Calibrating Model-Based Evaluation Metrics for Summarization

Hongye Liu, Dhanajit Brahma, Ricardo Henao
Duke University
{hongye.liu, dhanajit.brahma, ricardo.henao}@duke.edu

Abstract

Recent advances in summary evaluation are based on model-based metrics to assess quality dimensions, such as completeness, conciseness, and faithfulness. However, these methods often require large language models, and predicted scores are frequently miscalibrated, limiting their reliability. Moreover, evaluating the average quality across different summaries for a single document typically requires access to multiple reference summaries. Here, we propose a general framework that generates individual and average proxy scores without relying on reference summaries, human annotations, or expensive model-based metrics. We also propose *group isotonic regression binning* (GIRB)[1], a calibration method that adjusts the raw predictions to better align with ground-truth evaluation metrics. While we focus on continuous-value scenarios, such as summarization, the method is applicable to discrete-value tasks, such as question answering. Experiments on seven datasets demonstrate that our approach consistently outperforms existing baselines.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities in various natural language processing tasks, such as summarization and question answering (QA) (Minaee et al., 2024; Luo et al., 2025, 2026). However, these models are prone to issues such as hallucination, which results in incorrect or misleading outputs (Huang et al., 2023). Consequently, evaluating the outputs of LLMs is critical for their effective use (Lan et al., 2025a,b; Wang et al., 2026a,b). One direct approach to evaluation is through human judgment, but this method is costly and time-consuming (Zhang and Bansal, 2021; Liu and Henao, 2025).

To address this issue, *model-based evaluation metrics* have been proposed. For example, FineSurE (Song et al., 2024a) evaluates completeness, conciseness, and confidence, while UniEval (Lee et al., 2024) measures consistency, fluency, and coherence. Other more general metrics, such as BERTScore (Zhang et al., 2019) or BARTScore (Yuan et al., 2021), can also be used. However, summarization models typically do not produce evaluation scores, which requires additional model-based scoring methods. These methods often rely on LLMs, making evaluation computationally expensive. Furthermore, deep neural networks, including LLMs, are known to suffer from calibration issues (Guo et al., 2017; Parappan and Henao, 2026), a limitation that LLM-based evaluation methods also inherit (Manggala et al., 2024).

To address these challenges, our goal is to develop a simple approach that enables predictors to generate proxy scores calibrated against model-based scores. For example, under a well-calibrated model, if we collect a set of documents paired with summaries that receive a predicted conciseness score of 0.85, their average model-based conciseness score will also be 0.85. To this end, Manggala et al. (2024) proposed a group-wise calibration method for generative question answering systems, aiming to calibrate the confidence score of a question-answer pair to its final binary label, *i.e.*, whether the answer is correct. Their approach first clusters samples in the embedding space to perform adaptive binning, with the goal that each partition contains semantically similar question answer pairs that are "close" to each other in the embedding space. This design is motivated by the limitations of global calibrators, which can obscure important differences between content types. For example, two question-answer pairs may have similar raw model confidence scores (*e.g.*, 0.7), but one involves a factual geography question and the other a medical question; thus, a single global adjustment could miscalibrate one of them, whereas group-specific calibration can potentially correct

[1]The source code is available and can be accessed at https://github.com/Hyfred/Calibrating-Evaluation-Metrics.

Findings of the Association for Computational Linguistics: ACL 2026, pages 35285–35315

for such semantic differences. Within each cluster, the method bins the data by sorting its probabilistic predictions and then assigning the mean true label within each bin as the calibrated output. Although simple, this approach has the limitation that replacing all values in a bin with the mean creates a piecewise-constant function that may lose fine-grained information. This limitation suggests that alternative calibration strategies could be explored.

Moreover, Manggala et al. (2024) focus on calibrating discrete labels, such as binary correctness in QA. In contrast, tasks like summarization involve continuous metrics, *e.g.*, conciseness, where the goal is not merely to indicate whether a summary is concise, but to quantify how concise it is. However, in this context, the calibration for continuous metrics remains largely unexplored. In addition, if we wish to determine whether the summary of a document is above average relative to a set of summaries, a straightforward approach would involve generating multiple summary pairs and computing the average score across them. Here, the average is taken over summaries produced by different summarization systems. By comparing the score of the current summary with this average, we can assess whether it is better than average. However, generating multiple summaries and evaluating them repeatedly is computationally expensive. Therefore, beyond assessing the quality of a single summary, our goal also includes estimating both the score of an individual summary and the average score across multiple summaries in a single pass. This allows one to evaluate whether a given summary outperforms the average without accessing all reference summaries and to decide whether generating a higher scoring alternative is necessary.

We make the following contributions.

1. We propose a general framework that generates both individual and reliable *average* proxy scores in document summarization *in a single pass*, without needing reference summaries, human annotations, or expensive model-based metrics. Instead of iterating, the method directly produces reliable average scores, eliminating the need for multiple evaluation rounds.
2. We introduce *group isotonic regression binning* (GIRB), which performs group-level calibration to capture bias-causing semantic heterogeneity. Unlike traditional individual-based methods, GIRB adjusts the predictions based on group-wise monotonic trends, improving the consistency between predicted and true metrics.
3. Although our main focus is on summarization and other continuous-value tasks, the proposed framework is also applicable to discrete-value tasks like QA. Experiments across multiple pretrained models show that our method consistently outperforms existing baselines. We test our approach on both summarization and QA tasks, and find that, compared to other calibration methods, GIRB performs better across a variety of calibration error metrics.

2 Related Work

Summary Evaluation. Traditional summarization metrics, such as ROUGE, have long been used to evaluate summary quality. However, these metrics are often criticized for their poor correlation with human judgments on semantic fidelity and factuality. Recent advances have seen the emergence of LLM-based evaluators, which assess summary content at a finer granularity (Laban et al., 2023; Lu et al., 2023). For example, FineSurE (Song et al., 2024a) extracts atomic *keyfacts* from both the source and the summary, computing completeness and conciseness by aligning these keyfacts with the sentences in the summary. This results in interpretable dimension-specific scores. Meanwhile, UniSumEval (Lee et al., 2024) provides a unified and robust framework for benchmarking summarization systems that improves the precision and comprehensiveness of performance evaluations. Collectively, these works signal a shift from token-level to content-level evaluation, marking a significant step toward more reliable and actionable supervision in summarization. Our work builds on this trend by leveraging these evaluative scores as a ground truth signal for model calibration.

Calibration for Language Models. The calibration of LLMs has been a subject of growing interest, especially as LLMs continue to show remarkable capabilities in various tasks. However, LLMs often produce outputs with poorly calibrated confidence levels, which can be problematic for tasks that require accurate probabilistic estimates. The objective in reinforcement learning from human feedback (RLHF), *e.g.*, tends to prioritize adherence to user instructions in dialog over the need for well-calibrated predictions (Kadavath et al., 2022).

Liu et al. (2023); Winata et al. (2024) aim to develop evaluation metrics that better correlate with human judgments. Liu et al. (2023) improves alignment by modifying evaluation prompts dur-

ing score generation, while Winata et al. (2024) aggregates multiple metrics using a boosting-based approach. The latter assumes the availability of several metric scores, which leads to an increase in computational cost. Both works primarily assess performance using rank-based correlation measures such as Spearman's and Kendall's coefficients. Although both works use the term calibration, they do not adopt calibration in a formal statistical sense. Formal notions of calibration have been established in earlier work, including Dawid (1982); DeGroot and Fienberg (1983), as well as in recent studies on confidence calibration for LLMs.

Lin et al. (2022a) introduced the concept of verbalized confidence, which prompts LLMs to express their confidence directly, focusing on fine-tuning rather than zero-shot verbalized confidence. In contrast, Mielke et al. (2022) employed an external calibrator for white-box LLMs to adjust confidence levels. Other approaches focused on consistency measures to improve calibration (Lyu et al., 2024). The method proposed by Manggala et al. (2024) introduced a *post hoc* calibration technique that provides stronger calibration guarantees compared to previous approaches.

It is worth noting that prior work has largely studied calibration in discrete settings, such as question answering, where predictions are calibrated against binary labels, or used the term calibration without a formal definition. In contrast, we are the first to provide a formal definition of calibration tailored to the summarization task, together with a systematic study of calibration for summarization evaluation, where the target labels are continuous variables.

3 Methodology

In this section, we present our framework for generating calibrated proxy scores for summary evaluation. We first formalize the problem of predicting both individual and average quality scores across multiple evaluation dimensions without requiring reference summaries or expensive model-based metrics at inference time. We also motivate a *post hoc* calibration approach that produces reliable scores from a *lightweight* calibrator that preserves scalability. We then describe our method, group isotonic regression binning (GIRB), which calibrates the scores obtained from a LLM by performing clustering and group-based isotonic regression for *post hoc* calibration. Figure 2 provides an overview of the complete framework.

3.1 Problem Definition

Let $d \in \mathcal{D}$ be a source document, $s \in \mathcal{S}$ a generated summary, and for each document d we collect summaries $\{s^{(1)}, \ldots, s^{(N)}\}$ from different systems or LLMs. For each document-summary pair $(d, s^{(i)})$, we seek to evaluate each summary along K dimensions, *e.g.*, completeness, conciseness, faithfulness, *etc.* A direct way to do this is to use human judgments, which is expensive and time-consuming. To avoid this issue, we use model-based estimation.

Formally, let a scoring model produce a score $t_k^{(i)}$ for the summary $s^{(i)}$ on dimension k. We treat this score as the model-based ground-truth metric. For example, we can use FineSurE (Song et al., 2024a) for completeness, conciseness, and faithfulness, and UniEval (Lee et al., 2024) for consistency, fluency, and coherence. Other metrics such as BERTScore (Zhang et al., 2019) or BARTScore (Yuan et al., 2021) can also be used.

However, the problem is that the summarization models themselves do not produce evaluation scores and computing model-based scores can also be expensive, since they often rely on LLMs. Furthermore, deep neural networks are known to be miscalibrated (Guo et al., 2017; Parappan and Henao, 2026), and LLMs also inherit this limitation. Our goal is to design a simpler way to make a predictor h_k to generate a proxy score $y_k^{(i)}$ that is *calibrated* relative to the model-based ground-truth score $t_k^{(i)}$. For example, if a model is *perfectly* calibrated, of all summary-document pairs for which $y_k^{(i)} = 0.85$, their average FineSure score $t_k^{(i)}$ should also be 0.85. We use T, Y to denote the random variables corresponding to the model-based ground-truth score t and the proxy score y, respectively. Formally, a predictor $M : \mathcal{D} \times \mathcal{S} \to [0, 1]$ is calibrated if for all $p \in [0, 1]$, $\mathbb{E}[T \mid Y = p] = p$.

Importantly, we want to measure not only the quality of an individual summary but also whether it is better than the average summary quality. Given N summaries corresponding to one document, we can estimate both the individual score $t_k^{(i)}$ of the individual summary $s^{(i)}$ and the average score $\bar{t}_k$ for all N summaries. This allows one to determine whether a single summary is above or below average without needing access to all N summaries. Figure 1 illustrates judging the quality of an individual summary using the average score as a reference.

During inference, given a document-summary pair $(d, s^{(i)})$, our goal is to produce two types of

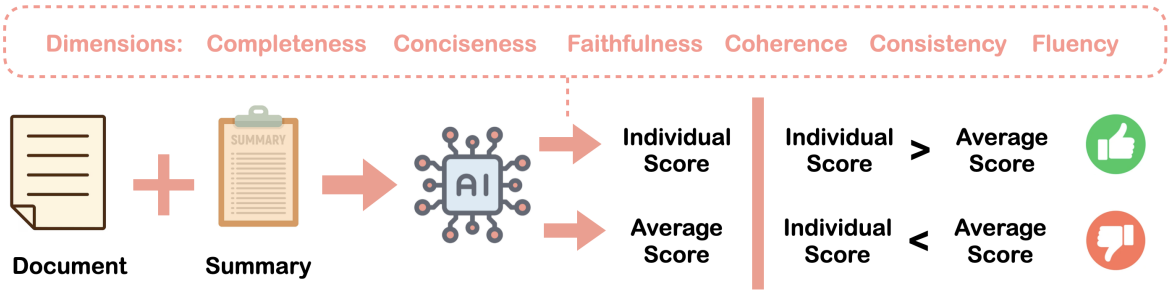

Figure 1: **Workflow Overview.** We input a document and its summary to produce two types of scores. The average score, reflecting the mean level across multiple summarization systems, serves as a reference. Comparing an individual score with this reference indicates summary quality: above average is good, below average is poor.

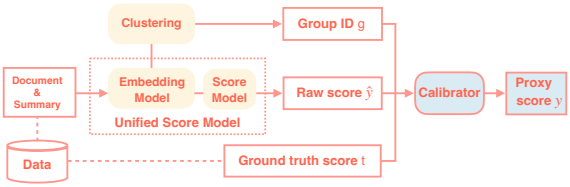

Figure 2: **Framework Overview.** A unified scoring model maps each document-summary pair to raw scores $\hat{y}$ via a shared embedding used for grouping. GIRB has two steps: i) embedding-space clustering to assign a group ID, and ii) *post hoc* calibration using the group ID, raw scores, and model-based ground truth.

proxy scores for K evaluation dimensions: the individual proxy score $y_k^{(i)}$ and the average proxy score $\bar{y}_k$, without requiring access to multiple summaries. Because there are K dimensions, this yields $2K$ target values per $(d, s^{(i)})$. This approach does not require multiple reference summaries, model-based metrics, or human annotations. The proxy scores serve as reliable stand-ins for true evaluation, providing both an estimate of the summary's quality and its relative standing compared to the expected average quality. This enables efficient and scalable evaluation of summaries in real-world settings.

3.2 Score Modeling and GIRB Calibration

Overview. For each document-summary pair $(d, s^{(i)})$ we can obtain model-based scores along K dimensions, which produce $2K$ values: one set of K individual scores $\{t_k^{(i)}\}_{k=1}^K$ and another set of K average scores $\{\bar{t}_k\}_{k=1}^K$. We propose the general framework in Figure 1 that generates both individual and *average* proxy scores in a *single pass*, without relying on reference summaries. Our objective is to train a model that, given only a single document and summary, can predict these $2K$ proxy scores. For simplicity, let $\hat{\mathbf{y}} \in \mathbb{R}^{2K}$ denote the raw multi-output predictions for one pair (d, s), where each component corresponds to either an individual or an average target for a specific dimension. Specifically, we introduce GIRB (groupwise isotonic regression with binning) as the combination of i) grouping through clustering on embeddings and ii) *post hoc* group-wise isotonic calibration applied to raw scores. GIRB calibrates the predictions of any score model, rather than being part of a unified score model itself.

Unified score model (embedding + score model). We treat the embedding encoder and score model as a single unified score model:

$$\hat{\mathbf{y}} = f_\theta(E_\psi(d, s)) = M_\Theta(d, s),$$

where $E_\psi(\cdot)$ is a pretrained or lightly fine-tuned embedding encoder (*e.g.*, a frozen LLM or sentence encoder), $f_\theta(\cdot)$ is a multi-output score model, and we use $M_\Theta(\cdot)$ to denote the unified score model. For ease of explanation, we describe them separately: *Embedding Model:* $E_\psi(d, s) \mapsto \mathbb{R}^m$, which produces a shared representation that is reused for both scoring and grouping. *Score Model:* A multi-output score model $f_\theta : \mathbb{R}^m \to \mathbb{R}^{2K}$, which can be instantiated with any function class; from simple linear mappings to expressive architectures (linear regression, MLP or transformer adapters).

Note that GIRB is not part of $M_\Theta(\cdot)$. It is a separate downstream model that post-calibrates its outputs (or that of any other score model) without changing its architecture or training objective.

Grouping via clustering on embeddings. We partition the embedding space into multiple clusters, grouping (d, s) pairs with similar semantic and structural properties. We define a fixed mapping $\phi : \mathcal{D} \times \mathcal{S} \to \mathcal{T}$ as $g = \phi(d, s)$, where $\mathcal{T}$ indexes clusters obtained by applying a clustering algorithm (*e.g.*, KMeans, KD-Tree) to $E_\psi(d, s)$. Thus, each (d, s) pair is assigned a group identifier $g = \phi(d, s)$. The same embedding feeding

the score model is used to assign groups, which ties modeling and calibration while keeping components modular. This grouping step is the first component of GIRB: it partitions the data into nearby regions in the embedding space, so each region can obtain its own simple, group-specific calibrator. We introduce clustering because the mapping from raw to calibrated score can differ across content types; thus a single global calibrator can over-correct some cases and under-correct others, whereas per-group mappings are poised to better match local patterns. Reusing the scoring embeddings for grouping keeps the system simple and avoids additional passes.

***Post hoc* group calibration.** Raw scores from a trained unified score model $M_\Theta(\cdot)$ can be miscalibrated, which means that predicted values may not correspond to the same expected ground truth value over the range of predictions; therefore, *post hoc* calibration aims to correct it without retraining $M_\Theta(\cdot)$. This calibration step is the second component of GIRB, namely within each group, we learn a mapping from the raw scores to the target scores so that the model predictions are better calibrated.

The *post hoc* calibration algorithm in Manggala et al. (2024) first sorts the prediction-target pairs according to their predicted values. Then it divides the sorted pairs into bins of equal size and uses the mean of the corresponding target values in each bin as the new calibrated value. This approach relies on a tunable hyperparameter, which is the number of points per bin, and produces a (non-smooth) piecewise constant function. The imputation of raw predictions with these bin-wise averages ensures that the predicted values match the observed outcomes on average within each bin. Consequently, the calibrated value of any prediction in a bin approximates the conditional expectation of the true target given the original prediction, satisfying the definition of calibration.

Instead of specifying fixed bins, we use isotonic regression (Barlow and Brunk, 1972) to learn a continuous piecewise linear function that increases monotonically relative to the raw scores. Isotonic regression adaptively merges adjacent violations to enforce monotonicity and thus, does not require a bin-size hyperparameter. For a group g and dimension k, given a set of n_g prediction-target pairs $(\hat{y}_k^{(j)}, t_k^{(j)})_{j=1}^{n_g}$ from a held-out calibration split, isotonic regression first sorts the pairs according to the predicted raw scores $\hat{y}_k^{(j)}$. Then it applies the pool-adjacent violators algorithm (PAVA) (De Leeuw et al., 2010), scanning the sequence from left to right to ensure that the targets $t_k^{(j)}$ are monotonic. Whenever a violation is detected, *i.e.*, $t_k^{(j)} > t_k^{(j+1)}$ for a non-decreasing fit, the adjacent points are merged into blocks and replaced with their weighted average. This process repeats until the entire sequence satisfies the monotonicity constraint. Then linear interpolation between adjacent blocks can be applied to construct a continuous isotonic function.

During inference, a new prediction score is mapped to the corresponding value on this isotonic function, resulting in the calibrated score. This means that isotonic regression fits a non-decreasing function $h : \mathbb{R} \to \mathbb{R}$ that maps uncalibrated predictions $\hat{y}$ to calibrated scores y. Under GIRB, we fit separate isotonic calibrators within each group: $y_k = h_{g,k}(\hat{y}_k)$, for $k = 1, \ldots, 2K$, where each $h_{g,k}(\cdot)$ is learned on a held-out calibration set restricted to group g. During inference, we compute the embedding, obtain $\hat{\mathbf{y}} = f_\theta(E_\psi(d, s))$, determine $g = \phi(d, s)$, and map each dimension k of $\hat{\mathbf{y}}$ using its corresponding calibrator $h_{g,k}$.

Remark. A global calibration approach can ignore important distinctions across different content types, as examples with similar raw confidence scores but different semantic characteristics may require different calibration adjustments. Group-wise calibration addresses this limitation through a two-stage process: first, it clusters examples in the embedding space to identify semantically similar groups, and then learns separate calibration mappings tailored to each group's specific characteristics. Consequently, the advantage of group-wise calibration is that it handles this local heterogeneity by first grouping the data via clustering in the embedding space and then learning a group-based mapping, rather than imposing a single global mapping. Since calibration acts on the output of a score model, GIRB is model-agnostic and can be used with any score model $f_\theta(\cdot)$; from simple linear models to more complex deep neural networks.

Training and inference. *Training:* Fit f_θ by minimizing a multi-output regression loss (*e.g.*, MSE) on $\hat{\mathbf{y}}$ against the $2K$ model-based target t per (d, s) pair; freeze ϕ after clustering using the training embeddings; then fit $h_{g,k}$ per group and per target on a held-out calibration split. *Inference:* Given one (d, s), compute $E_\psi(d, s)$, obtain $\hat{\mathbf{y}} = f_\theta(E_\psi(d, s))$, assign $g = \phi(d, s)$, and output

y by applying $h_{g,k}$ component-wise. This realizes both individual and average proxy scores in *single pass* without multiple reference summaries.

4 Experiments

Datasets. We evaluate the performance of our approach in two domains: summarization and question-answering (QA). For summarization, we use the FeedSum dataset (Song et al., 2024b), which covers documents from various domains. The dataset consists of model generated summaries from multiple systems, each annotated with fine-grained FineSurE scores (Song et al., 2024a) on three dimensions: completeness, conciseness, and faithfulness. To cover a wider range of evaluation dimensions, we further obtain consistency, fluency, and coherence using UniEval (Lee et al., 2024).

For QA, we follow Manggala et al. (2024) and evaluate in six QA datasets: SciQ (Welbl et al., 2017), TriviaQA (Joshi et al., 2017), MathQA (Amini et al., 2019), TruthfulQA (Lin et al., 2022b), MMLU (Hendrycks et al., 2021), and OpenBookQA (Mihaylov et al., 2018). Appendix Table 4 summarizes these datasets.

Baselines. In the summarization task, since our calibration targets are continuous-valued variables, calibration methods designed for discrete variables are not applicable, thus we compare to two basic settings: no calibration (None), and QA binning (QAB) (Manggala et al., 2024).

In the QA task, we consider the following baselines: no recalibration (None), histogram binning (B) (Gupta and Ramdas, 2021), Platt scaling (S) (Platt, 1999), scaling-binning (S-B) (Kumar et al., 2019), QA binning (QAB), scaling QA binning (S-QAB), hierarchical scaling QA binning (HS-QAB) (Manggala et al., 2024). These baselines constitute the state-of-the-art approaches in *post hoc* calibration.

