Towards Objective Fine-tuning: How LLMs' Prior Knowledge Causes Potential Poor Calibration?

Ziming Wang[1]*, **Zeyu Shi**[1]*, **Haoyi Zhou**[2,3]†, **Shiqi Gao**[1], **Qingyun Sun**[1], **Jianxin Li**[1,3]
[1]SKLCCSE, School of Computer Science and Engineering, Beihang University
[2]School of Software, Beihang University
[3]Zhongguancun Laboratory, Beijing
{wangzm412, szy_629, haoyi, gaoshiqi, sunqy, lijx}@buaa.edu.cn

Abstract

Fine-tuned Large Language Models (LLMs) often demonstrate poor calibration, with their confidence scores misaligned with actual performance. While calibration has been extensively studied in models trained from scratch, the impact of LLMs' prior knowledge on calibration during fine-tuning remains understudied. Our research reveals that LLMs' prior knowledge causes potential poor calibration due to the ubiquitous presence of known data in real-world fine-tuning, which appears harmful for calibration. Specifically, data aligned with LLMs' prior knowledge would induce overconfidence, while new knowledge improves calibration. Our findings expose a tension: LLMs' encyclopedic knowledge, while enabling task versatility, undermines calibration through unavoidable knowledge overlaps. To address this, we propose CogCalib, a cognition-aware framework that applies targeted learning strategies according to the model's prior knowledge. Experiments across 7 tasks using 3 LLM families prove that CogCalib significantly improves calibration while maintaining performance, achieving an average 57% reduction in ECE compared to standard fine-tuning in Llama3-8B. These improvements generalize well to out-of-domain tasks, enhancing the objectivity and reliability of domain-specific LLMs, and making them more trustworthy for critical human-AI interaction applications.

1 Introduction

Large Language Models (LLMs) have enabled powerful domain-specific applications through supervised fine-tuning (Zhuang et al., 2023; Imani et al., 2023; Yang et al., 2024a). However, fine-tuning often leads to poor-calibrating LLM, where models' predictive confidence fails to reflect their true performance, manifesting as overconfidence (Achiam et al., 2023; Zhu et al., 2023; Shen et al., 2024; Yang et al., 2023). This is particularly concerning in high-stakes scenarios where LLMs' incorrect yet confident predictions could lead to reliability and trustworthiness issues, such as medical diagnosis (Xu et al., 2024; Zhang et al., 2024; Wei et al., 2024) or safety-critical domain (Sarabadani, 2019).

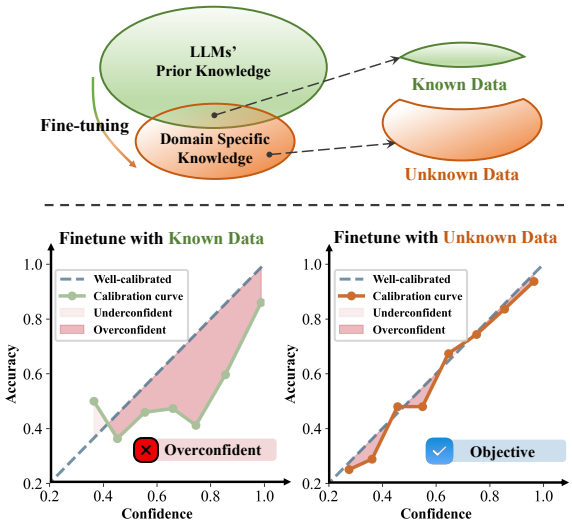


Figure 1: LLMs' prior knowledge leads to poor calibration. As LLMs grow stronger, lots of domain-specific fine-tuning data inevitably overlaps with the LLMs' prior knowledge. We reveal that data aligned with the model's prior knowledge (i.e., *known data*) tend to cause overconfidence, while data exhibiting bias (i.e., *unknown data*) contribute to better alignment between confidence and accuracy, resulting in more objective predictions.

Prior studies (Mukhoti et al., 2020; Wei et al., 2022; Guo et al., 2017a) investigating the causes of poor calibration mainly focus on simple models (ResNet) trained from scratch, where prior knowledge is absent. However, in the fine-tuning paradigm of LLMs, the training data is typically domain-incremental (Shi et al., 2024), encompassing both knowledge aligned with the pre-training corpus and novel domain-specific information (Gu-

* Equal contribution.
† Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14830–14853
July 27 - August 1, 2025

rurangan et al., 2020). This *knowledge bias* between LLMs' prior knowledge and fine-tuning knowledge has been shown as a critical factor affecting model adaptation (Gekhman et al., 2024; Kung et al., 2023; Huang et al., 2024; Seedat et al., 2023; Chen et al., 2024). Therefore, we try to extend previous research by investigating the underlying mechanisms of poor calibration specifically in the context of fine-tuning, particularly considering the impact of models' prior knowledge.

