underline{\mathbf{0.780}}_{\pm 0.003}$
	Platt	$0.665_{\pm 0.007}$	$0.649_{\pm 0.002}$	$0.665_{\pm 0.006}$	$0.669_{\pm 0.005}$	$0.660_{\pm 0.002}$	$0.668_{\pm 0.004}$	$0.666_{\pm 0.007}$
	ATS	$0.579_{\pm 0.009}$	$0.649_{\pm 0.002}$	$0.600_{\pm 0.005}$	$0.649_{\pm 0.005}$	$0.649_{\pm 0.006}$	$0.615_{\pm 0.003}$	$0.591_{\pm 0.009}$

in an ad hoc manner. We observed in Section 4.2 that temperature scaling improves discriminability for both formulations, with higher overall AUROCs achieved under SE_{conf}.

Here, we present the full results for the Qwen model across TriviaQA, NQ, and SQuAD. Complete tables are provided in Table 7 and Table 8 for SE_{conf} and $SE_{vanilla}$ respectively.

Table (8): **AUROC for $SE_{vanilla}$ Across Baseline and Calibration Methods for Qwen.** Mean and standard error of AUROC (↑) scores computed using $SE_{vanilla}$, where correctness is defined via an external decoding procedure (Kuhn et al., 2023). Results are reported for TriviaQA, NQ, and SQuAD across baseline and calibration methods. Bold values indicate the best result within each SC measure for a dataset, while underlined bold values denote the overall best result for that dataset.

		E-SC	L-SC	ML-SC	B-SC	IC-SC	T-SC	G-SC
TriviaQA	Base	$0.755_{\pm 0.002}$	$0.755_{\pm 0.002}$	$0.755_{\pm 0.002}$	$0.756_{\pm 0.002}$	$0.754_{\pm 0.002}$	$0.756_{\pm 0.002}$	$0.755_{\pm 0.002}$
	SE	$0.833_{\pm 0.001}$	$0.833_{\pm 0.001}$	$0.833_{\pm 0.001}$	$0.833_{\pm 0.001}$	$0.833_{\pm 0.001}$	$0.833_{\pm 0.001}$	$0.833_{\pm 0.001}$
	TS	$\mathbf{0.849_{\pm 0.001}}$	$\mathbf{0.844_{\pm 0.001}}$	$\mathbf{0.849_{\pm 0.002}}$	$\mathbf{0.848_{\pm 0.002}}$	$\underline{\mathbf{0.857}}_{\pm 0.002}$	$\mathbf{0.848_{\pm 0.002}}$	$\mathbf{0.848_{\pm 0.002}}$
	Platt	$0.832_{\pm 0.002}$	$0.832_{\pm 0.002}$	$0.832_{\pm 0.002}$	$0.833_{\pm 0.002}$	$0.832_{\pm 0.002}$	$0.833_{\pm 0.002}$	$0.832_{\pm 0.002}$
	ATS	$0.834_{\pm 0.001}$	$0.832_{\pm 0.002}$	$0.835_{\pm 0.001}$	$0.836_{\pm 0.001}$	$0.832_{\pm 0.002}$	$0.836_{\pm 0.001}$	$0.834_{\pm 0.001}$
NQ	Base	$0.690_{\pm 0.002}$	$0.689_{\pm 0.002}$	$0.691_{\pm 0.002}$	$0.692_{\pm 0.002}$	$0.689_{\pm 0.002}$	$0.692_{\pm 0.002}$	$0.691_{\pm 0.002}$
	SE	$0.749_{\pm 0.001}$	$\mathbf{0.745_{\pm 0.001}}$	$\mathbf{0.749_{\pm 0.001}}$	$0.749_{\pm 0.001}$	$0.745_{\pm 0.001}$	$0.749_{\pm 0.001}$	$\mathbf{0.750_{\pm 0.001}}$
	TS	$\underline{\mathbf{0.758}}_{\pm 0.001}$	$0.737_{\pm 0.002}$	$0.743_{\pm 0.002}$	$0.747_{\pm 0.002}$	$\mathbf{0.751_{\pm 0.004}}$	$0.747_{\pm 0.002}$	$0.745_{\pm 0.001}$
	Platt	$0.745_{\pm 0.001}$	$0.743_{\pm 0.003}$	$0.743_{\pm 0.002}$	$0.747_{\pm 0.003}$	$0.744_{\pm 0.003}$	$0.747_{\pm 0.003}$	$0.744_{\pm 0.002}$
	ATS	$0.740_{\pm 0.002}$	$0.743_{\pm 0.003}$	$0.742_{\pm 0.002}$	$\mathbf{0.750_{\pm 0.003}}$	$0.741_{\pm 0.002}$	$\mathbf{0.750_{\pm 0.002}}$	$0.740_{\pm 0.002}$
SQuAD	Base	$0.590_{\pm 0.001}$	$0.595_{\pm 0.001}$	$0.595_{\pm 0.001}$	$0.595_{\pm 0.002}$	$0.595_{\pm 0.002}$	$0.595_{\pm 0.002}$	$0.595_{\pm 0.001}$
	SE	$0.653_{\pm 0.002}$	$0.653_{\pm 0.001}$	$0.653_{\pm 0.002}$	$0.653_{\pm 0.002}$	$0.653_{\pm 0.002}$	$0.653_{\pm 0.001}$	$0.653_{\pm 0.002}$
	TS	$\mathbf{0.707_{\pm 0.007}}$	$\mathbf{0.717_{\pm 0.005}}$	$\mathbf{0.709_{\pm 0.007}}$	$\mathbf{0.710_{\pm 0.007}}$	$\underline{\mathbf{0.748}}_{\pm 0.003}$	$\mathbf{0.700_{\pm 0.001}}$	$\mathbf{0.707_{\pm 0.007}}$
	Platt	$0.653_{\pm 0.002}$	$0.652_{\pm 0.004}$	$0.654_{\pm 0.002}$	$0.653_{\pm 0.001}$	$0.649_{\pm 0.004}$	$0.653_{\pm 0.001}$	$0.653_{\pm 0.002}$
	ATS	$0.575_{\pm 0.007}$	$0.652_{\pm 0.004}$	$0.583_{\pm 0.008}$	$0.621_{\pm 0.003}$	$0.624_{\pm 0.002}$	$0.616_{\pm 0.004}$	$0.578_{\pm 0.009}$

F.5 Temperatures from Optimised Temperature Scaling

Table 9 gives the best final temperature settings after sweeping over hyperparameters based on the Brier score on the validation set, and these are used as the final settings for all results presented in this paper.