Evaluation Metrics. For the summarization task, we employ several performance metrics.

Expected Calibration Error (ECE): $\mathrm{ECE} = \mathbb{E}_{D,S}[\,|\mathbb{E}[T \mid Y] - Y|\,]$, where T and Y denote ground-truth scores and proxy scores, respectively.

Calibration Slope: Given a set of ground-truth scores t and proxy scores y, we first sort y and partition it into B bins according to quantiles. For each bin $b = 1, \ldots, B$, we compute the average model score and the average proxy score, and then fit a linear regression on these aggregated pairs: the slope quantifies the degree of linear alignment between proxy scores and model-based scores, with a slope close to 1 indicating good calibration and deviations from 1 suggesting miscalibration.

Group-Conditional Expected Calibration Error: The applicability of average calibration methods such as ECE is limited, since they average over all input pairs regardless of domain; inputs can span diverse domains such as news, politics or medicine. To address this, we also adopt group-based calibration metrics following Manggala et al. (2024), defining the group conditional expected calibration error (GECE) as $\mathrm{GECE} = \mathbb{E}_{D,S}[\,|\mathbb{E}[T \mid Y, \phi(D,S)] - Y|\,]$, where $\phi(D,S)$ denotes the group assignment function learned during training.

Brier Score: We also use the popular Brier score (Brier, 1950), which measures the accuracy of probabilistic predictions; although it can be decomposed into calibration and refinement components (Blattenberger and Lad, 1985), it does not directly assess calibration quality, so models with low squared error may still be poorly calibrated.

Accuracy: Our model predicts two types of scores: an individual and an average. We quantify whether the individual score is higher or lower than the average score, both in prediction and ground truth, thus framing it as a binary classification problem and reporting accuracy.

For QA tasks, we adopt the same calibration metrics (ECE, GECE, and Brier). We also evaluate selective QA performance using the Area Under the Accuracy-Confidence Curve (AUAC) (El-Yaniv and Wiener, 2010), where the x axis corresponds to confidence thresholds $\tau \in [0, 1]$, the y axis is the accuracy of predictions with confidence exceeding τ, and AUAC is defined as the area under this curve; a higher AUAC indicates that confidence scores better distinguish correct from incorrect answers, enabling more effective selective QA.

To make comparisons more intuitive (*larger values indicate better performance*), we transform the evaluation metrics into their higher-is-better variants. Thus, in Tables 1, 2, the summary Tables in Appendices C.2 and D.2, Figures 3 and 4, and the Figures in Appendices D.4 and C.4, we report the following transformed metrics: $\mathrm{ECE}' = 1-\mathrm{ECE}$, $\mathrm{GECE}' = 1-\mathrm{GECE}$, $\mathrm{Brier}' = 1-\mathrm{Brier}$, $\mathrm{Slope}' = 1 - |1-\mathrm{Slope}|$. Slope' attains 1 when $\mathrm{Slope} = 1$ and decreases symmetrically as Slope departs from 1. For each evaluation metric, we use (Ind) to denote the performance based on individual scores and (Avg) to denote the performance based on average scores.

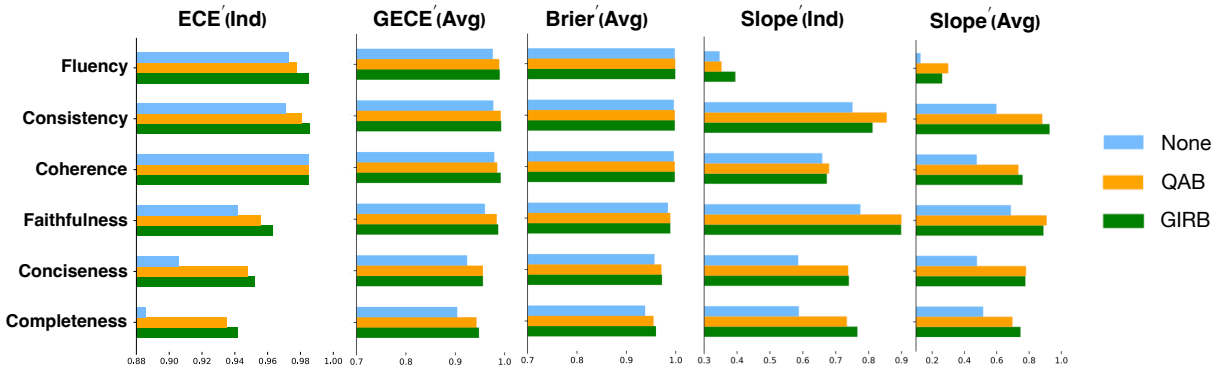

Figure 3: Results comparing calibration methods across dimensions and metrics. The left two plots show cases where our method achieves its best performance relative to others in terms of number of wins. The middle plot represent ties between our method and other methods. The right two plots show cases where our method performs slightly worse than its best case, yet still outperforms other methods.

Method	None		QAB		GIRB	
Metric	Gain	Wins	Gain	Wins	Gain	Wins
ECE′(Ind)	-	1	+0.022	1	**+0.027**	**6**
GECE′(Ind)	-	0	+0.028	2	**+0.030**	**5**
Brier′(Ind)	-	1	+0.005	2	**+0.008**	**5**
Slope′(Ind)	-	0	+0.147	**3**	**+0.161**	**3**
Accuracy (Ind)	-	**3**	-0.043	0	**+0.017**	**3**
ECE′(Avg)	-	0	+0.033	4	**+0.035**	**6**
GECE′(Avg)	-	0	+0.023	1	**+0.026**	**6**
Brier′(Avg)	-	0	+0.007	4	**+0.008**	**6**
Slope′(Avg)	-	0	**+0.612**	**3**	+0.594	**3**
Mean	-	0.56	+0.093	2.22	**+0.101**	**4.78**

Table 1: Summary of performance across calibration methods for each metric. For each metric, we report two indicators. **Gain** is the average improvement compared to "None", and **Wins** is the number of dimensions where a given method outperforms the others. In cases of tied first place, all top-ranking methods are counted as wins.

Training. For the summarization task dataset, we split the dataset into four subsets with a ratio of 30:30:30:10. The first subset is used to construct the clustering algorithm (KMeans), the second subset is used to train a linear regression model to predict raw scores, the third subset is reserved for post-hoc calibration training, and the fourth subset is used for testing. For each document-summary pair, we first obtain embedding representations using the Qwen3-8B embedding model (Zhang et al., 2025), following the `Instruct-DS-Prompt` format described in Appendix C.1. These embeddings are then fed into a multi-output linear regression layer to predict $2K$ dimensional model-based scores. We observed empirically that a simple linear layer performed similarly to more complex layers. The regression model is trained with batch size 500, learning rate 1×10^{-3}, and for 50 epochs.

For the QA task, in order to compare with the other baseline results, we strictly followed the settings of the environment and the hyper-parameters in Manggala et al. (2024). Following Manggala et al. (2024), we use KD-Tree for the clustering in GIRB. Note that unlike summarization task, we do not need to use the score model in QA task, and instead we use LLMs (Gemma and Mistral) with different prompts (`ling1s-topk` and `verb1s-topk`) to produce the scores for assessing the confidence of a given QA pair. We provide more details about other hyperparameters in Appendix A.

Summarization Results. Figure 3 illustrates the performance of different calibration methods over multiple evaluation metrics and dimensions. These visualizations allow direct comparison of each method under various settings.

To quantitatively compare the relative strengths of each method, we compute two indicators for each metric: the number of dimensions where a given method outperforms the others (Wins), and the corresponding relative improvement (Gain). In the summary Table 1, "Wins" indicates the number of dimensions where a method achieves the best performance. For example, for the ECE′ metric, if GIRB achieves the top performance in four dimensions, the corresponding cell shows "4". The main value in each cell represents the average relative improvement over all dimensions, calculated as $\frac{y_k - \hat{y}_k}{\hat{y}_k}$, where $\hat{y}_k$ is the uncalibrated score in dimension k, and y_k is the calibrated score. Positive values indicate improvements over the baseline, while negative values mean performance drops, summarizing each method's overall effectiveness.

In Figure 3, we show the two best performing metrics on the left, the two relatively weaker metrics on the right, and a middle metric between them. The ordering of good versus poor performance is

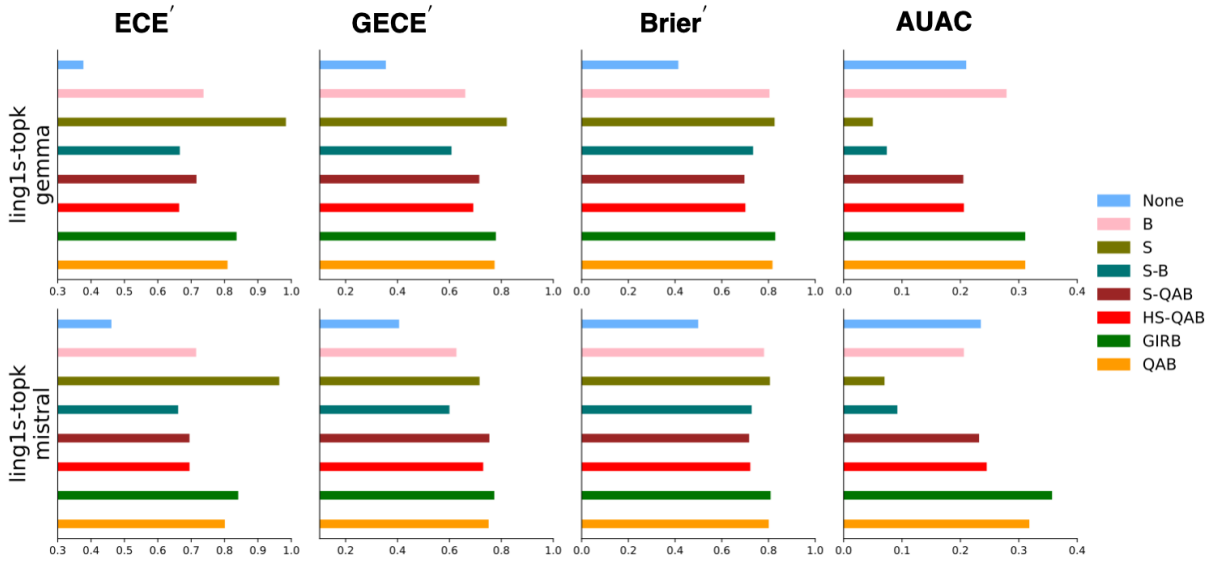

Figure 4: Results comparing calibration methods over all metrics in the MMLU dataset.

determined by the "Wins" in Table 1. As shown, QAB and GIRB outperform the uncalibrated baseline, consistently achieving higher scores on the conciseness and completeness dimensions.

In Table 1, GIRB achieves the highest number of wins in almost all metrics, except Slope′(Ind), Slope′(Avg), and Accuracy, where it ties with other methods. On average, GIRB wins 4.78 times, more than double that of the second-best method, QAB, which achieves 2.22.

By combining the bar plots with the summary table, we provide both qualitative and quantitative perspectives: the figures visualize dimension-level performance for each metric, while the table summarizes wins and average improvement. These complementary analyses offer a comprehensive understanding of the effectiveness of each calibration method in different evaluation criteria. Detailed results for all settings are provided in Appendix C.

Ablation Study. We denote a setting as a combination of embedding model, prompting strategy and clustering method. In the main results, we use Qwen, Prompt and KMeans, respectively. We perform an ablation study by varying one component at a time while keeping the others fixed to assess its impact. We conducted experiments with different clustering methods, prompting strategies and embedding models, as summarized in Appendix Tables 5, 6, and 7, respectively.

From these results, we observe that: i) Over all alternative settings, our method consistently outperforms the others in terms of the average number of wins. ii) Although the average Gain is not a strictly meaningful metric, as it averages values on different scales, our method (GIRB) still achieves the best overall performance, except in the KD-Tree setting. iii) When examining the contribution of individual components, the average number of wins follows the ranking: KMeans > KD-Tree, Instruct-DS-Prompt > DS-Prompt, and Qwen3-8B > DistilBERT > Gemma-2B, suggesting that our default configuration represents the optimal combination. Importantly, DistilBERT is not a large generative LLM but a lightweight embedding-based model. Its competitive performance indicates that our method does not strictly depend on large-scale LLMs, providing empirical evidence on the degree of dependence on embedding model size and demonstrating that the proposed approach remains effective even with substantially smaller models.

QA Results. We further validate our approach in a QA task, following the setup described in Manggala et al. (2024). Based on the questions and reference answers in the dataset, we use LLMs (Gemma and Mistral) to generate answers and associated confidence scores using specific prompts (`ling1s-topk` and `verb1s-topk`). Complete prompt specifications are provided in Appendix D.1. Subsequently, we employ another large language model, LLaMA, to compare the generated answers with the reference answers and determine their correctness, 0 or 1.

This study compares the effectiveness of different calibration methods in aligning confidence scores with answer correctness. To better illustrate the performance of each calibration method

	None		S-B		S		B		QAB		HS-QAB		S-QAB		GIRB	
Metric	Gain	Wins	Gain	Wins	Gain	Wins	Gain	Wins	Gain	Wins	Gain	Wins	Gain	Wins	Gain	Wins
ECE′	0	0	+0.695	0	**+1.424**	**4**	+0.864	0	+1.061	0	+0.739	0	+0.780	0	+1.145	0
GECE′	0	0	+0.654	0	+1.077	1	+0.814	0	+1.130	0	+0.968	0	+1.030	0	**+1.167**	**3**
Brier′	0	0	+0.700	0	+0.889	0	+0.845	0	+0.873	0	+0.648	0	+0.642	0	**+0.901**	**4**
AUAC	0	0	-0.606	0	-0.703	0	+0.270	0	+0.429	1	+0.039	0	+0.037	0	**+0.531**	**3**
Mean	0	0	+0.361	0	+0.672	0.2	+0.698	0	+0.873	0.2	+0.599	0	+0.622	0	**+0.936**	**2.5**

Table 2: Summary of performance across calibration methods for each metric in MMLU dataset. For each metric, we report two indicators: **Gain** is the average improvement compared to "None", and **Wins** is the number of settings (*e.g.*, ling1s-topk) where a given method outperforms the others. In cases of tied first place, all top-ranking methods are counted as wins.

Calibrator	ECE′	GECE′	Brier′
None	0.983	0.983	0.948
QAB	0.990	0.990	**0.953**
GIRB	**0.993**	**0.993**	0.940

Table 3: Performance of different calibration methods in human evaluation dataset.

over multiple evaluation metrics and settings, we present bar plots that allow for direct visual comparison. Figure 4 shows results for two setting, where answers and confidence scores were generated by Mistral and Gemma using `ling1s-topk` prompt. The complete set of results is provided in Appendix Figure 7. As shown, GIRB achieves the best performance in all metrics except ECE, where scaling binning (S-B) performs the best, and our method ranks second.

Following the same approach as in Table 1, Table 2 summarizes "Wins" and "Gains" indicating the number of settings in which a method achieves the best performance, along with the corresponding improvement magnitude. Considering the average number of wins across metrics, GIRB achieves 2.5, much higher than the second-best at 0.2. In terms of average gain, GIRB obtains 0.936, also surpassing the second-best 0.873. We provide detailed results for various settings in Appendix D.

Human Study. To further evaluate the effectiveness of the proposed calibration method, we leverage datasets with human annotations to examine whether the calibrated scores better align with human judgments. Specifically, we adopt the UniSumEval benchmark (Lee et al., 2024), which provides document-summary pairs annotated with both human evaluation scores and model-based scores. In our experiments, we use human faithfulness scores as ground truth score and calibrate model-based G-Eval+ scores using both QAB and GIRB. As shown in Table 3, calibrated scores provide a better proxy for human judgments, achieving higher ECE′, GECE′, and Brier′scores than without calibration. Moreover, GIRB consistently outperforms QAB in terms of ECE and GECE. Overall, these results demonstrate that our calibration method yields stronger agreement with human ratings and can be effectively generalized to calibrate the model-based scores using human annotations as the ground truth.

5 Discussion

The primary contribution of this work is the introduction of the GIRB framework, which addresses the persistent issue of miscalibration in model-based evaluation metrics. Our results show that GIRB enables reliable estimation of both individual-level summary quality scores and average-level summary quality, without requiring expensive reference summaries or human annotations. This capability has important practical implications for the deployment of LLMs in real-world settings, where quality assessment must be performed efficiently and at scale. In particular, GIRB allows systems to determine whether generated outputs meet predefined quality thresholds in a single, cost-effective pass.

Another finding is that calibration improves when semantic structure is incorporated through grouping. By clustering the embedding space, GIRB accounts for differences across content types (*e.g.*, news versus medical text). This approach breaks a complex global calibration problem into simpler local ones, allowing the model to better capture group-specific biases and improve alignment between proxy scores and ground-truth metrics.

Furthermore, isotonic regression offers a simple and effective alternative to prior binning-based methods such as QAB. Instead of relying on fixed-width bins and manual tuning, it learns a continuous and monotonic mapping from proxy scores to calibrated values. This leads to a smoother calibration function and better performance in continuous settings such as summarization. The consistent gains across datasets suggest that GIRB is both effective and robust, making it a practical solution for calibration in model-based evaluation.

Limitations

Despite these promising results, GIRB has several limitations. First, its performance depends on the choice of the clustering algorithm, as the quality of group assignments directly affects calibration. Furthermore, the effectiveness of the grouping stage is inherently tied to the representational power of the embedding model. If the embeddings fail to capture the semantic nuances required to distinguish between different bias distributions, the resulting calibration improvements may be limited. Second, the current calibration procedure relies on a sufficient amount of post-calibration data, which may not always be available in practice. This limitation highlights the need for more data-efficient calibration methods. Future work may also explore more adaptive or learned grouping strategies, as well as extend the proposed approach to other modalities and evaluation settings.

References

Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Richard E Barlow and Hugh D Brunk. 1972. The isotonic regression problem and its dual. *Journal of the American Statistical Association*, 67(337):140–147.

Gail Blattenberger and Frank Lad. 1985. Separating the brier score into calibration and refinement components: A graphical exposition. *The American Statistician*, 39(1):26–32.

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

A Philip Dawid. 1982. The well-calibrated bayesian. *Journal of the American statistical Association*, 77(379):605–610.

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

Morris H DeGroot and Stephen E Fienberg. 1983. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 32(1-2):12–22.

Ran El-Yaniv and Yair Wiener. 2010. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5).

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR.

Chirag Gupta and Aaditya Ramdas. 2021. Distribution-free calibration guarantees for histogram binning without sample splitting. In *International Conference on Machine Learning*, pages 3942–3952. PMLR.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, and Bing Qin. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv:2311.05232*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, and Eli Tran-Johnson. 2022. Language models (mostly) know what they know. *arXiv:2207.05221*.

Ananya Kumar, Percy S Liang, and Tengyu Ma. 2019. Verified uncertainty calibration. *Advances in Neural Information Processing Systems*, 32.

Philippe Laban, Wojciech Kryściński, Divyansh Agarwal, Alexander Richard Fabbri, Caiming Xiong, Shafiq Joty, and Chien-Sheng Wu. 2023. Summedits: Measuring llm ability at factual reasoning through the lens of summarization. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9662–9676.

Tian Lan, Jiang Li, Yemin Wang, Xu Liu, Xiangdong Su, and Guanglai Gao. 2025a. F^2bench: An open-ended fairness evaluation benchmark for llms with factuality considerations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2031–2046.

Tian Lan, Xiangdong Su, Xu Liu, Ruirui Wang, Ke Chang, Jiang Li, and Guanglai Gao. 2025b. Mcbe: A multi-task chinese bias evaluation benchmark for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6033–6056.

Yuho Lee, Taewon Yun, Jason Cai, Hang Su, and Hwanjun Song. 2024. Unisumeval: Towards unified, fine-grained, multi-dimensional summarization evaluation for llms. *arXiv preprint arXiv:2409.19898*.

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

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022b. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Hongye Liu and Ricardo Henao. 2025. Learning to substitute words with model-based score ranking. *arXiv preprint arXiv:2502.05933*.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2023. Calibrating llm-based evaluator. *arXiv preprint arXiv:2309.13308*.

Qingyu Lu, Liang Ding, Liping Xie, Kanjian Zhang, Derek F Wong, and Dacheng Tao. 2023. Toward human-like evaluation for natural language generation with error analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5892–5907.

Guanran Luo, Zhongquan Jian, Wentao Qiu, Meihong Wang, and Qingqiang Wu. 2025. DTCRS: Dynamic tree construction for recursive summarization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10948–10963, Vienna, Austria. Association for Computational Linguistics.

Guanran Luo, Wentao Qiu, Zhongquan Jian, Meihong Wang, and Qingqiang Wu. 2026. Gcot-decoding: Unlocking deep reasoning paths for universal question answering. *Preprint*, arXiv:2604.06794.

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. 2024. Calibrating large language models with sample consistency. *arXiv:2402.13904*.

Putra Manggala, Atalanti Mastakouri, Elke Kirschbaum, Shiva Prasad Kasiviswanathan, and Aaditya Ramdas. 2024. Qa-calibration of language model confidence scores. *arXiv preprint arXiv:2410.06615*.

Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Mohammed Fayiz Parappan and Ricardo Henao. 2026. Labels have human values: Value calibration of subjective tasks. *arXiv preprint arXiv:2601.06631*.

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

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024a. Finesure: Fine-grained summarization evaluation using llms. *arXiv preprint arXiv:2407.00908*.

Hwanjun Song, Taewon Yun, Yuho Lee, Jihwan Oh, Gihun Lee, Jason Cai, and Hang Su. 2024b. Learning to summarize from llm-generated feedback. *arXiv preprint arXiv:2410.13116*.

Yu Wang, Emmanuele Chersoni, and Chu-Ren Huang. 2026a. This one or that one? a study on accessibility via demonstratives with multimodal large language models. In *Language Resources and Evaluation Conference 2026*. European Language Resources Association (ELRA).

Zi-Han Wang, Lam Nguyen, Zhengyang Zhao, Mengyue Yang, Chengwei Qin, Yujiu Yang, and Linyi Yang. 2026b. Creativebench: Benchmarking and enhancing machine creativity via self-evolving challenges. *arXiv preprint arXiv:2603.11863*.

