Influences on LLM Calibration: A Study of Response Agreement, Loss Functions, and Prompt Styles

Yuxi Xia[1,2*], **Pedro Henrique Luz de Araujo**[1,2], **Klim Zaporojets**[3]
Benjamin Roth[1,4]
[1]Faculty of Computer Science, University of Vienna, Vienna, Austria
[2]UniVie Doctoral School Computer Science, Vienna, Austria
[3]Department of Computer Science, Aarhus University, Aarhus, Denmark
[4]Faculty of Philological and Cultural Studies, University of Vienna, Vienna, Austria
[*]yuxi.xia@univie.ac.at

Abstract

Calibration, the alignment between model confidence and prediction accuracy, is critical for the reliable deployment of large language models (LLMs). Existing works neglect to measure the generalization of their methods to other prompt styles and different sizes of LLMs. To address this, we define a controlled experimental setting covering 12 LLMs and four prompt styles. We additionally investigate if incorporating the response agreement of multiple LLMs and an appropriate loss function can improve calibration performance. Concretely, we build Calib-n, a novel framework that trains an auxiliary model for confidence estimation that aggregates responses from multiple LLMs to capture inter-model agreement. To optimize calibration, we integrate focal and AUC surrogate losses alongside binary cross-entropy. Experiments across four datasets demonstrate that both response agreement and focal loss improve calibration from baselines. We find that few-shot prompts are the most effective for auxiliary model-based methods, and auxiliary models demonstrate robust calibration performance across accuracy variations, outperforming LLMs' internal probabilities and verbalized confidences. [1]

1 Introduction

Improving the calibration of Large Language Models (LLMs), i.e., aligning the model's confidence with the accuracy of its predictions, can maintain their reliability, usability, and ethical deployment in domains like medicine, law, and education (Guo et al., 2017a; Jiang et al., 2021; Geng et al., 2024). As LLMs are increasingly integrated into decision-making processes, poor calibration can amplify risks of misinformation, propagate biases, and foster over-reliance among users (Raji et al., 2020).

Recent study (Ni et al., 2024) indicates that LLMs struggle to accurately express their internal confidence in natural language, particularly in the form of verbalized confidence (Tian et al., 2023). Liu et al. (2024) addresses this issue by training a linear layer to adjust the hidden states of the LLM's final layer for confidence prediction, but this approach only applies to LLMs with accessible weights. In contrast, Ulmer et al. (2024) introduces an auxiliary model for confidence estimation using only the generations of the target LLM but is constrained by its evaluation on just two LLMs and two prompt styles. These narrow experimental setups limit the generalization of their findings.

Our study tackles this limitation by comprehensively examining the generalization of different methods over 12 LLMs and four prompt styles. Additionally, we investigate new influence factors on the calibration of LLMs, which includes response agreements among LLMs, loss functions, and prompt styles. Concretely, we introduce **Calib-n** (illustrated in Fig. 1), a novel framework aggregating responses from multiple LLMs (n indicates the number of LLMs) to train a single auxiliary model for confidence estimation. Calib-n captures inter-model agreement, mitigating overfitting and reducing overconfidence associated with individual LLMs for calibration (Kim et al., 2023). Except for standard binary cross-entropy (BCE) loss, we incorporate focal (FL) (Lin et al., 2017) and AUC surrogate (AUC) (Yuan et al., 2021) loss functions that are effective for improving the calibration of non-transformer type of neural networks (Mukhoti et al., 2020; Moon et al., 2020). Finally, we systematically study the effects of prompt styles, testing four diverse prompt types: Verbalized (Tian et al., 2023), Chain-of-Thought (CoT) (Wei et al., 2022), Zero-shot, and Few-shot prompts.

Our experiments span four open-ended question answering datasets and 12 LLMs, including seven small models (2–9B parameters) and five

[1]Code and data are available at https://github.com/Yuuxii/Influences-on-LLM-Calibration.git.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3740–3761
July 27 - August 1, 2025

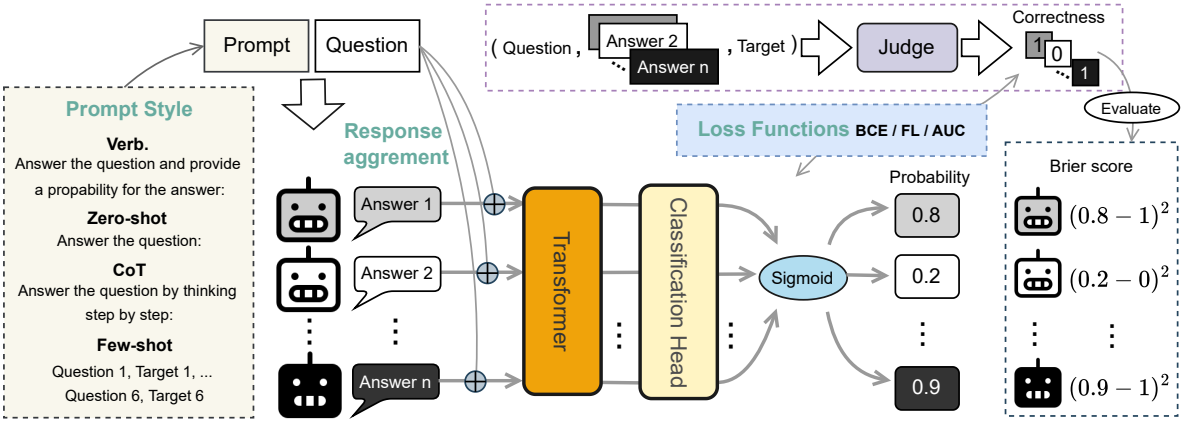


Figure 1: Overview of calibration training of Calib-n (n indicates the number of target LLMs that provide responses). We consider the effect of different **prompt styles** on calibration and design four diverse prompts to query n target LLMs to provide answers to a question. n joint strings of the question with each answer presenting the **response agreement** of LLMs are passed to the auxiliary model to generate probabilities for each answer. The auxiliary model is optimized with three **loss functions** respectively on the correctness of the LLM answers. Brier score (one of four metrics) is used to evaluate the calibration performance of the auxiliary model.

large models (27–72B parameters) from five distinct model families. The results indicate that no single method consistently outperforms all others across the various models, datasets, and prompt configurations. Yet, after aggregating the results and counting the number of overall wins for each method, we uncover new insights regarding the calibration factors of the analyzed settings:

- **Response Agreement:** By leveraging inter-model response agreement, Calib-n outperforms the state-of-the-art baselines.
- **Loss Function:** FL loss improves calibration compared to BCE and AUC losses, demonstrating its effectiveness for both Calib-1 (i.e., using responses from one LLM) and Calib-n. Calib-1 trained with FL yields the best results.
- **Prompt Style:** We find that the effectiveness of methods is highly influenced by the prompt styles, with few-shot prompts proving to be the most beneficial in improving calibration.
- **Accuracy-Calibration Correlation:** With accuracy variances caused by different dataset complexities, prompt styles and models, we find that auxiliary models maintain robust calibration performance across accuracy changes, in contrast to the fluctuating calibration of LLMs that rely on internal probabilities and verbalized confidences.

Our findings underscore the importance of reexamining calibration strategies for LLMs. Specifically, we suggest using response agreement (Calib-n) to prevent overconfidence and reduce computational costs by training a single auxiliary model for calibrating multiple target LLMs. In the case of one target LLM, we recommend incorporating focal loss over BCE (Calib-1 with FL).

2 Related Work

Calibration The concept of calibration of neural networks was introduced by Guo et al. (2017b). Lin et al. (2022) show that GPT-3 can learn to express uncertainty about its own answers without the use of logits. Later, Tian et al. (2023) demonstrate that verbalized confidence is generally better calibrated than the conditional probabilities w.r.t the consistency of the LLMs. However, Zhang et al. (2024) show that LLM probabilities and verbalized confidence tend to overly concentrate within a fixed range. Similarly, Ni et al. (2024) analyze and compare probabilistic and verbalized perceptions of the knowledge boundaries of LLM, highlighting their challenges in confidence estimation. To address these issues, Liu et al. (2024) train a linear layer to adjust the last layer's hidden states of LLMs for confidence generation. Ulmer et al. (2024) propose a method to estimate LLM confidence based only on textual input and output. However, none of these works perform a comprehensive analysis of **different influence factors in LLM calibration**. These factors can be loss functions, response agreement and prompt styles. Mukhoti et al. (2020) demonstrate that focal loss (Lin et al., 2017) can

improve the calibration of neural networks. AUC surrogate loss (Yuan et al., 2021) and correctness ranking loss (Moon et al., 2020) calibrate models by adjusting the ranking of logits to ensure positive samples are ranked higher than negative samples. Kim et al. (2023) propose that ensemble methods promoting prediction diversity can enhance calibration performance, particularly in scenarios with limited data. Min et al. (2022); Wang et al. (2023); Chen et al. (2023) showcase that LLMs output is highly dependent on the prompt and different prompt styles can significantly influence the performance of LLMs. However, these works either do not perform on LLMs or only study individual influence factors. Additionally, the impact of prompts and model agreement has never been explored in calibration. Our work comprehensively studies the influence of response agreement achieved with the ensemble method from Xia et al. (2024), three loss functions, and four prompt styles with 12 LLMs.

3 Methodology

Fig. 1 demonstrates our framework, which includes four prompt styles, assessment of the correctness of LLM generation, and calibration training of auxiliary models for confidence estimation and calibration evaluation for these models.

3.1 Prompt Styles

Prompt styles significantly impact LLMs' performance (Chen et al., 2023). Liu et al. (2024) use Few-shot prompts to query LLM answers for calibration experiments. Tian et al. (2023) employ verbalized prompts to elicit probability estimates from LLMs regarding their responses. Ulmer et al. (2024) evaluate their methods on both CoT and verbalized prompts. However, these studies either evaluate their methods against baselines using inconsistent prompt styles or fail to provide a comprehensive comparison across diverse prompt styles. For example, Liu et al. (2024) use verbalized prompts to generate verbalized probabilities as a baseline, but apply few-shot prompts to their method, which ignores the potential influence of prompts in calibration evaluation. Our work comprehensively studies the impact of prompt styles in LLM calibration and employs the most commonly used prompts: Verbalized (Verb.), Zero-shot, CoT, and Few-shot prompts. The detailed prompts are shown in Appendix A.1. Given a question q with one of the prompts, we process the answers generated by n target LLMs with regular expression-based text processing.

3.2 Correctness of LLM Generation

We employ a Judge model $\mathcal{J}$, Prometheus-8x7b-v2.0 (Kim et al., 2024), to assess the correctness of generated answers w.r.t target answers. We selected this model because it is open-source and its judgments have been shown to strongly correlate with those of human evaluators and large proprietary models (Kim et al., 2024). Given an input question q, a target answer y, and a generated answer a_i by the i-th target LLM $\mathcal{M}_i$, $\mathcal{J}$ is prompted to provide a binary correctness score c_i that reflects the semantical equivalence between a_i and y. The specific prompt is shown in Appendix A.2.

$$c_i = \mathcal{J}(a_i \overset{\text{semantic}}{=} y|q, y, a_i),\ c_i \in \{0, 1\} \quad (1)$$

3.3 Response Agreements of Multiple LLMs

Different from previous work (Liu et al., 2024; Ulmer et al., 2024; Tian et al., 2023) that estimate confidence using the information from a single target LLM, Calib-n leverages the responses from n target LLMs to train an auxiliary model for jointly estimating the confidence for each LLM. This setting is inspired by Kim et al. (2023), which suggests that combining predictions can mitigate overfitting and reduce overconfidence inherent to individual models for confidence estimation. By having access to the responses of n LLMs, the auxiliary model can infer cases of low consensus among models, signaling increased uncertainty. We verify the effectiveness of this setting by comprehensively comparing the results of using the generations produced by one LLM (**Calib-1**) to using the generations from multiple LLMs (**Calib-n**).

The auxiliary model, $f(\cdot)$, is composed of a transformer backbone (bert-base-uncased (Devlin et al., 2019)), a classification head ($n * 768 \rightarrow n$) and a sigmoid activation function, which outputs probabilities P for LLM answers $\{a_i\}_{i=1}^n$. The input to $f(\cdot)$ consists of n concatenated question–answer pairs of the form q[SEP]a_i (e.g., "What is the capital of France?[SEP]Paris"), denoted as $\{q + a_i\}_{i=1}^n$.

$$P = f(\{q + a_i\}_{i=1}^n) = \{p_i\}_{i=1}^n,\ p_i \in [0, 1] \quad (2)$$

Specifically, when the auxiliary model processes the input $\{q + a_1, q + a_2, \ldots, q + a_n\}$, each pair of question and LLM response $(q + a_i)$ is treated

as an independent input sequence. This enables the model to solve a binary classification task—predicting whether a LLM's response is correct w.r.t the target answer. The auxiliary model outputs logits for each LLM response, which are then transformed as corresponding confidence scores $\{p_1, p_2, \ldots, p_n\}$. This approach allows the auxiliary model to evaluate each LLM's response individually while still capturing inter-model agreement through the aggregation of results.

Training objective. The goal is to optimize the predicted probability to align with the correctness of the input answer. Given k questions, we minimize the average **BCE loss** of each LLM answer:

$$\mathcal{L}_{BCE} = -\frac{1}{k \cdot n} \sum_{j=1}^{k} \sum_{i=1}^{n} \Big(c_i^{(j)} \log(p_i^{(j)}) + (1 - c_i^{(j)}) \log(1 - p_i^{(j)}) \Big) \quad (3)$$

Evaluation. A target LLM $\mathcal{M}_i$ achieves a low calibration Brier score if p_i accurately reflects the reliability of a_i, which is the correctness c_i. Specifically, for all k generated answers of $\mathcal{M}_i$ regarding k questions, the Brier score (Brier, 1950) of $\mathcal{M}_i$ average squared error between all predicted probabilities and the correctness of these answers:

$$Brier(\mathcal{M}_i) = \frac{1}{k} \sum_{j=1}^{k} (p_i^{(j)} - c_i^{(j)})^2 \quad (4)$$

Following Tian et al. (2023), we also evaluate all methods with three other metrics (details in Section 4.3).

3.4 Loss Functions

To further improve the calibration, we experiment with focal and AUC losses in addition to BCE loss.

Focal loss (FL) (Lin et al., 2017; Mukhoti et al., 2020) focuses on hard-to-classify examples, reducing the weight of correctly classified samples and encouraging the model to focus on predictions with high BCE loss. This loss is commonly used in imbalanced datasets but can benefit calibration since it emphasizes predictions with a large discrepancy between confidence and correctness. The FL is defined as follows:

$$\mathcal{L}_{\text{FL}} = -\frac{1}{k} \sum_{j=1}^{k} \Big(\alpha (1 - e^{-\mathcal{L}_{BCE_j}})^{\gamma} \cdot \mathcal{L}_{BCE_j} \Big) \quad (5)$$

Where $\mathcal{L}_{BCE_j}$ is the BCE loss of the data sample j. We use the default parameters from Lin et al. (2017) to set $\alpha = 0.25$ and $\gamma = 2.0$.

AUC Surrogate loss (AUC) (Yuan et al., 2021) uses a logistic loss to maximize the differences between true and false answers' scores (logits x generated by the classification head). The equation is:

$$\mathcal{L}_{\text{AUC}} = \frac{1}{|T| \cdot |F|} \sum_{t \in T} \sum_{f \in F} \sigma (x_f - x_t) \quad (6)$$

Where T and F are the indexes set for all true and false answers respectively. σ stands for sigmoid function.

In the end, we propose the following methods to integrate the techniques mentioned above and validate their effectiveness with comprehensive experiments across different models, datasets and prompts.

(BCE)/(FL)/(AUC)Calib-1: calibration training using the generations from one target LLM and optimizing with BCE/FL/AUC loss function.

(BCE)/(FL)/(AUC)Calib-n: calibration training using the generations from n target LLM and optimizing with BCE/FL/AUC loss function.

(BCE)/(FL)/(AUC)Calib-n+PS: Platt Scaling (PS) (Platt, 1999) (explained in Section 4.4) rescales the probabilities of test data by learning on the probabilities of validation data. Those probabilities are generated by corresponding Calib-n models.

4 Experiments

To comprehensively compare confidence estimation methods, we include diverse datasets, LLMs, and state-of-the-art baselines in our experimental setting.

4.1 Datasets

We cover four open-ended quenstion-answering datasets: TriviaQA (Joshi et al., 2017), Sciq (Welbl et al., 2017), WikiQA (Yang et al., 2015), and NQ (Kwiatkowski et al., 2019). We adopt the setting from Liu et al. (2024), where 2k/1k samples are selected as training/test data for TriviaQA, Sciq, and NQ, and 1040/293 for WikiQA.

4.2 Models

We include 12 models from five families: Llama (Grattafiori et al., 2024; Touvron et al., 2023), Phi (Abdin et al., 2024), Gemma (Team et al., 2024), Qwen (Yang et al., 2024), and Mixtral (Jiang et al., 2024).[2] In our analyses, we cluster

[2]All models are available at https://huggingface.co/models.

the models based on their number of parameters:

Small models (2-9B parameters): Llama-2-7b-chat-hf (referred as Llama2-7b), Llama-3.1-8B-Instruct (Llama3.1-8b), Llama-3-8B-Instruct (Llama3-8b), Phi-3-small-128k-instruct (Phi3-7b), Phi-3.5-mini-instruct (Phi3-4b), gemma-2-2b-it (Gemma2-2b), gemma-2-9b-it (Gemma2-9b).

Large models (27-72B parameters): Qwen2-72B-Instruct (Qwen2-72b), Llama-3-70B-Instruct (Llama3-72b), Llama-3.1-70B-Instruct (Llama3.1-70b), Mixtral-8x7B-Instruct-v0.1 (Mixtral-8x7b), gemma-2-27b-it (Gemma2-27b).

Calib-n models are trained with all LLMs inside the same group and provide confidence scores for each model in that group.

4.3 Evaluation Metrics

We report four metrics for calibration evaluation. **ECE:** The expected calibration error (Guo et al., 2017b) is computed by partitioning the predictions into 10 bins based on their confidence and then taking the weighted (by the number of samples in a bin) average of the squared difference between bin average accuracy and confidence. **ECE-t:** The temperature-scaled expected calibration error (Tian et al., 2023) finds a single temperature scaling parameter β that minimizes the negative log-likelihood between model confidences and answer correctness. Then, β is used to scale the confidences before the ECE is computed. **Brier:** The Brier Score (Brier, 1950) is the average squared error between predictions' confidence and correctness (see Eq. 4). **AUC:** The area under the receiver operating characteristic curve used in Ulmer et al. (2024).

Aggregate analysis. We aggregate the performance of different confidence estimation methods by counting their number of **wins**: given each metric above, we count the number of times a method outperforms the others in all possible combinations of prompt style, dataset, and model.

4.4 Baseline Methods

We compare Calib-* with standard baselines and the state-of-the-art confidence estimation methods:

LLM Probabilities (LLM Prob.): The conditional sequence probability $P_\theta(y|x)$ of an answer y given an input x, according to the model parametrized by θ.

LLM Prob. + Platt scaling (PS): This method applies Platt scaling (Platt, 1999) to the previous baseline. That is, two scalars $a, b \in \mathcal{R}$ are used to scale the original LLM probability p: $p_{\text{ps}} = \sigma(ap + b)$, where σ is the sigmoid function. We obtain parameters a and b by minimizing the mean-squared error between model confidences and answer correctness.

Verbalized confidences (Verbalized %) (Tian et al., 2023): The probability of correctness expressed in models' (text) responses given the Verb. prompts.