Table (9): **Optimised Temperature Settings.** Best temperature settings for the temperature scaling (TS) method across models and datasets. Temperatures represent the final optimised after calibration training and hyperparameter selection and are those that are used to give the results shown in Figure 3.

Dataset	Model	E-SC	L-SC	ML-SC	B-SC	IC-SC	T-SC	G-SC
TriviaQA	Llama	1.26	1.63	1.26	1.21	1.21	1.21	1.21
	Qwen	1.80	2.07	1.32	1.56	1.32	1.56	1.56
	Mistral	1.18	1.18	1.00	1.00	1.00	1.00	1.00
NQ	Llama	2.00	2.00	1.47	1.47	1.47	1.71	1.71
	Qwen	2.41	2.41	2.41	2.03	2.41	2.41	2.41
	Mistral	1.70	1.70	1.28	1.28	1.28	1.28	1.28
SQuAD	Llama	1.38	1.38	1.38	1.40	1.38	1.38	1.38
	Qwen	1.75	2.20	1.59	1.69	1.69	1.59	1.59
	Mistral	1.09	1.09	1.09	1.09	1.09	1.27	1.27

F.6 Plots of Model Selective Accuracy

We compare the selective accuracy of recalibration methods and baselines across the different semantic confidence measures introduced in this paper for the Qwen model. The main results are shown in Figure 4 for four of the SC measures discussed within this work. For completeness, we include the full results for Qwen across all SC measures of confidence in Figure 8.

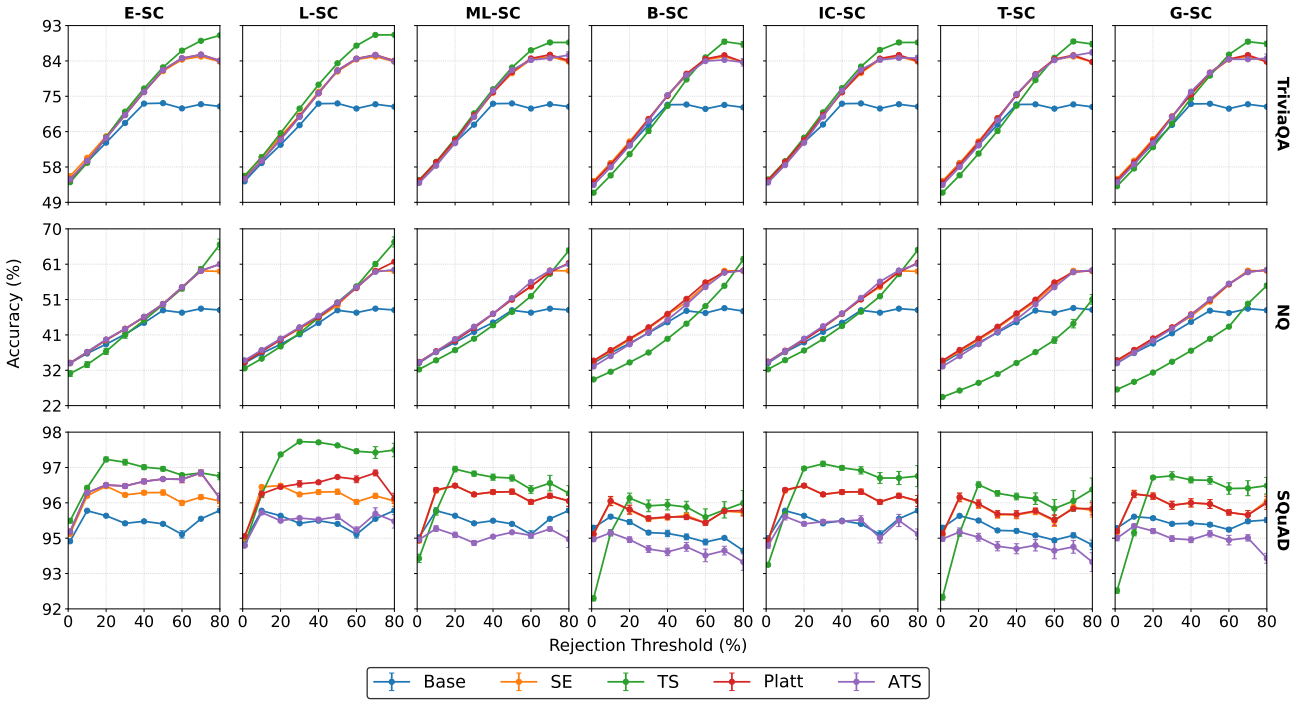


Figure (8): **Selective Accuracy for Qwen Across Varying Rejection Rates.** Mean selective accuracy (↑) curves of the Qwen model across datasets, comparing SC measures across baseline and recalibration. Error bars show one standard error across runs.

F.7 Model Brier Scores

As discussed in Section B.1, the Brier score decomposes into three components, uncertainty, resolution, and reliability, which relate to calibration and discrimination but differ from the literature standard $\widehat{\text{ACE}}$ and AUROC estimators we have used for our main results presentation. To complement the main results on semantic uncertainty quantification shown in Figure 3, we report Brier scores for the Llama model across baseline recalibration methods, and SC measures in Table 10.

For TriviaQA and NQ, temperature scaling (TS) consistently improves Brier scores, highlighting its effectiveness on these challenging closed-book tasks where base models are poorly calibrated and discriminative. In contrast, on SQuAD, TS often yields slightly worse Brier scores than the base and baseline models. This is because SQuAD base models are already well-calibrated, resulting in lower Brier scores overall and smaller absolute differences between methods.

These results reinforce our main conclusion: TS is most beneficial for datasets with poor base calibration, such as TriviaQA and NQ. On easier datasets like SQuAD, Brier scores show only minor absolute differences, yet AUROC results in Figure 3 reveal that TS can still enhance discriminability even when base calibration is strong.