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*.

Genta Indra Winata, David Anugraha, Lucky Susanto, Garry Kuwanto, and Derry Tanti Wijaya. 2024. Metametrics: Calibrating metrics for generation tasks using human preferences. *arXiv preprint arXiv:2410.02381*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Shiyue Zhang and Mohit Bansal. 2021. Finding a balanced degree of automation for summary evaluation. *arXiv preprint arXiv:2109.11503*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A Model Hyperparameters

For calibration analysis, we used quantile binning with 10 equal-frequency bins to compute ECE and Slope. In the summarization task, we employed KD-Tree with a maximum depth of 8 and KMeans with 256 clusters. The large language embedding models used in this task were Qwen-3-8B and Gemma-2B. In the QA task, we conducted a hyperparameter search over KD-Tree maximum depth values ranging from 0 to 10, using Mistral-7B and Gemma-7B models to produce score, and using LLaMA2-7B to produce the ground truth correctness label.

B Dataset Details

Table 4 provides the details about the summarization and QA datasets.

Table 4: Dataset summary.

Dataset	Size	Type
FeedSum	76945	Summarization
SciQ	11609	QA
TriviaQA	11313	QA
MathQA	29837	QA
TruthfulQA	790	QA
MMLU	20320	Multiple-choice QA
OpenBookQA	13869	Multiple-choice QA

C Summarization Task

C.1 Prompt for Obtaining Embedding in Document Summarization Task

We adopt two approaches to obtain embeddings from the document and its summary. The first approach directly concatenates the document and summary pair, which we refer to as `DS-Prompt`. The second approach simulates prompting a large language model, which we refer to as `Instruct-DS-Prompt`.

`DS-Prompt`

Document: <document>

Summary: <summary>

`Instruct-DS-Prompt`

Given a document and its summary, compute a score that estimates the summary's quality.

Document: <document>

Summary: <summary>

C.2 Summary Tables

This section presents the additional summary table results for the summarization task. For each metric, we report "Gain" (average improvement over the uncalibrated baseline) and "Wins" (number of dimensions where the method outperforms others), enabling comparison of calibration effectiveness across prompting strategies. Table 5 presents a summary of performance across calibration methods for each metric using KMeans and KD-Tree. Table 6 summarizes calibration method performance using `DS-Prompt` and `Instruct-DS-Prompt` settings. Table 7 reports the summary of performance across calibration methods using different embedding model settings.

C.3 Complete Results Tables

In this section, we present the complete set of results of all the experiments for the summarization task. We provide a comprehensive performance comparison of multiple calibration methods across various settings using the FeedSum dataset for the summarization task in Table 8. The performance comparison of different clustering methods and calibration methods across various dimensions is provided in Table 9. We compare the performance of different embedding models and calibration methods across various dimensions in Table 10. Table 11 shows the performance comparison of different prompting strategies and calibration methods across various dimensions.

C.4 Bar Plot Figures

Figure 5 provides a comparison of different calibration methods across different dimensions for various evaluation metrics.

D Question and Answering (QA) Task

D.1 Prompt for Obtaining Answer and Confidence in Question Answering task

We present the prompt templates used to elicit answers and confidence scores from the language models in our experiments. We employ two prompting strategies to extract confidence estimates alongside model responses: `verb1s-topk` and `ling1s-topk`.

`verb1s-topk`

Provide your best guess and the probability that it is correct (from 0.0 to 1.0) for the following question. Give only the guess and probability — no other words or explanations.

Table 5: Summary of performance across calibration methods for each metric. For each metric, we report two indicators: i. Gain: average improvement compared to "None", and ii. Wins: the number of dimensions where a given method outperforms the others using KMean and KD-Tree setting.

KMean							KD-Tree						
Method	None		QAB		GIRB		Method	None		QAB		GIRB	
Metric	Gain	Wins	Gain	Wins	Gain	Wins	Metric	Gain	Wins	Gain	Wins	Gain	Wins
ECE′(Ind)	-	1	+0.022	1	**+0.027**	**6**	ECE′(Ind)	-	1	+0.022	3	**+0.023**	**4**
GECE′(Ind)	-	0	+0.028	2	**+0.030**	**5**	GECE′(Ind)	-	0	+0.026	0	**+0.030**	**6**
Brier′(Ind)	-	1	+0.005	2	**+0.008**	**5**	Brier′(Ind)	-	2	+0.001	3	**+0.002**	**4**
Slope′(Ind)	-	0	+0.147	**3**	**+0.161**	**3**	Slope′(Ind)	-	1	**+0.255**	**4**	+0.126	1
Accuracy	-	**3**	-0.043	0	**+0.017**	**3**	Accuracy	-	1	+0.030	0	**+0.067**	**5**
ECE′(Avg)	-	0	+0.033	4	**+0.035**	**6**	ECE′(Avg)	-	0	+0.032	1	**+0.036**	**5**
GECE′(Avg)	-	0	+0.023	1	**+0.026**	**6**	GECE′(Avg)	-	0	+0.029	2	**+0.031**	**5**
Brier′(Avg)	-	0	+0.007	4	**+0.008**	**6**	Brier′(Avg)	-	0	+0.006	3	**+0.007**	**6**
Slope′(Avg)	-	0	**+0.612**	**3**	+0.594	**3**	Slope′(Avg)	-	0	**+1.075**	**6**	+0.899	0
Mean	-	0.56	+0.093	2.22	**+0.101**	**4.78**	Mean	-	0.56	**+0.164**	2.44	+0.136	**4**

Table 6: Summary of performance across calibration methods for each metric using either `DS-Prompt` or `Instruct-DS-Prompt` setting. For each metric, we report two indicators: i. Gain: average improvement compared to "None", and ii. Wins: the number of dimensions where a given method outperforms the others.

Instruct-DS-Prompt							DS-Prompt						
Method	None		QAB		GIRB		Method	None		QAB		GIRB	
Metric	Gain	Wins	Gain	Wins	Gain	Wins	Metric	Gain	Wins	Gain	Wins	Gain	Wins
ECE′(Ind)	-	1	+0.022	1	**+0.027**	**6**	ECE′(Ind)	-	0	**+0.039**	**4**	**+0.035**	**3**
GECE′(Ind)	-	0	+0.028	2	**+0.030**	**5**	GECE′(Ind)	-	0	**+0.043**	**3**	**+0.042**	**3**
Brier′(Ind)	-	1	+0.005	2	**+0.008**	**5**	Brier′(Ind)	-	1	+0.010	1	**+0.011**	**4**
Slope′(Ind)	-	0	+0.147	**3**	**+0.161**	**3**	Slope′(Ind)	-	0	**+0.556**	**5**	+0.443	1
Accuracy	-	**3**	-0.043	0	**+0.017**	**3**	Accuracy	-	**3**	-0.072	0	**+0.035**	**3**
ECE′(Avg)	-	0	+0.033	4	**+0.035**	**6**	ECE′(Avg)	-	0	+0.052	0	**+0.057**	**6**
GECE′(Avg)	-	0	+0.023	1	**+0.026**	**6**	GECE′(Avg)	-	0	+0.041	0	**+0.046**	**6**
Brier′(Avg)	-	0	+0.007	4	**+0.008**	**6**	Brier′(Avg)	-	0	+0.015	4	**+0.016**	**6**
Slope′(Avg)	-	0	**+0.612**	**3**	+0.594	**3**	Slope′(Avg)	-	0	**-5.673**	**5**	**-4.431**	1
Mean	-	0.56	+0.093	2.22	**+0.101**	**4.78**	Mean	-	0.44	-0.554	2.44	**-0.416**	**3.67**

Example: Guess: <most likely guess, as short as possible; not a complete sentence, just the guess!>
Probability: <a number between 0.0 and 1.0 representing how likely your guess is correct, with no extra commentary>.
The question is: <q>.

`ling1s-topk`

Provide your best guess for the following question, and indicate how confident you are using one of the following expressions: $EXPRESSION_LIST. Give only the guess and confidence — no other words or explanations.
Example: Guess: <most likely guess, as short as possible; not a complete sentence, just the guess!>
Confidence: <one short expression from the list, with no extra commentary>.
The question is: <q>.

D.2 Summary Results Tables

Tables 12, 13, 14, 15, and 16 present comprehensive summaries of calibration method performance across the MathQA, OpenBookQA, SciQ, TriviaQA, and TruthfulQA datasets, respectively. Each table reports two key performance indicators for every calibration metric: "Gain", which measures the average improvement relative to the uncalibrated baseline ("None"), and "Wins", which counts the number of dimensions in which a given calibration method achieves the best performance among all compared approaches. These summaries enable systematic comparison of calibration techniques across different datasets and identify which methods consistently deliver superior calibration performance. We observe that GIRB either outperforms other calibration approaches or performs comparably in terms of both "Wins" and "Gain" across all the datasets.

D.3 Complete Results Tables

Tables 17, 18, 19, 20, 21, and 22 provide detailed results of calibration method performance across various experimental settings for the MathQA, MMLU, OpenBookQA, SciQ, TriviaQA, and TruthfulQA datasets, respectively. Currently, there is no systematic exploration of isotonic regression models on QA datasets. To address this gap, we contribute an ablation study that investigates different variants of isotonic regression, including a version without group-level calibration (IRB). Moreover, inspired by the logic of S-QAB and HS-QAB in Manggala

Table 7: Summary of performance across calibration methods for each metric. For each metric, we report two indicators: i. Gain: average improvement compared to "None", and ii. Wins: the number of dimensions where a given method outperforms the others using different embedding model setting.

Qwen-8B							DistilBert							Gemma-2B						
Method	None		QAB		GIRB		Method	None		QAB		GIRB		Method	None		QAB		GIRB	
Metric	Gain	Wins	Gain	Wins	Gain	Wins	Metric	Gain	Wins	Gain	Wins	Gain	Wins	Metric	Gain	Wins	Gain	Wins	Gain	Wins
ECE′(Ind)	-	1	+0.022	1	**+0.027**	**6**	ECE′(Ind)	-	0	**+0.028**	**2**	**+0.029**	**5**	ECE′(Ind)	-	1	+0.015	2	**+0.018**	**4**
GECE′(Ind)	-	0	+0.028	2	**+0.030**	**5**	GECE′(Ind)	-	0	**+0.029**	**1**	**+0.032**	**5**	GECE′(Ind)	-	0	+0.023	2	**+0.026**	**6**
Brier′(Ind)	-	1	+0.005	2	**+0.008**	**5**	Brier′(Ind)	-	1	+0.006	**4**	**+0.008**	**4**	Brier′(Ind)	-	2	**+0.009**	**3**	**+0.009**	2
Slope′(Ind)	-	0	+0.147	**3**	**+0.161**	**3**	Slope′(Ind)	-	1	**+0.316**	**3**	+0.289	2	Slope′(Ind)	-	1	+0.077	**4**	**+0.089**	1
Accuracy	-	**3**	-0.043	0	**+0.017**	**3**	Accuracy	-	**1**	-0.043	0	**+0.033**	**5**	Accuracy	-	**4**	-0.096	0	**-0.013**	2
ECE′(Avg)	-	0	+0.033	4	**+0.035**	**6**	ECE′(Avg)	-	0	+0.041	3	**+0.042**	**5**	ECE′(Avg)	-	0	+0.039	1	**+0.042**	**6**
GECE′(Avg)	-	0	+0.023	1	**+0.026**	**6**	GECE′(Avg)	-	0	+0.033	1	**+0.036**	**6**	GECE′(Avg)	-	0	+0.028	1	**+0.031**	**6**
Brier′(Avg)	-	0	+0.007	4	**+0.008**	**6**	Brier′(Avg)	-	0	**+0.013**	3	**+0.013**	**5**	Brier′(Avg)	-	1	+0.012	3	**+0.013**	**6**
Slope′(Avg)	-	0	**+0.612**	**3**	+0.594	**3**	Slope′(Avg)	-	0	**+0.905**	**4**	**+0.887**	2	Slope′(Avg)	-	0	+0.731	**3**	**+0.738**	**3**
Mean	-	0.56	+0.093	2.22	**+0.101**	**4.78**	Mean	-	0.33	+0.148	2.33	**+0.152**	**4.33**	Mean	-	1	+0.093	2.11	**+0.106**	**4**

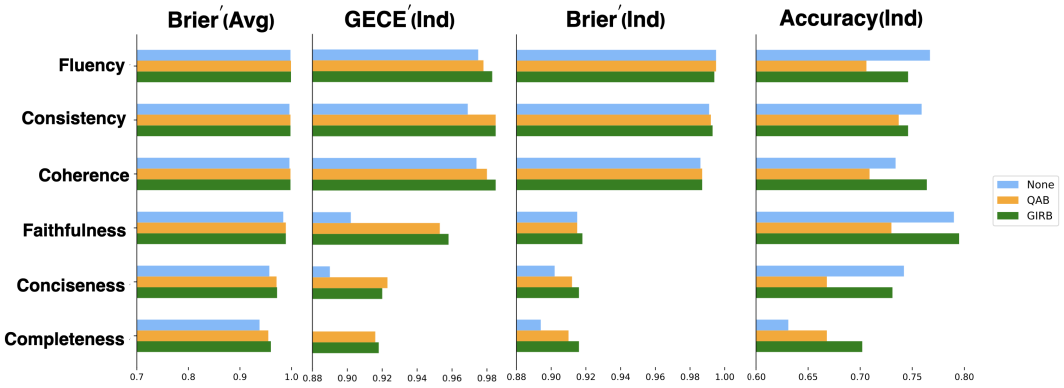

Figure 5: Bar Plot Comparison of Calibration Methods Across Dimensions and Metrics

et al. (2024), we derive the corresponding group-based versions, S-GIRB and HS-GIRB. These tables present the full range of performance metrics for each calibration approach under different configurations (LLM, prompting strategies), offering a comprehensive view of method effectiveness and variability. Together, they allow for in-depth comparison and analysis of calibration strategies across multiple benchmarks.

D.4 Bar Plot Figures

In this section, we present the bar plots in Figures 6, 7, 8, 9, 10, and 11 to show the comparison of the calibration method performance across the MathQA, MMLU, OpenBookQA, SciQ, TriviaQA, and TruthfulQA datasets, respectively.

Table 8: Performance of different calibration methods across various setting in FeedSum dataset.

Dimension	Calibrator	ECE	GECE	Brier	Slope	Accuracy
Completeness (Ind)	None	0.114 ± 0.034	0.125 ± 0.24	0.106 ± 0.040	0.588 ± 0.276	0.631 ± 0.122
Completeness (Ind)	QAB	0.065 ± 0.026	0.084 ± 0.022	0.090 ± 0.017	0.734 ± 0.188	0.668 ± 0.108
Completeness (Ind)	GIRB	$\mathbf{0.058 \pm 0.012}$	$\mathbf{0.082 \pm 0.031}$	$\mathbf{0.084 \pm 0.020}$	$\mathbf{0.766 \pm 0.311}$	$\mathbf{0.702 \pm 0.139}$
Conciseness (Ind)	None	0.094 ± 0.025	0.11 ± 0.023	0.098 ± 0.038	0.586 ± 0.274	$\mathbf{0.742 \pm 0.145}$
Conciseness (Ind)	QAB	0.052 ± 0.024	$\mathbf{0.077 \pm 0.021}$	0.088 ± 0.018	0.738 ± 0.265	0.668 ± 0.118
Conciseness (Ind)	GIRB	$\mathbf{0.048 \pm 0.023}$	0.080 ± 0.017	$\mathbf{0.084 \pm 0.020}$	$\mathbf{0.740 \pm 0.150}$	0.731 ± 0.119
Faithfulness (Ind)	None	0.058 ± 0.020	0.098 ± 0.25	0.085 ± 0.017	0.775 ± 0.267	0.790 ± 0.154
Faithfulness (Ind)	QAB	0.044 ± 0.012	0.047 ± 0.024	0.085 ± 0.029	$\mathbf{0.921 \pm 0.159}$	0.730 ± 0.089
Faithfulness (Ind)	GIRB	$\mathbf{0.037 \pm 0.017}$	$\mathbf{0.042 \pm 0.011}$	$\mathbf{0.082 \pm 0.028}$	0.899 ± 0.117	$\mathbf{0.795 \pm 0.118}$
Coherence (Ind)	None	$\mathbf{0.015 \pm 0.011}$	0.026 ± 0.022	0.014 ± 0.021	0.659 ± 0.225	0.734 ± 0.086
Coherence (Ind)	QAB	$\mathbf{0.015 \pm 0.010}$	0.020 ± 0.012	$\mathbf{0.013 \pm 0.012}$	$\mathbf{0.680 \pm 0.250}$	0.709 ± 0.115
Coherence (Ind)	GIRB	$\mathbf{0.015 \pm 0.012}$	$\mathbf{0.015 \pm 0.017}$	$\mathbf{0.013 \pm 0.011}$	0.673 ± 0.262	$\mathbf{0.764 \pm 0.104}$
Consistency (Ind)	None	0.029 ± 0.021	0.031 ± 0.21	0.009 ± 0.007	0.751 ± 0.194	$\mathbf{0.759 \pm 0.095}$
Consistency (Ind)	QAB	0.019 ± 0.015	$\mathbf{0.015 \pm 0.015}$	0.008 ± 0.014	$\mathbf{0.855 \pm 0.320}$	0.737 ± 0.122
Consistency (Ind)	GIRB	$\mathbf{0.014 \pm 0.008}$	$\mathbf{0.015 \pm 0.017}$	$\mathbf{0.007 \pm 0.010}$	0.812 ± 0.303	0.746 ± 0.187
Fluency (Ind)	None	0.027 ± 0.012	0.025 ± 0.019	$\mathbf{0.005 \pm 0.010}$	0.347 ± 0.064	$\mathbf{0.767 \pm 0.111}$
Fluency (Ind)	QAB	0.022 ± 0.021	0.022 ± 0.018	$\mathbf{0.005 \pm 0.013}$	0.353 ± 0.085	0.706 ± 0.108
Fluency (Ind)	GIRB	$\mathbf{0.015 \pm 0.021}$	$\mathbf{0.017 \pm 0.016}$	0.006 ± 0.010	$\mathbf{0.395 \pm 0.197}$	0.746 ± 0.189
Completeness (Avg)	None	0.125 ± 0.021	0.096 ± 0.018	0.062 ± 0.028	0.516 ± 0.242	-
Completeness (Avg)	QAB	$\mathbf{0.049 \pm 0.021}$	0.057 ± 0.020	0.045 ± 0.021	0.697 ± 0.296	-
Completeness (Avg)	GIRB	$\mathbf{0.049 \pm 0.016}$	$\mathbf{0.052 \pm 0.010}$	$\mathbf{0.040 \pm 0.020}$	$\mathbf{0.747 \pm 0.257}$	-
Conciseness (Avg)	None	0.084 ± 0.024	0.076 ± 0.017	0.043 ± 0.021	0.477 ± 0.107	-
Conciseness (Avg)	QAB	$\mathbf{0.034 \pm 0.019}$	$\mathbf{0.044 \pm 0.017}$	0.029 ± 0.020	$\mathbf{0.781 \pm 0.178}$	-
Conciseness (Avg)	GIRB	$\mathbf{0.034 \pm 0.017}$	$\mathbf{0.044 \pm 0.018}$	$\mathbf{0.028 \pm 0.011}$	0.777 ± 0.281	-
Faithfulness (Avg)	None	0.041 ± 0.014	0.04 ± 0.14	0.016 ± 0.015	0.687 ± 0.334	-
Faithfulness (Avg)	QAB	$\mathbf{0.018 \pm 0.021}$	0.016 ± 0.012	$\mathbf{0.011 \pm 0.007}$	$\mathbf{0.909 \pm 0.419}$	-
Faithfulness (Avg)	GIRB	$\mathbf{0.018 \pm 0.023}$	$\mathbf{0.013 \pm 0.018}$	$\mathbf{0.011 \pm 0.010}$	0.889 ± 0.131	-
Coherence (Avg)	None	0.019 ± 0.022	0.021 ± 0.07	0.004 ± 0.008	0.476 ± 0.150	-
Coherence (Avg)	QAB	0.015 ± 0.009	0.015 ± 0.008	$\mathbf{0.002 \pm 0.011}$	0.734 ± 0.361	-
Coherence (Avg)	GIRB	$\mathbf{0.008 \pm 0.010}$	$\mathbf{0.008 \pm 0.008}$	$\mathbf{0.002 \pm 0.008}$	$\mathbf{0.760 \pm 0.183}$	-
Consistency (Avg)	None	0.024 ± 0.012	0.023 ± 0.018	0.004 ± 0.007	0.598 ± 0.265	-
Consistency (Avg)	QAB	$\mathbf{0.007 \pm 0.012}$	0.008 ± 0.019	$\mathbf{0.002 \pm 0.019}$	0.882 ± 0.173	-
Consistency (Avg)	GIRB	$\mathbf{0.007 \pm 0.014}$	$\mathbf{0.007 \pm 0.019}$	$\mathbf{0.002 \pm 0.015}$	$\mathbf{0.927 \pm 0.439}$	-
Fluency (Avg)	None	0.027 ± 0.019	0.024 ± 0.009	0.002 ± 0.014	0.128 ± 0.048	-
Fluency (Avg)	QAB	0.018 ± 0.021	0.011 ± 0.010	$\mathbf{0.001 \pm 0.009}$	$\mathbf{0.300 \pm 0.122}$	-
Fluency (Avg)	GIRB	$\mathbf{0.011 \pm 0.010}$	$\mathbf{0.010 \pm 0.015}$	$\mathbf{0.001 \pm 0.019}$	0.262 ± 0.057	-

Table 9: The performance of different clustering methods and calibration methods across various dimensions.