APRICOT (Ulmer et al., 2024): A recent method for calibrating LLMs. It consists of clustering related questions and measuring the per-cluster accuracies given answers from a target LLM. Then the cluster accuracies are used as the references to train an auxiliary transformer model that outputs confidence values for the target LLM.

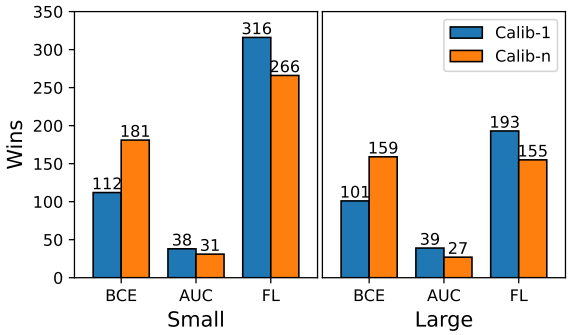


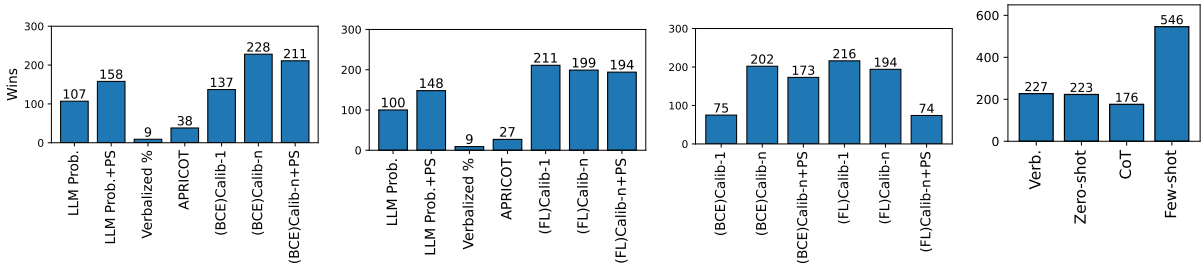
Figure 2: Comparison of Calib-1 and Calib-n methods based on the number of wins across different loss functions for calibrating small (left) and large (right) LLMs.

5 Results and Analysis

The detailed results (Table 1-9) indicate that no single method consistently outperforms all others across various models, datasets, and prompt configurations. Therefore, we first present the aggregated results, which summarize all the results, followed by a more detailed discussion.

5.1 Aggregated Result

What is the best loss function for improving calibration? In Fig. 2, we compare the performance of Calib-1 and Calib-n when optimizing with different loss functions. We observe that **FL loss wins in most settings**, followed by BCE loss. Additionally, we notice that BCE outperforms FL loss when applied in the Calib-n method for large models. This is because focal loss is designed for imbalanced datasets, while Calib-n which utilizes model responses from multiple models improves the bal-



(a) (BCE)Calib-* vs Baselines (b) (FL)Calib-* vs Baselines (c) (BCE) vs (FL)Calib-* (d) Prompts

Figure 3: The winning comparison results of different methods and prompts: 3a and 3b sub-figures present the superior results of Calib-* methods using BCE and FL loss respectively when against baselines. 3c shows the comparison results among all Calib-* methods, demonstrating that (FL)Calib-1 achieves the best overall performance. 3d compares the winning result of prompt styles and shows that few-shot prompting yields the most effective calibration across diverse configurations.

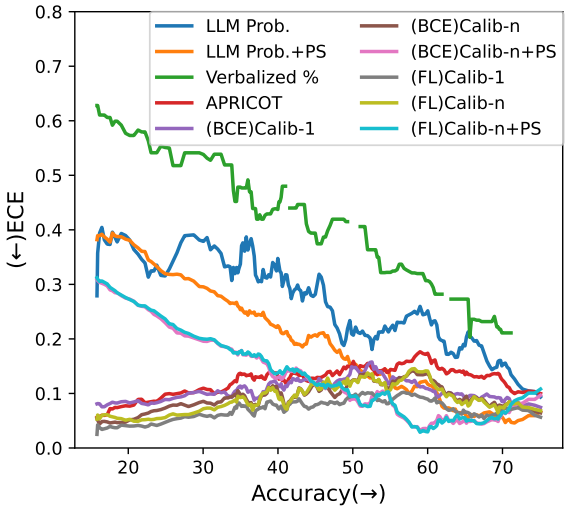


Figure 4: The correlation between accuracies achieved by different configurations (i.e., prompts, models, datasets) and corresponding ECE scores evaluated on different methods. The line of Verbalized % is not continuous because it can only be obtained using Verb. prompts and thus has fewer accuracy points than other methods. The result indicates that Calib-* and APRICOT are robust to accuracy variations. Different methods achieve the lowest ECE scores in different accuracy ranges.

ance of the training data (moderate the overall accuracy in this dataset) for the auxiliary model.

What is the best overall method? Fig. 3a and 3b present the winning comparison results of (BCE)Calib-* and (FL)Calib-* methods against baseline methods respectively. The results demonstrate that Calib-* methods gain more wins than baselines in both sub-figures. Verbalized confidences get the lowest number of wins. Applying Platt Scaling can further improve the calibration performance of LLM probabilities but this technique is not generalizable to enhance the calibration of Calib-n. Fig. 3a shows that (BCE)Calib-n gains the highest wins against the baseline methods. While (FL)Calib-1 accrues more wins in Fig. 3b. To identify the best method, we present Fig. 3c to compare the performance among our Calib-* methods, demonstrating that **(FL)Calib-1 exhibits the best overall performance**.

Which prompt style is most effective? The differences in prompt styles are well-known for their impact on the performance of LLMs. Fig. 3d showcases that prompt styles can also impact calibration performance. The results indicate that **few-shot is the most beneficial prompt** contributing to the highest wins among all other prompt styles.

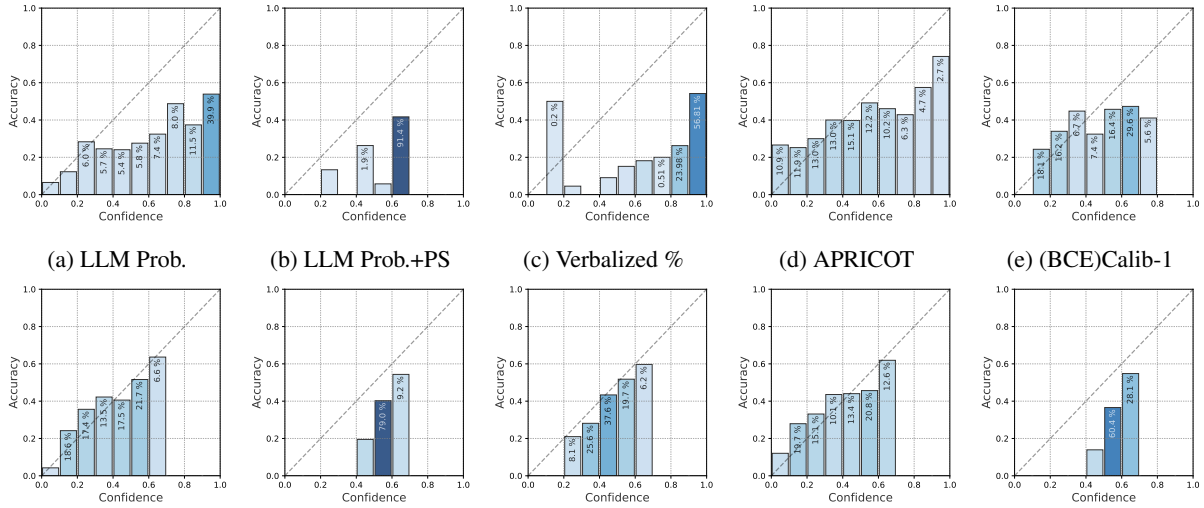
Which calibration methods maintain robustness to accuracy variations? Previous work (Zhang et al., 2024) reveals that confidence estimation methods like LLM probabilities and Verbalized confidence excessively concentrate on a fixed range (tend to be overconfident) and remain unchanged regardless of the dataset's complexity. To analyze this issue, we test all the confidence estimation methods with different datasets, prompts, and models. Each setting combination (e.g., using Zero-shot prompts to test the TriviaQA dataset on Gemma2-27b model) can result in one single accuracy value and multiple ECE scores—one for each confidence estimation method. We sort the accuracy values from all setting combinations and analyze the correlation between the accuracies and ECE scores. The results are presented in Fig. 4. We observe that ECE scores of LLM probabilities, LLM Prob.+PS, Verbalized confidences and (BCE)Calib-n+PS are highly correlated with accuracies, i.e., the ECE scores decrease when the

	Method	Verb.				Zero-shot				CoT				Few-shot			
		ECE↓	ECE-t↓	Brier ↓	AUC↑	ECE ↓	ECE-t↓	Brier ↓	AUC↑	ECE ↓	ECE-t↓	Brier ↓	AUC↑	ECE ↓	ECE-t↓	Brier ↓	AUC↑
Gemma2-27b	LLM Prob.	0.445	0.249	0.417	0.704	0.447	0.255	0.419	0.713	0.395	0.268	0.392	0.671	0.235	0.184	0.308	0.584
	LLM Prob.+PS	0.282	0.052	0.293	0.690	0.276	0.064	0.288	0.713	0.226	0.059	0.283	0.672	0.133	0.011	0.264	0.584
	Verbalized %	0.512	0.187	0.471	0.685	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.111	0.094	0.228	0.607	0.110	0.115	0.232	0.627	0.087	0.077	0.226	0.678	0.092	0.090	0.235	0.678
	(BCE)Calib-1	0.066	0.068	0.220	0.596	0.064	0.066	0.223	0.590	0.046	0.045	0.224	0.666	0.146	0.147	0.232	0.717
	(BCE)Calib-n	0.069	0.066	0.212	0.648	0.056	0.041	0.216	0.651	0.095	0.081	0.226	0.695	0.088	0.067	0.220	0.713
	(BCE)Calib-n+PS	0.202	0.062	0.255	0.637	0.198	0.046	0.257	0.651	0.170	0.076	0.255	0.695	0.103	0.122	0.241	0.713
	(FL)Calib-1	0.038	0.038	0.215	0.597	0.056	0.058	0.219	0.607	0.051	0.049	0.228	0.650	0.123	0.123	0.227	0.717
	(FL)Calib-n	0.083	0.083	0.217	0.642	0.070	0.057	0.219	0.645	0.084	0.083	0.226	0.692	0.085	0.080	0.220	0.710
	(FL)Calib-n+PS	0.203	0.057	0.254	0.642	0.191	0.056	0.255	0.645	0.167	0.079	0.254	0.692	0.101	0.120	0.241	0.710
Llama3-70b	LLM Prob.	0.445	0.302	0.430	0.712	0.458	0.259	0.426	0.747	0.412	0.251	0.416	0.612	0.336	0.117	0.364	0.532
	LLM Prob.+PS	0.355	0.113	0.350	0.703	0.266	0.132	0.291	0.747	0.224	0.023	0.289	0.612	0.224	0.009	0.298	0.532
	Verbalized %	0.406	0.084	0.380	0.698	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.137	0.130	0.255	0.596	0.111	0.116	0.239	0.635	0.114	0.117	0.244	0.656	0.106	0.068	0.236	0.683
	(BCE)Calib-1	0.075	0.071	0.231	0.610	0.066	0.067	0.225	0.641	0.066	0.045	0.232	0.662	0.119	0.121	0.232	0.709
	(BCE)Calib-n	0.058	0.062	0.223	0.644	0.040	0.025	0.215	0.670	0.096	0.088	0.229	0.690	0.061	0.062	0.213	0.730
	(BCE)Calib-n+PS	0.171	0.038	0.255	0.632	0.179	0.050	0.256	0.670	0.163	0.081	0.256	0.690	0.134	0.107	0.238	0.730
	(FL)Calib-1	0.079	0.065	0.228	0.625	0.080	0.080	0.229	0.630	0.057	0.037	0.234	0.647	0.105	0.111	0.225	0.713
	(FL)Calib-n	0.080	0.093	0.228	0.638	0.040	0.022	0.215	0.675	0.096	0.095	0.231	0.686	0.066	0.058	0.216	0.725
	(FL)Calib-n+PS	0.182	0.046	0.259	0.638	0.176	0.060	0.254	0.675	0.168	0.070	0.258	0.686	0.134	0.110	0.240	0.725
Llama3.1-70b	LLM Prob.	0.301	0.231	0.326	0.704	0.308	0.237	0.326	0.722	0.274	0.221	0.330	0.619	0.172	0.098	0.266	0.640
	LLM Prob.+PS	0.254	0.116	0.293	0.690	0.236	0.066	0.279	0.722	0.195	0.023	0.275	0.619	0.135	0.064	0.258	0.640
	Verbalized %	0.435	0.122	0.405	0.711	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.120	0.115	0.248	0.621	0.080	0.078	0.239	0.635	0.083	0.080	0.230	0.688	0.097	0.075	0.234	0.687
	(BCE)Calib-1	0.130	0.123	0.249	0.602	0.089	0.088	0.244	0.580	0.073	0.058	0.211	0.733	0.105	0.102	0.221	0.726
	(BCE)Calib-n	0.062	0.055	0.224	0.656	0.058	0.047	0.230	0.637	0.049	0.054	0.211	0.730	0.081	0.080	0.217	0.719
	(BCE)Calib-n+PS	0.157	0.015	0.258	0.606	0.143	0.078	0.254	0.637	0.138	0.121	0.246	0.730	0.084	0.099	0.237	0.719
	(FL)Calib-1	0.041	0.034	0.223	0.660	0.047	0.046	0.236	0.596	0.060	0.050	0.206	0.742	0.074	0.075	0.216	0.724
	(FL)Calib-n	0.077	0.079	0.227	0.657	0.055	0.049	0.229	0.643	0.055	0.053	0.210	0.732	0.068	0.072	0.217	0.714
	(FL)Calib-n+PS	0.169	0.051	0.257	0.657	0.147	0.060	0.255	0.643	0.142	0.117	0.248	0.732	0.092	0.104	0.238	0.714
Qwen2-72b	LLM Prob.	0.470	0.321	0.455	0.692	0.483	0.280	0.450	0.734	0.380	0.276	0.393	0.620	0.220	0.092	0.280	0.645
	LLM Prob.+PS	0.259	0.120	0.289	0.745	0.270	0.057	0.292	0.734	0.202	0.021	0.282	0.620	0.203	0.085	0.281	0.645
	Verbalized %	0.433	0.065	0.404	0.734	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.114	0.113	0.241	0.614	0.096	0.085	0.227	0.647	0.113	0.112	0.247	0.641	0.086	0.078	0.230	0.694
	(BCE)Calib-1	0.038	0.039	0.230	0.586	0.027	0.029	0.217	0.638	0.092	0.092	0.234	0.668	0.115	0.117	0.229	0.714
	(BCE)Calib-n	0.063	0.064	0.225	0.635	0.046	0.020	0.215	0.664	0.083	0.084	0.231	0.682	0.076	0.078	0.222	0.712
	(BCE)Calib-n+PS	0.166	0.039	0.255	0.625	0.185	0.047	0.256	0.664	0.137	0.084	0.252	0.682	0.117	0.091	0.244	0.712
	(FL)Calib-1	0.051	0.050	0.231	0.595	0.031	0.020	0.219	0.627	0.077	0.084	0.230	0.671	0.089	0.087	0.223	0.712
	(FL)Calib-n	0.095	0.095	0.233	0.622	0.051	0.040	0.217	0.664	0.083	0.086	0.230	0.683	0.071	0.074	0.221	0.712
	(FL)Calib-n+PS	0.183	0.032	0.261	0.622	0.188	0.048	0.257	0.664	0.135	0.083	0.251	0.683	0.116	0.090	0.244	0.712
Mixtral-8x7b	LLM Prob.	0.513	0.269	0.479	0.703	0.412	0.155	0.401	0.637	0.452	0.270	0.447	0.627	0.369	0.075	0.380	0.574
	LLM Prob.+PS	0.264	0.015	0.292	0.653	0.233	0.066	0.290	0.637	0.245	0.013	0.305	0.627	0.141	0.001	0.267	0.574
	Verbalized %	0.472	0.167	0.447	0.655	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.124	0.120	0.247	0.602	0.092	0.088	0.239	0.646	0.115	0.113	0.252	0.620	0.161	0.070	0.263	0.640
	(BCE)Calib-1	0.040	0.031	0.224	0.605	0.037	0.037	0.225	0.658	0.082	0.083	0.249	0.607	0.145	0.097	0.269	0.618
	(BCE)Calib-n	0.082	0.061	0.227	0.619	0.032	0.021	0.229	0.634	0.064	0.068	0.241	0.631	0.085	0.085	0.230	0.686
	(BCE)Calib-n+PS	0.188	0.022	0.260	0.610	0.143	0.070	0.255	0.634	0.116	0.060	0.254	0.631	0.101	0.078	0.244	0.686
	(FL)Calib-1	0.031	0.025	0.226	0.580	0.033	0.035	0.224	0.662	0.077	0.046	0.247	0.598	0.121	0.098	0.262	0.619
	(FL)Calib-n	0.093	0.083	0.231	0.618	0.024	0.025	0.227	0.643	0.078	0.076	0.242	0.631	0.072	0.077	0.230	0.682
	(FL)Calib-n+PS	0.183	0.034	0.256	0.618	0.147	0.074	0.255	0.643	0.129	0.054	0.257	0.631	0.104	0.068	0.244	0.682

Table 1: Test performance of our methods (Calib-*) and baseline methods on NQ dataset using four different prompts (Verb., Zero-shot, CoT, Few-shot). Calib-n is trained with the responses of all LLMs in the table. Only the Verb. prompt requests the LLM to provide a probability for a given answer and thus has the results for Verbalized % performance. We color the text with a scale normalized by the values gap in each column, with darker shades indicating better performance. The results of the other three datasets and seven models are shown in Appendix A.8.

accuracies get higher which verified the findings of Zhang et al. (2024). **APRICOT, Calib-1 and Calib-n are robust to the accuracy changes**, the ECE scores of these methods remain relatively stable across different accuracies. Verbalized confidence consistently shows the lowest performance across all accuracy levels.

What is the best method for different accuracy levels? The optimal goal of current state-of-the-art confidence estimation methods should not only focus on achieving a low calibration error at a narrow accuracy range but also analyze the performance of the method across different accuracy levels. We perform this analysis in Fig. 4. We observe that (FL)Calib-1 performs the best in low accuracy ranges up to 50%, and Calib-n+PS achieves the lowest ECE scores for accuracies between 50% and 70%. LLM Prob.+PS works the best for high accuracy (>70%) settings mainly because LLM probabilities are usually overconfident.



(a) LLM Prob. (b) LLM Prob.+PS (c) Verbalized % (d) APRICOT (e) (BCE)Calib-1

(f) (BCE)Calib-n (g) (BCE)Calib-n+PS (h) (FL)Calib-1 (i) (FL)Calib-n (j) (FL)Calib-n+PS

Figure 5: Reliability diagrams for our different methods using 10 bins each for Llama3.1-70b on NQ. The color and the percentage number within each bar present the proportion of total data samples contained in each bin. More figures of other models and datasets are shown in Appendix A.6.

5.2 Detailed Fraction Results

Table 1 presents the test results of our methods and baselines on the NQ dataset when evaluated on five large-size LLMs. We analyze the performance of all methods across prompts, models, and evaluation metrics. The results reveal that the effectiveness of methods is highly influenced by the prompt styles and models. However, our proposed methods (Calib-*) achieve better calibration than baselines on most prompts and models.