Table (10): **Brier Scores for SC Measures Across Baseline and Recalibration Methods for Llama.** Mean and standard error of Brier scores (↓) are reported for each SC measure, comparing baseline and recalibration methods for the Llama model.

		E-SC	L-SC	ML-SC	B-SC	IC-SC	T-SC	G-SC
TriviaQA	Base	$0.197_{\pm 0.001}$	$0.189_{\pm 0.001}$	$0.194_{\pm 0.001}$	$0.174_{\pm 0.001}$	$0.200_{\pm 0.001}$	$0.174_{\pm 0.001}$	$0.190_{\pm 0.001}$
	SE	$0.147_{\pm 0.001}$	$0.140_{\pm 0.001}$	$0.142_{\pm 0.001}$	$0.147_{\pm 0.001}$	$0.147_{\pm 0.001}$	$0.147_{\pm 0.001}$	$0.145_{\pm 0.001}$
	TS	$\mathbf{0.140_{\pm 0.002}}$	$\mathbf{0.135_{\pm 0.002}}$	$\mathbf{0.136_{\pm 0.003}}$	$0.150_{\pm 0.001}$	$\mathbf{0.130_{\pm 0.003}}$	$0.151_{\pm 0.002}$	$\mathbf{0.142_{\pm 0.002}}$
	Platt	$0.147_{\pm 0.001}$	$0.140_{\pm 0.001}$	$0.140_{\pm 0.001}$	$0.143_{\pm 0.001}$	$0.145_{\pm 0.000}$	$\mathbf{0.142_{\pm 0.001}}$	$0.144_{\pm 0.001}$
	ATS	$0.150_{\pm 0.001}$	$0.140_{\pm 0.001}$	$0.142_{\pm 0.001}$	$\mathbf{0.143_{\pm 0.001}}$	$0.146_{\pm 0.002}$	$0.143_{\pm 0.001}$	$0.147_{\pm 0.001}$
NQ	Base	$0.358_{\pm 0.001}$	$0.350_{\pm 0.002}$	$0.353_{\pm 0.001}$	$0.309_{\pm 0.000}$	$0.374_{\pm 0.001}$	$0.310_{\pm 0.000}$	$0.345_{\pm 0.001}$
	SE	$0.240_{\pm 0.002}$	$0.236_{\pm 0.002}$	$0.233_{\pm 0.002}$	$0.222_{\pm 0.002}$	$0.264_{\pm 0.003}$	$0.222_{\pm 0.002}$	$0.231_{\pm 0.002}$
	TS	$\mathbf{0.200_{\pm 0.001}}$	$\mathbf{0.175_{\pm 0.002}}$	$\mathbf{0.196_{\pm 0.002}}$	$\mathbf{0.207_{\pm 0.002}}$	$\mathbf{0.208_{\pm 0.000}}$	$\mathbf{0.206_{\pm 0.002}}$	$\mathbf{0.199_{\pm 0.002}}$
	Platt	$0.242_{\pm 0.001}$	$0.239_{\pm 0.002}$	$0.235_{\pm 0.001}$	$0.222_{\pm 0.001}$	$0.266_{\pm 0.001}$	$0.221_{\pm 0.001}$	$0.233_{\pm 0.001}$
	ATS	$0.235_{\pm 0.001}$	$0.238_{\pm 0.002}$	$0.228_{\pm 0.002}$	$0.222_{\pm 0.002}$	$0.261_{\pm 0.002}$	$0.221_{\pm 0.002}$	$0.228_{\pm 0.001}$
SQuAD	Base	$\mathbf{0.051_{\pm 0.000}}$	$\mathbf{0.047_{\pm 0.000}}$	$\mathbf{0.051_{\pm 0.000}}$	$\mathbf{0.050_{\pm 0.001}}$	$\mathbf{0.048_{\pm 0.001}}$	$\mathbf{0.050_{\pm 0.001}}$	$\mathbf{0.051_{\pm 0.000}}$
	SE	$0.056_{\pm 0.001}$	$0.055_{\pm 0.000}$	$0.056_{\pm 0.001}$	$0.060_{\pm 0.000}$	$0.052_{\pm 0.001}$	$0.060_{\pm 0.000}$	$0.055_{\pm 0.001}$
	TS	$0.067_{\pm 0.003}$	$0.065_{\pm 0.000}$	$0.070_{\pm 0.001}$	$0.077_{\pm 0.001}$	$0.058_{\pm 0.001}$	$0.077_{\pm 0.001}$	$0.068_{\pm 0.002}$
	Platt	$0.056_{\pm 0.001}$	$0.054_{\pm 0.001}$	$0.056_{\pm 0.001}$	$0.061_{\pm 0.001}$	$0.051_{\pm 0.001}$	$0.060_{\pm 0.000}$	$0.055_{\pm 0.001}$
	ATS	$0.058_{\pm 0.001}$	$0.055_{\pm 0.000}$	$0.056_{\pm 0.001}$	$0.051_{\pm 0.001}$	$0.049_{\pm 0.001}$	$0.051_{\pm 0.001}$	$0.057_{\pm 0.001}$

F.8 Validation of NLI-Based Semantic Clustering

Following prior work, we employ a DeBERTa-v2-XXLarge natural language inference (NLI) model to cluster semantically equivalent responses. To ensure that token-level likelihood adjustments propagate meaningfully to semantic clusters in our setting, we perform both model-based and manual audits of clustering quality using this methodology, corroborating the findings of Kuhn et al. (2023), who show that NLI-based clustering performs well for the short-form generative QA tasks studied both in their work and ours.

Verification Methodology We randomly sampled 100 examples from the Natural Questions (NQ) dataset and generated 10 responses per example using Llama-3.1-8B-Instruct. The resulting semantic clusters were evaluated using a two-stage verification process:

- **LLM Judge:** We prompted Claude 4.5 (Anthropic, 2025) to assess whether (i) each cluster contained only semantically equivalent responses and (ii) different clusters corresponded to distinct semantic meanings with no overlap.

- **Human Audit:** We manually reviewed the LLM annotations to verify the correctness of the semantic judgements.