Cluster_method	Dimension	Calibrator	ECE	GECE	Brier	Slope	Accuracy
KD-Tree	Completeness (Ind)	None	0.114 ± 0.034	0.112 ± 0.014	0.106 ± 0.044	0.588 ± 0.243	0.631 ± 0.105
KD-Tree	Completeness (Ind)	QAB	**0.058 ± 0.025**	0.071 ± 0.013	0.097 ± 0.032	**0.730 ± 0.169**	0.662 ± 0.146
KD-Tree	Completeness (Ind)	GIRB	**0.058 ± 0.012**	**0.070 ± 0.024**	**0.094 ± 0.025**	0.695 ± 0.288	**0.691 ± 0.131**
KD-Tree	Conciseness (Ind)	None	0.094 ± 0.025	0.117 ± 0.020	0.098 ± 0.033	0.586 ± 0.290	0.642 ± 0.107
KD-Tree	Conciseness (Ind)	QAB	0.056 ± 0.025	0.073 ± 0.014	0.096 ± 0.040	0.708 ± 0.196	0.688 ± 0.124
KD-Tree	Conciseness (Ind)	GIRB	**0.055 ± 0.023**	**0.067 ± 0.024**	**0.091 ± 0.021**	**0.716 ± 0.193**	**0.694 ± 0.186**
KD-Tree	Faithfulness (Ind)	None	0.058 ± 0.020	0.085 ± 0.017	0.085 ± 0.028	0.775 ± 0.344	0.690 ± 0.110
KD-Tree	Faithfulness (Ind)	QAB	**0.042 ± 0.012**	0.051 ± 0.017	**0.093 ± 0.022**	**0.917 ± 0.374**	0.743 ± 0.138
KD-Tree	Faithfulness (Ind)	GIRB	0.051 ± 0.019	**0.048 ± 0.020**	**0.093 ± 0.037**	0.763 ± 0.138	**0.777 ± 0.107**
KD-Tree	Coherence (Ind)	None	0.015 ± 0.011	0.026 ± 0.016	**0.014 ± 0.009**	**0.659 ± 0.216**	0.734 ± 0.124
KD-Tree	Coherence (Ind)	QAB	0.029 ± 0.012	0.025 ± 0.021	**0.014 ± 0.022**	0.506 ± 0.137	0.714 ± 0.122
KD-Tree	Coherence (Ind)	GIRB	**0.024 ± 0.013**	**0.018 ± 0.018**	**0.014 ± 0.014**	0.506 ± 0.137	**0.746 ± 0.091**
KD-Tree	Consistency (Ind)	None	0.029 ± 0.021	0.027 ± 0.019	0.009 ± 0.018	0.751 ± 0.341	0.659 ± 0.143
KD-Tree	Consistency (Ind)	QAB	0.021 ± 0.015	0.016 ± 0.017	**0.008 ± 0.017**	**0.957 ± 0.244**	0.714 ± 0.154
KD-Tree	Consistency (Ind)	GIRB	**0.014 ± 0.008**	**0.013 ± 0.015**	**0.008 ± 0.010**	0.757 ± 0.364	**0.731 ± 0.171**
KD-Tree	Fluency (Ind)	None	0.027 ± 0.012	0.026 ± 0.014	0.005 ± 0.018	0.347 ± 0.132	**0.767 ± 0.126**
KD-Tree	Fluency (Ind)	QAB	**0.010 ± 0.020**	0.018 ± 0.019	**0.004 ± 0.020**	**0.644 ± 0.155**	0.711 ± 0.111
KD-Tree	Fluency (Ind)	GIRB	**0.010 ± 0.019**	**0.011 ± 0.009**	0.005 ± 0.013	0.552 ± 0.216	0.746 ± 0.152
KD-Tree	Completeness (Avg)	None	0.125 ± 0.021	0.09 ± 0.012	0.062 ± 0.027	0.516 ± 0.226	-
KD-Tree	Completeness (Avg)	QAB	**0.041 ± 0.019**	**0.038 ± 0.016**	0.043 ± 0.025	**0.833 ± 0.140**	-
KD-Tree	Completeness (Avg)	GIRB	0.047 ± 0.016	0.041 ± 0.013	**0.041 ± 0.019**	0.742 ± 0.118	-
KD-Tree	Conciseness (Avg)	None	0.084 ± 0.024	0.081 ± 0.017	0.043 ± 0.012	0.477 ± 0.151	-
KD-Tree	Conciseness (Avg)	QAB	0.036 ± 0.019	0.025 ± 0.012	0.032 ± 0.011	**0.833 ± 0.212**	-
KD-Tree	Conciseness (Avg)	GIRB	**0.030 ± 0.016**	**0.023 ± 0.012**	**0.031 ± 0.018**	0.769 ± 0.148	-
KD-Tree	Faithfulness (Avg)	None	0.041 ± 0.014	0.036 ± 0.016	0.016 ± 0.016	0.687 ± 0.242	-
KD-Tree	Faithfulness (Avg)	QAB	0.024 ± 0.022	**0.016 ± 0.014**	0.016 ± 0.007	**0.808 ± 0.137**	-
KD-Tree	Faithfulness (Avg)	GIRB	**0.020 ± 0.023**	**0.016 ± 0.011**	**0.015 ± 0.014**	0.779 ± 0.131	-
KD-Tree	Coherence (Avg)	None	0.019 ± 0.022	0.022 ± 0.016	0.004 ± 0.013	0.476 ± 0.184	-
KD-Tree	Coherence (Avg)	QAB	0.016 ± 0.009	0.007 ± 0.008	**0.002 ± 0.009**	**0.774 ± 0.136**	-
KD-Tree	Coherence (Avg)	GIRB	**0.009 ± 0.010**	**0.006 ± 0.020**	**0.002 ± 0.007**	0.704 ± 0.190	-
KD-Tree	Consistency (Avg)	None	0.024 ± 0.012	0.021 ± 0.015	0.004 ± 0.013	0.598 ± 0.266	-
KD-Tree	Consistency (Avg)	QAB	0.015 ± 0.013	0.012 ± 0.016	**0.003 ± 0.015**	**0.881 ± 0.435**	-
KD-Tree	Consistency (Avg)	GIRB	**0.010 ± 0.014**	**0.007 ± 0.015**	**0.003 ± 0.020**	0.812 ± 0.214	-
KD-Tree	Fluency (Avg)	None	0.027 ± 0.019	0.023 ± 0.008	0.002 ± 0.010	0.128 ± 0.065	-
KD-Tree	Fluency (Avg)	QAB	0.012 ± 0.021	0.011 ± 0.010	**0.001 ± 0.009**	**0.616 ± 0.187**	-
KD-Tree	Fluency (Avg)	GIRB	**0.005 ± 0.009**	**0.004 ± 0.019**	**0.001 ± 0.006**	0.560 ± 0.102	-
KMeans	Completeness (Ind)	None	0.114 ± 0.034	0.125 ± 0.24	0.106 ± 0.040	0.588 ± 0.276	0.631 ± 0.122
KMeans	Completeness (Ind)	QAB	0.065 ± 0.026	0.084 ± 0.022	0.090 ± 0.017	0.734 ± 0.188	0.668 ± 0.108
KMeans	Completeness (Ind)	GIRB	**0.058 ± 0.012**	**0.082 ± 0.031**	**0.084 ± 0.020**	**0.766 ± 0.311**	**0.702 ± 0.139**
KMeans	Conciseness (Ind)	None	0.094 ± 0.025	0.11 ± 0.023	0.098 ± 0.038	0.586 ± 0.274	**0.742 ± 0.145**
KMeans	Conciseness (Ind)	QAB	0.052 ± 0.024	**0.077 ± 0.021**	0.088 ± 0.018	0.738 ± 0.265	0.668 ± 0.118
KMeans	Conciseness (Ind)	GIRB	**0.048 ± 0.023**	0.080 ± 0.017	**0.084 ± 0.020**	**0.740 ± 0.150**	0.731 ± 0.119
KMeans	Faithfulness (Ind)	None	0.058 ± 0.020	0.098 ± 0.25	0.085 ± 0.017	0.775 ± 0.267	0.790 ± 0.154
KMeans	Faithfulness (Ind)	QAB	0.044 ± 0.012	0.047 ± 0.024	0.085 ± 0.029	**0.921 ± 0.159**	0.730 ± 0.089
KMeans	Faithfulness (Ind)	GIRB	**0.037 ± 0.017**	**0.042 ± 0.011**	**0.082 ± 0.028**	0.899 ± 0.117	**0.795 ± 0.118**
KMeans	Coherence (Ind)	None	**0.015 ± 0.011**	0.026 ± 0.022	0.014 ± 0.021	0.659 ± 0.225	0.734 ± 0.086
KMeans	Coherence (Ind)	QAB	**0.015 ± 0.010**	0.020 ± 0.012	**0.013 ± 0.012**	**0.680 ± 0.250**	0.709 ± 0.115
KMeans	Coherence (Ind)	GIRB	**0.015 ± 0.012**	**0.015 ± 0.017**	**0.013 ± 0.011**	0.673 ± 0.262	**0.764 ± 0.104**
KMeans	Consistency (Ind)	None	0.029 ± 0.021	0.031 ± 0.21	0.009 ± 0.007	0.751 ± 0.194	**0.759 ± 0.095**
KMeans	Consistency (Ind)	QAB	0.019 ± 0.015	**0.015 ± 0.015**	0.008 ± 0.014	**0.855 ± 0.320**	0.737 ± 0.122
KMeans	Consistency (Ind)	GIRB	**0.014 ± 0.008**	**0.015 ± 0.017**	**0.007 ± 0.010**	0.812 ± 0.303	0.746 ± 0.187
KMeans	Fluency (Ind)	None	0.027 ± 0.012	0.025 ± 0.019	**0.005 ± 0.010**	0.347 ± 0.064	**0.767 ± 0.111**
KMeans	Fluency (Ind)	QAB	0.022 ± 0.021	0.022 ± 0.018	**0.005 ± 0.013**	0.353 ± 0.085	0.706 ± 0.108
KMeans	Fluency (Ind)	GIRB	**0.015 ± 0.021**	**0.017 ± 0.016**	0.006 ± 0.010	**0.395 ± 0.197**	0.746 ± 0.189
KMeans	Completeness (Avg)	None	0.125 ± 0.021	0.096 ± 0.018	0.062 ± 0.028	0.516 ± 0.242	-
KMeans	Completeness (Avg)	QAB	**0.049 ± 0.021**	0.057 ± 0.020	0.045 ± 0.021	0.697 ± 0.296	-
KMeans	Completeness (Avg)	GIRB	**0.049 ± 0.016**	**0.052 ± 0.010**	**0.040 ± 0.020**	**0.747 ± 0.257**	-
KMeans	Conciseness (Avg)	None	0.084 ± 0.024	0.076 ± 0.017	0.043 ± 0.021	0.477 ± 0.107	-
KMeans	Conciseness (Avg)	QAB	**0.034 ± 0.019**	**0.044 ± 0.017**	0.029 ± 0.020	**0.781 ± 0.178**	-
KMeans	Conciseness (Avg)	GIRB	**0.034 ± 0.017**	**0.044 ± 0.018**	**0.028 ± 0.011**	0.777 ± 0.281	-
KMeans	Faithfulness (Avg)	None	0.041 ± 0.014	0.04 ± 0.14	0.016 ± 0.015	0.687 ± 0.334	-
KMeans	Faithfulness (Avg)	QAB	**0.018 ± 0.021**	0.016 ± 0.012	**0.011 ± 0.007**	**0.909 ± 0.419**	-
KMeans	Faithfulness (Avg)	GIRB	**0.018 ± 0.023**	**0.013 ± 0.018**	**0.011 ± 0.010**	0.889 ± 0.131	-
KMeans	Coherence (Avg)	None	0.019 ± 0.022	0.021 ± 0.07	0.004 ± 0.008	0.476 ± 0.150	-
KMeans	Coherence (Avg)	QAB	0.015 ± 0.009	0.015 ± 0.008	**0.002 ± 0.011**	0.734 ± 0.361	-
KMeans	Coherence (Avg)	GIRB	**0.008 ± 0.010**	**0.008 ± 0.008**	**0.002 ± 0.008**	**0.760 ± 0.183**	-
KMeans	Consistency (Avg)	None	0.024 ± 0.012	0.023 ± 0.018	0.004 ± 0.007	0.598 ± 0.265	-
KMeans	Consistency (Avg)	QAB	**0.007 ± 0.012**	0.008 ± 0.019	**0.002 ± 0.019**	0.882 ± 0.173	-
KMeans	Consistency (Avg)	GIRB	**0.007 ± 0.014**	**0.007 ± 0.019**	**0.002 ± 0.015**	**0.927 ± 0.439**	-
KMeans	Fluency (Avg)	None	0.027 ± 0.019	0.024 ± 0.009	0.002 ± 0.014	0.128 ± 0.048	-
KMeans	Fluency (Avg)	QAB	0.018 ± 0.021	0.011 ± 0.010	**0.001 ± 0.009**	**0.300 ± 0.122**	-
KMeans	Fluency (Avg)	GIRB	**0.011 ± 0.010**	**0.010 ± 0.015**	**0.001 ± 0.019**	0.262 ± 0.057	-

Table 10: The performance of different embedding models and calibration methods across various dimensions.