Across Methods: LLM probabilities and Verbalized confidence show poor performance, exhibiting high ECE, ECE-t, and Brier scores across all prompts. Adding Platt Scaling (LLM Prob.+PS) improves calibration but is still outperformed by our proposed Calib-* methods. Among baselines, APRICOT performs better than Verbalized confidence, although it does not achieve the best results in most settings. Across all evaluation metrics, our Calib-* methods outperform others. For example, (BCE)Calib-n achieves the lowest ECE and Brier scores, particularly with the CoT and few-shot prompts, e.g., ECE of 0.049 on Llama3.1-70b. PS only helps decrease the ECE-t scores but can not enhance overall performance, this is because PS is similar to ECE-t which is a posthoc method for scaling probabilities. **Across Prompts**: Calibration quality improves with few-shot and CoT prompts compared to the verb. and zero-shot prompts. For instance, (BCE)Calib-1 achieves an ECE-t of 0.045 with the CoT prompts compared to 0.068 with the verb. prompts on Gemma2-27b. **Across Models**: Larger models (e.g., Qwen2-72b and Llama3.1-70b) generally exhibit better calibration when using Calib-* methods, reflecting their better alignment with calibration techniques. For instance, (FL)Calib-n+PS achieves an ECE of 0.071 on Qwen2-72b compared to 0.095 on smaller models like Mixtral-8x7b.

Reliability diagrams analysis. Fig. 5 shows that PS produces a narrow range of confidence scores, indicating limited diversity in the emitted confidence levels. LLM probabilities and Verbalized confidence often exhibit overconfidence, even when applied to datasets with low accuracy (<40%, reported in Fig. 8 in Appendix). In contrast, our Calib-* methods show a more conservative approach, aligning their confidence levels more closely with the true accuracy of the model, reflecting improved calibration and reliability.

6 Conclusion

Previous studies (Ulmer et al., 2024; Tian et al., 2023) have assessed calibration methods within limited settings, overlooking their generalization across diverse model sizes and prompt styles. This study addresses this limitation by conducting experiments on 12 LLMs with parameters ranging from 2B-72B and four prompt styles. We also comprehensively analyze the influence of response agreement and loss functions in LLM calibration. Experimental results show that both response agree-

ment and FL loss enhance calibration from baselines, with (FL)Calib-1 achieving the best performance. We also find that few-shot prompts improve LLM accuracy and calibration, and auxiliary-based methods show robust performance across diverse settings—maintaining stable calibration regardless of accuracy levels. These findings highlight the effectiveness of various calibration strategies and encourage future methods to re-evaluate the importance of our explored factors for achieving reliable confidence estimation.

Limitations

Although our work covers a wide range of factors, there are potentially more factors worth exploring. For example, we analyze a wide range of prompt types, including Few-shot and Chain-of-Thought, the influence of fine-grained prompt variations or automatically generated prompts remains unexplored. The interplay between prompt engineering and calibration could warrant deeper investigation.

Our study focuses on calibration performance metrics like ECE, ECE-t, and Brier scores. While these metrics are widely used, they may not fully capture all aspects of calibration quality, such as user-perceived confidence or task-specific utility.

Existing works use various ways of determining the accuracy of LLM generations. For instance, some works (Tian et al., 2023; Liu et al., 2024) use LLMs as a Judge, other works (Ulmer et al., 2024; Xiong et al., 2024) use certain metrics such as extract match or ROUGE score. While the optimal solution is underexplored, we choose the more commonly used and cost-efficient method.

Future work can address these limitations by testing broader LLMs and tasks, automating prompt optimization, and developing hybrid approaches that adapt to varying accuracy levels and application constraints.

Acknowledgments

This research has been funded by the Vienna Science and Technology Fund (WWTF)[10.47379/VRG19008] "Knowledge infused Deep Learning for Natural Language Processing", and co-funded by the European Union.

References

Marah Abdin et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *ArXiv preprint*, abs/2404.14219.

Glenn W. Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1 – 3.

Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2023. Unleashing the potential of prompt engineering in large language models: a comprehensive review.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Aaron Grattafiori et al. 2024. The llama 3 herd of models. *ArXiv preprint*, abs/2407.21783.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017a. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017b. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

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

Jaeyoung Kim, Dongbin Na, Sungchul Choi, and Sungbin Lim. 2023. Bag of tricks for in-distribution calibration of pretrained transformers. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 551–563, Dubrovnik, Croatia. Association for Computational Linguistics.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models.

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

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2999–3007. IEEE Computer Society.

Xin Liu, Muhammad Khalifa, and Lu Wang. 2024. Litcab: Lightweight language model calibration over short- and long-form responses. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jooyoung Moon, Jihyo Kim, Younghak Shin, and Sangheum Hwang. 2020. Confidence-aware learning for deep neural networks. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 7034–7044. PMLR.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip H. S. Torr, and Puneet K. Dokania. 2020. Calibrating deep neural networks using focal loss. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Shiyu Ni, Keping Bi, Lulu Yu, and Jiafeng Guo. 2024. Are large language models more honest in their probabilistic or verbalized confidence?

John Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74.

Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 33–44.

Gemma Team et al. 2024. Gemma 2: Improving open language models at a practical size. *ArXiv preprint*, abs/2408.00118.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu,

Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdoo Yun, and Seong Oh. 2024. Calibrating large language models using their generations only. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15440–15459, Bangkok, Thailand. Association for Computational Linguistics.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Label words are anchors: An information flow perspective for understanding in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9840–9855, Singapore. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

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

Yuxi Xia, Kilm Zaporojets, and Benjamin Roth. 2024. Black-box model ensembling for textual and visual question answering via information fusion.

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

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

Zhuoning Yuan, Yan Yan, Milan Sonka, and Tianbao Yang. 2021. Large-scale robust deep AUC maximization: A new surrogate loss and empirical studies on medical image classification. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 3020–3029. IEEE.

Min Zhang, Jianfeng He, Taoran Ji, and Chang-Tien Lu. 2024. Don't go to extremes: Revealing the excessive sensitivity and calibration limitations of LLMs in implicit hate speech detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12073–12086, Bangkok, Thailand. Association for Computational Linguistics.

A Appendix

A.1 Prompts for target LLM

A.1.1 Verb. prompt

> Provide your best guess and the probability that it is correct (0.0 to 1.0) for the following question. Give ONLY the guess and probability, no other words or explanation. For example: \n\n Guess: <most likely guess, as short as possible; not a complete sentence, just the guess!>\n "Probability: <the probability between 0.0 and 1.0 that your guess is correct, without any extra commentary whatsoever; just the probability!>\n\n The question is: [Question]

A.1.2 Zero-shot prompt

> Provide your best guess for the following question. Give ONLY the guess, no other words or explanations, as short as possible; not a complete sentence, just the guess! \n\n The question is: [Question]

A.1.3 CoT prompt

> Briefly answer the following question by thinking step by step. Give the final answer (start with 'Answer: ') with minimal words at the end. \n\n The question is: [Question]

A.1.4 Few-shot prompt

> user: [Question 1] assistant: [Target 1]
> user: [Question 2] assistant: [Target 2]
> ...
> user: [Question 6] assistant: [Target 6]
> user: [Question] assistant:

A.2 Prompt for Judge

> Task Description: \n An instruction (might include an Input inside it), a response to evaluate, a reference answer that gets a score of 1, and a score rubric representing a evaluation criteria are given.
> 1. Write detailed feedback that assesses the quality of the response strictly based on the given score rubric, not evaluating in general.
> 2. After writing feedback, write a score that is an integer between 0 and 1. You should refer to the score rubric.
> 3. The output format should look as follows: "Feedback: (write a feedback for criteria) [RESULT] (an integer number between 0 and 1)"
> 4. Please do not generate any other opening, closing, and explanations.
> The instruction to evaluate:[Question]
> Response to evaluate: [LLM Answer]
> Reference Answer (Score 1): [Target]
> Score Rubrics:\n Score 0: the response and reference answer to the instruction are not semantically equivalent.\n Score 1: the response and reference answer to the instruction are semantically equivalent.
> Feedback:

A.3 Technical Details

After a grid search of hyperparameters, we trained our auxiliary models (BERT-base, 110M parameters) using a learning rate of 1e-5 and a batch size of 16 for five epochs. All experiments, including LLM inferences, are performed on a maximum of 2 NVIDIA H100 GPUs. The training time for one epoch of 2k samples is around 200 seconds on one GPU, this time can be different depends on the dataset size and number of joint LLMs in training.

A.4 Aggregated Results Across Different Configurations

Fig. 6 shows the calibration performances of different methods across different configurations such as prompt styles, model sizes and datasets. We observe that the best method is highly dependent on these factors.

Prompt specific performance. The first row of Fig. 6 presents the prompt specific performance for different methods. We observe that Calib-n always

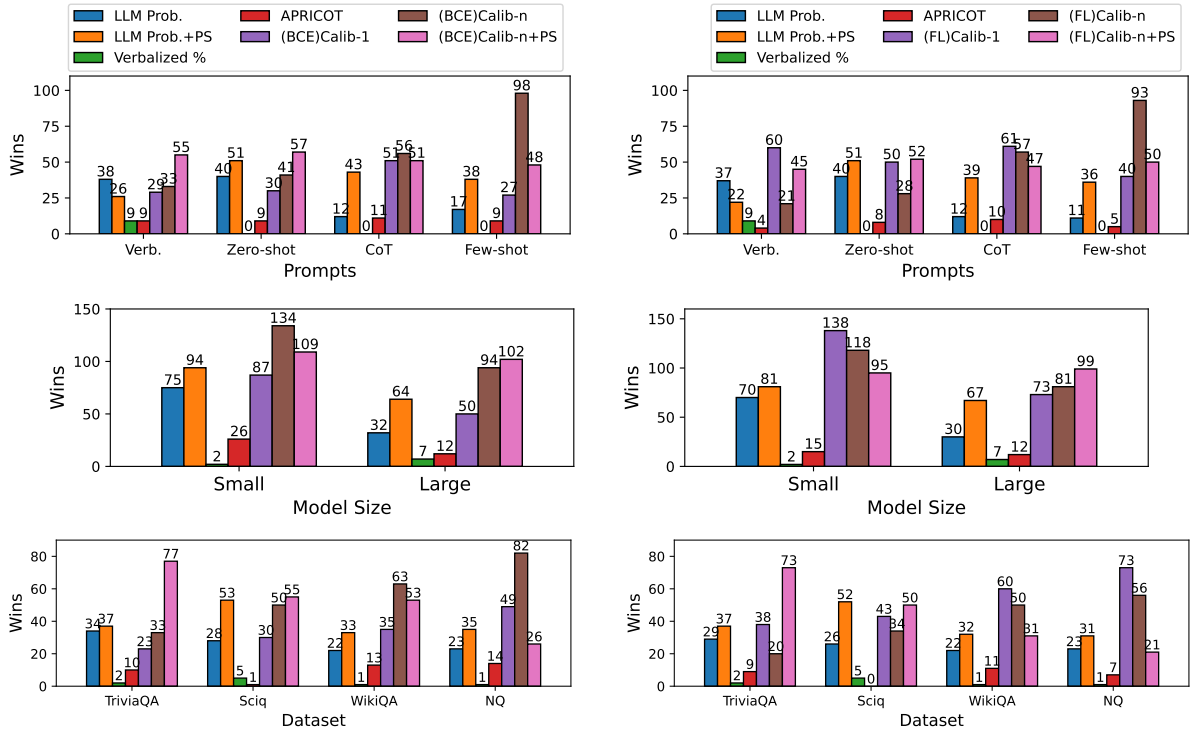


Figure 6: Performance Comparison results of different methods against baselines across three setting configurations (prompt styles, model sizes and datasets).

outperforms Calib-1 across different prompt styles when both are optimized with BCE loss. When FL loss is applied, Calib-n only outperforms Calib-1 in few-shot prompts. We hypothesize that the longer LLM answers generated with few-shot prompts usually lead to higher response agreement and thus enhance the performance of Calib-n.

Model size specific performance. The second row of Fig. 6 presents the model-size specific performance for different methods. We find that Calib-n outperforms Calib-1 on large-size models. However, (FL)Calib-1 performs better for small-size models. We also observe that the best method is model size dependent.

Dataset specific performance. The second row of Fig. 6 presents the dataset specific performance for different methods. (BCE)Calib-n consistently achieves better performances than (BCE)Calib-1 over all datasets. In contrast, (FL)Calib-1 always outperform (FL)Calib-n.

A.5 Accuracy Statistics of LLMs

We present Fig. 7 and 8 to demonstrate the accuracy performance of LLMs across different prompts and datasets. Large size models typically yield better performance than small size models. Few shot prompts improve the performance more than other prompt styles. Most LLMs achieve their highest accuracy on the Sciq dataset, while the NQ dataset proves to be the most challenging.

A.6 Reliability Diagrams

Similar to Fig. 5, Fig. 9-11 show the reliability diagrams of other three datasets for Llama3.1-70 with Verb. prompts. We find that for high accuracy (>50%) datasets (TriviaQA and Sciq), Calib-* using BCE and FL loss is less conservative and more likely to predict high confidence.

A.7 Out-domain Generalization

Table 2 presents the generalization results for the best two settings in our paper ((FL)Calib-1 and (BCE)Calib-n). We observe performance drops in the out-of-distribution (OOD) setting, but no catastrophic degradation. In some settings, the OOD auxiliary models perform on par with or better than the in-domain models.

A.8 Detailed Results of LLM Calibration

Similar to Table 1, Table 3 -9 show the rest of the detailed calibration results of different methods.

Method	Train	Test	ECE↓	ECE-t↓	Brier↓	AUC↑
(FL)Calib-1	TriviaQA	TriviaQA	0.042	0.053	0.181	0.708
(FL)Calib-1	WikiQA+NQ+Sciq	TriviaQA	0.097	0.090	**0.171**	**0.733**
(BCE)Calib-n	TriviaQA	TriviaQA	0.068	0.065	0.178	0.715
(BCE)Calib-n	WikiQA+NQ+Sciq	TriviaQA	0.089	0.074	**0.173**	0.707
(FL)Calib-1	WikiQA	WikiQA	0.073	0.068	0.193	0.574
(FL)Calib-1	TriviaQA+NQ+Sciq	WikiQA	0.250	0.100	0.280	**0.628**
(BCE)Calib-n	WikiQA	WikiQA	0.085	0.033	0.194	0.576
(BCE)Calib-n	TriviaQA+NQ+Sciq	WikiQA	0.311	0.084	0.317	**0.616**
(FL)Calib-1	NQ	NQ	0.074	0.075	0.216	0.724
(FL)Calib-1	TriviaQA+Sciq+WikiQA	NQ	0.149	0.118	0.249	0.682
(BCE)Calib-n	NQ	NQ	0.081	0.080	0.217	0.719
(BCE)Calib-n	TriviaQA+Sciq+WikiQA	NQ	0.119	0.101	0.239	0.686
(FL)Calib-1	Sciq	Sciq	0.023	0.017	0.144	0.738
(FL)Calib-1	TriviaQA+WikiQA+NQ	Sciq	0.057	0.053	0.152	0.707
(BCE)Calib-n	Sciq	Sciq	0.038	0.042	0.146	0.745
(BCE)Calib-n	TriviaQA+WikiQA+NQ	Sciq	**0.018**	**0.021**	0.147	0.696

Table 2: Generalization results test on Llama3.1-70b model using few-shot prompt. We bold the results when the out-domain outperforms in-domain performance.

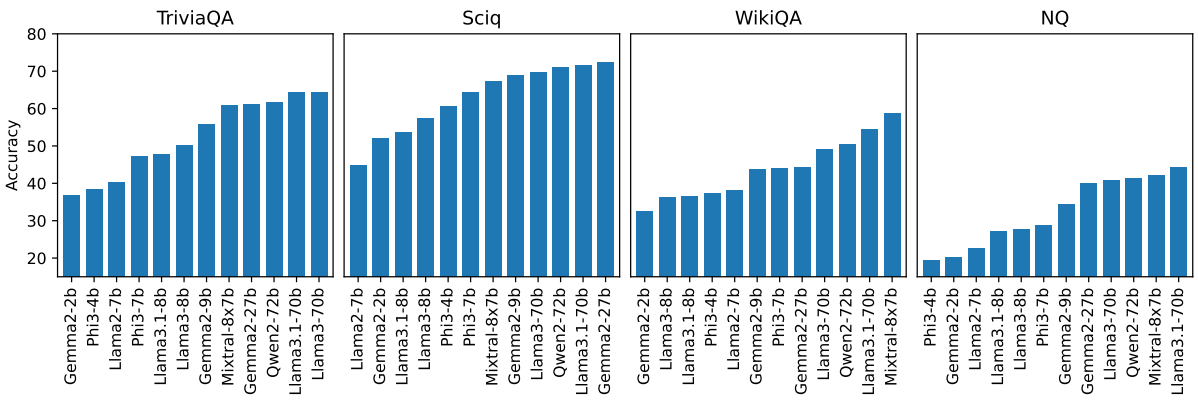


Figure 7: Model performance ranking across different datasets, performance is averaged over four prompt styles.

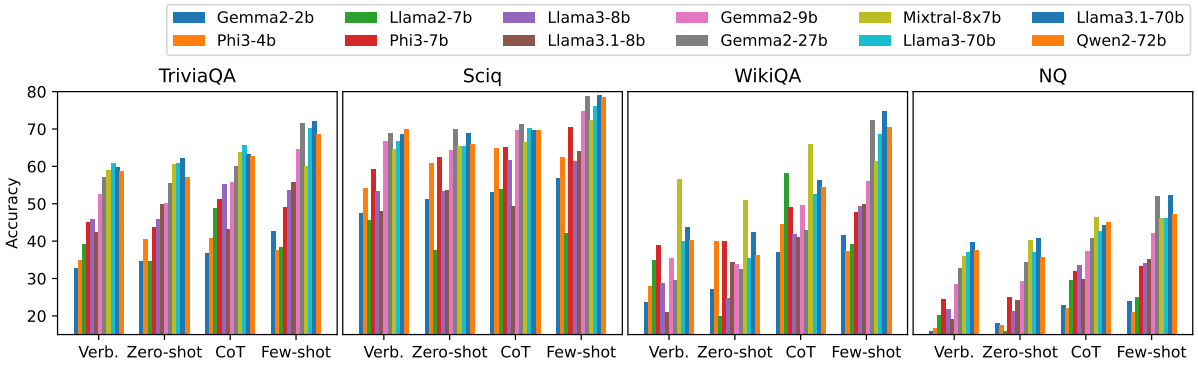
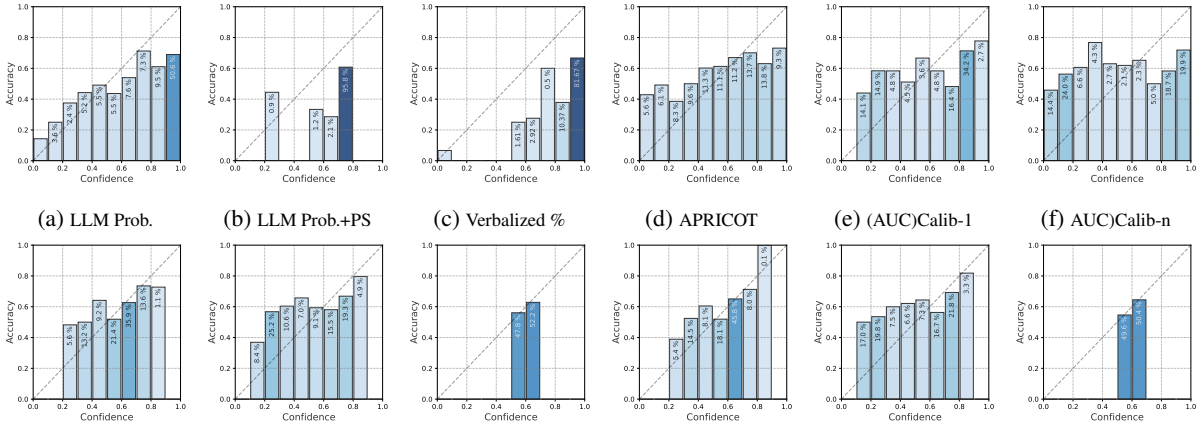


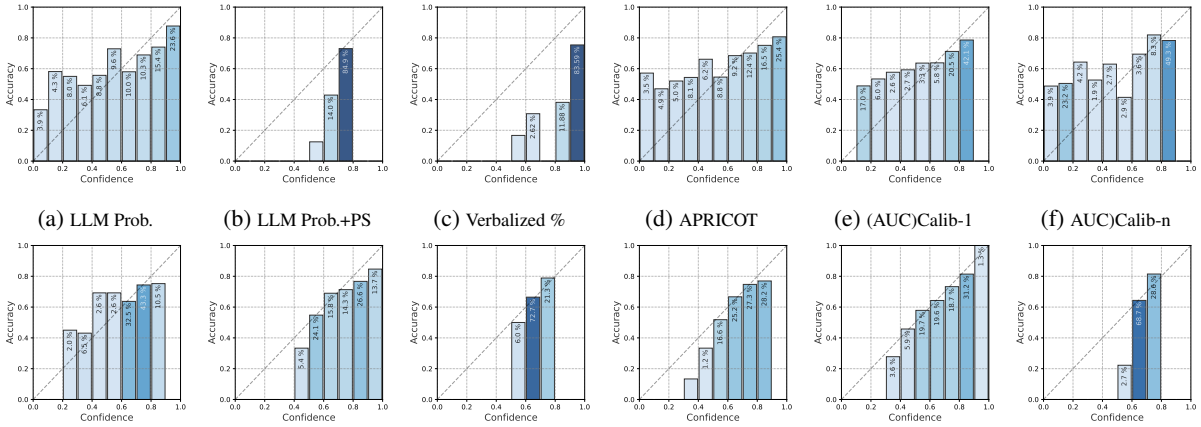
Figure 8: Model performance across different prompts and datasets.