Results and Error Analysis This audit yielded a clustering accuracy of 94%. The remaining 6% of errors were primarily false negatives produced by the NLI model, where responses conveying the same lack of information, often through verbose refusals or meta-level statements (e.g., "I am not sure, but..."), were incorrectly split into separate clusters.

F.9 Validation of Lexical vs. NLI Correctness Evaluation

In our main evaluation, we employ a hybrid strategy to quantify uncertainty: we use a DeBERTa-based NLI model to cluster model generations (where variability is high) but rely on standard lexical metrics to determine the correctness of the final response against ground-truth answers. We adopt this methodology to remain strictly comparable with prior work in semantic uncertainty quantification (Kuhn et al., 2023; Farquhar et al., 2024) and to avoid the prohibitive computational cost of performing NLI pairwise comparisons for every correctness check.

Ablation Study: Lexical vs. NLI Correctness. To validate the robustness of this approach, we conducted an ablation study comparing our standard lexical evaluation against a purely NLI-based correctness check. Specifically, we used the Llama-3.1-8B-Instruct model on the TriviaQA dataset to generate responses and independently assessed their correctness using both the standard lexical criteria (see Section F.9) and the DeBERTa-based NLI model used for clustering. In the NLI setting, a response is deemed correct if it entails, and is entailed by, a ground-truth answer within the context of the question, following the same bidirectional entailment logic used for clustering responses.

Results. We observed a 91% agreement rate between the standard lexical methodology and the NLI-based evaluation when determining response correctness relative to the ground truth. This high level of consistency corroborates that the literature standard approach aligns well with the NLI approach whilst being computationally much cheaper, highlighting its practical utility and robustness for practitioners.

F.10 Model Calibration Results Are Robust to the Choice of Metric

As noted in Section B.1, bin-based metrics such as ECE and ACE can be sensitive to the choice of binning scheme. This has motivated the development of alternative, bin-free calibration metrics, such as MCB-CORP, which is non-parametric and comes with strong statistical guarantees (Dimitriadis et al., 2021).

To evaluate the robustness of our findings, we compare ACE, ECE and MCB-CORP for the Qwen model in Table 11. Consistent with Figure 3, we observe that TS consistently improves semantic calibration on the more challenging closed-book datasets (TriviaQA and NQ), while its effect on the open-book dataset SQuAD is negligible, reflecting the already strong calibration of base models here.

We further quantify robustness in Table 13 by reporting the mean absolute rank change (volatility) when replacing ACE with MCB-CORP. Volatility is minimal for TriviaQA and NQ and slightly higher for SQuAD, where methods perform similarly. Overall, these results confirm that our main conclusions are independent of the specific calibration metric chosen.

G Extended Analysis: Inductive Bias and Overfitting in Semantic Calibration

Our results demonstrate that scalar TS consistently outperforms more expressive methods, such as ATS and Platt Scaling, in the context of semantic UQ. In this section, we expand on the theoretical and practical reasons for this finding, focusing on two key mechanisms: the preservation of semantic hierarchies and the prevention of overfitting to semantically irrelevant tokens.

Rank Preservation (TS vs Platt). A critical advantage of scalar TS is its strict preservation of the base model's likelihood-based token rankings at any given token position. TS applies a strictly monotonic transformation, guaranteeing that if the model assigns a higher logit to one token over another, this preference is preserved after temperature scaling. In contrast, more expressive methods like Platt Scaling introduce shift

Table (11): **Calibration Metric Comparison of SC Measures Across Methods for Qwen.** Mean and standard error of calibration error for SC measures under $\widehat{\text{ACE}}$ (bin-based) (↓) and MCB-CORP (bin-free) (↓) across baseline and recalibration methods on TriviaQA, NQ, and SQuAD. **Bold** entries indicate the best-performing method for each SC measure within a dataset, allowing direct comparison between bin-based and bin-free calibration metrics.