Embedding_model	Dimension	Calibrator	ECE	GECE	Brier	Slope	Accuracy
DistilBERT	Completeness (Ind)	None	0.116 ± 0.034	0.121 ± 0.018	0.112 ± 0.029	0.598 ± 0.122	0.629 ± 0.129
DistilBERT	Completeness (Ind)	QAB	0.059 ± 0.025	**0.062 ± 0.024**	**0.091 ± 0.026**	0.754 ± 0.233	0.607 ± 0.111
DistilBERT	Completeness (Ind)	GIRB	**0.054 ± 0.013**	0.067 ± 0.030	**0.091 ± 0.032**	**0.809 ± 0.222**	**0.641 ± 0.100**
DistilBERT	Conciseness (Ind)	None	0.118 ± 0.026	0.104 ± 0.026	0.108 ± 0.044	0.550 ± 0.173	**0.696 ± 0.111**
DistilBERT	Conciseness (Ind)	QAB	0.070 ± 0.027	0.076 ± 0.027	0.102 ± 0.016	0.562 ± 0.204	0.622 ± 0.163
DistilBERT	Conciseness (Ind)	GIRB	**0.069 ± 0.024**	**0.075 ± 0.016**	**0.091 ± 0.012**	**0.663 ± 0.105**	0.669 ± 0.106
DistilBERT	Faithfulness (Ind)	None	0.088 ± 0.027	0.11 ± 0.016	0.112 ± 0.024	0.534 ± 0.096	0.656 ± 0.118
DistilBERT	Faithfulness (Ind)	QAB	**0.054 ± 0.013**	0.065 ± 0.016	0.109 ± 0.018	**0.675 ± 0.129**	0.618 ± 0.118
DistilBERT	Faithfulness (Ind)	GIRB	0.066 ± 0.020	**0.058 ± 0.025**	**0.108 ± 0.028**	0.636 ± 0.128	**0.681 ± 0.151**
DistilBERT	Coherence (Ind)	None	0.034 ± 0.014	0.03 ± 0.019	0.016 ± 0.013	**0.457 ± 0.214**	0.607 ± 0.103
DistilBERT	Coherence (Ind)	QAB	**0.024 ± 0.011**	0.032 ± 0.022	**0.015 ± 0.017**	0.392 ± 0.149	0.641 ± 0.128
DistilBERT	Coherence (Ind)	GIRB	**0.024 ± 0.013**	**0.027 ± 0.018**	0.016 ± 0.021	0.424 ± 0.086	**0.684 ± 0.120**
DistilBERT	Consistency (Ind)	None	0.027 ± 0.021	0.034 ± 0.013	0.012 ± 0.019	0.508 ± 0.101	0.618 ± 0.169
DistilBERT	Consistency (Ind)	QAB	0.023 ± 0.015	0.025 ± 0.016	**0.010 ± 0.009**	**0.699 ± 0.135**	0.590 ± 0.115
DistilBERT	Consistency (Ind)	GIRB	**0.019 ± 0.009**	**0.019 ± 0.018**	**0.010 ± 0.019**	0.674 ± 0.122	**0.631 ± 0.170**
DistilBERT	Fluency (Ind)	None	0.019 ± 0.011	0.037 ± 0.017	**0.005 ± 0.012**	0.208 ± 0.051	0.615 ± 0.136
DistilBERT	Fluency (Ind)	QAB	0.020 ± 0.021	0.020 ± 0.008	**0.005 ± 0.017**	**0.440 ± 0.148**	0.573 ± 0.060
DistilBERT	Fluency (Ind)	GIRB	**0.014 ± 0.020**	**0.015 ± 0.010**	0.006 ± 0.020	0.360 ± 0.065	**0.635 ± 0.113**
DistilBERT	Completeness (Avg)	None	0.125 ± 0.021	0.102 ± 0.021	0.065 ± 0.025	0.556 ± 0.199	-
DistilBERT	Completeness (Avg)	QAB	**0.034 ± 0.018**	0.038 ± 0.028	**0.029 ± 0.022**	**0.884 ± 0.403**	-
DistilBERT	Completeness (Avg)	GIRB	0.036 ± 0.015	**0.037 ± 0.019**	0.031 ± 0.021	0.831 ± 0.127	-
DistilBERT	Conciseness (Avg)	None	0.093 ± 0.025	0.077 ± 0.022	0.045 ± 0.024	0.429 ± 0.192	-
DistilBERT	Conciseness (Avg)	QAB	0.035 ± 0.019	0.029 ± 0.021	0.025 ± 0.021	**0.813 ± 0.160**	-
DistilBERT	Conciseness (Avg)	GIRB	**0.030 ± 0.016**	**0.025 ± 0.023**	**0.023 ± 0.008**	0.805 ± 0.216	-
DistilBERT	Faithfulness (Avg)	None	0.051 ± 0.016	0.044 ± 0.021	0.019 ± 0.011	0.585 ± 0.211	-
DistilBERT	Faithfulness (Avg)	QAB	0.018 ± 0.021	0.008 ± 0.020	0.010 ± 0.020	**0.958 ± 0.291**	-
DistilBERT	Faithfulness (Avg)	GIRB	**0.014 ± 0.022**	**0.007 ± 0.012**	**0.009 ± 0.016**	0.933 ± 0.133	-
DistilBERT	Coherence (Avg)	None	0.020 ± 0.022	0.023 ± 0.020	0.004 ± 0.012	0.437 ± 0.183	-
DistilBERT	Coherence (Avg)	QAB	0.006 ± 0.008	**0.005 ± 0.011**	**0.002 ± 0.010**	0.804 ± 0.258	-
DistilBERT	Coherence (Avg)	GIRB	**0.005 ± 0.010**	**0.005 ± 0.014**	**0.002 ± 0.017**	**0.811 ± 0.140**	-
DistilBERT	Consistency (Avg)	None	0.025 ± 0.012	0.021 ± 0.010	0.006 ± 0.018	0.496 ± 0.163	-
DistilBERT	Consistency (Avg)	QAB	**0.007 ± 0.012**	0.013 ± 0.018	**0.003 ± 0.013**	**0.872 ± 0.200**	-
DistilBERT	Consistency (Avg)	GIRB	**0.007 ± 0.014**	**0.007 ± 0.008**	**0.003 ± 0.013**	0.841 ± 0.117	-
DistilBERT	Fluency (Avg)	None	0.012 ± 0.018	0.02 ± 0.016	0.005 ± 0.018	0.295 ± 0.068	-
DistilBERT	Fluency (Avg)	QAB	**0.003 ± 0.020**	0.010 ± 0.022	0.002 ± 0.007	0.800 ± 0.149	-
DistilBERT	Fluency (Avg)	GIRB	**0.003 ± 0.009**	**0.003 ± 0.013**	**0.001 ± 0.016**	**0.827 ± 0.291**	-
Gemma2b	Completeness (Ind)	None	0.110 ± 0.024	0.105 ± 0.022	0.109 ± 0.025	0.623 ± 0.264	0.703 ± 0.102
Gemma2b	Completeness (Ind)	QAB	**0.060 ± 0.027**	**0.052 ± 0.027**	0.081 ± 0.022	**0.807 ± 0.135**	0.672 ± 0.180
Gemma2b	Completeness (Ind)	GIRB	0.063 ± 0.019	**0.052 ± 0.025**	**0.077 ± 0.023**	0.796 ± 0.280	**0.747 ± 0.111**
Gemma2b	Conciseness (Ind)	None	0.093 ± 0.018	0.127 ± 0.013	0.114 ± 0.028	0.456 ± 0.132	**0.686 ± 0.102**
Gemma2b	Conciseness (Ind)	QAB	0.068 ± 0.029	0.078 ± 0.019	**0.094 ± 0.027**	**0.688 ± 0.163**	0.619 ± 0.166
Gemma2b	Conciseness (Ind)	GIRB	**0.062 ± 0.013**	**0.075 ± 0.033**	0.095 ± 0.023	0.622 ± 0.233	0.666 ± 0.147
Gemma2b	Faithfulness (Ind)	None	0.078 ± 0.019	0.092 ± 0.043	0.112 ± 0.042	0.612 ± 0.172	**0.728 ± 0.135**
Gemma2b	Faithfulness (Ind)	QAB	**0.064 ± 0.027**	0.083 ± 0.037	**0.111 ± 0.043**	**0.638 ± 0.224**	0.631 ± 0.131
Gemma2b	Faithfulness (Ind)	GIRB	**0.064 ± 0.026**	**0.076 ± 0.024**	0.113 ± 0.021	0.610 ± 0.103	0.660 ± 0.140
Gemma2b	Coherence (Ind)	None	**0.020 ± 0.017**	0.023 ± 0.007	**0.014 ± 0.016**	**0.530 ± 0.118**	**0.708 ± 0.125**
Gemma2b	Coherence (Ind)	QAB	0.031 ± 0.020	**0.019 ± 0.019**	0.015 ± 0.021	0.279 ± 0.061	0.635 ± 0.163
Gemma2b	Coherence (Ind)	GIRB	0.026 ± 0.018	**0.019 ± 0.020**	0.017 ± 0.014	0.341 ± 0.081	0.702 ± 0.122
Gemma2b	Consistency (Ind)	None	0.025 ± 0.015	0.027 ± 0.021	0.009 ± 0.013	0.625 ± 0.106	**0.649 ± 0.089**
Gemma2b	Consistency (Ind)	QAB	0.020 ± 0.008	0.020 ± 0.014	**0.008 ± 0.007**	**0.909 ± 0.263**	0.618 ± 0.090
Gemma2b	Consistency (Ind)	GIRB	**0.015 ± 0.016**	**0.016 ± 0.008**	**0.008 ± 0.015**	0.737 ± 0.129	0.678 ± 0.102
Gemma2b	Fluency (Ind)	None	0.009 ± 0.014	0.011 ± 0.007	**0.004 ± 0.015**	0.692 ± 0.343	**0.717 ± 0.128**
Gemma2b	Fluency (Ind)	QAB	0.010 ± 0.011	0.011 ± 0.007	0.005 ± 0.012	0.440 ± 0.191	0.609 ± 0.157
Gemma2b	Fluency (Ind)	GIRB	**0.007 ± 0.017**	**0.008 ± 0.014**	0.005 ± 0.019	**0.741 ± 0.150**	0.678 ± 0.196
Gemma2b	Completeness (Avg)	None	0.116 ± 0.018	0.088 ± 0.015	0.060 ± 0.025	0.566 ± 0.271	-
Gemma2b	Completeness (Avg)	QAB	0.032 ± 0.011	0.027 ± 0.017	0.024 ± 0.021	**0.919 ± 0.309**	-
Gemma2b	Completeness (Avg)	GIRB	**0.027 ± 0.023**	**0.021 ± 0.007**	**0.022 ± 0.020**	0.911 ± 0.369	-
Gemma2b	Conciseness (Avg)	None	0.096 ± 0.035	0.069 ± 0.027	0.047 ± 0.027	0.434 ± 0.114	-
Gemma2b	Conciseness (Avg)	QAB	**0.031 ± 0.014**	0.032 ± 0.017	0.024 ± 0.015	**0.801 ± 0.251**	-
Gemma2b	Conciseness (Avg)	GIRB	**0.031 ± 0.020**	**0.031 ± 0.015**	**0.023 ± 0.018**	0.781 ± 0.140	-
Gemma2b	Faithfulness (Avg)	None	0.043 ± 0.027	0.038 ± 0.016	0.015 ± 0.018	0.635 ± 0.156	-
Gemma2b	Faithfulness (Avg)	QAB	0.014 ± 0.008	0.013 ± 0.011	0.010 ± 0.016	0.904 ± 0.171	-
Gemma2b	Faithfulness (Avg)	GIRB	**0.013 ± 0.013**	**0.012 ± 0.008**	**0.008 ± 0.020**	**0.923 ± 0.294**	-
Gemma2b	Coherence (Avg)	None	0.022 ± 0.022	0.02 ± 0.011	0.003 ± 0.009	0.465 ± 0.119	-
Gemma2b	Coherence (Avg)	QAB	0.014 ± 0.014	0.012 ± 0.014	**0.001 ± 0.019**	**0.807 ± 0.135**	-
Gemma2b	Coherence (Avg)	GIRB	**0.007 ± 0.009**	**0.005 ± 0.015**	**0.001 ± 0.014**	0.783 ± 0.202	-
Gemma2b	Consistency (Avg)	None	0.024 ± 0.012	0.022 ± 0.013	0.004 ± 0.008	0.552 ± 0.132	-
Gemma2b	Consistency (Avg)	QAB	0.005 ± 0.011	0.005 ± 0.021	**0.002 ± 0.011**	0.923 ± 0.377	-
Gemma2b	Consistency (Avg)	GIRB	**0.004 ± 0.016**	**0.004 ± 0.010**	**0.002 ± 0.006**	**0.926 ± 0.280**	-
Gemma2b	Fluency (Avg)	None	0.015 ± 0.021	0.013 ± 0.019	**0.001 ± 0.020**	0.324 ± 0.099	-
Gemma2b	Fluency (Avg)	QAB	0.006 ± 0.010	**0.004 ± 0.010**	**0.001 ± 0.013**	0.675 ± 0.302	-
Gemma2b	Fluency (Avg)	GIRB	**0.004 ± 0.010**	**0.004 ± 0.007**	**0.001 ± 0.017**	**0.714 ± 0.134**	-
Qwen8b	Completeness (Ind)	None	0.114 ± 0.034	0.125 ± 0.24	0.106 ± 0.040	0.588 ± 0.276	0.631 ± 0.122
Qwen8b	Completeness (Ind)	QAB	0.065 ± 0.026	0.084 ± 0.022	0.090 ± 0.017	0.734 ± 0.188	0.668 ± 0.108
Qwen8b	Completeness (Ind)	GIRB	**0.058 ± 0.012**	**0.082 ± 0.031**	**0.084 ± 0.020**	**0.766 ± 0.311**	**0.702 ± 0.139**
Qwen8b	Conciseness (Ind)	None	0.094 ± 0.025	0.11 ± 0.023	0.098 ± 0.038	0.586 ± 0.274	**0.742 ± 0.145**
Qwen8b	Conciseness (Ind)	QAB	0.052 ± 0.024	**0.077 ± 0.021**	0.088 ± 0.018	0.738 ± 0.265	0.668 ± 0.118
Qwen8b	Conciseness (Ind)	GIRB	**0.048 ± 0.023**	0.080 ± 0.017	**0.084 ± 0.020**	**0.740 ± 0.150**	0.731 ± 0.119
Qwen8b	Faithfulness (Ind)	None	0.058 ± 0.020	0.098 ± 0.25	0.085 ± 0.017	0.775 ± 0.267	0.790 ± 0.154
Qwen8b	Faithfulness (Ind)	QAB	0.044 ± 0.012	0.047 ± 0.024	0.085 ± 0.029	**0.921 ± 0.159**	0.730 ± 0.089
Qwen8b	Faithfulness (Ind)	GIRB	**0.037 ± 0.017**	**0.042 ± 0.011**	**0.082 ± 0.028**	0.899 ± 0.117	**0.795 ± 0.118**
Qwen8b	Coherence (Ind)	None	**0.015 ± 0.011**	0.026 ± 0.022	0.014 ± 0.021	0.659 ± 0.225	0.734 ± 0.086
Qwen8b	Coherence (Ind)	QAB	**0.015 ± 0.010**	0.020 ± 0.012	**0.013 ± 0.012**	**0.680 ± 0.250**	0.709 ± 0.115
Qwen8b	Coherence (Ind)	GIRB	**0.015 ± 0.012**	**0.015 ± 0.017**	**0.013 ± 0.011**	0.673 ± 0.262	**0.764 ± 0.104**
Qwen8b	Consistency (Ind)	None	0.029 ± 0.021	0.031 ± 0.21	0.009 ± 0.007	0.751 ± 0.194	**0.759 ± 0.095**
Qwen8b	Consistency (Ind)	QAB	0.019 ± 0.015	**0.015 ± 0.015**	0.008 ± 0.014	**0.855 ± 0.320**	0.737 ± 0.122
Qwen8b	Consistency (Ind)	GIRB	**0.014 ± 0.008**	**0.015 ± 0.017**	**0.007 ± 0.010**	0.812 ± 0.303	0.746 ± 0.187
Qwen8b	Fluency (Ind)	None	0.027 ± 0.012	0.025 ± 0.019	**0.005 ± 0.010**	0.347 ± 0.064	**0.767 ± 0.111**
Qwen8b	Fluency (Ind)	QAB	0.022 ± 0.021	0.022 ± 0.018	**0.005 ± 0.013**	0.353 ± 0.085	0.706 ± 0.108
Qwen8b	Fluency (Ind)	GIRB	**0.015 ± 0.021**	**0.017 ± 0.016**	0.006 ± 0.010	**0.395 ± 0.197**	0.746 ± 0.189
Qwen8b	Completeness (Avg)	None	0.125 ± 0.021	0.096 ± 0.018	0.062 ± 0.028	0.516 ± 0.242	-
Qwen8b	Completeness (Avg)	QAB	**0.049 ± 0.021**	0.057 ± 0.020	0.045 ± 0.021	0.697 ± 0.296	-
Qwen8b	Completeness (Avg)	GIRB	**0.049 ± 0.016**	**0.052 ± 0.010**	**0.040 ± 0.020**	**0.747 ± 0.257**	-
Qwen8b	Conciseness (Avg)	None	0.084 ± 0.024	0.076 ± 0.017	0.043 ± 0.021	0.477 ± 0.107	-
Qwen8b	Conciseness (Avg)	QAB	**0.034 ± 0.019**	**0.044 ± 0.017**	0.029 ± 0.020	**0.781 ± 0.178**	-
Qwen8b	Conciseness (Avg)	GIRB	**0.034 ± 0.017**	**0.044 ± 0.018**	**0.028 ± 0.011**	0.777 ± 0.281	-
Qwen8b	Faithfulness (Avg)	None	0.041 ± 0.014	0.04 ± 0.14	0.016 ± 0.015	0.687 ± 0.334	-
Qwen8b	Faithfulness (Avg)	QAB	**0.018 ± 0.021**	0.016 ± 0.012	**0.011 ± 0.007**	**0.909 ± 0.419**	-
Qwen8b	Faithfulness (Avg)	GIRB	**0.018 ± 0.023**	**0.013 ± 0.018**	**0.011 ± 0.010**	0.889 ± 0.131	-
Qwen8b	Coherence (Avg)	None	0.019 ± 0.022	0.021 ± 0.07	0.004 ± 0.008	0.476 ± 0.150	-
Qwen8b	Coherence (Avg)	QAB	0.015 ± 0.009	0.015 ± 0.008	**0.002 ± 0.011**	0.734 ± 0.361	-
Qwen8b	Coherence (Avg)	GIRB	**0.008 ± 0.010**	**0.008 ± 0.008**	**0.002 ± 0.008**	**0.760 ± 0.183**	-
Qwen8b	Consistency (Avg)	None	0.024 ± 0.012	0.023 ± 0.018	0.004 ± 0.007	0.598 ± 0.265	-
Qwen8b	Consistency (Avg)	QAB	**0.007 ± 0.012**	0.008 ± 0.019	**0.002 ± 0.019**	0.882 ± 0.173	-
Qwen8b	Consistency (Avg)	GIRB	**0.007 ± 0.014**	**0.007 ± 0.019**	**0.002 ± 0.015**	**0.927 ± 0.439**	-
Qwen8b	Fluency (Avg)	None	0.027 ± 0.019	0.024 ± 0.009	0.002 ± 0.014	0.128 ± 0.048	-
Qwen8b	Fluency (Avg)	QAB	0.018 ± 0.021	0.011 ± 0.010	**0.001 ± 0.009**	**0.300 ± 0.122**	-
Qwen8b	Fluency (Avg)	GIRB	**0.011 ± 0.010**	**0.010 ± 0.015**	**0.001 ± 0.019**	0.262 ± 0.057	-

Table 11: The performance of different prompting strategy and calibration methods across various dimensions.

Strategy	Dimension	Calibrator	ECE	GECE	Brier	Slope	Accuracy
DS-Prompt	Completeness (Ind)	None	0.130 ± 0.036	0.114 ± 0.042	0.119 ± 0.047	0.495 ± 0.207	0.699 ± 0.110
DS-Prompt	Completeness (Ind)	QAB	**0.057 ± 0.025**	**0.046 ± 0.011**	0.089 ± 0.031	**0.780 ± 0.180**	0.644 ± 0.141
DS-Prompt	Completeness (Ind)	GIRB	0.070 ± 0.014	0.060 ± 0.023	**0.085 ± 0.024**	0.750 ± 0.310	**0.681 ± 0.128**
DS-Prompt	Conciseness (Ind)	None	0.123 ± 0.027	0.127 ± 0.024	0.120 ± 0.036	0.414 ± 0.208	**0.714 ± 0.103**
DS-Prompt	Conciseness (Ind)	QAB	**0.070 ± 0.027**	**0.071 ± 0.014**	0.102 ± 0.042	**0.584 ± 0.165**	0.594 ± 0.197
DS-Prompt	Conciseness (Ind)	GIRB	0.072 ± 0.024	0.073 ± 0.025	**0.098 ± 0.022**	0.567 ± 0.157	0.680 ± 0.183
DS-Prompt	Faithfulness (Ind)	None	0.058 ± 0.020	0.097 ± 0.047	**0.101 ± 0.031**	0.704 ± 0.313	0.646 ± 0.097
DS-Prompt	Faithfulness (Ind)	QAB	**0.038 ± 0.011**	**0.055 ± 0.017**	0.105 ± 0.023	**0.812 ± 0.333**	0.591 ± 0.100
DS-Prompt	Faithfulness (Ind)	GIRB	0.053 ± 0.019	0.061 ± 0.023	0.105 ± 0.039	0.682 ± 0.126	**0.686 ± 0.189**
DS-Prompt	Coherence (Ind)	None	0.057 ± 0.017	0.05 ± 0.027	0.020 ± 0.010	0.196 ± 0.077	0.579 ± 0.099
DS-Prompt	Coherence (Ind)	QAB	0.039 ± 0.014	0.032 ± 0.022	**0.016 ± 0.023**	**0.366 ± 0.105**	0.599 ± 0.194
DS-Prompt	Coherence (Ind)	GIRB	**0.035 ± 0.014**	**0.025 ± 0.018**	0.019 ± 0.014	0.261 ± 0.073	**0.686 ± 0.084**
DS-Prompt	Consistency (Ind)	None	0.052 ± 0.024	0.05 ± 0.017	0.012 ± 0.019	0.368 ± 0.170	**0.671 ± 0.130**
DS-Prompt	Consistency (Ind)	QAB	0.024 ± 0.015	0.029 ± 0.018	0.009 ± 0.017	**0.712 ± 0.187**	0.588 ± 0.133
DS-Prompt	Consistency (Ind)	GIRB	**0.018 ± 0.009**	**0.024 ± 0.017**	**0.008 ± 0.010**	0.676 ± 0.326	0.659 ± 0.157
DS-Prompt	Fluency (Ind)	None	0.031 ± 0.013	0.042 ± 0.010	0.006 ± 0.018	0.311 ± 0.120	0.641 ± 0.200
DS-Prompt	Fluency (Ind)	QAB	**0.014 ± 0.020**	0.012 ± 0.018	0.005 ± 0.020	0.434 ± 0.108	0.635 ± 0.104
DS-Prompt	Fluency (Ind)	GIRB	**0.014 ± 0.020**	**0.010 ± 0.009**	**0.004 ± 0.013**	**0.509 ± 0.201**	**0.677 ± 0.143**
DS-Prompt	Completeness (Avg)	None	0.117 ± 0.020	0.082 ± 0.023	0.062 ± 0.027	0.520 ± 0.227	-
DS-Prompt	Completeness (Avg)	QAB	0.032 ± 0.017	0.032 ± 0.016	0.026 ± 0.022	**0.898 ± 0.150**	-
DS-Prompt	Completeness (Avg)	GIRB	**0.027 ± 0.015**	**0.027 ± 0.010**	**0.024 ± 0.015**	0.877 ± 0.137	-
DS-Prompt	Conciseness (Avg)	None	0.097 ± 0.025	0.091 ± 0.006	0.049 ± 0.013	0.398 ± 0.129	-
DS-Prompt	Conciseness (Avg)	QAB	0.032 ± 0.018	0.032 ± 0.013	0.024 ± 0.009	**0.816 ± 0.208**	-
DS-Prompt	Conciseness (Avg)	GIRB	**0.028 ± 0.016**	**0.028 ± 0.013**	**0.022 ± 0.016**	0.798 ± 0.153	-
DS-Prompt	Faithfulness (Avg)	None	0.055 ± 0.016	0.045 ± 0.013	0.018 ± 0.016	0.631 ± 0.223	-
DS-Prompt	Faithfulness (Avg)	QAB	0.019 ± 0.021	0.019 ± 0.014	**0.008 ± 0.006**	0.932 ± 0.156	-
DS-Prompt	Faithfulness (Avg)	GIRB	**0.016 ± 0.022**	**0.016 ± 0.011**	**0.008 ± 0.012**	**0.928 ± 0.153**	-
DS-Prompt	Coherence (Avg)	None	0.053 ± 0.028	0.051 ± 0.009	0.008 ± 0.013	0.146 ± 0.061	-
DS-Prompt	Coherence (Avg)	QAB	0.018 ± 0.009	0.018 ± 0.009	**0.002 ± 0.009**	0.680 ± 0.122	-
DS-Prompt	Coherence (Avg)	GIRB	**0.011 ± 0.011**	**0.011 ± 0.021**	**0.002 ± 0.007**	**0.688 ± 0.186**	-
DS-Prompt	Consistency (Avg)	None	0.049 ± 0.015	0.045 ± 0.020	0.008 ± 0.014	0.276 ± 0.133	-
DS-Prompt	Consistency (Avg)	QAB	0.016 ± 0.014	0.016 ± 0.017	**0.003 ± 0.015**	**0.778 ± 0.386**	-
DS-Prompt	Consistency (Avg)	GIRB	**0.010 ± 0.014**	**0.010 ± 0.015**	**0.003 ± 0.020**	0.765 ± 0.202	-
DS-Prompt	Fluency (Avg)	None	0.045 ± 0.021	0.046 ± 0.030	0.005 ± 0.011	-0.013 ± 0.019	-
DS-Prompt	Fluency (Avg)	QAB	0.015 ± 0.021	0.015 ± 0.011	**0.001 ± 0.009**	**0.530 ± 0.163**	-
DS-Prompt	Fluency (Avg)	GIRB	**0.009 ± 0.010**	**0.009 ± 0.019**	**0.001 ± 0.006**	0.432 ± 0.082	-
Instruct-DS-Prompt	Completeness (Ind)	None	0.114 ± 0.034	0.125 ± 0.24	0.106 ± 0.040	0.588 ± 0.276	0.631 ± 0.122
Instruct-DS-Prompt	Completeness (Ind)	QAB	0.065 ± 0.026	0.084 ± 0.022	0.090 ± 0.017	0.734 ± 0.188	0.668 ± 0.108
Instruct-DS-Prompt	Completeness (Ind)	GIRB	**0.058 ± 0.012**	**0.082 ± 0.031**	**0.084 ± 0.020**	**0.766 ± 0.311**	**0.702 ± 0.139**
Instruct-DS-Prompt	Conciseness (Ind)	None	0.094 ± 0.025	0.11 ± 0.023	0.098 ± 0.038	0.586 ± 0.274	**0.742 ± 0.145**
Instruct-DS-Prompt	Conciseness (Ind)	QAB	0.052 ± 0.024	**0.077 ± 0.021**	0.088 ± 0.018	0.738 ± 0.265	0.668 ± 0.118
Instruct-DS-Prompt	Conciseness (Ind)	GIRB	**0.048 ± 0.023**	0.080 ± 0.017	**0.084 ± 0.020**	**0.740 ± 0.150**	0.731 ± 0.119
Instruct-DS-Prompt	Faithfulness (Ind)	None	0.058 ± 0.020	0.098 ± 0.25	0.085 ± 0.017	0.775 ± 0.267	0.790 ± 0.154
Instruct-DS-Prompt	Faithfulness (Ind)	QAB	0.044 ± 0.012	0.047 ± 0.024	0.085 ± 0.029	**0.921 ± 0.159**	0.730 ± 0.089
Instruct-DS-Prompt	Faithfulness (Ind)	GIRB	**0.037 ± 0.017**	**0.042 ± 0.011**	**0.082 ± 0.028**	0.899 ± 0.117	**0.795 ± 0.118**
Instruct-DS-Prompt	Coherence (Ind)	None	**0.015 ± 0.011**	0.026 ± 0.022	0.014 ± 0.021	0.659 ± 0.225	0.734 ± 0.086
Instruct-DS-Prompt	Coherence (Ind)	QAB	**0.015 ± 0.010**	0.020 ± 0.012	**0.013 ± 0.012**	**0.680 ± 0.250**	0.709 ± 0.115
Instruct-DS-Prompt	Coherence (Ind)	GIRB	**0.015 ± 0.012**	**0.015 ± 0.017**	**0.013 ± 0.011**	0.673 ± 0.262	**0.764 ± 0.104**
Instruct-DS-Prompt	Consistency (Ind)	None	0.029 ± 0.021	0.031 ± 0.21	0.009 ± 0.007	0.751 ± 0.194	**0.759 ± 0.095**
Instruct-DS-Prompt	Consistency (Ind)	QAB	0.019 ± 0.015	**0.015 ± 0.015**	0.008 ± 0.014	**0.855 ± 0.320**	0.737 ± 0.122
Instruct-DS-Prompt	Consistency (Ind)	GIRB	**0.014 ± 0.008**	**0.015 ± 0.017**	**0.007 ± 0.010**	0.812 ± 0.303	0.746 ± 0.187
Instruct-DS-Prompt	Fluency (Ind)	None	0.027 ± 0.012	0.025 ± 0.019	**0.005 ± 0.010**	0.347 ± 0.064	**0.767 ± 0.111**
Instruct-DS-Prompt	Fluency (Ind)	QAB	0.022 ± 0.021	0.022 ± 0.018	**0.005 ± 0.013**	0.353 ± 0.085	0.706 ± 0.108
Instruct-DS-Prompt	Fluency (Ind)	GIRB	**0.015 ± 0.021**	**0.017 ± 0.016**	0.006 ± 0.010	**0.395 ± 0.197**	0.746 ± 0.189
Instruct-DS-Prompt	Completeness (Avg)	None	0.125 ± 0.021	0.096 ± 0.018	0.062 ± 0.028	0.516 ± 0.242	-
Instruct-DS-Prompt	Completeness (Avg)	QAB	**0.049 ± 0.021**	0.057 ± 0.020	0.045 ± 0.021	0.697 ± 0.296	-
Instruct-DS-Prompt	Completeness (Avg)	GIRB	**0.049 ± 0.016**	**0.052 ± 0.010**	**0.040 ± 0.020**	**0.747 ± 0.257**	-
Instruct-DS-Prompt	Conciseness (Avg)	None	0.084 ± 0.024	0.076 ± 0.017	0.043 ± 0.021	0.477 ± 0.107	-
Instruct-DS-Prompt	Conciseness (Avg)	QAB	**0.034 ± 0.019**	**0.044 ± 0.017**	0.029 ± 0.020	**0.781 ± 0.178**	-
Instruct-DS-Prompt	Conciseness (Avg)	GIRB	**0.034 ± 0.017**	**0.044 ± 0.018**	**0.028 ± 0.011**	0.777 ± 0.281	-
Instruct-DS-Prompt	Faithfulness (Avg)	None	0.041 ± 0.014	0.04 ± 0.14	0.016 ± 0.015	0.687 ± 0.334	-
Instruct-DS-Prompt	Faithfulness (Avg)	QAB	**0.018 ± 0.021**	0.016 ± 0.012	**0.011 ± 0.007**	**0.909 ± 0.419**	-
Instruct-DS-Prompt	Faithfulness (Avg)	GIRB	**0.018 ± 0.023**	**0.013 ± 0.018**	**0.011 ± 0.010**	0.889 ± 0.131	-
Instruct-DS-Prompt	Coherence (Avg)	None	0.019 ± 0.022	0.021 ± 0.07	0.004 ± 0.008	0.476 ± 0.150	-
Instruct-DS-Prompt	Coherence (Avg)	QAB	0.015 ± 0.009	0.015 ± 0.008	**0.002 ± 0.011**	0.734 ± 0.361	-
Instruct-DS-Prompt	Coherence (Avg)	GIRB	**0.008 ± 0.010**	**0.008 ± 0.008**	**0.002 ± 0.008**	**0.760 ± 0.183**	-
Instruct-DS-Prompt	Consistency (Avg)	None	0.024 ± 0.012	0.023 ± 0.018	0.004 ± 0.007	0.598 ± 0.265	-
Instruct-DS-Prompt	Consistency (Avg)	QAB	**0.007 ± 0.012**	0.008 ± 0.019	**0.002 ± 0.019**	0.882 ± 0.173	-
Instruct-DS-Prompt	Consistency (Avg)	GIRB	**0.007 ± 0.014**	**0.007 ± 0.019**	**0.002 ± 0.015**	**0.927 ± 0.439**	-
Instruct-DS-Prompt	Fluency (Avg)	None	0.027 ± 0.019	0.024 ± 0.009	0.002 ± 0.014	0.128 ± 0.048	-
Instruct-DS-Prompt	Fluency (Avg)	QAB	0.018 ± 0.021	0.011 ± 0.010	**0.001 ± 0.009**	**0.300 ± 0.122**	-
Instruct-DS-Prompt	Fluency (Avg)	GIRB	**0.011 ± 0.010**	**0.010 ± 0.015**	**0.001 ± 0.019**	0.262 ± 0.057	-

Table 12: Summary of performance across calibration methods for each metric in MathQA dataset. For each metric, we report two indicators: i. Gain: average improvement compared to "None", and ii. Wins: the number of setting where a given method outperforms the others.