(a) LLM Prob. (b) LLM Prob.+PS (c) Verbalized % (d) APRICOT (e) (AUC)Calib-1 (f) AUC)Calib-n

(g) (BCE)Calib-1 (h) (BCE)Calib-n (i) (BCE)Calib-n+PS (j) (FL)Calib-1 (k) (FL)Calib-n (l) (FL)Calib-n+PS

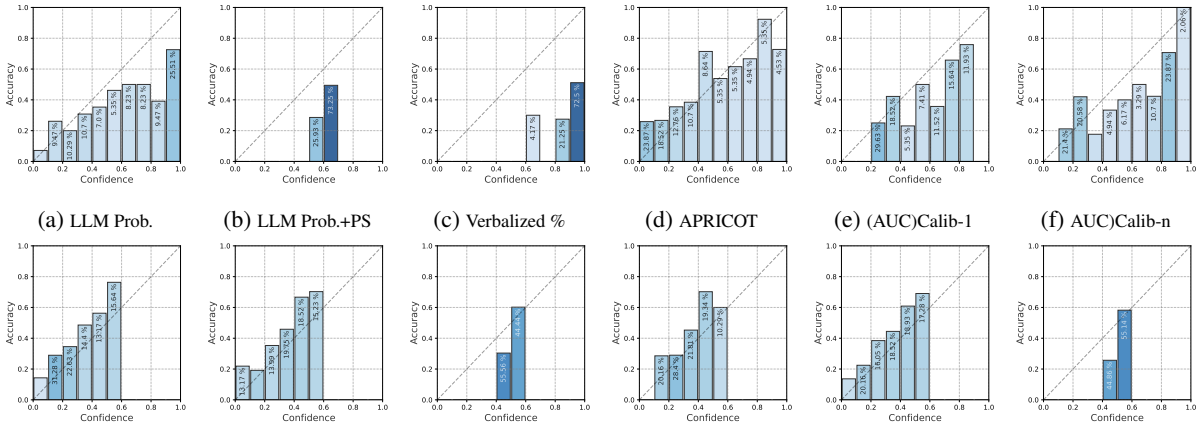
Figure 9: Reliability diagrams for Llama3.1-70b on TriviaQA with Verb. prompts.



(a) LLM Prob. (b) LLM Prob.+PS (c) Verbalized % (d) APRICOT (e) (AUC)Calib-1 (f) AUC)Calib-n

(g) (BCE)Calib-1 (h) (BCE)Calib-n (i) (BCE)Calib-n+PS (j) (FL)Calib-1 (k) (FL)Calib-n (l) (FL)Calib-n+PS

Figure 10: Reliability diagrams for Llama3.1-70b on Sciq with Verb. prompts.



(a) LLM Prob. (b) LLM Prob.+PS (c) Verbalized % (d) APRICOT (e) (AUC)Calib-1 (f) AUC)Calib-n

(g) (BCE)Calib-1 (h) (BCE)Calib-n (i) (BCE)Calib-n+PS (j) (FL)Calib-1 (k) (FL)Calib-n (l) (FL)Calib-n+PS

Figure 11: Reliability diagrams for Llama3.1-70b on WikiQA with Verb. prompts.

	Method	Verb. ECE↓	Verb. ECE-t↓	Verb. Brier ↓	Verb. AUC↑	Zero-shot ECE ↓	Zero-shot ECE-t↓	Zero-shot Brier ↓	Zero-shot AUC↑	CoT ECE ↓	CoT ECE-t↓	CoT Brier ↓	CoT AUC↑	Few-shot ECE ↓	Few-shot ECE-t↓	Few-shot Brier ↓	Few-shot AUC↑
Llama2-7b	LLM Prob.	0.437	0.239	0.401	0.737	0.296	0.170	0.291	0.746	0.314	0.180	0.324	0.693	0.419	0.042	0.400	0.643
	LLM Prob.+PS	0.256	0.065	0.299	0.613	0.246	0.092	0.266	0.746	0.165	0.113	0.261	0.708	0.249	0.054	0.295	0.643
	Verblized %	0.395	0.049	0.384	0.618	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.140	0.087	0.236	0.685	0.146	0.109	0.232	0.658	0.141	0.122	0.253	0.673	0.118	0.073	0.232	0.667
	(AUC)Calib-1	0.179	0.180	0.267	0.636	0.163	0.165	0.257	0.620	0.239	0.238	0.302	0.599	0.328	0.201	0.326	0.693
	(AUC)Calib-n	0.259	0.250	0.308	0.610	0.256	0.256	0.298	0.635	0.226	0.222	0.298	0.634	0.251	0.202	0.286	0.717
	(BCE)Calib-1	0.092	0.036	0.227	0.664	0.117	0.045	0.234	0.612	0.089	0.080	0.252	0.617	0.116	0.093	0.228	0.679
	(BCE)Calib-n	0.124	0.091	0.240	0.637	0.102	0.085	0.225	0.632	0.098	0.076	0.245	0.654	0.093	0.081	0.216	0.705
	(BCE)Calib-n+PS	0.134	0.032	0.246	0.629	0.178	0.037	0.249	0.632	0.099	0.056	0.246	0.654	0.162	0.065	0.242	0.705
	(FL)Calib-1	0.073	0.041	0.227	0.652	0.084	0.051	0.223	0.606	0.091	0.086	0.254	0.602	0.086	0.078	0.221	0.683
	(FL)Calib-n	0.121	0.093	0.243	0.617	0.098	0.079	0.226	0.620	0.093	0.065	0.241	0.662	0.098	0.080	0.216	0.702
	(FL)Calib-n+PS	0.138	0.034	0.249	0.617	0.156	0.030	0.243	0.620	0.105	0.068	0.245	0.662	0.164	0.067	0.241	0.702
Llama3-8b	LLM Prob.	0.274	0.165	0.277	0.751	0.336	0.222	0.325	0.739	0.246	0.181	0.296	0.638	0.251	0.048	0.286	0.691
	LLM Prob.+PS	0.192	0.037	0.279	0.623	0.188	0.129	0.264	0.739	0.130	0.058	0.254	0.638	0.160	0.073	0.264	0.691
	Verblized %	0.385	0.089	0.381	0.642	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.172	0.089	0.259	0.663	0.152	0.140	0.257	0.661	0.169	0.127	0.263	0.657	0.131	0.073	0.240	0.701
	(AUC)Calib-1	0.248	0.235	0.302	0.636	0.277	0.254	0.316	0.609	0.250	0.229	0.308	0.591	0.198	0.153	0.251	0.727
	(AUC)Calib-n	0.239	0.176	0.288	0.665	0.248	0.206	0.304	0.629	0.220	0.223	0.289	0.635	0.257	0.208	0.275	0.757
	(BCE)Calib-1	0.123	0.098	0.252	0.637	0.124	0.086	0.256	0.626	0.127	0.083	0.264	0.606	0.095	0.068	0.217	0.738
	(BCE)Calib-n	0.168	0.070	0.256	0.676	0.148	0.073	0.261	0.649	0.140	0.079	0.260	0.647	0.060	0.049	0.209	0.741
	(BCE)Calib-n+PS	0.100	0.051	0.243	0.635	0.088	0.052	0.244	0.649	0.068	0.048	0.237	0.647	0.100	0.111	0.233	0.741
	(FL)Calib-1	0.090	0.076	0.238	0.660	0.112	0.098	0.252	0.626	0.102	0.081	0.258	0.603	0.075	0.047	0.212	0.740
	(FL)Calib-n	0.165	0.070	0.268	0.618	0.149	0.075	0.260	0.640	0.135	0.058	0.255	0.655	0.052	0.049	0.208	0.743
	(FL)Calib-n+PS	0.094	0.034	0.245	0.618	0.080	0.048	0.243	0.640	0.071	0.054	0.237	0.655	0.108	0.109	0.232	0.743
Llama3.1-8b	LLM Prob.	0.255	0.129	0.256	0.786	0.167	0.127	0.234	0.750	0.298	0.154	0.315	0.685	0.107	0.037	0.204	0.774
	LLM Prob.+PS	0.223	0.050	0.286	0.610	0.169	0.140	0.237	0.810	0.209	0.053	0.277	0.685	0.158	0.117	0.238	0.775
	Verblized %	0.440	0.175	0.429	0.604	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.154	0.122	0.245	0.661	0.126	0.066	0.226	0.725	0.115	0.061	0.207	0.771	0.119	0.089	0.232	0.703
	(AUC)Calib-1	0.251	0.251	0.292	0.687	0.256	0.260	0.291	0.699	0.236	0.166	0.245	0.789	0.126	0.137	0.201	0.778
	(AUC)Calib-n	0.203	0.189	0.258	0.705	0.215	0.186	0.268	0.706	0.190	0.165	0.240	0.733	0.255	0.203	0.255	0.780
	(BCE)Calib-1	0.107	0.083	0.220	0.692	0.121	0.134	0.237	0.703	0.083	0.057	0.178	0.806	0.062	0.067	0.187	0.777
	(BCE)Calib-n	0.126	0.069	0.226	0.704	0.127	0.063	0.232	0.723	0.074	0.042	0.197	0.779	0.041	0.038	0.185	0.783
	(BCE)Calib-n+PS	0.146	0.060	0.244	0.643	0.119	0.122	0.238	0.723	0.176	0.140	0.234	0.779	0.135	0.155	0.224	0.783
	(FL)Calib-1	0.096	0.089	0.217	0.686	0.088	0.112	0.225	0.708	0.065	0.047	0.180	0.796	0.055	0.052	0.189	0.772
	(FL)Calib-n	0.117	0.070	0.238	0.640	0.122	0.077	0.231	0.720	0.084	0.053	0.193	0.790	0.042	0.032	0.186	0.780
	(FL)Calib-n+PS	0.148	0.057	0.244	0.640	0.116	0.117	0.240	0.720	0.178	0.157	0.234	0.790	0.137	0.152	0.224	0.780
Gemma2-2b	LLM Prob.	0.185	0.114	0.209	0.801	0.195	0.111	0.211	0.822	0.142	0.107	0.222	0.751	0.209	0.225	0.303	0.620
	LLM Prob.+PS	0.272	0.025	0.286	0.635	0.244	0.140	0.253	0.822	0.203	0.099	0.250	0.751	0.155	0.025	0.260	0.620
	Verblized %	0.526	0.099	0.493	0.667	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.150	0.073	0.221	0.691	0.143	0.129	0.236	0.658	0.148	0.097	0.232	0.685	0.145	0.093	0.247	0.671
	(AUC)Calib-1	0.189	0.203	0.268	0.637	0.208	0.208	0.273	0.633	0.206	0.216	0.280	0.654	0.291	0.205	0.317	0.677
	(AUC)Calib-n	0.240	0.251	0.296	0.644	0.269	0.270	0.311	0.637	0.251	0.234	0.303	0.672	0.394	0.303	0.393	0.690
	(BCE)Calib-1	0.120	0.081	0.221	0.635	0.103	0.070	0.226	0.639	0.114	0.046	0.226	0.671	0.092	0.101	0.233	0.665
	(BCE)Calib-n	0.094	0.099	0.213	0.653	0.104	0.094	0.227	0.629	0.073	0.074	0.213	0.687	0.128	0.106	0.240	0.685
	(BCE)Calib-n+PS	0.184	0.052	0.242	0.651	0.185	0.037	0.253	0.629	0.176	0.081	0.249	0.687	0.163	0.066	0.257	0.685
	(FL)Calib-1	0.049	0.027	0.203	0.658	0.068	0.049	0.219	0.637	0.067	0.035	0.216	0.680	0.084	0.090	0.238	0.645
	(FL)Calib-n	0.088	0.072	0.211	0.644	0.098	0.096	0.227	0.621	0.065	0.062	0.209	0.697	0.129	0.105	0.240	0.683
	(FL)Calib-n+PS	0.195	0.055	0.246	0.644	0.178	0.033	0.248	0.621	0.160	0.082	0.244	0.697	0.162	0.067	0.257	0.683
Gemma2-9b	LLM Prob.	0.244	0.180	0.265	0.767	0.276	0.207	0.289	0.762	0.217	0.165	0.269	0.688	0.176	0.177	0.260	0.589
	LLM Prob.+PS	0.126	0.042	0.260	0.679	0.177	0.119	0.258	0.762	0.120	0.072	0.245	0.689	0.042	0.039	0.225	0.589
	Verblized %	0.406	0.118	0.395	0.697	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.168	0.106	0.264	0.675	0.183	0.144	0.279	0.633	0.179	0.142	0.271	0.655	0.164	0.111	0.246	0.661
	(AUC)Calib-1	0.248	0.210	0.310	0.590	0.247	0.228	0.304	0.603	0.238	0.239	0.298	0.604	0.160	0.164	0.240	0.670
	(AUC)Calib-n	0.284	0.193	0.327	0.604	0.266	0.205	0.315	0.617	0.218	0.225	0.289	0.635	0.266	0.229	0.283	0.685
	(BCE)Calib-1	0.146	0.106	0.268	0.606	0.101	0.097	0.258	0.598	0.090	0.053	0.249	0.613	0.099	0.087	0.225	0.652
	(BCE)Calib-n	0.230	0.078	0.304	0.609	0.176	0.088	0.279	0.618	0.145	0.080	0.260	0.648	0.068	0.078	0.218	0.679
	(BCE)Calib-n+PS	0.048	0.039	0.241	0.625	0.069	0.038	0.243	0.618	0.056	0.054	0.237	0.648	0.035	0.037	0.218	0.679
	(FL)Calib-1	0.126	0.102	0.264	0.600	0.102	0.103	0.258	0.602	0.090	0.074	0.251	0.621	0.098	0.073	0.224	0.657
	(FL)Calib-n	0.224	0.061	0.299	0.609	0.175	0.082	0.279	0.610	0.141	0.061	0.256	0.653	0.070	0.079	0.218	0.678
	(FL)Calib-n+PS	0.039	0.039	0.243	0.609	0.082	0.043	0.244	0.610	0.067	0.048	0.238	0.653	0.034	0.034	0.218	0.678
Phi3-4b	LLM Prob.	0.331	0.140	0.287	0.832	0.321	0.137	0.277	0.832	0.301	0.155	0.289	0.776	0.135	0.041	0.233	0.674
	LLM Prob.+PS	0.273	0.051	0.266	0.801	0.232	0.177	0.266	0.832	0.223	0.135	0.267	0.776	0.212	0.040	0.272	0.674
	Verblized %	0.472	0.066	0.437	0.753	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.139	0.068	0.205	0.728	0.125	0.114	0.226	0.712	0.164	0.116	0.249	0.674	0.161	0.070	0.251	0.636
	(AUC)Calib-1	0.183	0.176	0.236	0.711	0.215	0.231	0.290	0.655	0.218	0.220	0.285	0.636	0.277	0.188	0.295	0.717
	(AUC)Calib-n	0.219	0.222	0.269	0.696	0.277	0.287	0.319	0.637	0.238	0.217	0.296	0.667	0.276	0.195	0.289	0.731
	(BCE)Calib-1	0.133	0.081	0.210	0.712	0.103	0.049	0.229	0.682	0.122	0.059	0.243	0.648	0.156	0.046	0.243	0.673
	(BCE)Calib-n	0.098	0.097	0.199	0.714	0.110	0.108	0.238	0.641	0.084	0.065	0.224	0.677	0.053	0.049	0.203	0.728
	(BCE)Calib-n+PS	0.199	0.060	0.239	0.708	0.151	0.041	0.249	0.641	0.139	0.071	0.246	0.677	0.160	0.086	0.240	0.728
	(FL)Calib-1	0.107	0.089	0.203	0.706	0.084	0.040	0.222	0.693	0.103	0.057	0.236	0.648	0.124	0.052	0.232	0.666
	(FL)Calib-n	0.101	0.093	0.201	0.699	0.104	0.099	0.239	0.631	0.070	0.057	0.222	0.678	0.052	0.046	0.203	0.726
	(FL)Calib-n+PS	0.205	0.064	0.241	0.699	0.167	0.025	0.261	0.631	0.148	0.082	0.248	0.678	0.161	0.087	0.240	0.726
Phi3-7b	LLM Prob.	0.234	0.176	0.261	0.783	0.276	0.186	0.272	0.798	0.182	0.171	0.269	0.681	0.085	0.057	0.242	0.636
	LLM Prob.+PS	0.227	0.053	0.278	0.727	0.196	0.128	0.256	0.798	0.131	0.067	0.250	0.681	0.123	0.043	0.256	0.636
	Verblized %	0.369	0.150	113.60(	0.724	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.126	0.121	0.247	0.672	0.120	0.097	0.239	0.683	0.174	0.138	0.268	0.660	0.196	0.073	0.274	0.650
	(AUC)Calib-1	0.249	0.238	0.302	0.612	0.292	0.281	0.330	0.568	0.249	0.250	0.303	0.621	0.185	0.206	0.273	0.681
	(AUC)Calib-n	0.260	0.230	0.314	0.602	0.283	0.274	0.325	0.609	0.191	0.198	0.276	0.653	0.263	0.223	0.297	0.699
	(BCE)Calib-1	0.104	0.084	0.249	0.623	0.111	0.098	0.252	0.593	0.082	0.072	0.245	0.641	0.106	0.090	0.247	0.655
	(BCE)Calib-n	0.148	0.083	0.262	0.636	0.137	0.104	0.256	0.623	0.137	0.059	0.252	0.665	0.091	0.075	0.230	0.690
	(BCE)Calib-n+PS	0.092	0.061	0.244	0.638	0.114	0.047	0.248	0.623	0.091	0.065	0.240	0.665	0.094	0.062	0.241	0.690
	(FL)Calib-1	0.090	0.092	0.247	0.615	0.062	0.052	0.242	0.602	0.077	0.065	0.245	0.639	0.089	0.054	0.234	0.674
	(FL)Calib-n	0.148	0.082	0.262	0.624	0.130	0.109	0.258	0.609	0.139	0.047	0.250	0.667	0.088	0.068	0.230	0.689
	(FL)Calib-n+PS	0.099	0.038	0.246	0.624	0.103	0.041	0.248	0.609	0.095	0.080	0.240	0.667	0.090	0.069	0.241	0.689

Table 3: Test results of small-size models on TriviaQA dataset.