		$\widehat{\text{ACE}}$							CORP-MCB						
		E-SC	L-SC	ML-SC	B-SC	IC-SC	T-SC	G-SC	E-SC	L-SC	ML-SC	B-SC	IC-SC	T-SC	G-SC
TriviaQA	Base	$0.308_{\pm 0.001}$	$0.299_{\pm 0.001}$	$0.304_{\pm 0.000}$	$0.240_{\pm 0.000}$	$0.320_{\pm 0.001}$	$0.240_{\pm 0.000}$	$0.293_{\pm 0.001}$	$0.105_{\pm 0.001}$	$0.096_{\pm 0.001}$	$0.101_{\pm 0.001}$	$0.062_{\pm 0.000}$	$0.114_{\pm 0.001}$	$0.062_{\pm 0.000}$	$0.093_{\pm 0.000}$
	SE	$0.165_{\pm 0.002}$	$0.159_{\pm 0.001}$	$0.158_{\pm 0.001}$	$0.097_{\pm 0.002}$	$0.194_{\pm 0.001}$	$0.099_{\pm 0.002}$	$0.145_{\pm 0.002}$	$0.035_{\pm 0.001}$	$0.028_{\pm 0.001}$	$0.031_{\pm 0.000}$	$0.015_{\pm 0.000}$	$0.045_{\pm 0.001}$	$0.015_{\pm 0.000}$	$0.027_{\pm 0.001}$
	TS	$\mathbf{0.080_{\pm 0.001}}$	$\mathbf{0.066_{\pm 0.005}}$	$\mathbf{0.069_{\pm 0.002}}$	$\mathbf{0.082_{\pm 0.003}}$	$\mathbf{0.085_{\pm 0.004}}$	$\mathbf{0.081_{\pm 0.003}}$	$\mathbf{0.067_{\pm 0.002}}$	$\mathbf{0.010_{\pm 0.000}}$	$\mathbf{0.006_{\pm 0.001}}$	$\mathbf{0.009_{\pm 0.000}}$	$\mathbf{0.011_{\pm 0.001}}$	$\mathbf{0.011_{\pm 0.001}}$	$\mathbf{0.010_{\pm 0.001}}$	$\mathbf{0.008_{\pm 0.000}}$
	Platt	$0.170_{\pm 0.001}$	$0.168_{\pm 0.002}$	$0.163_{\pm 0.002}$	$0.100_{\pm 0.002}$	$0.203_{\pm 0.001}$	$0.101_{\pm 0.003}$	$0.150_{\pm 0.001}$	$0.035_{\pm 0.000}$	$0.032_{\pm 0.001}$	$0.032_{\pm 0.001}$	$0.015_{\pm 0.000}$	$0.050_{\pm 0.001}$	$0.015_{\pm 0.000}$	$0.028_{\pm 0.000}$
	ATS	$0.204_{\pm 0.003}$	$0.168_{\pm 0.002}$	$0.178_{\pm 0.002}$	$0.132_{\pm 0.003}$	$0.219_{\pm 0.001}$	$0.133_{\pm 0.002}$	$0.183_{\pm 0.003}$	$0.050_{\pm 0.001}$	$0.032_{\pm 0.001}$	$0.037_{\pm 0.001}$	$0.020_{\pm 0.001}$	$0.057_{\pm 0.001}$	$0.021_{\pm 0.001}$	$0.040_{\pm 0.001}$
NQ	Base	$0.487_{\pm 0.001}$	$0.473_{\pm 0.001}$	$0.480_{\pm 0.001}$	$0.399_{\pm 0.001}$	$0.496_{\pm 0.001}$	$0.399_{\pm 0.001}$	$0.468_{\pm 0.001}$	$0.253_{\pm 0.001}$	$0.237_{\pm 0.002}$	$0.245_{\pm 0.002}$	$0.181_{\pm 0.001}$	$0.258_{\pm 0.002}$	$0.181_{\pm 0.001}$	$0.235_{\pm 0.001}$
	SE	$0.319_{\pm 0.002}$	$0.318_{\pm 0.002}$	$0.311_{\pm 0.002}$	$0.228_{\pm 0.002}$	$0.354_{\pm 0.002}$	$0.228_{\pm 0.002}$	$0.294_{\pm 0.002}$	$0.121_{\pm 0.001}$	$0.118_{\pm 0.001}$	$0.116_{\pm 0.001}$	$0.072_{\pm 0.001}$	$0.142_{\pm 0.001}$	$0.072_{\pm 0.001}$	$0.106_{\pm 0.001}$
	TS	$\mathbf{0.155_{\pm 0.002}}$	$\mathbf{0.084_{\pm 0.002}}$	$\mathbf{0.085_{\pm 0.003}}$	$\mathbf{0.085_{\pm 0.001}}$	$\mathbf{0.192_{\pm 0.002}}$	$\mathbf{0.083_{\pm 0.002}}$	$\mathbf{0.155_{\pm 0.002}}$	$\mathbf{0.034_{\pm 0.001}}$	$\mathbf{0.013_{\pm 0.001}}$	$\mathbf{0.018_{\pm 0.001}}$	$\mathbf{0.012_{\pm 0.000}}$	$\mathbf{0.043_{\pm 0.001}}$	$\mathbf{0.011_{\pm 0.000}}$	$\mathbf{0.039_{\pm 0.003}}$
	Platt	$0.314_{\pm 0.001}$	$0.318_{\pm 0.001}$	$0.307_{\pm 0.001}$	$0.231_{\pm 0.002}$	$0.357_{\pm 0.000}$	$0.232_{\pm 0.002}$	$0.289_{\pm 0.001}$	$0.117_{\pm 0.001}$	$0.119_{\pm 0.001}$	$0.113_{\pm 0.001}$	$0.074_{\pm 0.001}$	$0.143_{\pm 0.001}$	$0.074_{\pm 0.001}$	$0.102_{\pm 0.001}$
	ATS	$0.315_{\pm 0.004}$	$0.318_{\pm 0.001}$	$0.280_{\pm 0.002}$	$0.236_{\pm 0.004}$	$0.353_{\pm 0.002}$	$0.237_{\pm 0.003}$	$0.295_{\pm 0.004}$	$0.120_{\pm 0.002}$	$0.118_{\pm 0.000}$	$0.097_{\pm 0.001}$	$0.074_{\pm 0.002}$	$0.139_{\pm 0.001}$	$0.074_{\pm 0.002}$	$0.108_{\pm 0.002}$
SQuAD	Base	$0.051_{\pm 0.001}$	$0.047_{\pm 0.000}$	$0.051_{\pm 0.001}$	$0.048_{\pm 0.000}$	$0.048_{\pm 0.000}$	$0.048_{\pm 0.000}$	$0.051_{\pm 0.001}$	$\mathbf{0.003_{\pm 0.000}}$	$\mathbf{0.004_{\pm 0.000}}$	$\mathbf{0.003_{\pm 0.000}}$	$0.004_{\pm 0.000}$	$0.005_{\pm 0.000}$	$0.004_{\pm 0.000}$	$\mathbf{0.003_{\pm 0.000}}$
	SE	$\mathbf{0.050_{\pm 0.001}}$	$0.048_{\pm 0.001}$	$0.044_{\pm 0.001}$	$0.046_{\pm 0.000}$	$0.048_{\pm 0.002}$	$\mathbf{0.046_{\pm 0.000}}$	$0.049_{\pm 0.001}$	$0.004_{\pm 0.000}$	$0.006_{\pm 0.000}$	$0.004_{\pm 0.000}$	$0.004_{\pm 0.000}$	$\mathbf{0.005_{\pm 0.000}}$	$0.004_{\pm 0.000}$	$0.004_{\pm 0.000}$
	TS	$0.058_{\pm 0.001}$	$0.060_{\pm 0.001}$	$0.053_{\pm 0.002}$	$0.051_{\pm 0.001}$	$0.049_{\pm 0.003}$	$0.052_{\pm 0.002}$	$0.055_{\pm 0.001}$	$0.008_{\pm 0.000}$	$0.011_{\pm 0.000}$	$0.007_{\pm 0.001}$	$0.007_{\pm 0.000}$	$0.007_{\pm 0.000}$	$0.007_{\pm 0.001}$	$0.008_{\pm 0.000}$
	Platt	$0.050_{\pm 0.001}$	$\mathbf{0.047_{\pm 0.000}}$	$\mathbf{0.044_{\pm 0.001}}$	$\mathbf{0.046_{\pm 0.000}}$	$\mathbf{0.047_{\pm 0.002}}$	$0.046_{\pm 0.000}$	$\mathbf{0.048_{\pm 0.001}}$	$0.004_{\pm 0.000}$	$0.005_{\pm 0.000}$	$0.004_{\pm 0.000}$	$\mathbf{0.004_{\pm 0.000}}$	$0.005_{\pm 0.001}$	$0.004_{\pm 0.000}$	$0.004_{\pm 0.000}$
	ATS	$0.056_{\pm 0.001}$	$\mathbf{0.047_{\pm 0.000}}$	$0.055_{\pm 0.001}$	$0.053_{\pm 0.001}$	$0.053_{\pm 0.001}$	$0.053_{\pm 0.001}$	$0.056_{\pm 0.001}$	$0.004_{\pm 0.000}$	$0.005_{\pm 0.000}$	$0.004_{\pm 0.000}$	$0.005_{\pm 0.000}$	$0.005_{\pm 0.000}$	$\mathbf{0.004_{\pm 0.000}}$	$0.004_{\pm 0.000}$