	None		S-B		S		B		QAB		HS-QAB		S-QAB		GIRB	
Metric	Gain	Wins	Gain	Wins	Gain	Wins	Gain	Wins	Gain	Wins	Gain	Wins	Gain	Wins	Gain	Wins
ECE′	0	0	-0.061	0	**+0.344**	**4**	-0.014	0	+0.145	0	+0.017	0	+0.059	0	+0.185	0
GECE′	0	0	+0.281	0	+0.590	0	+0.329	0	+0.696	0	+0.747	0	**+0.828**	**3**	+0.749	**1**
Brier′	0	0	+1.038	0	+1.190	**2**	+1.116	0	+1.134	0	+1.129	0	+1.144	0	**+1.173**	**2**
AUAC	0	0	-0.230	0	-0.200	0	+0.017	0	+0.036	0	+0.009	0	-0.003	0	**+0.043**	**4**

Table 13: Summary of performance across calibration methods for each metric in OpenBookQA dataset. For each metric, we report two indicators: i. Gain: average improvement compared to "None", and ii. Wins: the number of setting where a given method outperforms the others.

	None		S-B		S		B		QAB		HS-QAB		S-QAB		GIRB	
Metric	Gain	Wins	Gain	Wins	Gain	Wins	Gain	Wins	Gain	Wins	Gain	Wins	Gain	Wins	Gain	Wins
ECE′	0	0	+0.388	0	**+1.085**	**4**	+0.532	0	+0.605	0	+0.447	0	+0.483	0	+0.745	0
GECE′	0	0	+0.260	0	+0.759	1	+0.341	0	+0.645	0	+0.642	0	**+0.754**	**2**	+0.728	1
Brier′	0	0	+0.431	0	**+0.636**	1	+0.603	0	+0.591	0	+0.400	0	+0.413	0	**+0.636**	**3**
AUAC	0	1	-0.523	0	-0.657	0	-0.179	0	-0.041	0	-0.053	0	-0.053	1	**+0.030**	**2**

Table 14: Summary of performance across calibration methods for each metric in SciQ dataset. For each metric, we report two indicators: i. Gain: average improvement compared to "None", and ii. Wins: the number of setting where a given method outperforms the others.

	None		S-B		S		B		QAB		HS-QAB		S-QAB		GIRB	
Metric	Gain	Wins	Gain	Wins	Gain	Wins	Gain	Wins	Gain	Wins	Gain	Wins	Gain	Wins	Gain	Wins
ECE′	0	0	+0.748	0	**+1.588**	**4**	+0.901	0	+1.090	0	+0.776	0	+0.837	0	+1.164	0
GECE′	0	0	+0.567	0	**+1.085**	**2**	+0.648	0	+0.990	1	+0.916	0	+1.027	1	+1.006	0
Brier′	0	0	+0.777	0	+0.995	0	+0.937	0	+0.936	0	+0.693	0	+0.693	0	**+1.009**	**4**
AUAC	0	0	-0.610	0	-0.660	0	-0.039	0	+0.068	0	+0.024	1	+0.015	0	**+0.127**	**3**

Table 15: Summary of performance across calibration methods for each metric in TriviaQA dataset. For each metric, we report two indicators: i. Gain: average improvement compared to "None", and ii. Wins: the number of setting where a given method outperforms the others.

	None		S-B		S		B		QAB		HS-QAB		S-QAB		GIRB	
Metric	Gain	Wins	Gain	Wins	Gain	Wins	Gain	Wins	Gain	Wins	Gain	Wins	Gain	Wins	Gain	Wins
ECE′	0	0	+1.567	0	**+2.718**	**4**	+2.127	0	+2.262	0	+1.488	0	+1.510	0	+2.317	0
GECE′	0	0	+1.541	0	**+2.263**	**2**	+2.035	0	+2.206	1	+1.675	0	+1.696	0	+2.216	1
Brier′	0	0	+1.397	0	**+1.802**	**2**	+1.759	0	+1.744	0	+1.225	0	+1.214	0	+1.763	**2**
AUAC	0	0	-0.668	0	-0.824	0	-0.220	0	+0.193	0	+0.002	0	+0.008	0	**+0.345**	**4**

Table 16: Summary of performance across calibration methods for each metric in TruthfulQA dataset. For each metric, we report two indicators: i. Gain: average improvement compared to "None", and ii. Wins: the number of settings where a given method outperforms the others.

	None		S-B		S		B		QAB		HS-QAB		S-QAB		GIRB	
Metric	Gain	Wins	Gain	Wins	Gain	Wins	Gain	Wins	Gain	Wins	Gain	Wins	Gain	Wins	Gain	Wins
ECE′	0	0	+1.008	0	**+2.132**	**4**	+1.936	0	+1.901	0	+0.974	0	+0.924	0	+1.896	0
GECE′	0	0	+0.981	0	**+2.073**	**3**	+2.014	0	+1.984	0	+0.994	0	+0.954	0	+2.010	1
Brier′	0	0	+0.886	0	**+1.480**	**2**	+1.464	0	+1.416	0	+0.796	0	+0.776	0	+1.433	2
AUAC	0	1	-0.554	0	-0.909	0	-0.651	0	+0.110	1	+0.053	1	+0.076	0	**+0.226**	**1**

Table 17: Performance of different calibration methods across various setting in MathQA dataset.

Dataset	Prompt	LLM	Calibrator	ECE	GECE	Brier	AUAC
MathQA	ling1s-topk	gemma	QAB	0.223 ± 0.01	0.298 ± 0.006	0.247 ± 0.006	0.636 ± 0.009
MathQA	ling1s-topk	gemma	GIRB	0.187 ± 0.005	0.264 ± 0.012	0.252 ± 0.006	$\mathbf{0.639 \pm 0.007}$
MathQA	ling1s-topk	gemma	HS-QAB	0.313 ± 0.081	0.264 ± 0.013	0.31 ± 0.071	0.628 ± 0.008
MathQA	ling1s-topk	gemma	HS-GIRB	0.283 ± 0.09	0.228 ± 0.013	0.31 ± 0.071	0.628 ± 0.008
MathQA	ling1s-topk	gemma	S-QAB	0.24 ± 0.188	0.218 ± 0.02	0.317 ± 0.113	0.613 ± 0.013
MathQA	ling1s-topk	gemma	S-GIRB	0.223 ± 0.203	$\mathbf{0.173 \pm 0.023}$	0.317 ± 0.113	0.613 ± 0.013
MathQA	ling1s-topk	gemma	S-B	0.381 ± 0.116	0.461 ± 0.018	0.32 ± 0.118	0.44 ± 0.024
MathQA	ling1s-topk	gemma	S	$\mathbf{0.014 \pm 0.005}$	0.283 ± 0.008	0.224 ± 0.004	0.446 ± 0.005
MathQA	ling1s-topk	gemma	B	0.314 ± 0.008	0.447 ± 0.003	0.25 ± 0.005	0.629 ± 0.007
MathQA	ling1s-topk	gemma	IRB	0.021 ± 0.001	0.223 ± 0.017	$\mathbf{0.221 \pm 0.003}$	0.63 ± 0.085
MathQA	ling1s-topk	gemma	None	0.24 ± 0.011	0.26 ± 0.005	0.282 ± 0.008	0.611 ± 0.006
MathQA	verb1s-topk	gemma	QAB	0.197 ± 0.007	0.291 ± 0.004	0.246 ± 0.002	0.648 ± 0.005
MathQA	verb1s-topk	gemma	GIRB	0.172 ± 0.006	0.273 ± 0.004	$\mathbf{0.225 \pm 0.003}$	$\mathbf{0.653 \pm 0.007}$
MathQA	verb1s-topk	gemma	HS-QAB	0.295 ± 0.048	0.266 ± 0.011	0.703 ± 0.795	0.622 ± 0.006
MathQA	verb1s-topk	gemma	HS-GIRB	0.359 ± 0.066	0.326 ± 0.026	0.703 ± 0.796	0.623 ± 0.006
MathQA	verb1s-topk	gemma	S-QAB	0.272 ± 0.092	$\mathbf{0.225 \pm 0.008}$	0.688 ± 0.821	0.625 ± 0.013
MathQA	verb1s-topk	gemma	S-GIRB	0.276 ± 0.117	0.312 ± 0.076	0.688 ± 0.822	0.626 ± 0.014
MathQA	verb1s-topk	gemma	S-B	0.332 ± 0.068	0.457 ± 0.004	0.268 ± 0.061	0.584 ± 0.078
MathQA	verb1s-topk	gemma	S	0.068 ± 0.009	0.328 ± 0.004	0.228 ± 0.003	0.651 ± 0.044
MathQA	verb1s-topk	gemma	B	0.317 ± 0.003	0.447 ± 0.002	0.253 ± 0.001	0.631 ± 0.005
MathQA	verb1s-topk	gemma	IRB	$\mathbf{0.022 \pm 0.006}$	0.276 ± 0.005	0.255 ± 0.002	0.61 ± 0.054
MathQA	verb1s-topk	gemma	None	0.454 ± 0.049	0.8 ± 0.272	1.98 ± 2.989	0.625 ± 0.007
MathQA	ling1s-topk	mistral	QAB	0.199 ± 0.007	0.289 ± 0.003	0.243 ± 0.006	0.646 ± 0.007
MathQA	ling1s-topk	mistral	GIRB	0.17 ± 0.008	0.267 ± 0.004	0.247 ± 0.007	$\mathbf{0.65 \pm 0.005}$
MathQA	ling1s-topk	mistral	HS-QAB	0.276 ± 0.036	0.277 ± 0.029	0.275 ± 0.03	0.631 ± 0.009
MathQA	ling1s-topk	mistral	HS-GIRB	0.335 ± 0.064	0.368 ± 0.036	0.275 ± 0.031	0.632 ± 0.009
MathQA	ling1s-topk	mistral	S-QAB	0.262 ± 0.092	0.249 ± 0.069	0.27 ± 0.069	0.621 ± 0.016
MathQA	ling1s-topk	mistral	S-GIRB	0.254 ± 0.117	0.41 ± 0.05	0.27 ± 0.069	0.62 ± 0.016
MathQA	ling1s-topk	mistral	S-B	0.335 ± 0.08	0.473 ± 0.041	0.27 ± 0.072	0.402 ± 0.089
MathQA	ling1s-topk	mistral	S	0.024 ± 0.006	0.315 ± 0.002	$\mathbf{0.223 \pm 0.003}$	0.45 ± 0.005
MathQA	ling1s-topk	mistral	B	0.31 ± 0.003	0.443 ± 0.002	0.249 ± 0.001	0.635 ± 0.004
MathQA	ling1s-topk	mistral	IRB	$\mathbf{0.017 \pm 0.009}$	$\mathbf{0.249 \pm 0.016}$	0.223 ± 0.003	0.52 ± 0.069
MathQA	ling1s-topk	mistral	None	0.189 ± 0.011	0.337 ± 0.003	0.277 ± 0.006	0.62 ± 0.005
MathQA	verb1s-topk	mistral	QAB	0.196 ± 0.004	0.295 ± 0.007	0.243 ± 0.005	0.65 ± 0.012
MathQA	verb1s-topk	mistral	GIRB	0.173 ± 0.007	$\mathbf{0.272 \pm 0.005}$	$\mathbf{0.223 \pm 0.003}$	$\mathbf{0.654 \pm 0.006}$
MathQA	verb1s-topk	mistral	HS-QAB	0.283 ± 0.043	0.287 ± 0.045	0.28 ± 0.034	0.63 ± 0.008
MathQA	verb1s-topk	mistral	HS-GIRB	0.365 ± 0.073	0.368 ± 0.074	0.28 ± 0.035	0.63 ± 0.006
MathQA	verb1s-topk	mistral	S-QAB	0.275 ± 0.092	0.279 ± 0.094	0.272 ± 0.069	0.624 ± 0.025
MathQA	verb1s-topk	mistral	S-GIRB	0.309 ± 0.167	0.32 ± 0.177	0.272 ± 0.069	0.624 ± 0.025
MathQA	verb1s-topk	mistral	S-B	0.343 ± 0.078	0.487 ± 0.058	0.271 ± 0.072	0.493 ± 0.146
MathQA	verb1s-topk	mistral	S	0.146 ± 0.004	0.424 ± 0.002	0.224 ± 0.003	0.446 ± 0.007
MathQA	verb1s-topk	mistral	B	0.314 ± 0.007	0.443 ± 0.005	0.25 ± 0.005	0.636 ± 0.012
MathQA	verb1s-topk	mistral	IRB	$\mathbf{0.025 \pm 0.003}$	0.297 ± 0.016	0.223 ± 0.003	0.571 ± 0.053
MathQA	verb1s-topk	mistral	None	0.269 ± 0.008	0.422 ± 0.009	0.895 ± 0.651	0.634 ± 0.008

Table 18: Performance of different calibration methods across various setting in MMLU dataset.

Dataset	Prompt	LLM	Calibrator	ECE	GECE	Brier	AUAC
MMLU	ling1s-topk	gemma	QAB	0.191 ± 0.006	0.226 ± 0.006	0.183 ± 0.01	$\mathbf{0.311 \pm 0.039}$
MMLU	ling1s-topk	gemma	GIRB	0.164 ± 0.013	0.221 ± 0.013	$\mathbf{0.171 \pm 0.012}$	$\mathbf{0.311 \pm 0.028}$
MMLU	ling1s-topk	gemma	HS-QAB	0.336 ± 0.105	0.308 ± 0.012	0.299 ± 0.094	0.206 ± 0.011
MMLU	ling1s-topk	gemma	HS-GIRB	0.328 ± 0.104	0.298 ± 0.013	0.299 ± 0.095	0.206 ± 0.011
MMLU	ling1s-topk	gemma	S-QAB	0.284 ± 0.17	0.285 ± 0.02	0.303 ± 0.135	0.205 ± 0.018
MMLU	ling1s-topk	gemma	S-GIRB	0.277 ± 0.173	0.275 ± 0.023	0.303 ± 0.135	0.206 ± 0.018
MMLU	ling1s-topk	gemma	S-B	0.334 ± 0.154	0.392 ± 0.033	0.266 ± 0.151	0.074 ± 0.013
MMLU	ling1s-topk	gemma	S	$\mathbf{0.016 \pm 0.008}$	$\mathbf{0.179 \pm 0.007}$	0.174 ± 0.008	0.05 ± 0.002
MMLU	ling1s-topk	gemma	B	0.263 ± 0.007	0.339 ± 0.005	0.196 ± 0.009	0.279 ± 0.077
MMLU	ling1s-topk	gemma	IRB	0.022 ± 0.006	0.233 ± 0.023	0.174 ± 0.007	0.255 ± 0.092
MMLU	ling1s-topk	gemma	None	0.623 ± 0.013	0.645 ± 0.008	0.586 ± 0.01	0.21 ± 0.009
MMLU	verb1s-topk	gemma	QAB	0.188 ± 0.005	0.235 ± 0.002	0.185 ± 0.003	0.343 ± 0.02
MMLU	verb1s-topk	gemma	GIRB	0.153 ± 0.004	$\mathbf{0.22 \pm 0.006}$	$\mathbf{0.168 \pm 0.004}$	$\mathbf{0.394 \pm 0.015}$
MMLU	verb1s-topk	gemma	HS-QAB	0.311 ± 0.103	0.331 ± 0.1	0.281 ± 0.087	0.237 ± 0.015
MMLU	verb1s-topk	gemma	HS-GIRB	0.327 ± 0.106	0.355 ± 0.101	0.281 ± 0.087	0.238 ± 0.015
MMLU	verb1s-topk	gemma	S-QAB	0.311 ± 0.145	0.321 ± 0.138	0.283 ± 0.126	0.257 ± 0.02
MMLU	verb1s-topk	gemma	S-GIRB	0.295 ± 0.162	0.36 ± 0.163	0.282 ± 0.126	0.258 ± 0.018
MMLU	verb1s-topk	gemma	S-B	0.336 ± 0.142	0.445 ± 0.115	0.26 ± 0.136	0.101 ± 0.053
MMLU	verb1s-topk	gemma	S	0.077 ± 0.007	0.263 ± 0.003	0.177 ± 0.002	0.087 ± 0.014
MMLU	verb1s-topk	gemma	B	0.255 ± 0.004	0.327 ± 0.004	0.189 ± 0.003	0.378 ± 0.049
MMLU	verb1s-topk	gemma	IRB	$\mathbf{0.025 \pm 0.004}$	0.235 ± 0.006	0.17 ± 0.003	0.278 ± 0.013
MMLU	verb1s-topk	gemma	None	0.62 ± 0.003	0.645 ± 0.003	0.567 ± 0.003	0.236 ± 0.004
MMLU	ling1s-topk	mistral	QAB	0.199 ± 0.005	0.249 ± 0.005	0.199 ± 0.006	0.318 ± 0.048
MMLU	ling1s-topk	mistral	GIRB	0.159 ± 0.009	0.227 ± 0.007	$\mathbf{0.191 \pm 0.009}$	$\mathbf{0.357 \pm 0.023}$
MMLU	ling1s-topk	mistral	HS-QAB	0.305 ± 0.095	0.27 ± 0.016	0.278 ± 0.078	0.245 ± 0.023
MMLU	ling1s-topk	mistral	HS-GIRB	0.319 ± 0.098	0.282 ± 0.024	0.278 ± 0.079	0.245 ± 0.022
MMLU	ling1s-topk	mistral	S-QAB	0.305 ± 0.144	0.246 ± 0.021	0.283 ± 0.121	0.232 ± 0.033
MMLU	ling1s-topk	mistral	S-GIRB	0.292 ± 0.156	0.255 ± 0.059	0.283 ± 0.121	0.232 ± 0.034
MMLU	ling1s-topk	mistral	S-B	0.339 ± 0.134	0.4 ± 0.019	0.272 ± 0.125	0.092 ± 0.014
MMLU	ling1s-topk	mistral	S	0.036 ± 0.003	0.284 ± 0.006	0.194 ± 0.005	0.07 ± 0.002
MMLU	ling1s-topk	mistral	B	0.285 ± 0.007	0.373 ± 0.003	0.219 ± 0.006	0.206 ± 0.018
MMLU	ling1s-topk	mistral	IRB	$\mathbf{0.02 \pm 0.006}$	$\mathbf{0.203 \pm 0.007}$	0.193 ± 0.005	0.123 ± 0.022
MMLU	ling1s-topk	mistral	None	0.539 ± 0.008	0.594 ± 0.006	0.501 ± 0.006	0.235 ± 0.006
MMLU	verb1s-topk	mistral	QAB	0.189 ± 0.007	0.239 ± 0.007	0.19 ± 0.003	0.334 ± 0.032
MMLU	verb1s-topk	mistral	GIRB	0.157 ± 0.004	$\mathbf{0.227 \pm 0.011}$	$\mathbf{0.178 \pm 0.003}$	$\mathbf{0.34 \pm 0.015}$
MMLU	verb1s-topk	mistral	HS-QAB	0.316 ± 0.092	0.27 ± 0.016	0.285 ± 0.081	0.264 ± 0.021
MMLU	verb1s-topk	mistral	HS-GIRB	0.328 ± 0.097	0.285 ± 0.022	0.285 ± 0.081	0.264 ± 0.022
MMLU	verb1s-topk	mistral	S-QAB	0.307 ± 0.143	0.238 ± 0.016	0.285 ± 0.122	0.256 ± 0.032
MMLU	verb1s-topk	mistral	S-GIRB	0.301 ± 0.159	0.249 ± 0.062	0.285 ± 0.122	0.257 ± 0.032
MMLU	verb1s-topk	mistral	S-B	0.331 ± 0.137	0.393 ± 0.022	0.259 ± 0.13	0.095 ± 0.013
MMLU	verb1s-topk	mistral	S	0.064 ± 0.009	0.297 ± 0.005	0.184 ± 0.004	0.066 ± 0.006
MMLU	verb1s-topk	mistral	B	0.274 ± 0.007	0.363 ± 0.006	0.202 ± 0.006	0.298 ± 0.036
MMLU	verb1s-topk	mistral	IRB	$\mathbf{0.02 \pm 0.007}$	0.235 ± 0.015	0.184 ± 0.004	0.093 ± 0.026
MMLU	verb1s-topk	mistral	None	0.635 ± 0.007	0.674 ± 0.006	0.601 ± 0.005	0.234 ± 0.009

Table 19: Performance of different calibration methods across various setting in OpenBookQA dataset.