	Method	Verb.				Zero-shot				CoT				Few-shot			
		ECE↓	ECE-t↓	Brier ↓	AUC↑	ECE ↓	ECE-t↓	Brier ↓	AUC↑	ECE ↓	ECE-t↓	Brier ↓	AUC↑	ECE ↓	ECE-t↓	Brier ↓	AUC↑
Gemma2-27b	LLM Prob.	0.294	0.251	0.308	0.700	0.315	0.227	0.312	0.729	0.257	0.254	0.307	0.593	0.136	0.173	0.226	0.563
	LLM Prob.+PS	0.100	0.016	0.249	0.679	0.144	0.091	0.250	0.729	0.087	0.018	0.242	0.594	0.013	0.020	0.200	0.563
	Verblized %	0.347	0.122	0.345	0.689	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.195	0.168	0.288	0.605	0.183	0.152	0.277	0.622	0.175	0.133	0.268	0.639	0.096	0.095	0.209	0.637
	(AUC)Calib-1	0.243	0.209	0.302	0.596	0.217	0.204	0.295	0.588	0.242	0.203	0.298	0.587	0.171	0.152	0.241	0.630
	(AUC)Calib-n	0.304	0.219	0.337	0.589	0.284	0.220	0.321	0.604	0.261	0.213	0.310	0.585	0.217	0.211	0.241	0.626
	(BCE)Calib-1	0.135	0.104	0.262	0.625	0.107	0.109	0.259	0.594	0.097	0.081	0.252	0.591	0.098	0.085	0.207	0.621
	(BCE)Calib-n	0.163	0.114	0.270	0.598	0.120	0.098	0.257	0.603	0.102	0.097	0.250	0.591	0.107	0.101	0.212	0.605
	(BCE)Calib-n+PS	0.018	0.029	0.240	0.594	0.055	0.033	0.244	0.603	0.022	0.007	0.236	0.591	0.048	0.012	0.201	0.605
	(FL)Calib-1	0.118	0.059	0.254	0.621	0.111	0.093	0.258	0.593	0.080	0.049	0.240	0.589	0.095	0.053	0.206	0.614
	(FL)Calib-n	0.162	0.122	0.275	0.593	0.105	0.077	0.250	0.623	0.109	0.097	0.251	0.594	0.121	0.109	0.216	0.601
	(FL)Calib-n+PS	0.031	0.015	0.240	0.593	0.042	0.042	0.242	0.623	0.013	0.012	0.235	0.594	0.051	0.016	0.201	0.601
Llama3-70b	LLM Prob.	0.308	0.278	0.324	0.659	0.308	0.295	0.327	0.649	0.254	0.230	0.289	0.569	0.149	0.095	0.227	0.601
	LLM Prob.+PS	0.109	0.091	0.239	0.665	0.098	0.040	0.241	0.649	0.048	0.002	0.225	0.569	0.028	0.029	0.207	0.601
	Verblized %	0.282	0.069	0.298	0.673	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.178	0.150	0.278	0.589	0.176	0.134	0.279	0.593	0.184	0.124	0.267	0.610	0.118	0.094	0.213	0.677
	(AUC)Calib-1	0.244	0.219	0.306	0.575	0.209	0.172	0.285	0.611	0.210	0.179	0.286	0.577	0.145	0.154	0.210	0.680
	(AUC)Calib-n	0.314	0.210	0.344	0.578	0.283	0.190	0.318	0.611	0.282	0.227	0.311	0.570	0.147	0.149	0.204	0.706
	(BCE)Calib-1	0.113	0.084	0.252	0.595	0.114	0.101	0.252	0.614	0.137	0.098	0.254	0.559	0.081	0.062	0.194	0.697
	(BCE)Calib-n	0.176	0.124	0.277	0.590	0.148	0.087	0.262	0.608	0.121	0.107	0.247	0.578	0.123	0.072	0.209	0.693
	(BCE)Calib-n+PS	0.021	0.001	0.234	0.596	0.030	0.029	0.233	0.608	0.023	0.036	0.223	0.578	0.101	0.077	0.202	0.693
	(FL)Calib-1	0.098	0.104	0.257	0.571	0.123	0.088	0.254	0.609	0.124	0.099	0.255	0.573	0.095	0.062	0.199	0.687
	(FL)Calib-n	0.180	0.126	0.282	0.584	0.138	0.061	0.253	0.625	0.136	0.092	0.251	0.578	0.139	0.063	0.214	0.692
	(FL)Calib-n+PS	0.012	0.010	0.234	0.584	0.039	0.042	0.232	0.625	0.028	0.031	0.223	0.578	0.108	0.073	0.203	0.692
Llama3.1-70b	LLM Prob.	0.203	0.215	0.271	0.659	0.212	0.234	0.267	0.670	0.168	0.150	0.254	0.619	0.065	0.064	0.172	0.716
	LLM Prob.+PS	0.163	0.096	0.262	0.657	0.073	0.046	0.225	0.670	0.039	0.048	0.224	0.619	0.105	0.087	0.188	0.716
	Verblized %	0.306	0.131	0.311	0.661	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.149	0.123	0.261	0.611	0.164	0.134	0.266	0.605	0.161	0.109	0.244	0.653	0.087	0.082	0.183	0.692
	(AUC)Calib-1	0.206	0.195	0.278	0.609	0.216	0.201	0.283	0.587	0.219	0.188	0.272	0.596	0.113	0.120	0.177	0.714
	(AUC)Calib-n	0.310	0.225	0.340	0.585	0.294	0.223	0.327	0.586	0.269	0.229	0.307	0.572	0.157	0.162	0.195	0.725
	(BCE)Calib-1	0.066	0.067	0.244	0.583	0.070	0.053	0.236	0.592	0.047	0.027	0.225	0.620	0.072	0.072	0.179	0.710
	(BCE)Calib-n	0.172	0.129	0.274	0.590	0.147	0.106	0.266	0.583	0.100	0.094	0.245	0.591	0.068	0.065	0.178	0.715
	(BCE)Calib-n+PS	0.010	0.011	0.237	0.584	0.031	0.001	0.233	0.583	0.004	0.004	0.228	0.591	0.115	0.096	0.191	0.715
	(FL)Calib-1	0.055	0.064	0.241	0.602	0.092	0.090	0.248	0.580	0.048	0.033	0.225	0.624	0.042	0.053	0.181	0.708
	(FL)Calib-n	0.177	0.125	0.278	0.586	0.135	0.092	0.261	0.592	0.110	0.083	0.246	0.603	0.067	0.063	0.181	0.714
	(FL)Calib-n+PS	0.001	0.001	0.236	0.586	0.024	0.006	0.232	0.592	0.022	0.014	0.227	0.603	0.126	0.098	0.192	0.714
Qwen2-72b	LLM Prob.	0.305	0.253	0.316	0.708	0.341	0.261	0.340	0.697	0.228	0.222	0.282	0.603	0.074	0.062	0.207	0.661
	LLM Prob.+PS	0.116	0.053	0.242	0.682	0.137	0.056	0.253	0.697	0.068	0.046	0.231	0.603	0.067	0.053	0.209	0.661
	Verblized %	0.314	0.077	0.317	0.684	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.194	0.156	0.281	0.602	0.176	0.136	0.279	0.615	0.145	0.134	0.251	0.632	0.095	0.090	0.218	0.632
	(AUC)Calib-1	0.224	0.203	0.292	0.591	0.172	0.170	0.271	0.579	0.168	0.174	0.260	0.605	0.146	0.175	0.222	0.672
	(AUC)Calib-n	0.293	0.234	0.334	0.584	0.274	0.220	0.316	0.605	0.260	0.231	0.303	0.586	0.154	0.168	0.216	0.691
	(BCE)Calib-1	0.099	0.080	0.249	0.594	0.062	0.015	0.238	0.625	0.045	0.025	0.225	0.627	0.094	0.077	0.208	0.674
	(BCE)Calib-n	0.150	0.130	0.269	0.592	0.112	0.096	0.257	0.607	0.091	0.087	0.242	0.597	0.120	0.060	0.216	0.680
	(BCE)Calib-n+PS	0.031	0.014	0.238	0.595	0.034	0.034	0.240	0.607	0.014	0.013	0.229	0.597	0.088	0.059	0.208	0.680
	(FL)Calib-1	0.120	0.092	0.259	0.584	0.069	0.015	0.239	0.619	0.060	0.020	0.228	0.625	0.101	0.070	0.210	0.676
	(FL)Calib-n	0.158	0.130	0.273	0.591	0.096	0.079	0.249	0.622	0.104	0.095	0.248	0.588	0.136	0.083	0.222	0.678
	(FL)Calib-n+PS	0.011	0.007	0.238	0.591	0.048	0.056	0.239	0.622	0.013	0.016	0.230	0.588	0.080	0.059	0.209	0.678
Mixtral-8x7b	LLM Prob.	0.342	0.273	0.342	0.683	0.280	0.212	0.303	0.654	0.296	0.246	0.315	0.585	0.194	0.085	0.276	0.535
	LLM Prob.+PS	0.088	0.015	0.244	0.607	0.142	0.075	0.252	0.655	0.075	0.015	0.234	0.585	0.131	0.004	0.257	0.535
	Verblized %	0.318	0.052	0.324	0.624	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.215	0.166	0.300	0.584	0.153	0.153	0.261	0.603	0.149	0.132	0.262	0.606	0.242	0.141	0.319	0.531
	(AUC)Calib-1	0.215	0.210	0.290	0.581	0.157	0.144	0.253	0.647	0.142	0.146	0.255	0.583	0.105	0.117	0.257	0.560
	(AUC)Calib-n	0.305	0.263	0.341	0.565	0.291	0.292	0.327	0.552	0.192	0.196	0.274	0.573	0.184	0.192	0.271	0.597
	(BCE)Calib-1	0.093	0.079	0.253	0.582	0.061	0.057	0.226	0.658	0.057	0.051	0.232	0.588	0.172	0.052	0.271	0.567
	(BCE)Calib-n	0.154	0.142	0.270	0.580	0.121	0.134	0.260	0.561	0.085	0.055	0.239	0.576	0.143	0.064	0.256	0.594
	(BCE)Calib-n+PS	0.018	0.011	0.239	0.587	0.037	0.042	0.238	0.561	0.026	0.021	0.228	0.576	0.015	0.020	0.236	0.594
	(FL)Calib-1	0.071	0.044	0.243	0.580	0.068	0.062	0.227	0.660	0.054	0.024	0.230	0.593	0.165	0.054	0.271	0.555
	(FL)Calib-n	0.169	0.160	0.279	0.568	0.106	0.110	0.254	0.572	0.099	0.043	0.239	0.603	0.151	0.070	0.260	0.589
	(FL)Calib-n+PS	0.025	0.014	0.240	0.568	0.007	0.010	0.236	0.572	0.045	0.020	0.228	0.603	0.023	0.018	0.237	0.589

Table 4: Test results of large-size models on TriviaQA dataset.

	Method	Verb.				Zero-shot				CoT				Few-shot			
		ECE↓	ECE-t↓	Brier ↓	AUC↑	ECE ↓	ECE-t↓	Brier ↓	AUC↑	ECE ↓	ECE-t↓	Brier ↓	AUC↑	ECE ↓	ECE-t↓	Brier ↓	AUC↑
Llama2-7b	LLM Prob.	0.323	0.196	0.329	0.712	0.238	0.152	0.270	0.735	0.240	0.147	0.286	0.661	0.398	0.045	0.394	0.604
	LLM Prob.+PS	0.313	0.012	0.341	0.599	0.234	0.093	0.268	0.736	0.164	0.087	0.253	0.722	0.203	0.022	0.282	0.606
	Verblized %	0.369	0.026	0.375	0.616	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.130	0.118	0.259	0.630	0.118	0.108	0.231	0.666	0.120	0.125	0.244	0.675	0.121	0.088	0.237	0.682
	(AUC)Calib-1	0.216	0.216	0.284	0.619	0.255	0.255	0.297	0.652	0.222	0.162	0.275	0.654	0.207	0.207	0.259	0.728
	(AUC)Calib-n	0.304	0.308	0.340	0.654	0.297	0.294	0.324	0.674	0.228	0.224	0.283	0.679	0.215	0.205	0.256	0.735
	(BCE)Calib-1	0.065	0.063	0.242	0.624	0.081	0.080	0.222	0.673	0.088	0.085	0.233	0.659	0.073	0.071	0.216	0.718
	(BCE)Calib-n	0.170	0.132	0.259	0.661	0.189	0.139	0.260	0.689	0.136	0.104	0.238	0.694	0.106	0.104	0.214	0.735
	(BCE)Calib-n+PS	0.126	0.032	0.256	0.615	0.194	0.066	0.257	0.689	0.140	0.089	0.253	0.694	0.153	0.092	0.242	0.735
	(FL)Calib-1	0.040	0.012	0.236	0.636	0.069	0.067	0.222	0.661	0.057	0.080	0.229	0.660	0.086	0.048	0.218	0.710
	(FL)Calib-n	0.198	0.173	0.278	0.617	0.184	0.131	0.257	0.689	0.126	0.107	0.235	0.689	0.094	0.092	0.211	0.734
	(FL)Calib-n+PS	0.136	0.039	0.259	0.617	0.203	0.065	0.261	0.689	0.123	0.084	0.251	0.689	0.165	0.097	0.240	0.734
Llama3-8b	LLM Prob.	0.162	0.108	0.229	0.752	0.232	0.172	0.265	0.737	0.201	0.182	0.277	0.603	0.165	0.045	0.255	0.612
	LLM Prob.+PS	0.177	0.057	0.277	0.621	0.136	0.104	0.246	0.737	0.074	0.028	0.236	0.603	0.095	0.042	0.243	0.612
	Verblized %	0.321	0.070	0.340	0.626	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.185	0.187	0.286	0.587	0.146	0.146	0.259	0.636	0.163	0.147	0.254	0.639	0.115	0.111	0.225	0.691
	(AUC)Calib-1	0.233	0.229	0.288	0.646	0.224	0.223	0.290	0.615	0.201	0.201	0.269	0.634	0.212	0.211	0.245	0.730
	(AUC)Calib-n	0.257	0.256	0.290	0.697	0.280	0.276	0.315	0.654	0.224	0.218	0.270	0.650	0.202	0.208	0.238	0.753
	(BCE)Calib-1	0.118	0.120	0.239	0.673	0.111	0.094	0.252	0.616	0.094	0.097	0.233	0.638	0.106	0.103	0.211	0.733
	(BCE)Calib-n	0.122	0.126	0.233	0.704	0.147	0.147	0.255	0.660	0.112	0.098	0.229	0.669	0.088	0.088	0.201	0.752
	(BCE)Calib-n+PS	0.110	0.022	0.256	0.600	0.091	0.042	0.246	0.660	0.081	0.067	0.227	0.669	0.103	0.115	0.214	0.752
	(FL)Calib-1	0.112	0.105	0.237	0.676	0.094	0.088	0.249	0.610	0.075	0.079	0.230	0.644	0.089	0.084	0.207	0.731
	(FL)Calib-n	0.203	0.213	0.289	0.598	0.136	0.135	0.251	0.661	0.091	0.087	0.226	0.658	0.075	0.071	0.199	0.750
	(FL)Calib-n+PS	0.094	0.029	0.253	0.598	0.087	0.050	0.245	0.661	0.045	0.061	0.227	0.658	0.104	0.117	0.215	0.750
Llama3.1-8b	LLM Prob.	0.145	0.103	0.223	0.762	0.086	0.072	0.235	0.682	0.263	0.157	0.302	0.674	0.070	0.031	0.212	0.689
	LLM Prob.+PS	0.252	0.098	0.308	0.588	0.180	0.147	0.229	0.803	0.186	0.085	0.269	0.674	0.100	0.049	0.222	0.689
	Verblized %	0.415	0.184	0.413	0.587	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.182	0.179	0.273	0.590	0.108	0.103	0.216	0.734	0.092	0.084	0.224	0.721	0.101	0.082	0.202	0.738
	(AUC)Calib-1	0.251	0.211	0.276	0.729	0.240	0.221	0.268	0.753	0.233	0.197	0.253	0.761	0.180	0.178	0.214	0.740
	(AUC)Calib-n	0.223	0.224	0.258	0.755	0.200	0.193	0.243	0.782	0.223	0.182	0.251	0.763	0.192	0.193	0.224	0.758
	(BCE)Calib-1	0.084	0.106	0.199	0.770	0.118	0.133	0.207	0.771	0.089	0.091	0.196	0.781	0.082	0.090	0.192	0.744
	(BCE)Calib-n	0.118	0.116	0.206	0.758	0.117	0.104	0.198	0.788	0.106	0.073	0.192	0.793	0.081	0.082	0.192	0.756
	(BCE)Calib-n+PS	0.173	0.008	0.263	0.587	0.166	0.121	0.243	0.788	0.126	0.158	0.236	0.793	0.115	0.127	0.206	0.756
	(FL)Calib-1	0.068	0.091	0.198	0.757	0.093	0.101	0.202	0.759	0.074	0.077	0.192	0.774	0.061	0.069	0.190	0.741
	(FL)Calib-n	0.240	0.245	0.295	0.584	0.119	0.103	0.196	0.787	0.110	0.074	0.197	0.780	0.075	0.068	0.191	0.755
	(FL)Calib-n+PS	0.166	0.017	0.261	0.584	0.155	0.126	0.240	0.787	0.151	0.150	0.235	0.780	0.110	0.117	0.207	0.755
Gemma2-2b	LLM Prob.	0.060	0.065	0.196	0.774	0.069	0.067	0.192	0.791	0.104	0.097	0.236	0.691	0.141	0.145	0.260	0.633
	LLM Prob.+PS	0.179	0.001	0.279	0.586	0.156	0.151	0.229	0.791	0.083	0.080	0.237	0.691	0.060	0.033	0.238	0.633
	Verblized %	0.424	0.108	0.420	0.629	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.158	0.172	0.277	0.602	0.164	0.159	0.266	0.632	0.150	0.153	0.261	0.641	0.110	0.105	0.228	0.704
	(AUC)Calib-1	0.183	0.184	0.270	0.631	0.221	0.218	0.292	0.604	0.161	0.161	0.259	0.654	0.225	0.225	0.266	0.708
	(AUC)Calib-n	0.277	0.282	0.319	0.662	0.300	0.294	0.331	0.630	0.236	0.229	0.293	0.674	0.258	0.258	0.282	0.716
	(BCE)Calib-1	0.061	0.070	0.240	0.635	0.109	0.110	0.255	0.604	0.076	0.073	0.235	0.659	0.129	0.134	0.225	0.716
	(BCE)Calib-n	0.142	0.124	0.247	0.680	0.167	0.167	0.265	0.640	0.119	0.100	0.240	0.687	0.122	0.109	0.224	0.723
	(BCE)Calib-n+PS	0.143	0.041	0.261	0.630	0.131	0.021	0.257	0.640	0.098	0.073	0.245	0.687	0.111	0.099	0.233	0.723
	(FL)Calib-1	0.024	0.036	0.235	0.637	0.113	0.109	0.256	0.613	0.066	0.066	0.236	0.653	0.092	0.088	0.215	0.716
	(FL)Calib-n	0.177	0.175	0.271	0.627	0.173	0.169	0.265	0.640	0.103	0.074	0.235	0.682	0.111	0.096	0.221	0.720
	(FL)Calib-n+PS	0.137	0.034	0.260	0.627	0.099	0.013	0.251	0.640	0.094	0.081	0.245	0.682	0.102	0.107	0.232	0.720
Gemma2-9b	LLM Prob.	0.164	0.165	0.242	0.655	0.163	0.146	0.227	0.722	0.138	0.120	0.215	0.669	0.097	0.093	0.195	0.614
	LLM Prob.+PS	0.056	0.023	0.220	0.675	0.098	0.075	0.214	0.722	0.064	0.059	0.202	0.669	0.053	0.013	0.187	0.614
	Verblized %	0.269	0.101	0.276	0.698	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.153	0.149	0.250	0.590	0.141	0.135	0.248	0.607	0.145	0.144	0.231	0.642	0.096	0.087	0.179	0.693
	(AUC)Calib-1	0.119	0.132	0.241	0.596	0.119	0.123	0.244	0.620	0.200	0.175	0.245	0.648	0.186	0.167	0.214	0.690
	(AUC)Calib-n	0.262	0.229	0.296	0.620	0.282	0.248	0.312	0.618	0.176	0.172	0.226	0.672	0.170	0.175	0.191	0.726
	(BCE)Calib-1	0.078	0.044	0.223	0.590	0.083	0.063	0.233	0.598	0.057	0.052	0.200	0.659	0.091	0.081	0.173	0.700
	(BCE)Calib-n	0.116	0.115	0.244	0.637	0.142	0.137	0.247	0.631	0.060	0.062	0.193	0.688	0.059	0.061	0.165	0.732
	(BCE)Calib-n+PS	0.041	0.036	0.218	0.615	0.016	0.020	0.222	0.631	0.082	0.077	0.200	0.688	0.108	0.086	0.177	0.732
	(FL)Calib-1	0.037	0.032	0.218	0.597	0.060	0.044	0.224	0.617	0.039	0.033	0.200	0.651	0.068	0.058	0.170	0.698
	(FL)Calib-n	0.142	0.131	0.258	0.611	0.126	0.137	0.245	0.634	0.055	0.050	0.195	0.678	0.056	0.052	0.165	0.728
	(FL)Calib-n+PS	0.030	0.023	0.217	0.611	0.029	0.030	0.221	0.634	0.091	0.077	0.203	0.678	0.113	0.080	0.178	0.728
Phi3-4b	LLM Prob.	0.187	0.121	0.226	0.796	0.186	0.121	0.225	0.785	0.142	0.119	0.220	0.720	0.039	0.020	0.225	0.624
	LLM Prob.+PS	0.154	0.092	0.201	0.835	0.150	0.138	0.226	0.785	0.104	0.054	0.216	0.720	0.022	0.051	0.230	0.624
	Verblized %	0.322	0.034	0.324	0.703	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.064	0.059	0.204	0.755	0.143	0.132	0.247	0.672	0.127	0.133	0.238	0.643	0.125	0.120	0.231	0.685
	(AUC)Calib-1	0.205	0.147	0.228	0.810	0.161	0.151	0.240	0.709	0.140	0.121	0.239	0.711	0.237	0.174	0.262	0.732
	(AUC)Calib-n	0.178	0.182	0.216	0.814	0.238	0.182	0.278	0.686	0.182	0.181	0.244	0.690	0.203	0.184	0.240	0.740
	(BCE)Calib-1	0.072	0.089	0.177	0.828	0.086	0.083	0.217	0.711	0.068	0.065	0.210	0.697	0.132	0.116	0.219	0.723
	(BCE)Calib-n	0.079	0.077	0.174	0.826	0.110	0.103	0.232	0.691	0.085	0.065	0.212	0.705	0.092	0.075	0.205	0.746
	(BCE)Calib-n+PS	0.169	0.163	0.228	0.823	0.072	0.082	0.224	0.691	0.075	0.079	0.213	0.705	0.109	0.112	0.212	0.746
	(FL)Calib-1	0.040	0.045	0.168	0.827	0.062	0.063	0.215	0.703	0.041	0.042	0.202	0.707	0.100	0.097	0.212	0.720
	(FL)Calib-n	0.078	0.086	0.178	0.821	0.096	0.082	0.229	0.695	0.060	0.041	0.208	0.703	0.082	0.067	0.203	0.746
	(FL)Calib-n+PS	0.173	0.163	0.227	0.821	0.082	0.088	0.223	0.695	0.090	0.069	0.215	0.703	0.115	0.113	0.212	0.746
Phi3-7b	LLM Prob.	0.133	0.123	0.219	0.737	0.118	0.128	0.217	0.730	0.122	0.125	0.236	0.625	0.092	0.021	0.205	0.638
	LLM Prob.+PS	0.086	0.052	0.234	0.691	0.108	0.091	0.216	0.729	0.022	0.034	0.219	0.625	0.081	0.050	0.207	0.638
	Verblized %	0.314	0.055	0.324	0.670	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.113	0.109	0.236	0.689	0.129	0.137	0.243	0.654	0.130	0.137	0.227	0.654	0.101	0.099	0.209	0.654
	(AUC)Calib-1	0.152	0.159	0.236	0.712	0.190	0.201	0.260	0.666	0.217	0.195	0.255	0.676	0.229	0.202	0.256	0.676
	(AUC)Calib-n	0.235	0.206	0.275	0.697	0.268	0.227	0.296	0.664	0.174	0.172	0.233	0.697	0.194	0.203	0.222	0.706
	(BCE)Calib-1	0.071	0.071	0.216	0.704	0.088	0.081	0.224	0.674	0.094	0.065	0.210	0.692	0.119	0.128	0.207	0.666
	(BCE)Calib-n	0.099	0.092	0.228	0.708	0.100	0.095	0.232	0.682	0.064	0.067	0.205	0.707	0.090	0.088	0.192	0.711
	(BCE)Calib-n+PS	0.084	0.084	0.226	0.706	0.050	0.058	0.222	0.682	0.072	0.077	0.212	0.707	0.094	0.079	0.193	0.711
	(FL)Calib-1	0.054	0.062	0.215	0.701	0.076	0.085	0.223	0.664	0.063	0.042	0.205	0.691	0.125	0.124	0.208	0.662
	(FL)Calib-n	0.127	0.100	0.237	0.700	0.107	0.097	0.233	0.681	0.067	0.058	0.205	0.700	0.073	0.077	0.190	0.711
	(FL)Calib-n+PS	0.082	0.076	0.227	0.700	0.053	0.057	0.222	0.681	0.100	0.089	0.214	0.700	0.093	0.080	0.193	0.711