Table (12): **$\widehat{\text{ACE}}$ vs $\widehat{\text{ECE}}$ Comparison of SC Measures Across Methods for Qwen.** Mean and standard error of calibration error for SC measures under $\widehat{\text{ACE}}$ (bin-based) (↓) and $\widehat{\text{ECE}}$ (↓) across baseline and recalibration methods on TriviaQA, NQ, and SQuAD. **Bold** entries indicate the best-performing method for each SC measure within a dataset.

	Method	ACE							ECE						
		E-SC	L-SC	ML-SC	B-SC	IC-SC	T-SC	G-SC	E-SC	L-SC	ML-SC	B-SC	IC-SC	T-SC	G-SC
TriviaQA	Base	$0.308_{\pm 0.001}$	$0.299_{\pm 0.001}$	$0.304_{\pm 0.000}$	$0.240_{\pm 0.000}$	$0.320_{\pm 0.001}$	$0.240_{\pm 0.000}$	$0.293_{\pm 0.001}$	$0.308_{\pm 0.001}$	$0.299_{\pm 0.001}$	$0.304_{\pm 0.000}$	$0.240_{\pm 0.000}$	$0.321_{\pm 0.001}$	$0.240_{\pm 0.000}$	$0.293_{\pm 0.001}$
	SE	$0.165_{\pm 0.002}$	$0.159_{\pm 0.001}$	$0.158_{\pm 0.001}$	$0.097_{\pm 0.002}$	$0.194_{\pm 0.001}$	$0.099_{\pm 0.002}$	$0.145_{\pm 0.002}$	$0.165_{\pm 0.002}$	$0.158_{\pm 0.001}$	$0.158_{\pm 0.001}$	$0.100_{\pm 0.002}$	$0.194_{\pm 0.001}$	$0.101_{\pm 0.002}$	$0.145_{\pm 0.002}$
	TS	$\mathbf{0.080_{\pm 0.001}}$	$\mathbf{0.066_{\pm 0.005}}$	$\mathbf{0.069_{\pm 0.002}}$	$\mathbf{0.082_{\pm 0.003}}$	$\mathbf{0.085_{\pm 0.004}}$	$\mathbf{0.081_{\pm 0.003}}$	$\mathbf{0.067_{\pm 0.002}}$	$\mathbf{0.080_{\pm 0.001}}$	$\mathbf{0.066_{\pm 0.004}}$	$\mathbf{0.069_{\pm 0.002}}$	$\mathbf{0.086_{\pm 0.004}}$	$\mathbf{0.087_{\pm 0.004}}$	$\mathbf{0.085_{\pm 0.005}}$	$\mathbf{0.069_{\pm 0.002}}$
	Platt	$0.170_{\pm 0.001}$	$0.168_{\pm 0.002}$	$0.163_{\pm 0.002}$	$0.100_{\pm 0.002}$	$0.203_{\pm 0.001}$	$0.101_{\pm 0.003}$	$0.150_{\pm 0.001}$	$0.170_{\pm 0.001}$	$0.168_{\pm 0.002}$	$0.163_{\pm 0.002}$	$0.101_{\pm 0.002}$	$0.203_{\pm 0.001}$	$0.102_{\pm 0.002}$	$0.150_{\pm 0.001}$
	ATS	$0.204_{\pm 0.003}$	$0.168_{\pm 0.002}$	$0.178_{\pm 0.002}$	$0.132_{\pm 0.003}$	$0.219_{\pm 0.001}$	$0.133_{\pm 0.002}$	$0.183_{\pm 0.003}$	$0.204_{\pm 0.003}$	$0.168_{\pm 0.002}$	$0.178_{\pm 0.002}$	$0.132_{\pm 0.003}$	$0.219_{\pm 0.001}$	$0.133_{\pm 0.002}$	$0.183_{\pm 0.003}$
NQ	Base	$0.487_{\pm 0.001}$	$0.473_{\pm 0.001}$	$0.480_{\pm 0.001}$	$0.399_{\pm 0.001}$	$0.496_{\pm 0.001}$	$0.399_{\pm 0.001}$	$0.468_{\pm 0.001}$	$0.486_{\pm 0.001}$	$0.472_{\pm 0.002}$	$0.479_{\pm 0.001}$	$0.399_{\pm 0.001}$	$0.496_{\pm 0.001}$	$0.399_{\pm 0.001}$	$0.468_{\pm 0.001}$
	SE	$0.319_{\pm 0.002}$	$0.318_{\pm 0.002}$	$0.311_{\pm 0.002}$	$0.228_{\pm 0.002}$	$0.354_{\pm 0.002}$	$0.228_{\pm 0.002}$	$0.294_{\pm 0.002}$	$0.319_{\pm 0.002}$	$0.318_{\pm 0.002}$	$0.311_{\pm 0.002}$	$0.228_{\pm 0.002}$	$0.353_{\pm 0.002}$	$0.228_{\pm 0.002}$	$0.294_{\pm 0.002}$
	TS	$\mathbf{0.155_{\pm 0.002}}$	$\mathbf{0.084_{\pm 0.002}}$	$\mathbf{0.085_{\pm 0.003}}$	$\mathbf{0.085_{\pm 0.001}}$	$\mathbf{0.192_{\pm 0.002}}$	$\mathbf{0.083_{\pm 0.002}}$	$\mathbf{0.155_{\pm 0.002}}$	$\mathbf{0.155_{\pm 0.002}}$	$\mathbf{0.081_{\pm 0.002}}$	$\mathbf{0.085_{\pm 0.003}}$	$\mathbf{0.082_{\pm 0.003}}$	$\mathbf{0.192_{\pm 0.002}}$	$\mathbf{0.081_{\pm 0.003}}$	$\mathbf{0.155_{\pm 0.002}}$