We reveal that LLMs' extensive prior knowledge, while enabling remarkable few-shot generalization, paradoxically causes their poor calibration in fine-tuning paradigms. Through empirical analysis, we discover that during fine-tuning, data aligned with the model's prior knowledge (i.e. *known data*) tend to cause overconfidence, while data exhibiting knowledge bias (i.e. *unknown data*) contribute to better calibration as shown in Figure 1. This disparity stems from the distinct learning dynamics between known and unknown data: the model quickly assimilates known data, leading to continued confidence growth even after accuracy plateaus. In contrast, unknown data, inherently more challenging for LLMs to learn (Zhang and Wu, 2024; Gekhman et al., 2024), results in more synchronized increases in both accuracy and confidence. This phenomenon becomes increasingly problematic as LLMs' prior knowledge expands, making it nearly impossible to avoid overlap between fine-tuning data and their prior knowledge.

However, existing approaches (Yang et al., 2023; Shen et al., 2024; Liu et al., 2023) are insufficient to handle this issue, primarily because they rely on post-hoc calibration methods. They typically introduce additional learnable modules after fine-tuning to reconstruct the mapping between model outputs and probabilities, which incurs extra computational overhead during deployment. On the other hand, the influence of LLMs' prior knowledge on calibration provides a promising opportunity to address poor calibration during fine-tuning.

Therefore, we introduce CogCalib, a real-time fine-tuning calibration framework compatible with various training-based calibration methods. Specifically, CogCalib dynamically evaluates knowledge bias during fine-tuning and applies targeted learning strategies accordingly, regulating confidence fitting and maintaining task learning. Moreover, CogCalib introduces no additional computational overhead during deployment.

We conduct comprehensive experiments across 7 commonly used downstream tasks (including multiple-choice and open-ended QA tasks) using 3 popular LLM families, to demonstrate CogCalib's effectiveness. CogCalib successfully preserves fine-tuning performance while achieving substantial improvements in calibration across all tasks and models, without incurring additional computational overhead during deployment. For instance, Llama3-8B achieves average ECE reductions of 55.92% and 65.02% compared to TS and SFT on multiple-choice QA tasks. Notably, these improvements generalize well to out-of-domain tasks, indicating that models trained with CogCalib consistently demonstrate enhanced objectivity across diverse scenarios. The main contributions of our work can be summarized as follows:

- As far as we know, we are the first to reveal the neglected negative impacts of LLMs' prior knowledge on calibration during fine-tuning. Specifically, data aligned with the model's prior knowledge tends to induce overconfidence, while new knowledge is beneficial for calibration.
- We propose CogCalib, a real-time calibration framework that employs distinct learning strategies for data with different knowledge biases during fine-tuning, aiming to achieve more objective fine-tuning.
- We conduct extensive experiments on domain-specific multiple-choice and open-ended QA tasks with multiple models, using different fine-tuning methods, which demonstrate the effectiveness and generality of CogCalib in enhancing calibration.

2 Related Works

2.1 Confidence Calibration

Confidence calibration methods can be categorized into three main approaches (Gawlikowski et al., 2023): post-processing adjustments (Guo et al., 2017b), training-based optimization (Szegedy et al., 2016), and uncertainty estimation (Lakshminarayanan et al., 2017). For LLMs specifically, recent efficient post-processing techniques have emerged, including Bayesian LoRA (Yang et al., 2023), LLM-oriented temperature scaling (Shen et al., 2024), and distribution adjustment methods (Liu et al., 2023). While these approaches address calibration computational complexity, they