Table 5: Test results of small-size models on Sciq dataset.

	Method	Verb.				Zero-shot				CoT				Few-shot			
		ECE↓	ECE-t↓	Brier ↓	AUC↑	ECE ↓	ECE-t↓	Brier ↓	AUC↑	ECE ↓	ECE-t↓	Brier ↓	AUC↑	ECE ↓	ECE-t↓	Brier ↓	AUC↑
Gemma2-27b	LLM Prob.	0.193	0.181	0.236	0.705	0.204	0.179	0.229	0.713	0.161	0.158	0.227	0.630	0.080	0.088	0.172	0.616
	LLM Prob.+PS	0.077	0.038	0.216	0.658	0.068	0.081	0.199	0.713	0.028	0.035	0.199	0.630	0.080	0.026	0.170	0.616
	Verblized Prob.	0.224	0.071	0.247	0.685	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.140	0.139	0.239	0.594	0.165	0.163	0.236	0.595	0.111	0.111	0.215	0.618	0.092	0.097	0.171	0.621
	(AUC)Calib-1	0.127	0.137	0.245	0.606	0.118	0.152	0.254	0.589	0.122	0.122	0.212	0.624	0.147	0.140	0.181	0.683
	(AUC)Calib-n	0.201	0.188	0.275	0.598	0.231	0.207	0.289	0.583	0.217	0.207	0.250	0.637	0.168	0.150	0.193	0.684
	(BCE)Calib-1	0.101	0.099	0.223	0.598	0.098	0.098	0.221	0.571	0.059	0.059	0.200	0.621	0.065	0.045	0.157	0.675
	(BCE)Calib-n	0.095	0.103	0.222	0.585	0.113	0.114	0.224	0.577	0.065	0.060	0.197	0.643	0.052	0.037	0.153	0.673
	(BCE)Calib-n+PS	0.021	0.024	0.213	0.578	0.026	0.023	0.208	0.577	0.054	0.061	0.199	0.643	0.092	0.060	0.162	0.673
	(FL)Calib-1	0.081	0.077	0.219	0.599	0.066	0.075	0.214	0.594	0.058	0.059	0.199	0.625	0.038	0.037	0.155	0.662
	(FL)Calib-n	0.064	0.073	0.216	0.597	0.094	0.098	0.221	0.576	0.061	0.063	0.195	0.649	0.040	0.041	0.155	0.664
	(FL)Calib-n+PS	0.041	0.042	0.211	0.597	0.020	0.017	0.208	0.576	0.058	0.063	0.199	0.649	0.092	0.060	0.163	0.664
Llama3-70b	LLM Prob.	0.172	0.177	0.251	0.626	0.212	0.223	0.261	0.660	0.182	0.168	0.233	0.664	0.062	0.042	0.184	0.571
	LLM Prob.+PS	0.104	0.018	0.219	0.680	0.044	0.020	0.219	0.661	0.033	0.050	0.204	0.664	0.048	0.000	0.182	0.571
	Verblized Prob.	0.203	0.021	0.244	0.685	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.155	0.156	0.250	0.580	0.139	0.142	0.244	0.614	0.108	0.112	0.215	0.630	0.105	0.084	0.181	0.715
	(AUC)Calib-1	0.156	0.168	0.254	0.575	0.144	0.151	0.254	0.599	0.170	0.177	0.223	0.655	0.172	0.092	0.214	0.738
	(AUC)Calib-n	0.203	0.188	0.271	0.610	0.243	0.227	0.295	0.602	0.237	0.228	0.253	0.645	0.209	0.100	0.227	0.766
	(BCE)Calib-1	0.094	0.087	0.232	0.536	0.122	0.120	0.243	0.563	0.073	0.066	0.201	0.654	0.057	0.064	0.163	0.738
	(BCE)Calib-n	0.081	0.091	0.224	0.612	0.123	0.122	0.236	0.601	0.074	0.071	0.199	0.660	0.038	0.031	0.158	0.744
	(BCE)Calib-n+PS	0.036	0.016	0.219	0.593	0.041	0.017	0.222	0.601	0.068	0.077	0.202	0.660	0.113	0.081	0.173	0.744
	(FL)Calib-1	0.052	0.051	0.225	0.557	0.079	0.071	0.226	0.615	0.055	0.053	0.198	0.655	0.049	0.021	0.159	0.739
	(FL)Calib-n	0.067	0.073	0.220	0.613	0.110	0.115	0.233	0.598	0.072	0.067	0.197	0.665	0.062	0.021	0.162	0.741
	(FL)Calib-n+PS	0.047	0.022	0.218	0.613	0.037	0.014	0.223	0.598	0.060	0.083	0.201	0.665	0.132	0.081	0.176	0.741
Llama3.1-70b	LLM Prob.	0.133	0.119	0.220	0.688	0.120	0.101	0.205	0.704	0.125	0.119	0.213	0.671	0.062	0.047	0.154	0.692
	LLM Prob.+PS	0.054	0.009	0.212	0.698	0.078	0.063	0.199	0.705	0.073	0.070	0.202	0.671	0.105	0.058	0.165	0.692
	Verblized Prob.	0.231	0.066	0.246	0.709	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.136	0.129	0.236	0.636	0.146	0.147	0.239	0.574	0.134	0.132	0.215	0.644	0.076	0.070	0.160	0.700
	(AUC)Calib-1	0.113	0.118	0.231	0.650	0.167	0.093	0.253	0.622	0.157	0.112	0.246	0.658	0.138	0.117	0.178	0.732
	(AUC)Calib-n	0.179	0.144	0.250	0.664	0.232	0.189	0.289	0.605	0.209	0.203	0.245	0.681	0.188	0.131	0.206	0.740
	(BCE)Calib-1	0.037	0.039	0.212	0.602	0.082	0.081	0.214	0.617	0.098	0.095	0.199	0.674	0.041	0.040	0.145	0.751
	(BCE)Calib-n	0.054	0.061	0.204	0.666	0.089	0.089	0.221	0.597	0.056	0.057	0.193	0.691	0.038	0.042	0.146	0.745
	(BCE)Calib-n+PS	0.023	0.038	0.213	0.619	0.033	0.036	0.211	0.597	0.079	0.087	0.202	0.691	0.129	0.103	0.161	0.745
	(FL)Calib-1	0.036	0.035	0.206	0.636	0.057	0.064	0.213	0.624	0.079	0.092	0.198	0.666	0.023	0.017	0.144	0.738
	(FL)Calib-n	0.027	0.040	0.200	0.669	0.079	0.082	0.220	0.592	0.065	0.053	0.193	0.690	0.036	0.038	0.147	0.742
	(FL)Calib-n+PS	0.051	0.070	0.210	0.669	0.040	0.029	0.212	0.592	0.078	0.085	0.202	0.690	0.133	0.099	0.163	0.742
Qwen2-72b	LLM Prob.	0.177	0.167	0.231	0.666	0.231	0.185	0.254	0.743	0.191	0.172	0.242	0.623	0.093	0.021	0.166	0.687
	LLM Prob.+PS	0.065	0.117	0.205	0.697	0.056	0.056	0.216	0.743	0.037	0.052	0.207	0.623	0.120	0.045	0.176	0.687
	Verblized Prob.	0.211	0.051	0.232	0.715	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.145	0.146	0.227	0.623	0.149	0.150	0.241	0.607	0.125	0.121	0.222	0.613	0.094	0.102	0.184	0.653
	(AUC)Calib-1	0.189	0.163	0.255	0.616	0.156	0.153	0.244	0.601	0.139	0.163	0.216	0.665	0.210	0.131	0.237	0.721
	(AUC)Calib-n	0.187	0.172	0.267	0.638	0.230	0.220	0.290	0.608	0.223	0.199	0.246	0.688	0.209	0.119	0.234	0.750
	(BCE)Calib-1	0.068	0.079	0.208	0.620	0.101	0.100	0.233	0.596	0.088	0.097	0.200	0.680	0.077	0.060	0.160	0.717
	(BCE)Calib-n	0.043	0.057	0.203	0.638	0.114	0.102	0.232	0.610	0.063	0.072	0.194	0.687	0.052	0.035	0.151	0.750
	(BCE)Calib-n+PS	0.035	0.035	0.207	0.604	0.058	0.011	0.221	0.610	0.079	0.076	0.202	0.687	0.114	0.092	0.165	0.750
	(FL)Calib-1	0.054	0.055	0.205	0.625	0.065	0.061	0.226	0.584	0.074	0.072	0.197	0.679	0.062	0.045	0.162	0.715
	(FL)Calib-n	0.051	0.053	0.205	0.623	0.099	0.093	0.229	0.611	0.053	0.062	0.194	0.683	0.061	0.025	0.152	0.746
	(FL)Calib-n+PS	0.046	0.048	0.206	0.623	0.050	0.012	0.221	0.611	0.070	0.088	0.202	0.683	0.126	0.095	0.167	0.746
Mixtral-8x7b	LLM Prob.	0.261	0.216	0.286	0.675	0.204	0.163	0.256	0.660	0.258	0.214	0.289	0.582	0.097	0.071	0.209	0.528
	LLM Prob.+PS	0.058	0.024	0.228	0.603	0.052	0.024	0.223	0.660	0.066	0.005	0.225	0.582	0.011	0.001	0.199	0.528
	Verblized Prob.	0.273	0.059	0.288	0.611	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.157	0.152	0.254	0.609	0.120	0.117	0.230	0.640	0.098	0.101	0.218	0.661	0.101	0.095	0.211	0.605
	(AUC)Calib-1	0.206	0.231	0.263	0.631	0.183	0.188	0.237	0.687	0.175	0.156	0.240	0.677	0.343	0.097	0.339	0.636
	(AUC)Calib-n	0.199	0.182	0.277	0.634	0.177	0.157	0.266	0.654	0.276	0.098	0.297	0.636	0.273	0.136	0.295	0.698
	(BCE)Calib-1	0.071	0.078	0.224	0.633	0.079	0.087	0.211	0.678	0.079	0.085	0.213	0.671	0.087	0.083	0.196	0.654
	(BCE)Calib-n	0.053	0.059	0.220	0.647	0.060	0.041	0.219	0.632	0.014	0.016	0.207	0.660	0.053	0.053	0.184	0.701
	(BCE)Calib-n+PS	0.050	0.008	0.226	0.609	0.036	0.014	0.221	0.632	0.061	0.053	0.215	0.660	0.094	0.070	0.192	0.701
	(FL)Calib-1	0.072	0.091	0.224	0.642	0.043	0.043	0.205	0.684	0.051	0.067	0.210	0.669	0.075	0.056	0.194	0.676
	(FL)Calib-n	0.042	0.042	0.221	0.631	0.038	0.029	0.217	0.631	0.032	0.020	0.208	0.657	0.049	0.044	0.187	0.696
	(FL)Calib-n+PS	0.051	0.004	0.224	0.631	0.033	0.012	0.221	0.631	0.052	0.060	0.215	0.657	0.095	0.066	0.194	0.696

Table 6: Test results of large-size models on Sciq dataset.