	Platt	$0.314_{\pm 0.001}$	$0.318_{\pm 0.001}$	$0.307_{\pm 0.001}$	$0.231_{\pm 0.002}$	$0.357_{\pm 0.000}$	$0.232_{\pm 0.002}$	$0.289_{\pm 0.001}$	$0.313_{\pm 0.001}$	$0.318_{\pm 0.001}$	$0.306_{\pm 0.001}$	$0.230_{\pm 0.002}$	$0.356_{\pm 0.000}$	$0.232_{\pm 0.002}$	$0.288_{\pm 0.001}$
	ATS	$0.315_{\pm 0.004}$	$0.318_{\pm 0.001}$	$0.280_{\pm 0.002}$	$0.236_{\pm 0.004}$	$0.353_{\pm 0.002}$	$0.237_{\pm 0.003}$	$0.295_{\pm 0.004}$	$0.315_{\pm 0.004}$	$0.317_{\pm 0.001}$	$0.280_{\pm 0.002}$	$0.235_{\pm 0.004}$	$0.352_{\pm 0.002}$	$0.236_{\pm 0.003}$	$0.295_{\pm 0.004}$
SQuAD	Base	$0.051_{\pm 0.001}$	$0.047_{\pm 0.000}$	$0.051_{\pm 0.001}$	$0.048_{\pm 0.000}$	$0.048_{\pm 0.000}$	$0.048_{\pm 0.000}$	$0.051_{\pm 0.001}$	$\mathbf{0.053_{\pm 0.001}}$	$0.051_{\pm 0.001}$	$0.053_{\pm 0.000}$	$0.054_{\pm 0.000}$	$0.056_{\pm 0.000}$	$0.054_{\pm 0.000}$	$0.053_{\pm 0.001}$
	SE	$\mathbf{0.050_{\pm 0.001}}$	$0.048_{\pm 0.001}$	$0.044_{\pm 0.001}$	$0.046_{\pm 0.000}$	$0.048_{\pm 0.002}$	$\mathbf{0.046_{\pm 0.000}}$	$0.049_{\pm 0.001}$	$0.053_{\pm 0.001}$	$0.051_{\pm 0.001}$	$\mathbf{0.050_{\pm 0.002}}$	$\mathbf{0.049_{\pm 0.001}}$	$0.050_{\pm 0.001}$	$\mathbf{0.049_{\pm 0.001}}$	$0.053_{\pm 0.001}$
	TS	$0.058_{\pm 0.001}$	$0.060_{\pm 0.001}$	$0.053_{\pm 0.002}$	$0.051_{\pm 0.001}$	$0.049_{\pm 0.003}$	$0.052_{\pm 0.002}$	$0.055_{\pm 0.001}$	$0.064_{\pm 0.001}$	$0.060_{\pm 0.001}$	$0.056_{\pm 0.002}$	$0.054_{\pm 0.000}$	$\mathbf{0.049_{\pm 0.002}}$	$0.053_{\pm 0.002}$	$0.062_{\pm 0.001}$
	Platt	$0.050_{\pm 0.001}$	$\mathbf{0.047_{\pm 0.000}}$	$\mathbf{0.044_{\pm 0.001}}$	$\mathbf{0.046_{\pm 0.000}}$	$\mathbf{0.047_{\pm 0.002}}$	$0.046_{\pm 0.000}$	$\mathbf{0.048_{\pm 0.001}}$	$0.053_{\pm 0.001}$	$\mathbf{0.049_{\pm 0.001}}$	$0.050_{\pm 0.002}$	$0.049_{\pm 0.001}$	$0.050_{\pm 0.002}$	$0.050_{\pm 0.002}$	$\mathbf{0.052_{\pm 0.001}}$
	ATS	$0.056_{\pm 0.001}$	$\mathbf{0.047_{\pm 0.000}}$	$0.055_{\pm 0.001}$	$0.053_{\pm 0.001}$	$0.053_{\pm 0.001}$	$0.053_{\pm 0.001}$	$0.056_{\pm 0.001}$	$0.057_{\pm 0.001}$	$\mathbf{0.049_{\pm 0.001}}$	$0.057_{\pm 0.001}$	$0.058_{\pm 0.001}$	$0.057_{\pm 0.001}$	$0.057_{\pm 0.001}$	$0.056_{\pm 0.001}$

Table (13): **Ranking Volatility of Calibration Metrics Across Methods for Qwen.** Mean absolute rank change (volatility) of methods on TriviaQA, NQ, and SQuAD when switching from ACE (bin-based) to MCB-CORP (bin-free). Higher values indicate greater sensitivity of method rankings to the choice of calibration metric. **Bold** entries denote the lowest mean volatility per dataset, corresponding to the most stable ranking of methods.

	TriviaQA	NQ	SQuAD
Base	0.00	0.00	1.43
SE	0.29	0.14	1.29
TS	0.00	0.00	0.71
Platt	0.29	0.00	1.29
ATS	0.00	0.14	2.14
Mean	**0.11**	**0.06**	**1.37**

parameters or token-specific scaling. While theoretically more flexible, these affine transformations can distort the relative ranking of tokens. In the semantic setting, this is particularly dangerous: an affine transformation might inadvertently suppress a *semantically crucial* token while upweighting a generic, high-frequency token. By design, TS is constrained to respect the model's original rank-ordering, making it inherently robust to such semantic distortions.