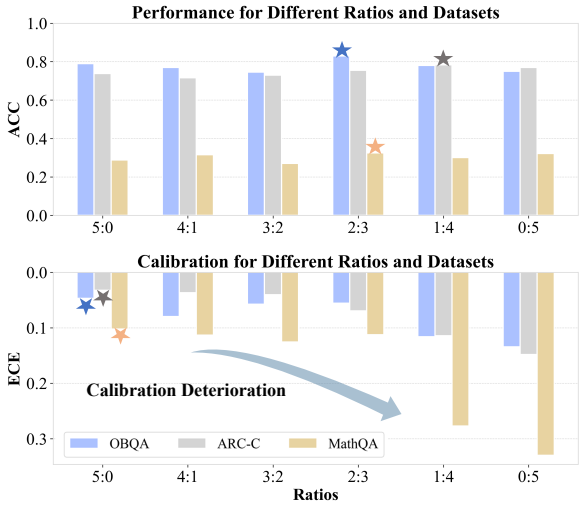


Figure 2: Accuracy and ECE of Llama3-8B fine-tuned with different knowledge biases. We fine-tune Llama3-8B using OBQA, with ARC-C and MathQA as OOD tests. The ratio varies from 5:0 to 0:5 (unknown data:known data), with equal dataset sizes. Calibration deteriorates as the knowledge bias lowers, while a higher knowledge bias helps improve calibration.

do not investigate the underlying causes of calibration degradation. Previous studies have identified negative log-likelihood (NLL) overfitting as a key factor in poor calibration (Mukhoti et al., 2020; Wei et al., 2022; Guo et al., 2017a). However, these findings were based on models without prior knowledge, whereas the unique pre-trained nature of LLMs (Shi et al., 2024) necessitates a fresh examination of calibration.

2.2 Impact of Knowledge Bias in Fine-tuning

Fine-tuning LLMs presents several critical challenges, including hallucinations (Huang et al., 2025), generalization problems (He et al., 2021), and calibration degradation (Zhu et al., 2023), with the bias between LLMs' prior knowledge and fine-tuning knowledge emerging as a contributing factor (Kung et al., 2023; Huang et al., 2024; Seedat et al., 2023; Yang et al., 2024b). Gekhman et al. (Gekhman et al., 2024) demonstrate that introducing new knowledge during fine-tuning can trigger hallucinations. Effective generalization can be achieved through knowledge selection strategies based on knowledge bias (Albalak et al., 2024; Chen et al., 2024). Regarding calibration, the impact of knowledge bias during fine-tuning warrants further investigation.

3 Prior Knowledge Affects Calibration?

In this section, we investigate how LLMs' prior knowledge affects calibration during fine-tuning. To quantify the overlap between fine-tuning data and the model's prior knowledge, we first reintroduce the concept of *knowledge bias*, which represents the discrepancy between the model's prior knowledge domain and the downstream task knowledge domain. Following the framework proposed by SliCK (Gekhman et al., 2024) (details shown in Appendix G.1), we categorize the data into two distinct types: *known data* that aligns with the model's prior knowledge, and *unknown data* that deviates from this knowledge base. Finally, we simulate varying knowledge bias by adjusting the ratio between unknown and known data in the fine-tuning dataset, where a higher proportion of known data indicates a lower knowledge bias (i.e., greater alignment with prior knowledge). The details of data construction are shown in Appendix B.1.

3.1 Minimal Bias, Maximal Overconfidence

To simulate varying knowledge biases, we construct fine-tuning datasets with six ratios of unknown to known data in OBQA. While Figure 2 reveals irregular performance trends across knowledge bias levels, the calibration exhibits a clear directional pattern: **lower knowledge bias consistently degrades calibration, whereas higher bias improves it**, a phenomenon persistent across both in-domain and out-of-domain. Notably, the introduction of even a small fraction of known data leads to calibration deterioration (from pure unknown data to 4:1 ratio). This suggests that the model's pre-existing knowledge dominance for calibration begins immediately upon exposure to aligned data. We observed the same phenomenon across other models and datasets (further experimental results and analyses are provided in the Appendix B.2).

Furthermore, our tracking for accuracy and confidence reveals divergent learning dynamics: in low-bias fine-tuning (Figure 3a), accuracy of OBQA test set plateaus early (200 steps) while confidence escalates continuously, creating widening calibration error. Conversely, high-bias conditions maintain synchronized accuracy-confidence growth, minimizing discrepancies — a pattern potentially rooted in gradual new knowledge assimilation (Zhang and Wu, 2024; Gekhman et al., 2024). The different confidence patterns persist in OOD detection (Berger et al., 2021) (Figure 3b): low-