	Method	Verb.				Zero-shot				CoT				Few-shot			
		ECE↓	ECE-t↓	Brier ↓	AUC↑	ECE ↓	ECE-t↓	Brier ↓	AUC↑	ECE ↓	ECE-t↓	Brier ↓	AUC↑	ECE ↓	ECE-t↓	Brier ↓	AUC↑
Llama2-7b	LLM Prob.	0.426	0.207	0.392	0.692	0.369	0.140	0.301	0.751	0.248	0.153	0.291	0.620	0.409	0.117	0.402	0.578
	LLM Prob.+PS	0.269	0.039	0.274	0.720	0.344	0.069	0.264	0.752	0.155	0.090	0.248	0.716	0.265	0.044	0.306	0.581
	Verblized Prob.	0.432	0.047	0.398	0.666	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.180	0.117	0.256	0.590	0.139	0.070	0.169	0.693	0.153	0.117	0.262	0.634	0.145	0.075	0.231	0.688
	(AUC)Calib-1	0.216	0.227	0.283	0.534	0.178	0.145	0.212	0.641	0.233	0.190	0.274	0.651	0.273	0.121	0.297	0.663
	(AUC)Calib-n	0.233	0.243	0.298	0.615	0.245	0.204	0.251	0.653	0.186	0.171	0.266	0.651	0.226	0.233	0.267	0.730
	(BCE)Calib-1	0.158	0.100	0.252	0.565	0.086	0.029	0.157	0.677	0.147	0.061	0.249	0.648	0.146	0.057	0.242	0.646
	(BCE)Calib-n	0.112	0.089	0.229	0.616	0.034	0.022	0.146	0.697	0.128	0.058	0.233	0.684	0.068	0.050	0.216	0.693
	(BCE)Calib-n+PS	0.133	0.015	0.239	0.588	0.246	0.068	0.214	0.697	0.049	0.120	0.242	0.684	0.126	0.107	0.242	0.693
	(FL)Calib-1	0.113	0.072	0.237	0.574	0.058	0.025	0.152	0.674	0.133	0.081	0.236	0.649	0.073	0.024	0.230	0.648
	(FL)Calib-n	0.109	0.089	0.232	0.589	0.040	0.038	0.148	0.686	0.141	0.090	0.229	0.696	0.086	0.054	0.213	0.704
	(FL)Calib-n+PS	0.143	0.022	0.242	0.589	0.256	0.049	0.219	0.686	0.046	0.151	0.241	0.696	0.121	0.125	0.241	0.704
Llama3-8b	LLM Prob.	0.344	0.125	0.290	0.795	0.422	0.196	0.359	0.756	0.357	0.182	0.355	0.667	0.271	0.121	0.326	0.529
	LLM Prob.+PS	0.305	0.134	0.288	0.664	0.316	0.083	0.271	0.756	0.191	0.053	0.271	0.667	0.184	0.009	0.283	0.529
	Verblized Prob.	0.494	0.089	0.436	0.694	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.150	0.089	0.210	0.639	0.106	0.112	0.185	0.649	0.156	0.151	0.258	0.598	0.148	0.060	0.229	0.739
	(AUC)Calib-1	0.133	0.136	0.210	0.693	0.142	0.123	0.210	0.634	0.138	0.123	0.257	0.602	0.238	0.060	0.261	0.756
	(AUC)Calib-n	0.104	0.106	0.194	0.721	0.161	0.178	0.235	0.654	0.252	0.253	0.305	0.570	0.177	0.166	0.215	0.809
	(BCE)Calib-1	0.126	0.053	0.190	0.692	0.097	0.050	0.179	0.675	0.028	0.022	0.240	0.586	0.149	0.085	0.228	0.748
	(BCE)Calib-n	0.101	0.035	0.182	0.710	0.082	0.059	0.177	0.641	0.044	0.064	0.231	0.641	0.108	0.076	0.214	0.762
	(BCE)Calib-n+PS	0.195	0.065	0.221	0.627	0.226	0.018	0.225	0.641	0.124	0.060	0.253	0.641	0.144	0.148	0.237	0.762
	(FL)Calib-1	0.074	0.063	0.184	0.691	0.068	0.054	0.176	0.622	0.019	0.011	0.240	0.578	0.132	0.090	0.241	0.667
	(FL)Calib-n	0.085	0.050	0.185	0.661	0.038	0.055	0.171	0.665	0.032	0.038	0.227	0.656	0.110	0.067	0.213	0.766
	(FL)Calib-n+PS	0.214	0.072	0.225	0.661	0.231	0.031	0.233	0.665	0.134	0.075	0.254	0.656	0.144	0.165	0.237	0.766
Llama3.1-8b	LLM Prob.	0.306	0.119	0.244	0.788	0.176	0.148	0.246	0.704	0.276	0.159	0.319	0.617	0.183	0.096	0.273	0.594
	LLM Prob.+PS	0.344	0.029	0.276	0.638	0.234	0.079	0.253	0.733	0.189	0.019	0.271	0.617	0.166	0.010	0.269	0.594
	Verblized Prob.	0.629	0.143	0.560	0.676	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.098	0.053	0.148	0.756	0.148	0.061	0.218	0.691	0.123	0.089	0.234	0.683	0.151	0.088	0.259	0.652
	(AUC)Calib-1	0.209	0.129	0.206	0.727	0.212	0.159	0.254	0.679	0.298	0.127	0.315	0.672	0.027	0.062	0.244	0.588
	(AUC)Calib-n	0.182	0.142	0.191	0.760	0.170	0.186	0.255	0.685	0.294	0.287	0.329	0.546	0.262	0.255	0.282	0.700
	(BCE)Calib-1	0.047	0.037	0.135	0.760	0.106	0.063	0.210	0.681	0.076	0.031	0.217	0.710	0.135	0.054	0.265	0.584
	(BCE)Calib-n	0.054	0.050	0.133	0.769	0.115	0.041	0.215	0.660	0.075	0.034	0.222	0.685	0.104	0.047	0.244	0.654
	(BCE)Calib-n+PS	0.255	0.052	0.212	0.676	0.167	0.039	0.233	0.660	0.102	0.090	0.244	0.685	0.052	0.091	0.242	0.654
	(FL)Calib-1	0.030	0.040	0.134	0.766	0.062	0.034	0.203	0.681	0.064	0.026	0.218	0.702	0.096	0.032	0.252	0.593
	(FL)Calib-n	0.046	0.049	0.143	0.674	0.086	0.050	0.211	0.645	0.119	0.061	0.223	0.672	0.112	0.052	0.241	0.664
	(FL)Calib-n+PS	0.264	0.051	0.218	0.674	0.163	0.024	0.236	0.645	0.134	0.078	0.252	0.672	0.087	0.088	0.243	0.664
Gemma2-2b	LLM Prob.	0.204	0.089	0.203	0.769	0.210	0.092	0.196	0.825	0.184	0.141	0.261	0.667	0.107	0.090	0.225	0.665
	LLM Prob.+PS	0.309	0.025	0.272	0.625	0.277	0.125	0.249	0.825	0.208	0.047	0.264	0.667	0.189	0.056	0.252	0.665
	Verblized Prob.	0.623	0.155	0.571	0.622	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.163	0.073	0.193	0.665	0.124	0.107	0.201	0.663	0.191	0.141	0.262	0.552	0.133	0.092	0.240	0.676
	(AUC)Calib-1	0.309	0.122	0.281	0.603	0.333	0.050	0.290	0.698	0.192	0.093	0.273	0.573	0.280	0.158	0.289	0.718
	(AUC)Calib-n	0.248	0.243	0.285	0.598	0.218	0.197	0.267	0.647	0.274	0.249	0.326	0.555	0.234	0.261	0.288	0.711
	(BCE)Calib-1	0.094	0.033	0.184	0.613	0.108	0.059	0.192	0.685	0.096	0.041	0.234	0.560	0.118	0.043	0.236	0.678
	(BCE)Calib-n	0.082	0.071	0.175	0.643	0.072	0.048	0.187	0.680	0.064	0.077	0.230	0.581	0.049	0.055	0.219	0.691
	(BCE)Calib-n+PS	0.217	0.004	0.223	0.639	0.209	0.074	0.228	0.680	0.178	0.008	0.262	0.581	0.152	0.115	0.252	0.691
	(FL)Calib-1	0.051	0.031	0.175	0.623	0.074	0.047	0.188	0.683	0.050	0.048	0.229	0.573	0.104	0.048	0.233	0.679
	(FL)Calib-n	0.063	0.036	0.174	0.656	0.043	0.049	0.184	0.669	0.054	0.060	0.228	0.595	0.062	0.068	0.219	0.688
	(FL)Calib-n+PS	0.227	0.011	0.227	0.656	0.207	0.067	0.230	0.669	0.184	0.003	0.264	0.595	0.129	0.113	0.249	0.688
Gemma2-9b	LLM Prob.	0.316	0.203	0.308	0.753	0.338	0.162	0.303	0.799	0.275	0.182	0.303	0.674	0.177	0.166	0.268	0.618
	LLM Prob.+PS	0.221	0.022	0.271	0.739	0.238	0.113	0.258	0.799	0.121	0.083	0.254	0.676	0.121	0.056	0.249	0.618
	Verblized Prob.	0.531	0.147	0.490	0.732	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.186	0.085	0.243	0.672	0.114	0.054	0.209	0.703	0.190	0.122	0.282	0.593	0.145	0.119	0.249	0.654
	(AUC)Calib-1	0.131	0.120	0.246	0.638	0.152	0.119	0.234	0.701	0.124	0.118	0.262	0.549	0.199	0.145	0.265	0.684
	(AUC)Calib-n	0.106	0.110	0.220	0.691	0.179	0.187	0.247	0.696	0.245	0.249	0.317	0.509	0.236	0.247	0.271	0.703
	(BCE)Calib-1	0.159	0.055	0.231	0.674	0.123	0.032	0.216	0.693	0.116	0.024	0.262	0.563	0.079	0.075	0.224	0.701
	(BCE)Calib-n	0.176	0.042	0.230	0.714	0.145	0.024	0.224	0.690	0.128	0.059	0.264	0.565	0.132	0.087	0.235	0.689
	(BCE)Calib-n+PS	0.150	0.081	0.229	0.691	0.118	0.004	0.230	0.690	0.033	0.018	0.249	0.565	0.072	0.092	0.234	0.689
	(FL)Calib-1	0.122	0.066	0.222	0.677	0.092	0.048	0.214	0.676	0.096	0.025	0.255	0.575	0.094	0.055	0.220	0.702
	(FL)Calib-n	0.161	0.057	0.224	0.712	0.114	0.032	0.216	0.690	0.119	0.042	0.260	0.573	0.137	0.077	0.234	0.693
	(FL)Calib-n+PS	0.170	0.065	0.230	0.712	0.140	0.022	0.233	0.690	0.038	0.028	0.249	0.573	0.090	0.088	0.235	0.693
Phi3-4b	LLM Prob.	0.374	0.187	0.332	0.742	0.245	0.123	0.253	0.760	0.209	0.100	0.249	0.739	0.181	0.084	0.243	0.665
	LLM Prob.+PS	0.310	0.054	0.279	0.768	0.192	0.117	0.257	0.760	0.170	0.096	0.257	0.739	0.223	0.086	0.277	0.665
	Verblized Prob.	0.546	0.088	0.489	0.749	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.135	0.095	0.195	0.631	0.141	0.076	0.238	0.675	0.133	0.135	0.258	0.609	0.103	0.048	0.228	0.661
	(AUC)Calib-1	0.200	0.182	0.255	0.572	0.196	0.059	0.273	0.601	0.127	0.120	0.240	0.639	0.314	0.123	0.313	0.678
	(AUC)Calib-n	0.186	0.202	0.256	0.616	0.313	0.311	0.345	0.529	0.205	0.204	0.284	0.596	0.256	0.267	0.297	0.694
	(BCE)Calib-1	0.064	0.054	0.191	0.614	0.124	0.004	0.251	0.584	0.056	0.043	0.242	0.607	0.083	0.077	0.226	0.652
	(BCE)Calib-n	0.100	0.071	0.189	0.645	0.163	0.063	0.262	0.575	0.061	0.052	0.236	0.638	0.054	0.065	0.220	0.665
	(BCE)Calib-n+PS	0.205	0.058	0.230	0.653	0.085	0.044	0.245	0.575	0.098	0.016	0.251	0.638	0.151	0.054	0.248	0.665
	(FL)Calib-1	0.052	0.029	0.189	0.584	0.090	0.037	0.239	0.625	0.039	0.011	0.242	0.596	0.082	0.055	0.221	0.662
	(FL)Calib-n	0.071	0.070	0.185	0.652	0.125	0.056	0.251	0.590	0.053	0.046	0.233	0.647	0.048	0.081	0.216	0.676
	(FL)Calib-n+PS	0.208	0.052	0.231	0.652	0.101	0.055	0.247	0.590	0.122	0.013	0.253	0.647	0.154	0.058	0.248	0.676
Phi3-7b	LLM Prob.	0.165	0.138	0.230	0.751	0.194	0.139	0.235	0.744	0.206	0.153	0.287	0.640	0.077	0.082	0.247	0.603
	LLM Prob.+PS	0.193	0.139	0.257	0.720	0.182	0.113	0.251	0.744	0.136	0.043	0.258	0.640	0.124	0.031	0.260	0.603
	Verblized Prob.	0.444	0.097	0.415	0.702	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.143	0.091	0.227	0.717	0.131	0.077	0.242	0.642	0.094	0.132	0.247	0.641	0.132	0.122	0.254	0.628
	(AUC)Calib-1	0.164	0.168	0.263	0.587	0.205	0.197	0.285	0.521	0.174	0.175	0.277	0.564	0.107	0.075	0.260	0.585
	(AUC)Calib-n	0.178	0.180	0.257	0.680	0.252	0.259	0.310	0.583	0.209	0.204	0.293	0.574	0.288	0.271	0.316	0.695
	(BCE)Calib-1	0.125	0.041	0.237	0.649	0.124	0.092	0.260	0.478	0.114	0.068	0.254	0.570	0.086	0.050	0.248	0.613
	(BCE)Calib-n	0.174	0.056	0.242	0.682	0.166	0.076	0.265	0.576	0.113	0.070	0.252	0.605	0.078	0.046	0.239	0.666
	(BCE)Calib-n+PS	0.156	0.103	0.234	0.692	0.086	0.003	0.244	0.576	0.098	0.010	0.250	0.605	0.070	0.084	0.246	0.666
	(FL)Calib-1	0.096	0.029	0.232	0.646	0.081	0.091	0.253	0.490	0.079	0.082	0.248	0.606	0.072	0.045	0.246	0.615
	(FL)Calib-n	0.162	0.050	0.238	0.687	0.132	0.100	0.258	0.565	0.107	0.044	0.248	0.607	0.067	0.035	0.236	0.676
	(FL)Calib-n+PS	0.150	0.105	0.236	0.687	0.096	0.011	0.247	0.565	0.092	0.016	0.250	0.607	0.071	0.090	0.246	0.676

Table 7: Test results of small-size models on WikiQA dataset.

	Method	Verb.				Zero-shot				CoT				Few-shot			
		ECE↓	ECE-t↓	Brier ↓	AUC↑	ECE ↓	ECE-t↓	Brier ↓	AUC↑	ECE ↓	ECE-t↓	Brier ↓	AUC↑	ECE ↓	ECE-t↓	Brier ↓	AUC↑
Gemma2-27b	LLM Prob.	0.423	0.258	0.382	0.775	0.448	0.259	0.410	0.730	0.359	0.222	0.356	0.654	0.158	0.164	0.226	0.501
	LLM Prob.+PS	0.285	0.101	0.284	0.712	0.234	0.078	0.261	0.730	0.160	0.095	0.261	0.656	0.057	0.065	0.204	0.501
	Verblized Prob.	0.578	0.132	0.519	0.752	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.114	0.070	0.186	0.749	0.127	0.071	0.201	0.734	0.152	0.095	0.244	0.679	0.109	0.103	0.216	0.619
	(AUC)Calib-1	0.171	0.157	0.240	0.650	0.104	0.112	0.217	0.709	0.098	0.108	0.236	0.692	0.075	0.063	0.204	0.621
	(AUC)Calib-n	0.170	0.157	0.230	0.705	0.085	0.086	0.192	0.752	0.200	0.169	0.261	0.640	0.093	0.124	0.205	0.615
	(BCE)Calib-1	0.099	0.039	0.191	0.725	0.129	0.038	0.204	0.735	0.132	0.059	0.230	0.719	0.052	0.052	0.186	0.660
	(BCE)Calib-n	0.042	0.052	0.182	0.718	0.102	0.052	0.187	0.778	0.084	0.086	0.230	0.650	0.092	0.026	0.198	0.618
	(BCE)Calib-n+PS	0.174	0.131	0.228	0.721	0.193	0.120	0.226	0.778	0.109	0.090	0.250	0.650	0.066	0.050	0.201	0.618
	(FL)Calib-1	0.070	0.041	0.192	0.689	0.103	0.074	0.197	0.745	0.105	0.067	0.227	0.715	0.040	0.028	0.185	0.669
	(FL)Calib-n	0.064	0.049	0.183	0.715	0.095	0.069	0.186	0.780	0.114	0.098	0.224	0.674	0.102	0.043	0.203	0.612
	(FL)Calib-n+PS	0.184	0.104	0.231	0.715	0.189	0.133	0.226	0.780	0.094	0.112	0.246	0.674	0.122	0.052	0.206	0.612
Llama3-70b	LLM Prob.	0.359	0.245	0.353	0.727	0.424	0.234	0.385	0.752	0.319	0.239	0.351	0.582	0.125	0.111	0.230	0.512
	LLM Prob.+PS	0.207	0.136	0.278	0.673	0.219	0.071	0.263	0.752	0.091	0.008	0.254	0.582	0.024	0.043	0.214	0.512
	Verblized Prob.	0.439	0.098	0.415	0.686	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.181	0.077	0.234	0.700	0.151	0.081	0.207	0.744	0.174	0.099	0.274	0.614	0.125	0.114	0.239	0.620
	(AUC)Calib-1	0.127	0.133	0.245	0.650	0.111	0.114	0.228	0.690	0.121	0.119	0.251	0.628	0.087	0.057	0.199	0.699
	(AUC)Calib-n	0.156	0.158	0.242	0.679	0.122	0.124	0.216	0.719	0.181	0.115	0.264	0.638	0.123	0.120	0.212	0.669
	(BCE)Calib-1	0.165	0.069	0.244	0.669	0.131	0.041	0.213	0.729	0.147	0.095	0.261	0.626	0.042	0.030	0.201	0.666
	(BCE)Calib-n	0.123	0.095	0.221	0.717	0.108	0.024	0.201	0.746	0.125	0.055	0.244	0.668	0.072	0.036	0.200	0.670
	(BCE)Calib-n+PS	0.140	0.125	0.237	0.672	0.179	0.098	0.229	0.746	0.029	0.048	0.241	0.668	0.039	0.075	0.211	0.670
	(FL)Calib-1	0.147	0.069	0.241	0.667	0.095	0.054	0.206	0.729	0.132	0.072	0.255	0.632	0.045	0.019	0.202	0.662
	(FL)Calib-n	0.109	0.049	0.219	0.715	0.097	0.034	0.202	0.737	0.111	0.049	0.240	0.675	0.081	0.016	0.208	0.631
	(FL)Calib-n+PS	0.155	0.135	0.236	0.715	0.176	0.076	0.230	0.737	0.019	0.068	0.241	0.675	0.032	0.077	0.211	0.631
Llama3.1-70b	LLM Prob.	0.174	0.179	0.250	0.729	0.229	0.182	0.273	0.706	0.156	0.177	0.260	0.634	0.085	0.074	0.194	0.570
	LLM Prob.+PS	0.168	0.023	0.269	0.687	0.152	0.073	0.250	0.706	0.037	0.063	0.237	0.634	0.078	0.004	0.191	0.570
	Verblized Prob.	0.453	0.130	0.432	0.708	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.135	0.103	0.228	0.718	0.195	0.131	0.277	0.619	0.162	0.108	0.264	0.644	0.142	0.120	0.204	0.589
	(AUC)Calib-1	0.093	0.092	0.227	0.695	0.098	0.098	0.242	0.622	0.209	0.203	0.270	0.573	0.148	0.152	0.223	0.566
	(AUC)Calib-n	0.150	0.151	0.234	0.704	0.154	0.151	0.247	0.657	0.278	0.228	0.316	0.532	0.111	0.126	0.197	0.598
	(BCE)Calib-1	0.140	0.054	0.236	0.694	0.159	0.041	0.253	0.644	0.150	0.077	0.266	0.582	0.064	0.062	0.193	0.565
	(BCE)Calib-n	0.130	0.033	0.228	0.716	0.155	0.049	0.241	0.690	0.144	0.092	0.270	0.584	0.085	0.033	0.194	0.576
	(BCE)Calib-n+PS	0.122	0.072	0.239	0.685	0.132	0.089	0.237	0.690	0.027	0.059	0.244	0.584	0.087	0.000	0.194	0.576
	(FL)Calib-1	0.117	0.070	0.234	0.689	0.123	0.009	0.249	0.612	0.142	0.037	0.261	0.587	0.073	0.068	0.193	0.574
	(FL)Calib-n	0.120	0.050	0.228	0.712	0.141	0.057	0.243	0.668	0.145	0.120	0.265	0.598	0.090	0.051	0.198	0.575
	(FL)Calib-n+PS	0.123	0.105	0.239	0.712	0.110	0.067	0.240	0.668	0.015	0.057	0.243	0.598	0.091	0.000	0.194	0.575
Qwen2-72b	LLM Prob.	0.403	0.249	0.389	0.693	0.500	0.235	0.459	0.750	0.301	0.246	0.345	0.542	0.127	0.133	0.225	0.551
	LLM Prob.+PS	0.196	0.100	0.272	0.716	0.227	0.032	0.273	0.749	0.079	0.002	0.253	0.542	0.049	0.037	0.209	0.551
	Verblized Prob.	0.480	0.042	0.449	0.745	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.159	0.120	0.251	0.645	0.118	0.096	0.226	0.665	0.169	0.135	0.267	0.628	0.070	0.070	0.215	0.581
	(AUC)Calib-1	0.185	0.169	0.283	0.563	0.139	0.145	0.248	0.615	0.102	0.110	0.250	0.612	0.066	0.101	0.208	0.608
	(AUC)Calib-n	0.270	0.279	0.318	0.557	0.194	0.201	0.264	0.638	0.172	0.152	0.271	0.629	0.126	0.160	0.218	0.635
	(BCE)Calib-1	0.108	0.044	0.247	0.564	0.118	0.029	0.235	0.631	0.103	0.081	0.248	0.627	0.065	0.063	0.207	0.592
	(BCE)Calib-n	0.106	0.092	0.249	0.593	0.095	0.061	0.226	0.648	0.118	0.056	0.243	0.662	0.062	0.025	0.199	0.635
	(BCE)Calib-n+PS	0.098	0.023	0.246	0.586	0.119	0.078	0.237	0.648	0.047	0.046	0.240	0.662	0.042	0.082	0.206	0.635
	(FL)Calib-1	0.068	0.056	0.244	0.553	0.072	0.042	0.226	0.631	0.095	0.043	0.248	0.621	0.069	0.064	0.211	0.581
	(FL)Calib-n	0.105	0.091	0.246	0.593	0.082	0.074	0.228	0.636	0.107	0.088	0.238	0.669	0.069	0.020	0.200	0.628
	(FL)Calib-n+PS	0.119	0.027	0.251	0.593	0.124	0.061	0.239	0.636	0.063	0.067	0.240	0.669	0.040	0.069	0.205	0.628
Mixtral-8x7b	LLM Prob.	0.306	0.264	0.340	0.590	0.297	0.137	0.317	0.664	0.268	0.195	0.295	0.614	0.222	0.093	0.286	0.529
	LLM Prob.+PS	0.099	0.036	0.254	0.574	0.113	0.091	0.255	0.664	0.061	0.018	0.227	0.614	0.078	0.039	0.243	0.529
	Verblized Prob.	0.327	0.151	0.336	0.615	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.123	0.107	0.258	0.617	0.165	0.126	0.274	0.584	0.080	0.085	0.228	0.612	0.106	0.084	0.242	0.617
	(AUC)Calib-1	0.187	0.187	0.284	0.522	0.098	0.113	0.258	0.588	0.049	0.030	0.220	0.609	0.030	0.028	0.220	0.665
	(AUC)Calib-n	0.252	0.252	0.301	0.546	0.254	0.228	0.304	0.561	0.173	0.191	0.274	0.484	0.221	0.165	0.269	0.639
	(BCE)Calib-1	0.119	0.032	0.263	0.520	0.099	0.038	0.257	0.576	0.046	0.037	0.214	0.638	0.054	0.047	0.221	0.666
	(BCE)Calib-n	0.170	0.076	0.278	0.567	0.106	0.101	0.267	0.552	0.121	0.104	0.253	0.533	0.068	0.017	0.229	0.636
	(BCE)Calib-n+PS	0.015	0.022	0.247	0.516	0.038	0.002	0.249	0.552	0.048	0.014	0.229	0.533	0.060	0.052	0.233	0.636
	(FL)Calib-1	0.110	0.014	0.257	0.551	0.088	0.042	0.257	0.561	0.013	0.014	0.217	0.613	0.064	0.055	0.219	0.683
	(FL)Calib-n	0.163	0.078	0.275	0.566	0.096	0.087	0.264	0.552	0.115	0.150	0.248	0.554	0.061	0.055	0.226	0.645
	(FL)Calib-n+PS	0.024	0.025	0.244	0.566	0.045	0.007	0.249	0.552	0.040	0.006	0.228	0.554	0.030	0.077	0.232	0.645