Dataset	Prompt	LLM	Calibrator	ECE	GECE	Brier	AUAC
OpenBookQA	ling1s-topk	gemma	QAB	0.267 ± 0.011	0.283 ± 0.011	0.246 ± 0.014	0.302 ± 0.031
OpenBookQA	ling1s-topk	gemma	GIRB	0.209 ± 0.026	0.255 ± 0.01	$\mathbf{0.215 \pm 0.011}$	0.313 ± 0.041
OpenBookQA	ling1s-topk	gemma	HS-QAB	0.34 ± 0.106	0.277 ± 0.029	0.337 ± 0.095	0.299 ± 0.027
OpenBookQA	ling1s-topk	gemma	HS-GIRB	0.322 ± 0.114	0.256 ± 0.03	0.338 ± 0.095	0.3 ± 0.027
OpenBookQA	ling1s-topk	gemma	S-QAB	0.297 ± 0.161	0.233 ± 0.032	0.326 ± 0.127	0.324 ± 0.036
OpenBookQA	ling1s-topk	gemma	S-GIRB	0.281 ± 0.172	0.201 ± 0.041	0.326 ± 0.128	$\mathbf{0.325 \pm 0.036}$
OpenBookQA	ling1s-topk	gemma	S-B	0.37 ± 0.122	0.457 ± 0.03	0.321 ± 0.129	0.136 ± 0.028
OpenBookQA	ling1s-topk	gemma	S	$\mathbf{0.023 \pm 0.006}$	$\mathbf{0.163 \pm 0.011}$	0.217 ± 0.003	0.102 ± 0.007
OpenBookQA	ling1s-topk	gemma	B	0.303 ± 0.018	0.431 ± 0.011	0.241 ± 0.009	0.225 ± 0.034
OpenBookQA	ling1s-topk	gemma	IRB	0.046 ± 0.021	0.249 ± 0.026	0.218 ± 0.003	0.164 ± 0.053
OpenBookQA	ling1s-topk	gemma	None	0.56 ± 0.015	0.562 ± 0.024	0.537 ± 0.012	0.294 ± 0.021
OpenBookQA	verb1s-topk	gemma	QAB	0.267 ± 0.007	0.294 ± 0.004	0.231 ± 0.002	0.284 ± 0.031
OpenBookQA	verb1s-topk	gemma	GIRB	0.207 ± 0.023	0.257 ± 0.007	$\mathbf{0.206 \pm 0.004}$	$\mathbf{0.31 \pm 0.034}$
OpenBookQA	verb1s-topk	gemma	HS-QAB	0.341 ± 0.124	0.302 ± 0.026	0.332 ± 0.116	0.286 ± 0.019
OpenBookQA	verb1s-topk	gemma	HS-GIRB	0.347 ± 0.137	0.314 ± 0.028	0.332 ± 0.116	0.286 ± 0.019
OpenBookQA	verb1s-topk	gemma	S-QAB	0.339 ± 0.142	0.255 ± 0.032	0.325 ± 0.134	0.288 ± 0.029
OpenBookQA	verb1s-topk	gemma	S-GIRB	0.314 ± 0.165	0.266 ± 0.079	0.325 ± 0.134	0.29 ± 0.03
OpenBookQA	verb1s-topk	gemma	S-B	0.363 ± 0.137	0.455 ± 0.033	0.315 ± 0.14	0.136 ± 0.024
OpenBookQA	verb1s-topk	gemma	S	$\mathbf{0.064 \pm 0.009}$	0.268 ± 0.011	0.234 ± 0.007	0.097 ± 0.003
OpenBookQA	verb1s-topk	gemma	B	0.296 ± 0.007	0.412 ± 0.003	0.23 ± 0.012	0.232 ± 0.018
OpenBookQA	verb1s-topk	gemma	IRB	0.044 ± 0.018	$\mathbf{0.237 \pm 0.009}$	0.207 ± 0.008	0.168 ± 0.017
OpenBookQA	verb1s-topk	gemma	None	0.576 ± 0.016	0.589 ± 0.013	0.547 ± 0.012	0.291 ± 0.013
OpenBookQA	ling1s-topk	mistral	QAB	0.277 ± 0.015	0.299 ± 0.007	0.252 ± 0.012	0.264 ± 0.04
OpenBookQA	ling1s-topk	mistral	GIRB	0.202 ± 0.019	0.264 ± 0.01	$\mathbf{0.222 \pm 0.015}$	$\mathbf{0.324 \pm 0.025}$
OpenBookQA	ling1s-topk	mistral	HS-QAB	0.332 ± 0.103	0.29 ± 0.018	0.328 ± 0.095	0.269 ± 0.044
OpenBookQA	ling1s-topk	mistral	HS-GIRB	0.353 ± 0.109	0.296 ± 0.032	0.327 ± 0.095	0.27 ± 0.044
OpenBookQA	ling1s-topk	mistral	S-QAB	0.331 ± 0.132	$\mathbf{0.238 \pm 0.029}$	0.329 ± 0.118	0.219 ± 0.058
OpenBookQA	ling1s-topk	mistral	S-GIRB	0.307 ± 0.154	0.248 ± 0.085	0.328 ± 0.119	0.216 ± 0.059
OpenBookQA	ling1s-topk	mistral	S-B	0.376 ± 0.115	0.465 ± 0.027	0.325 ± 0.119	0.151 ± 0.025
OpenBookQA	ling1s-topk	mistral	S	0.061 ± 0.009	0.283 ± 0.009	0.223 ± 0.009	0.11 ± 0.005
OpenBookQA	ling1s-topk	mistral	B	0.305 ± 0.014	0.432 ± 0.007	0.247 ± 0.014	0.286 ± 0.044
OpenBookQA	ling1s-topk	mistral	IRB	$\mathbf{0.048 \pm 0.019}$	0.255 ± 0.021	0.223 ± 0.009	0.151 ± 0.024
OpenBookQA	ling1s-topk	mistral	None	0.456 ± 0.016	0.51 ± 0.013	0.442 ± 0.009	0.308 ± 0.013
OpenBookQA	verb1s-topk	mistral	QAB	0.271 ± 0.017	0.299 ± 0.009	0.251 ± 0.018	0.293 ± 0.022
OpenBookQA	verb1s-topk	mistral	GIRB	0.206 ± 0.02	0.257 ± 0.012	0.249 ± 0.022	0.282 ± 0.016
OpenBookQA	verb1s-topk	mistral	HS-QAB	0.353 ± 0.101	0.308 ± 0.03	0.344 ± 0.093	0.275 ± 0.016
OpenBookQA	verb1s-topk	mistral	HS-GIRB	0.352 ± 0.111	0.311 ± 0.031	0.343 ± 0.093	0.279 ± 0.016
OpenBookQA	verb1s-topk	mistral	S-QAB	0.335 ± 0.129	0.258 ± 0.036	0.336 ± 0.114	0.296 ± 0.017
OpenBookQA	verb1s-topk	mistral	S-GIRB	0.314 ± 0.148	$\mathbf{0.256 \pm 0.077}$	0.336 ± 0.114	0.297 ± 0.017
OpenBookQA	verb1s-topk	mistral	S-B	0.368 ± 0.114	0.459 ± 0.029	0.322 ± 0.12	0.146 ± 0.026
OpenBookQA	verb1s-topk	mistral	S	0.06 ± 0.013	0.264 ± 0.007	$\mathbf{0.219 \pm 0.009}$	0.1 ± 0.003
OpenBookQA	verb1s-topk	mistral	B	0.31 ± 0.016	0.423 ± 0.004	0.24 ± 0.014	0.238 ± 0.038
OpenBookQA	verb1s-topk	mistral	IRB	$\mathbf{0.05 \pm 0.015}$	0.287 ± 0.008	$\mathbf{0.219 \pm 0.009}$	0.201 ± 0.096
OpenBookQA	verb1s-topk	mistral	None	0.57 ± 0.017	0.609 ± 0.013	0.559 ± 0.019	$\mathbf{0.3 \pm 0.011}$

Table 20: Performance of different calibration methods across various setting in SciQ dataset.

Dataset	Prompt	LLM	Calibrator	ECE	GECE	Brier	AUAC
SciQ	ling1s-topk	gemma	QAB	0.23 ± 0.007	0.281 ± 0.006	0.218 ± 0.005	0.253 ± 0.029
SciQ	ling1s-topk	gemma	GIRB	0.177 ± 0.007	0.237 ± 0.009	$\mathbf{0.195 \pm 0.006}$	0.28 ± 0.013
SciQ	ling1s-topk	gemma	HS-QAB	0.315 ± 0.113	0.27 ± 0.022	0.301 ± 0.1	$\mathbf{0.286 \pm 0.03}$
SciQ	ling1s-topk	gemma	HS-GIRB	0.284 ± 0.12	0.229 ± 0.022	0.301 ± 0.1	0.286 ± 0.03
SciQ	ling1s-topk	gemma	S-QAB	0.25 ± 0.186	0.231 ± 0.023	0.297 ± 0.138	0.278 ± 0.043
SciQ	ling1s-topk	gemma	S-GIRB	0.234 ± 0.197	0.177 ± 0.03	0.297 ± 0.138	0.277 ± 0.043
SciQ	ling1s-topk	gemma	S-B	0.351 ± 0.14	0.433 ± 0.025	0.291 ± 0.14	0.105 ± 0.017
SciQ	ling1s-topk	gemma	S	$\mathbf{0.009 \pm 0.004}$	$\mathbf{0.135 \pm 0.006}$	0.198 ± 0.004	0.08 ± 0.003
SciQ	ling1s-topk	gemma	B	0.287 ± 0.005	0.405 ± 0.007	0.22 ± 0.005	0.246 ± 0.037
SciQ	ling1s-topk	gemma	IRB	0.024 ± 0.011	0.203 ± 0.026	0.199 ± 0.004	0.096 ± 0.014
SciQ	ling1s-topk	gemma	None	0.625 ± 0.007	0.607 ± 0.01	0.59 ± 0.006	0.27 ± 0.01
SciQ	verb1s-topk	gemma	QAB	0.212 ± 0.005	$\mathbf{0.27 \pm 0.004}$	0.22 ± 0.002	0.321 ± 0.009
SciQ	verb1s-topk	gemma	GIRB	0.195 ± 0.014	0.277 ± 0.009	$\mathbf{0.194 \pm 0.004}$	$\mathbf{0.325 \pm 0.011}$
SciQ	verb1s-topk	gemma	HS-QAB	0.349 ± 0.094	0.316 ± 0.013	0.332 ± 0.08	0.272 ± 0.013
SciQ	verb1s-topk	gemma	HS-GIRB	0.364 ± 0.092	0.335 ± 0.017	0.332 ± 0.08	0.271 ± 0.013
SciQ	verb1s-topk	gemma	S-QAB	0.345 ± 0.135	0.272 ± 0.015	0.336 ± 0.116	0.275 ± 0.016
SciQ	verb1s-topk	gemma	S-GIRB	0.339 ± 0.143	0.282 ± 0.053	0.336 ± 0.116	0.275 ± 0.016
SciQ	verb1s-topk	gemma	S-B	0.344 ± 0.138	0.422 ± 0.023	0.286 ± 0.134	0.109 ± 0.021
SciQ	verb1s-topk	gemma	S	$\mathbf{0.03 \pm 0.012}$	0.272 ± 0.006	0.198 ± 0.003	0.097 ± 0.004
SciQ	verb1s-topk	gemma	B	0.289 ± 0.009	0.395 ± 0.008	0.222 ± 0.004	0.245 ± 0.022
SciQ	verb1s-topk	gemma	IRB	0.044 ± 0.009	0.315 ± 0.015	0.199 ± 0.004	0.153 ± 0.023
SciQ	verb1s-topk	gemma	None	0.676 ± 0.01	0.681 ± 0.009	0.654 ± 0.01	0.267 ± 0.009
SciQ	ling1s-topk	mistral	QAB	0.218 ± 0.012	0.274 ± 0.005	0.227 ± 0.009	0.275 ± 0.03
SciQ	ling1s-topk	mistral	GIRB	0.195 ± 0.015	0.272 ± 0.006	$\mathbf{0.196 \pm 0.01}$	$\mathbf{0.305 \pm 0.025}$
SciQ	ling1s-topk	mistral	HS-QAB	0.333 ± 0.087	0.298 ± 0.011	0.316 ± 0.074	0.274 ± 0.012
SciQ	ling1s-topk	mistral	HS-GIRB	0.351 ± 0.09	0.323 ± 0.016	0.315 ± 0.075	0.275 ± 0.012
SciQ	ling1s-topk	mistral	S-QAB	0.327 ± 0.126	$\mathbf{0.265 \pm 0.014}$	0.315 ± 0.108	0.268 ± 0.017
SciQ	ling1s-topk	mistral	S-GIRB	0.318 ± 0.135	0.282 ± 0.056	0.315 ± 0.108	0.269 ± 0.017
SciQ	ling1s-topk	mistral	S-B	0.348 ± 0.129	0.429 ± 0.019	0.284 ± 0.126	0.105 ± 0.013
SciQ	ling1s-topk	mistral	S	0.042 ± 0.013	0.287 ± 0.002	0.198 ± 0.008	0.095 ± 0.001
SciQ	ling1s-topk	mistral	B	0.291 ± 0.01	0.401 ± 0.005	0.221 ± 0.009	0.258 ± 0.029
SciQ	ling1s-topk	mistral	IRB	$\mathbf{0.038 \pm 0.012}$	0.267 ± 0.016	0.197 ± 0.008	0.132 ± 0.049
SciQ	ling1s-topk	mistral	None	0.531 ± 0.022	0.569 ± 0.005	0.492 ± 0.017	0.273 ± 0.004
SciQ	verb1s-topk	mistral	QAB	0.219 ± 0.012	0.271 ± 0.006	0.232 ± 0.008	0.32 ± 0.028
SciQ	verb1s-topk	mistral	GIRB	0.2 ± 0.009	0.281 ± 0.009	$\mathbf{0.194 \pm 0.008}$	$\mathbf{0.323 \pm 0.012}$
SciQ	verb1s-topk	mistral	HS-QAB	0.346 ± 0.08	0.312 ± 0.01	0.332 ± 0.068	0.288 ± 0.02
SciQ	verb1s-topk	mistral	HS-GIRB	0.357 ± 0.082	0.326 ± 0.016	0.332 ± 0.068	0.289 ± 0.02
SciQ	verb1s-topk	mistral	S-QAB	0.331 ± 0.124	0.269 ± 0.015	0.332 ± 0.104	0.289 ± 0.025
SciQ	verb1s-topk	mistral	S-GIRB	0.324 ± 0.131	0.275 ± 0.046	0.332 ± 0.104	0.289 ± 0.025
SciQ	verb1s-topk	mistral	S-B	0.347 ± 0.12	0.429 ± 0.02	0.289 ± 0.118	0.107 ± 0.013
SciQ	verb1s-topk	mistral	S	0.054 ± 0.009	$\mathbf{0.255 \pm 0.004}$	0.206 ± 0.003	0.1 ± 0.003
SciQ	verb1s-topk	mistral	B	0.293 ± 0.009	0.394 ± 0.004	0.23 ± 0.005	0.304 ± 0.026
SciQ	verb1s-topk	mistral	IRB	$\mathbf{0.033 \pm 0.007}$	0.299 ± 0.01	0.205 ± 0.002	0.156 ± 0.023
SciQ	verb1s-topk	mistral	None	0.646 ± 0.007	0.662 ± 0.006	0.627 ± 0.008	0.284 ± 0.007

Table 21: Performance of different calibration methods across various setting in TriviaQA dataset.

Dataset	Prompt	LLM	Calibrator	ECE	GECE	Brier	AUAC
TriviaQA	ling1s-topk	gemma	QAB	0.159 ± 0.004	0.187 ± 0.007	0.127 ± 0.007	0.141 ± 0.041
TriviaQA	ling1s-topk	gemma	GIRB	0.138 ± 0.006	0.171 ± 0.004	$\mathbf{0.105 \pm 0.009}$	$\mathbf{0.156 \pm 0.042}$
TriviaQA	ling1s-topk	gemma	HS-QAB	0.348 ± 0.176	0.304 ± 0.039	0.288 ± 0.159	0.128 ± 0.009
TriviaQA	ling1s-topk	gemma	HS-GIRB	0.336 ± 0.181	0.287 ± 0.04	0.288 ± 0.159	0.128 ± 0.009
TriviaQA	ling1s-topk	gemma	S-QAB	0.321 ± 0.224	0.3 ± 0.053	0.29 ± 0.195	0.129 ± 0.011
TriviaQA	ling1s-topk	gemma	S-GIRB	0.31 ± 0.232	0.282 ± 0.059	0.29 ± 0.195	0.129 ± 0.011
TriviaQA	ling1s-topk	gemma	S-B	0.332 ± 0.22	0.348 ± 0.066	0.244 ± 0.215	0.043 ± 0.01
TriviaQA	ling1s-topk	gemma	S	0.037 ± 0.005	$\mathbf{0.118 \pm 0.007}$	0.111 ± 0.005	0.016 ± 0.001
TriviaQA	ling1s-topk	gemma	B	0.188 ± 0.008	0.224 ± 0.006	0.126 ± 0.005	0.073 ± 0.011
TriviaQA	ling1s-topk	gemma	IRB	$\mathbf{0.017 \pm 0.004}$	0.166 ± 0.014	0.111 ± 0.005	0.022 ± 0.003
TriviaQA	ling1s-topk	gemma	None	0.756 ± 0.007	0.755 ± 0.007	0.694 ± 0.005	0.122 ± 0.007
TriviaQA	verb1s-topk	gemma	QAB	0.142 ± 0.009	$\mathbf{0.176 \pm 0.006}$	0.124 ± 0.009	0.152 ± 0.069
TriviaQA	verb1s-topk	gemma	GIRB	0.13 ± 0.008	0.179 ± 0.007	0.133 ± 0.008	$\mathbf{0.168 \pm 0.059}$
TriviaQA	verb1s-topk	gemma	HS-QAB	0.353 ± 0.153	0.353 ± 0.153	0.293 ± 0.137	0.117 ± 0.013
TriviaQA	verb1s-topk	gemma	HS-GIRB	0.359 ± 0.153	0.36 ± 0.153	0.293 ± 0.137	0.117 ± 0.013
TriviaQA	verb1s-topk	gemma	S-QAB	0.362 ± 0.194	0.365 ± 0.193	0.301 ± 0.182	0.113 ± 0.024
TriviaQA	verb1s-topk	gemma	S-GIRB	0.354 ± 0.201	0.357 ± 0.204	0.301 ± 0.183	0.113 ± 0.024
TriviaQA	verb1s-topk	gemma	S-B	0.327 ± 0.213	0.372 ± 0.209	0.233 ± 0.209	0.045 ± 0.037
TriviaQA	verb1s-topk	gemma	S	0.03 ± 0.003	0.182 ± 0.005	0.106 ± 0.006	0.024 ± 0.005
TriviaQA	verb1s-topk	gemma	B	0.177 ± 0.004	0.211 ± 0.006	0.119 ± 0.007	0.108 ± 0.015
TriviaQA	verb1s-topk	gemma	IRB	$\mathbf{0.025 \pm 0.006}$	0.193 ± 0.007	$\mathbf{0.105 \pm 0.006}$	0.086 ± 0.068
TriviaQA	verb1s-topk	gemma	None	0.788 ± 0.008	0.793 ± 0.008	0.735 ± 0.008	0.122 ± 0.009
TriviaQA	ling1s-topk	mistral	QAB	0.159 ± 0.003	0.206 ± 0.005	0.143 ± 0.007	0.165 ± 0.014
TriviaQA	ling1s-topk	mistral	GIRB	0.148 ± 0.008	0.205 ± 0.007	$\mathbf{0.115 \pm 0.007}$	$\mathbf{0.198 \pm 0.019}$
TriviaQA	ling1s-topk	mistral	HS-QAB	0.351 ± 0.144	0.302 ± 0.029	0.296 ± 0.126	0.157 ± 0.01
TriviaQA	ling1s-topk	mistral	HS-GIRB	0.362 ± 0.143	0.313 ± 0.033	0.296 ± 0.126	0.158 ± 0.01
TriviaQA	ling1s-topk	mistral	S-QAB	0.352 ± 0.182	0.284 ± 0.038	0.299 ± 0.166	0.156 ± 0.017
TriviaQA	ling1s-topk	mistral	S-GIRB	0.345 ± 0.187	0.298 ± 0.061	0.299 ± 0.166	0.156 ± 0.017
TriviaQA	ling1s-topk	mistral	S-B	0.341 ± 0.203	0.356 ± 0.05	0.251 ± 0.196	0.049 ± 0.011
TriviaQA	ling1s-topk	mistral	S	$\mathbf{0.021 \pm 0.007}$	0.214 ± 0.004	0.125 ± 0.007	0.03 ± 0.001
TriviaQA	ling1s-topk	mistral	B	0.204 ± 0.006	0.261 ± 0.006	0.14 ± 0.006	0.131 ± 0.031
TriviaQA	ling1s-topk	mistral	IRB	$\mathbf{0.021 \pm 0.007}$	$\mathbf{0.18 \pm 0.012}$	0.125 ± 0.007	0.046 ± 0.031
TriviaQA	ling1s-topk	mistral	None	0.645 ± 0.007	0.655 ± 0.008	0.575 ± 0.006	0.158 ± 0.008
TriviaQA	verb1s-topk	mistral	QAB	0.159 ± 0.008	0.191 ± 0.006	$\mathbf{0.148 \pm 0.007}$	0.2 ± 0.038
TriviaQA	verb1s-topk	mistral	GIRB	0.147 ± 0.007	0.194 ± 0.007	0.157 ± 0.006	$\mathbf{0.222 \pm 0.031}$
TriviaQA	verb1s-topk	mistral	HS-QAB	0.366 ± 0.143	0.32 ± 0.033	0.316 ± 0.13	0.152 ± 0.014
TriviaQA	verb1s-topk	mistral	HS-GIRB	0.369 ± 0.144	0.327 ± 0.035	0.316 ± 0.13	0.153 ± 0.014
TriviaQA	verb1s-topk	mistral	S-QAB	0.361 ± 0.184	0.301 ± 0.041	0.316 ± 0.167	0.16 ± 0.018
TriviaQA	verb1s-topk	mistral	S-GIRB	0.355 ± 0.19	0.306 ± 0.059	0.316 ± 0.168	0.161 ± 0.019
TriviaQA	verb1s-topk	mistral	S-B	0.34 ± 0.199	0.347 ± 0.053	0.252 ± 0.193	0.045 ± 0.011
TriviaQA	verb1s-topk	mistral	S	0.052 ± 0.003	$\mathbf{0.191 \pm 0.003}$	0.128 ± 0.008	0.028 ± 0.001
TriviaQA	verb1s-topk	mistral	B	0.193 ± 0.005	0.242 ± 0.003	0.139 ± 0.006	0.122 ± 0.029
TriviaQA	verb1s-topk	mistral	IRB	$\mathbf{0.033 \pm 0.006}$	0.203 ± 0.008	0.126 ± 0.007	0.061 ± 0.027
TriviaQA	verb1s-topk	mistral	None	0.736 ± 0.012	0.749 ± 0.008	0.696 ± 0.011	0.151 ± 0.007

Table 22: Performance of different calibration methods across various setting in TruthfulQA dataset.