Overparameterisation and Filler Tokens (TS vs. ATS). The comparison with ATS highlights a critical failure mode driven by model capacity. ATS introduces transformer heads with millions of parameters, creating a high susceptibility to overparameterisation given the relatively small size of standard QA calibration sets. We argue that ATS exploits this capacity to overfit to *semantically irrelevant* features. Since generic filler tokens such as stop words and punctuation are far more abundant than semantically loaded terms, ATS can drastically reduce the token-level NLL by aggressively optimising temperatures for these frequent but semantically irrelevant tokens. Empirically, we find that ATS frequently achieves near-zero training loss despite its poor generalisation to semantic UQ, indicating that this overfitting mechanism is indeed likely at play. Indeed, this is further supported from our observations that TS achieves a worse loss during training, yet achieves much better downstream semantic UQ performance.

In contrast, we argue that Scalar TS provides a superior inductive bias for semantic calibration. By enforcing a single global constraint across the entire generation, TS acts as a powerful regulariser against the overfitting behaviour observed in ATS. Specifically, the method lacks the capacity to minimise loss by merely fitting frequent, non-semantic filler tokens. Instead, TS forces the calibration to reflect the uncertainty of the sequence as a holistic unit, thereby better capturing the overall meaning conveyed. In this context, TS represents a robust Occam-style model selection (MacKay, 2003), where the simplest sufficient constraint yields superior semantic generalisation.

H Comparisons of Semantic Measures of Confidence

Semantic Confidence Measure Distribution Plots. To highlight differences between SC measures, and to illustrate the effect of the temperature parameter on their distributions, we plot the semantic measure distributions for the Base, SE, and TS methods of the Llama model on the NQ dataset in Figure 9.

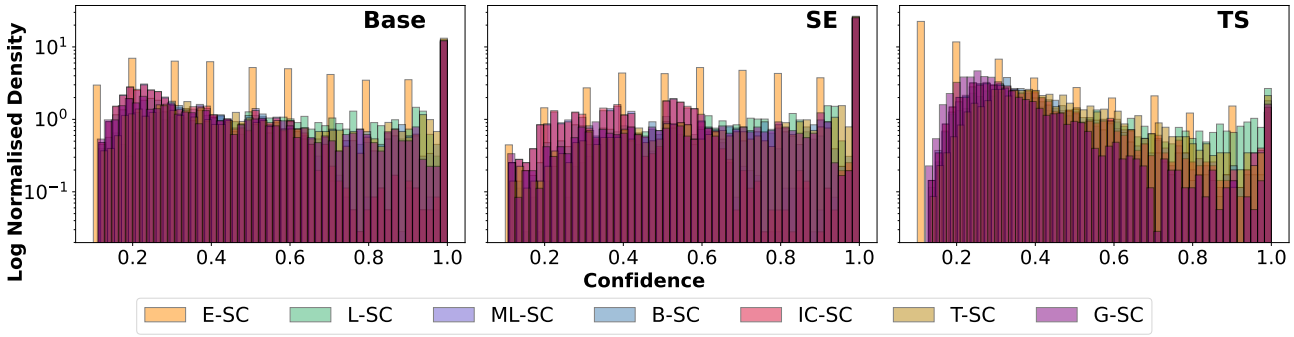


Figure (9): **SC Distribution Plots.** Distributions of semantic confidence measures for the Base, SE, and TS methods on the NQ dataset with the Llama model.

Correlation Between Semantic Measures Across Temperature Settings. To complement the distribution plots in Figure 9, we report pairwise Pearson correlations between the semantic confidence (SC) measures under the Base, SE, and TS settings.

Correlation Between Semantic Measures Across Temperature Settings. To complement the distribution plots show in Figure 9, we report pairwise Pearson correlations between the semantic confidence (SC) measures under the Base, SE, and TS settings.

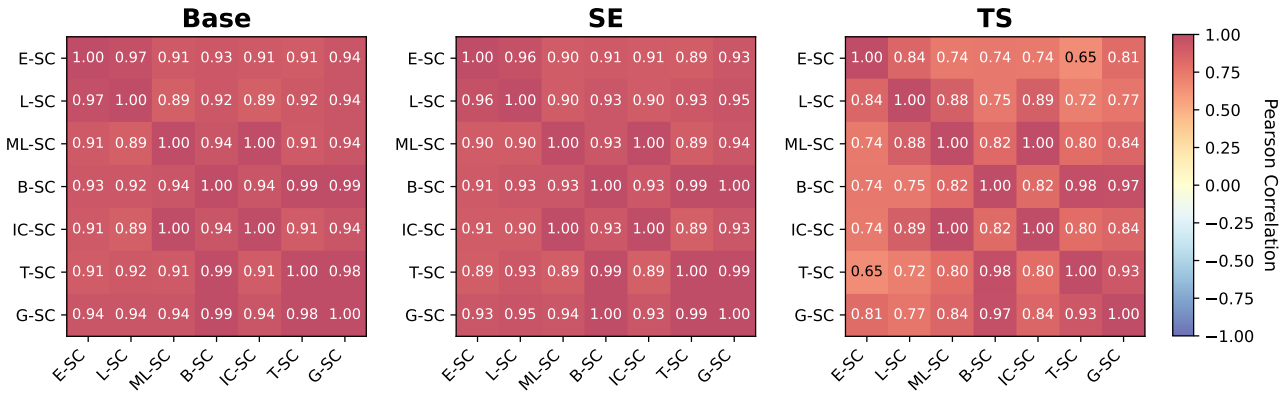


Figure (10): **Correlation matrices of semantic confidence measures.** Pairwise Pearson correlations between SC measures for the Base, SE, and TS methods on the NQ dataset with the Llama model. While correlations are very high under Base and SE settings, indicating redundancy between measures, optimised Temperature Scaling (TS) substantially lowers correlations, revealing greater diversity in how measures capture semantic uncertainty.

The results in Figure 10 show that at Base and SE temperatures, semantic confidence measures are almost perfectly correlated (coefficients > 0.9), behaving as near reparameterisations of one another. Under optimised Temperature Scaling (TS), correlations drop substantially (down to 0.65), indicating that TS increases the diversity of the measures and highlights their distinct behaviour. Crucially, this reduced alignment is consistent with the differences observed in downstream calibration and discrimination metrics, confirming that the measures are not redundant but capture complementary aspects of semantic uncertainty.