Table 8: Test results of large-size models on WikiQA dataset.

	Method	Verb.				Zero-shot				CoT				Few-shot			
		ECE↓	ECE-t↓	Brier ↓	AUC↑	ECE ↓	ECE-t↓	Brier ↓	AUC↑	ECE ↓	ECE-t↓	Brier ↓	AUC↑	ECE ↓	ECE-t↓	Brier ↓	AUC↑
Llama2-7b	LLM Prob.	0.554	0.248	0.480	0.700	0.436	0.182	0.356	0.719	0.488	0.182	0.445	0.647	0.543	0.047	0.477	0.628
	LLM Prob.+PS	0.513	0.020	0.420	0.640	0.397	0.032	0.283	0.721	0.328	0.050	0.297	0.680	0.333	0.002	0.297	0.628
	Verblized Prob.	0.549	0.056	0.467	0.638	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.066	0.066	0.155	0.655	0.050	0.055	0.123	0.712	0.086	0.076	0.200	0.646	0.071	0.044	0.180	0.681
	(AUC)Calib-1	0.212	0.221	0.258	0.643	0.176	0.178	0.208	0.649	0.223	0.211	0.271	0.654	0.193	0.193	0.240	0.686
	(AUC)Calib-n	0.266	0.249	0.284	0.676	0.248	0.228	0.257	0.680	0.290	0.235	0.319	0.651	0.200	0.197	0.231	0.722
	(BCE)Calib-1	0.043	0.050	0.156	0.651	0.047	0.049	0.128	0.670	0.052	0.060	0.191	0.657	0.093	0.060	0.181	0.687
	(BCE)Calib-n	0.052	0.052	0.156	0.675	0.048	0.041	0.131	0.714	0.076	0.072	0.197	0.657	0.060	0.057	0.172	0.727
	(BCE)Calib-n+PS	0.278	0.048	0.232	0.667	0.305	0.013	0.221	0.714	0.233	0.037	0.246	0.657	0.229	0.062	0.227	0.727
	(FL)Calib-1	0.026	0.039	0.153	0.653	0.060	0.038	0.130	0.660	0.050	0.049	0.190	0.655	0.061	0.043	0.176	0.691
	(FL)Calib-n	0.051	0.049	0.156	0.663	0.068	0.037	0.133	0.710	0.048	0.048	0.193	0.644	0.060	0.054	0.170	0.730
	(FL)Calib-n+PS	0.270	0.050	0.228	0.663	0.307	0.022	0.223	0.710	0.227	0.021	0.245	0.644	0.228	0.052	0.228	0.730
Llama3-8b	LLM Prob.	0.398	0.150	0.328	0.755	0.435	0.199	0.366	0.737	0.415	0.211	0.392	0.653	0.408	0.046	0.380	0.639
	LLM Prob.+PS	0.353	0.042	0.280	0.712	0.378	0.059	0.298	0.737	0.290	0.051	0.298	0.653	0.318	0.011	0.322	0.639
	Verblized Prob.	0.517	0.084	0.437	0.686	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.065	0.055	0.156	0.689	0.075	0.072	0.169	0.639	0.073	0.081	0.220	0.641	0.089	0.059	0.213	0.682
	(AUC)Calib-1	0.191	0.199	0.236	0.656	0.145	0.146	0.201	0.619	0.306	0.120	0.304	0.679	0.213	0.220	0.274	0.706
	(AUC)Calib-n	0.222	0.232	0.253	0.676	0.218	0.227	0.264	0.654	0.280	0.273	0.323	0.671	0.233	0.233	0.267	0.734
	(BCE)Calib-1	0.044	0.043	0.158	0.642	0.076	0.073	0.169	0.638	0.014	0.013	0.211	0.647	0.093	0.084	0.206	0.702
	(BCE)Calib-n	0.055	0.056	0.156	0.680	0.056	0.055	0.164	0.668	0.067	0.064	0.209	0.683	0.081	0.074	0.197	0.733
	(BCE)Calib-n+PS	0.260	0.055	0.224	0.683	0.257	0.063	0.228	0.668	0.196	0.076	0.250	0.683	0.184	0.110	0.240	0.733
	(FL)Calib-1	0.049	0.035	0.159	0.628	0.073	0.068	0.173	0.605	0.029	0.022	0.213	0.639	0.064	0.065	0.202	0.702
	(FL)Calib-n	0.039	0.037	0.153	0.682	0.046	0.046	0.163	0.664	0.053	0.049	0.210	0.666	0.061	0.061	0.194	0.738
	(FL)Calib-n+PS	0.262	0.059	0.226	0.682	0.269	0.054	0.234	0.664	0.193	0.075	0.252	0.666	0.183	0.117	0.240	0.738
Llama3.1-8b	LLM Prob.	0.335	0.104	0.255	0.808	0.303	0.146	0.272	0.737	0.428	0.182	0.395	0.630	0.278	0.029	0.277	0.716
	LLM Prob.+PS	0.363	0.038	0.279	0.621	0.340	0.068	0.280	0.747	0.320	0.050	0.306	0.630	0.283	0.059	0.296	0.716
	Verblized Prob.	0.597	0.132	0.518	0.684	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.068	0.067	0.143	0.646	0.074	0.078	0.175	0.674	0.071	0.055	0.194	0.699	0.071	0.026	0.200	0.732
	(AUC)Calib-1	0.168	0.189	0.221	0.681	0.215	0.226	0.266	0.607	0.401	0.218	0.355	0.712	0.175	0.166	0.224	0.760
	(AUC)Calib-n	0.228	0.224	0.247	0.692	0.242	0.248	0.282	0.641	0.276	0.259	0.306	0.675	0.212	0.207	0.242	0.775
	(BCE)Calib-1	0.054	0.056	0.137	0.695	0.060	0.050	0.174	0.640	0.075	0.071	0.190	0.719	0.080	0.056	0.196	0.751
	(BCE)Calib-n	0.043	0.040	0.136	0.704	0.047	0.044	0.168	0.683	0.056	0.051	0.186	0.729	0.064	0.055	0.184	0.777
	(BCE)Calib-n+PS	0.282	0.019	0.218	0.636	0.238	0.043	0.229	0.683	0.221	0.086	0.244	0.729	0.193	0.138	0.238	0.777
	(FL)Calib-1	0.044	0.034	0.137	0.692	0.040	0.042	0.174	0.629	0.058	0.057	0.187	0.715	0.055	0.029	0.190	0.757
	(FL)Calib-n	0.063	0.063	0.143	0.632	0.043	0.042	0.167	0.684	0.038	0.034	0.183	0.728	0.053	0.050	0.184	0.769
	(FL)Calib-n+PS	0.286	0.030	0.221	0.632	0.256	0.044	0.238	0.684	0.210	0.106	0.241	0.728	0.196	0.133	0.236	0.769
Gemma2-2b	LLM Prob.	0.279	0.124	0.229	0.758	0.290	0.130	0.244	0.755	0.261	0.152	0.260	0.701	0.240	0.134	0.256	0.646
	LLM Prob.+PS	0.382	0.013	0.273	0.705	0.367	0.049	0.270	0.756	0.318	0.052	0.266	0.701	0.301	0.026	0.265	0.646
	Verblized Prob.	0.628	0.099	0.551	0.700	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.056	0.039	0.127	0.694	0.059	0.048	0.142	0.693	0.086	0.042	0.171	0.687	0.071	0.021	0.156	0.755
	(AUC)Calib-1	0.215	0.156	0.208	0.716	0.185	0.174	0.220	0.666	0.225	0.199	0.263	0.642	0.167	0.183	0.219	0.763
	(AUC)Calib-n	0.269	0.211	0.269	0.714	0.254	0.224	0.269	0.692	0.289	0.259	0.312	0.663	0.267	0.241	0.267	0.785
	(BCE)Calib-1	0.043	0.032	0.127	0.683	0.065	0.060	0.146	0.650	0.054	0.032	0.171	0.643	0.071	0.030	0.156	0.765
	(BCE)Calib-n	0.054	0.041	0.130	0.688	0.050	0.043	0.144	0.705	0.072	0.068	0.174	0.688	0.096	0.069	0.166	0.782
	(BCE)Calib-n+PS	0.309	0.015	0.222	0.679	0.293	0.041	0.227	0.705	0.259	0.065	0.235	0.688	0.265	0.094	0.234	0.782
	(FL)Calib-1	0.025	0.034	0.126	0.665	0.040	0.036	0.145	0.645	0.026	0.027	0.169	0.639	0.032	0.033	0.154	0.761
	(FL)Calib-n	0.053	0.037	0.130	0.688	0.064	0.033	0.143	0.708	0.044	0.045	0.169	0.682	0.089	0.056	0.163	0.776
	(FL)Calib-n+PS	0.312	0.019	0.223	0.688	0.294	0.052	0.228	0.708	0.272	0.068	0.244	0.682	0.273	0.102	0.240	0.776
Gemma2-9b	LLM Prob.	0.390	0.230	0.368	0.716	0.396	0.201	0.355	0.746	0.374	0.232	0.377	0.641	0.298	0.207	0.341	0.594
	LLM Prob.+PS	0.306	0.095	0.289	0.696	0.309	0.100	0.285	0.746	0.243	0.046	0.284	0.641	0.213	0.029	0.284	0.594
	Verblized Prob.	0.547	0.196	0.496	0.695	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.096	0.099	0.210	0.626	0.084	0.073	0.206	0.646	0.097	0.088	0.234	0.638	0.077	0.076	0.226	0.689
	(AUC)Calib-1	0.210	0.218	0.279	0.594	0.192	0.203	0.265	0.604	0.227	0.223	0.299	0.621	0.240	0.236	0.287	0.727
	(AUC)Calib-n	0.256	0.257	0.287	0.642	0.219	0.230	0.275	0.649	0.270	0.268	0.311	0.662	0.273	0.272	0.296	0.740
	(BCE)Calib-1	0.098	0.087	0.206	0.597	0.048	0.040	0.203	0.616	0.053	0.061	0.223	0.637	0.122	0.120	0.224	0.722
	(BCE)Calib-n	0.108	0.075	0.209	0.629	0.093	0.040	0.201	0.674	0.092	0.064	0.224	0.667	0.067	0.072	0.208	0.742
	(BCE)Calib-n+PS	0.178	0.035	0.230	0.631	0.180	0.064	0.231	0.674	0.155	0.052	0.247	0.667	0.148	0.107	0.246	0.742
	(FL)Calib-1	0.057	0.064	0.203	0.604	0.046	0.042	0.204	0.599	0.024	0.024	0.223	0.634	0.108	0.097	0.221	0.718
	(FL)Calib-n	0.091	0.063	0.206	0.621	0.066	0.051	0.196	0.677	0.083	0.044	0.224	0.655	0.067	0.082	0.210	0.736
	(FL)Calib-n+PS	0.182	0.030	0.232	0.621	0.199	0.071	0.239	0.677	0.157	0.048	0.249	0.655	0.152	0.096	0.248	0.736
Phi3-4b	LLM Prob.	0.467	0.145	0.367	0.788	0.436	0.143	0.337	0.767	0.385	0.143	0.319	0.734	0.306	0.026	0.244	0.714
	LLM Prob.+PS	0.389	0.017	0.279	0.768	0.400	0.061	0.292	0.767	0.368	0.059	0.296	0.734	0.378	0.077	0.303	0.714
	Verblized Prob.	0.593	0.077	0.505	0.751	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.068	0.038	0.131	0.722	0.066	0.040	0.133	0.758	0.061	0.059	0.162	0.718	0.070	0.028	0.147	0.761
	(AUC)Calib-1	0.134	0.148	0.185	0.705	0.129	0.128	0.175	0.732	0.147	0.161	0.204	0.709	0.158	0.153	0.187	0.753
	(AUC)Calib-n	0.230	0.215	0.252	0.712	0.273	0.228	0.278	0.699	0.275	0.209	0.282	0.703	0.148	0.146	0.172	0.803
	(BCE)Calib-1	0.078	0.049	0.131	0.717	0.050	0.031	0.130	0.756	0.055	0.043	0.161	0.705	0.078	0.050	0.150	0.760
	(BCE)Calib-n	0.030	0.032	0.126	0.729	0.062	0.038	0.138	0.731	0.072	0.065	0.171	0.694	0.038	0.038	0.134	0.806
	(BCE)Calib-n+PS	0.297	0.030	0.216	0.715	0.297	0.045	0.224	0.731	0.278	0.060	0.242	0.694	0.283	0.113	0.226	0.806
	(FL)Calib-1	0.027	0.029	0.124	0.725	0.024	0.016	0.128	0.750	0.020	0.012	0.158	0.704	0.037	0.035	0.144	0.760
	(FL)Calib-n	0.040	0.036	0.129	0.712	0.079	0.034	0.139	0.727	0.057	0.050	0.166	0.687	0.041	0.035	0.133	0.812
	(FL)Calib-n+PS	0.297	0.022	0.216	0.712	0.309	0.054	0.232	0.727	0.261	0.079	0.235	0.687	0.252	0.107	0.214	0.812
Phi3-7b	LLM Prob.	0.294	0.166	0.272	0.747	0.303	0.148	0.272	0.758	0.312	0.159	0.310	0.693	0.165	0.036	0.228	0.679
	LLM Prob.+PS	0.358	0.023	0.303	0.713	0.333	0.081	0.280	0.758	0.286	0.065	0.286	0.693	0.258	0.074	0.280	0.679
	Verblized Prob.	0.484	0.198	0.441	0.710	-	-	-	-	-	-	-	-	-	-	-	-
	APRICOT	0.078	0.073	0.176	0.696	0.079	0.066	0.178	0.702	0.084	0.089	0.212	0.679	0.093	0.067	0.208	0.690
	(AUC)Calib-1	0.207	0.209	0.256	0.650	0.224	0.233	0.272	0.656	0.207	0.204	0.274	0.659	0.180	0.171	0.242	0.692
	(AUC)Calib-n	0.256	0.253	0.280	0.657	0.263	0.265	0.297	0.639	0.253	0.248	0.295	0.671	0.256	0.256	0.286	0.717
	(BCE)Calib-1	0.051	0.057	0.177	0.660	0.030	0.018	0.173	0.688	0.063	0.059	0.209	0.663	0.096	0.078	0.207	0.700
	(BCE)Calib-n	0.069	0.058	0.176	0.670	0.052	0.046	0.176	0.682	0.084	0.063	0.205	0.676	0.105	0.103	0.207	0.711
	(BCE)Calib-n+PS	0.223	0.021	0.227	0.666	0.227	0.035	0.231	0.682	0.194	0.081	0.243	0.676	0.200	0.076	0.247	0.711
	(FL)Calib-1	0.044	0.039	0.177	0.647	0.022	0.007	0.173	0.682	0.035	0.034	0.206	0.655	0.074	0.063	0.205	0.697
	(FL)Calib-n	0.059	0.056	0.177	0.658	0.047	0.052	0.176	0.676	0.074	0.052	0.208	0.655	0.090	0.089	0.203	0.714
	(FL)Calib-n+PS	0.224	0.021	0.228	0.658	0.241	0.032	0.238	0.676	0.191	0.081	0.247	0.655	0.192	0.075	0.244	0.714

Table 9: Test results of small-size models on NQ dataset.