Dataset	Prompt	LLM	Calibrator	ECE	GECE	Brier	AUAC
TruthfulQA	ling1s-topk	gemma	QAB	0.12 ± 0.024	0.146 ± 0.03	0.092 ± 0.028	0.045 ± 0.056
TruthfulQA	ling1s-topk	gemma	GIRB	0.12 ± 0.023	0.14 ± 0.026	0.097 ± 0.029	0.055 ± 0.065
TruthfulQA	ling1s-topk	gemma	HS-QAB	0.463 ± 0.175	0.48 ± 0.093	0.374 ± 0.175	0.062 ± 0.028
TruthfulQA	ling1s-topk	gemma	HS-GIRB	0.463 ± 0.176	0.479 ± 0.094	0.373 ± 0.176	0.063 ± 0.029
TruthfulQA	ling1s-topk	gemma	S-QAB	0.482 ± 0.205	0.492 ± 0.098	0.384 ± 0.206	0.059 ± 0.033
TruthfulQA	ling1s-topk	gemma	S-GIRB	0.481 ± 0.206	0.493 ± 0.1	0.383 ± 0.206	0.059 ± 0.033
TruthfulQA	ling1s-topk	gemma	S-B	0.388 ± 0.283	0.443 ± 0.094	0.286 ± 0.277	0.024 ± 0.008
TruthfulQA	ling1s-topk	gemma	S	0.026 ± 0.005	$\mathbf{0.095 \pm 0.012}$	0.058 ± 0.008	0.005 ± 0.002
TruthfulQA	ling1s-topk	gemma	B	0.091 ± 0.024	0.12 ± 0.013	0.062 ± 0.013	0.014 ± 0.014
TruthfulQA	ling1s-topk	gemma	IRB	$\mathbf{0.023 \pm 0.008}$	0.107 ± 0.014	$\mathbf{0.057 \pm 0.008}$	0.007 ± 0.004
TruthfulQA	ling1s-topk	gemma	None	0.721 ± 0.041	0.73 ± 0.023	0.65 ± 0.039	$\mathbf{0.065 \pm 0.021}$
TruthfulQA	verb1s-topk	gemma	QAB	0.126 ± 0.027	0.131 ± 0.019	0.089 ± 0.033	0.105 ± 0.087
TruthfulQA	verb1s-topk	gemma	GIRB	0.124 ± 0.034	0.127 ± 0.014	$\mathbf{0.069 \pm 0.034}$	$\mathbf{0.106 \pm 0.074}$
TruthfulQA	verb1s-topk	gemma	HS-QAB	0.377 ± 0.238	0.44 ± 0.078	0.296 ± 0.224	0.071 ± 0.027
TruthfulQA	verb1s-topk	gemma	HS-GIRB	0.391 ± 0.229	0.447 ± 0.079	0.297 ± 0.225	0.072 ± 0.027
TruthfulQA	verb1s-topk	gemma	S-QAB	0.398 ± 0.27	0.446 ± 0.079	0.315 ± 0.257	0.079 ± 0.03
TruthfulQA	verb1s-topk	gemma	S-GIRB	0.397 ± 0.271	0.454 ± 0.087	0.315 ± 0.257	0.079 ± 0.03
TruthfulQA	verb1s-topk	gemma	S-B	0.396 ± 0.279	0.427 ± 0.099	0.289 ± 0.277	0.034 ± 0.012
TruthfulQA	verb1s-topk	gemma	S	0.073 ± 0.01	0.128 ± 0.015	0.07 ± 0.03	0.006 ± 0.002
TruthfulQA	verb1s-topk	gemma	B	0.113 ± 0.021	0.135 ± 0.017	0.073 ± 0.027	0.026 ± 0.03
TruthfulQA	verb1s-topk	gemma	IRB	$\mathbf{0.047 \pm 0.022}$	$\mathbf{0.112 \pm 0.013}$	0.07 ± 0.031	0.036 ± 0.05
TruthfulQA	verb1s-topk	gemma	None	0.752 ± 0.045	0.764 ± 0.023	0.689 ± 0.038	0.069 ± 0.02
TruthfulQA	ling1s-topk	mistral	QAB	0.16 ± 0.033	0.157 ± 0.017	0.118 ± 0.025	$\mathbf{0.104 \pm 0.052}$
TruthfulQA	ling1s-topk	mistral	GIRB	0.168 ± 0.046	0.153 ± 0.021	0.132 ± 0.032	0.09 ± 0.051
TruthfulQA	ling1s-topk	mistral	HS-QAB	0.439 ± 0.182	0.392 ± 0.08	0.343 ± 0.18	0.068 ± 0.025
TruthfulQA	ling1s-topk	mistral	HS-GIRB	0.435 ± 0.188	0.396 ± 0.079	0.341 ± 0.179	0.066 ± 0.025
TruthfulQA	ling1s-topk	mistral	S-QAB	0.444 ± 0.211	0.414 ± 0.093	0.344 ± 0.208	0.075 ± 0.036
TruthfulQA	ling1s-topk	mistral	S-GIRB	0.439 ± 0.215	0.426 ± 0.098	0.343 ± 0.206	0.076 ± 0.036
TruthfulQA	ling1s-topk	mistral	S-B	0.404 ± 0.267	0.44 ± 0.088	0.3 ± 0.256	0.034 ± 0.012
TruthfulQA	ling1s-topk	mistral	S	0.071 ± 0.017	0.129 ± 0.012	$\mathbf{0.077 \pm 0.023}$	0.005 ± 0.002
TruthfulQA	ling1s-topk	mistral	B	0.143 ± 0.027	0.144 ± 0.014	0.091 ± 0.027	0.024 ± 0.021
TruthfulQA	ling1s-topk	mistral	IRB	$\mathbf{0.063 \pm 0.018}$	$\mathbf{0.123 \pm 0.014}$	0.08 ± 0.021	0.008 ± 0.003
TruthfulQA	ling1s-topk	mistral	None	0.565 ± 0.042	0.605 ± 0.026	0.482 ± 0.038	0.069 ± 0.018
TruthfulQA	verb1s-topk	mistral	QAB	0.132 ± 0.021	0.181 ± 0.017	0.096 ± 0.026	0.079 ± 0.052
TruthfulQA	verb1s-topk	mistral	GIRB	0.134 ± 0.013	0.166 ± 0.016	$\mathbf{0.079 \pm 0.026}$	0.134 ± 0.057
TruthfulQA	verb1s-topk	mistral	HS-QAB	0.371 ± 0.252	0.408 ± 0.075	0.308 ± 0.24	$\mathbf{0.137 \pm 0.033}$
TruthfulQA	verb1s-topk	mistral	HS-GIRB	0.377 ± 0.244	0.415 ± 0.074	0.306 ± 0.241	0.137 ± 0.033
TruthfulQA	verb1s-topk	mistral	S-QAB	0.381 ± 0.264	0.417 ± 0.083	0.304 ± 0.25	0.128 ± 0.051
TruthfulQA	verb1s-topk	mistral	S-GIRB	0.376 ± 0.267	0.429 ± 0.091	0.304 ± 0.25	0.128 ± 0.051
TruthfulQA	verb1s-topk	mistral	S-B	0.41 ± 0.277	0.442 ± 0.088	0.307 ± 0.272	0.047 ± 0.013
TruthfulQA	verb1s-topk	mistral	S	$\mathbf{0.083 \pm 0.026}$	0.158 ± 0.011	0.088 ± 0.038	0.014 ± 0.003
TruthfulQA	verb1s-topk	mistral	B	0.148 ± 0.021	0.179 ± 0.017	0.095 ± 0.033	0.05 ± 0.021
TruthfulQA	verb1s-topk	mistral	IRB	0.085 ± 0.024	$\mathbf{0.155 \pm 0.019}$	0.088 ± 0.037	0.075 ± 0.095
TruthfulQA	verb1s-topk	mistral	None	0.71 ± 0.055	0.723 ± 0.028	0.629 ± 0.049	0.11 ± 0.02

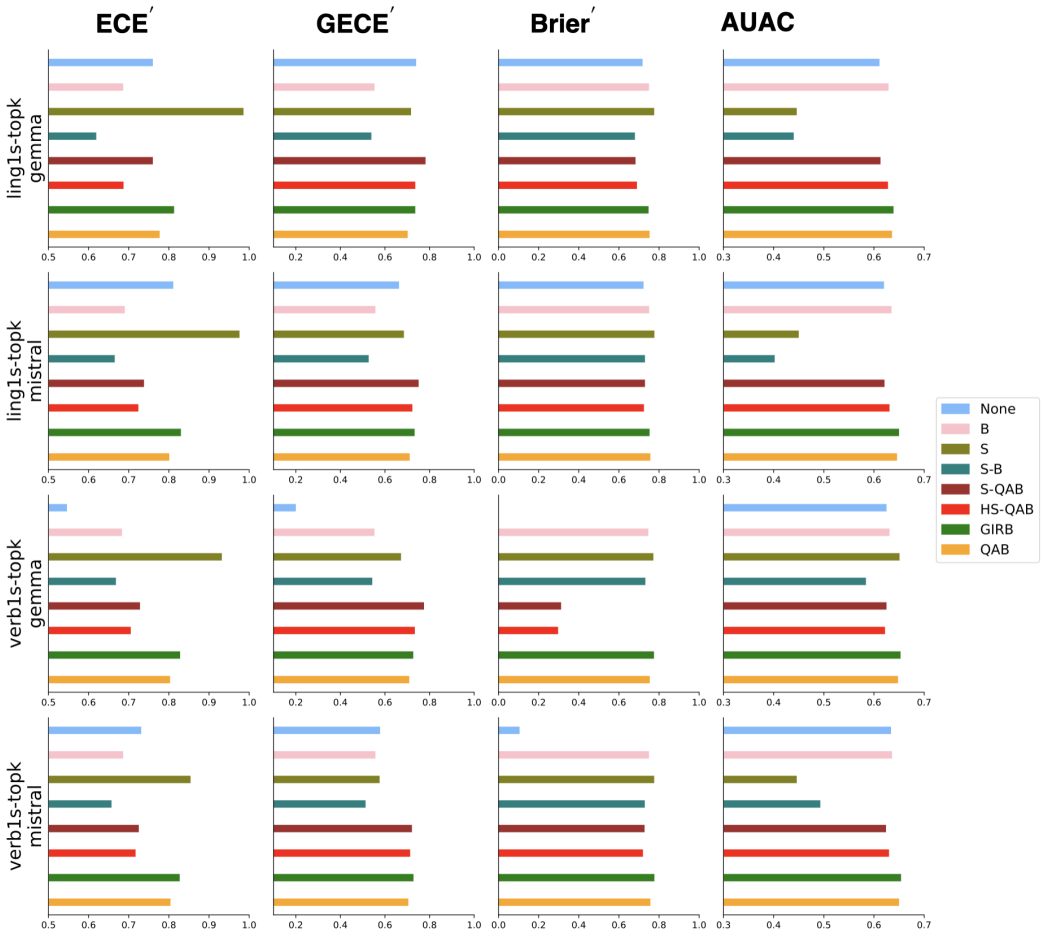

Figure 6: Bar Plot Comparison of Calibration Methods Across Settings in MathQA Dataset.

Figure 7: Bar Plot Comparison of Calibration Methods Across Settings in MMLU Dataset.

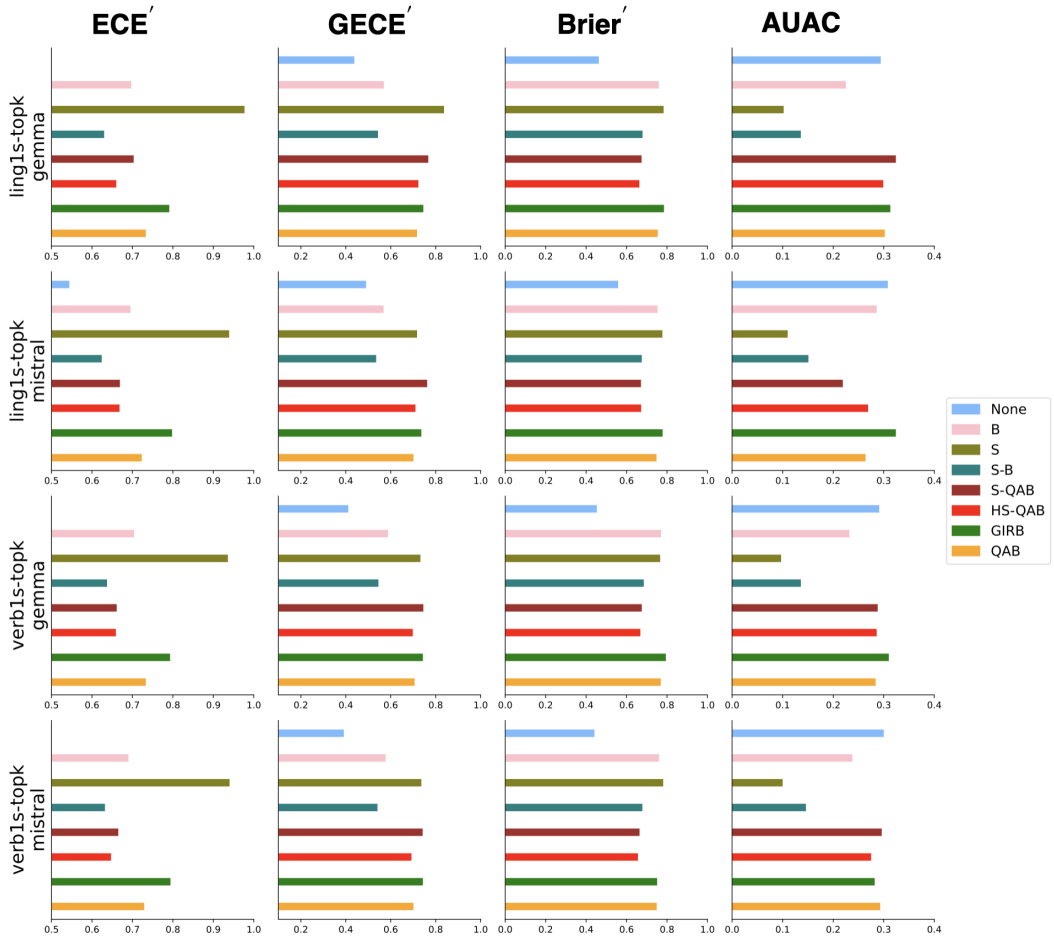

Figure 8: Bar Plot Comparison of Calibration Methods Across Settings in OpenBookQA Dataset.

Figure 9: Bar Plot Comparison of Calibration Methods Across Settings in SciQ Dataset.

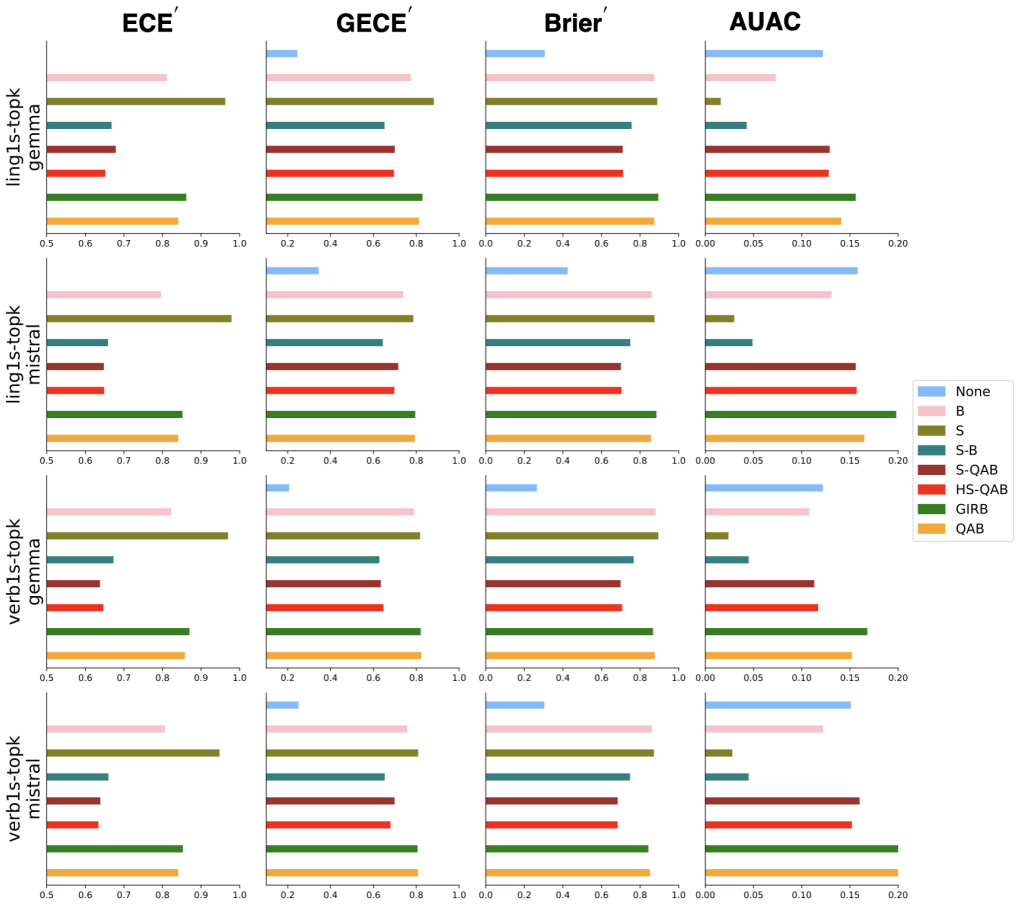

Figure 10: Bar Plot Comparison of Calibration Methods Across Settings in TriviaQA Dataset.

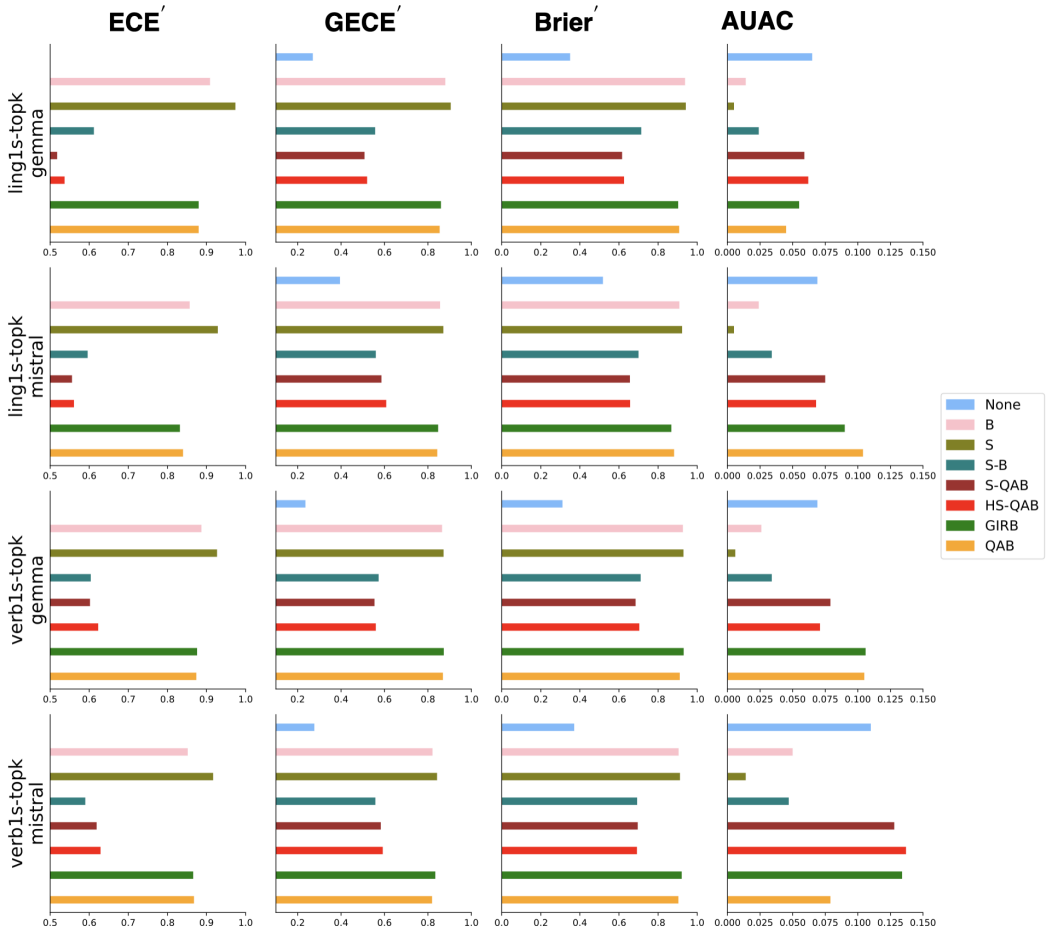

Figure 11: Bar Plot Comparison of Calibration Methods Across Settings in TruthfulQA Dataset.