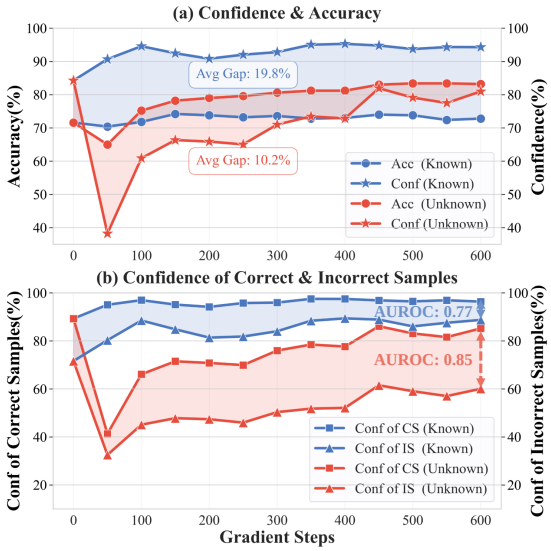


Figure 3: (a) Accuracy and confidence of Llama3-8B during fine-tuning on known and unknown data sampled from OBQA. The asymmetric fitting rates between accuracy and confidence in known fine-tuning result in model overconfidence. Conversely, unknown fine-tuning exhibits synchronized fitting of both, minimizing their disparity. (b) Average confidence of correct and incorrect predictions. Unknown fine-tuning yields distinct confidence separation between correct and incorrect samples, facilitating OOD detection.

bias models show compressed confidence distributions for correct and incorrect predictions (both correct/incorrect > 85%), whereas high-bias models develop discriminative confidence gaps (AUROC 0.85 vs 0.77 at step 600), enhancing OOD detection. More results, including standard deviations are shown in Appendix B.2.

These findings collectively demonstrate how LLMs' prior knowledge induces poor calibration: **pre-existing knowledge enables rapid confidence inflation on aligned data, while insufficient exposure to new knowledge prevents calibration improvement — a harmful interaction amplified by the near-ubiquitous presence of known data in the real-world fine-tuning.** Additionally, we examine this phenomenon in realistic scenarios (details are shown in Figure 16 of Appendix B.2).

3.2 Analysis and a Potential Solution

An intuitive explanation is that known samples closely align with pre-trained models' prior distribution, while unknown samples represent the target distribution. Therefore, fine-tuning in low-bias scenarios leads to rapid confidence overfitting. In contrast, high-bias fine-tuning requires the model to adjust its decision boundaries to accommodate new distributions, resulting in better calibration.

A potential solution is to increase the bias between the fine-tuning data and the model's prior knowledge, specifically by eliminating known data. However, we reveal that simple bias adjustment through data removal is insufficient, as it fails to consistently improve calibration performance across different datasets. As shown in Figure 4, removing 25% of low-bias data improves calibration in ARC-C but degrades it in OBQA. This inconsistency may stem from the inherent characteristics of different datasets, making it challenging to find a universal optimal adjustment ratio. Additionally, accuracy consistently improves with known data reduction, aligning with findings new knowledge enhances task performance (Kung et al., 2023; Swayamdipta et al., 2020). This necessitates methods that decouple knowledge retention from calibration, preserving fine-tuning performance while improving calibration.

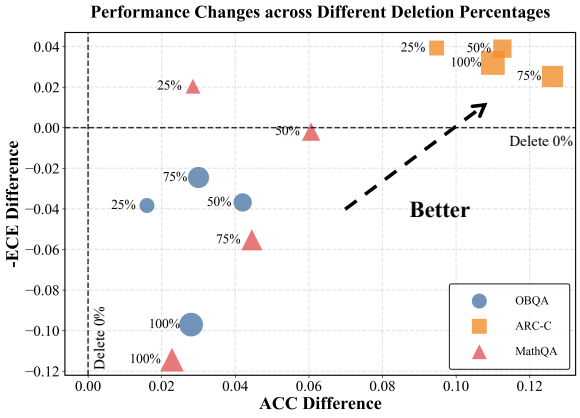


Figure 4: Differences in ACC and ECE compared to baseline (delete 0%) under various percentages of known data deletion. The results from in-domain tests indicate that simple bias adjustment fails to achieve consistent calibration improvements across all tasks.

4 Cognition-aware Calibration

In this section, we propose CogCalib, a **Cog**nition-aware **Calib**ration framework for fine-tuning, designed to achieve an optimal balance between fine-tuning performance and calibration. CogCalib is motivated by the above observation that known and unknown data exhibit distinct fitting characteristics during the fine-tuning process, necessitating different learning strategies. To develop an effective solution, we mainly address two challenges: (1) *How to evaluate knowledge bias, particularly as*

the model's internal states are continuously evolving? (2) *What specific learning strategies should be applied to achieve objective fine-tuning?*

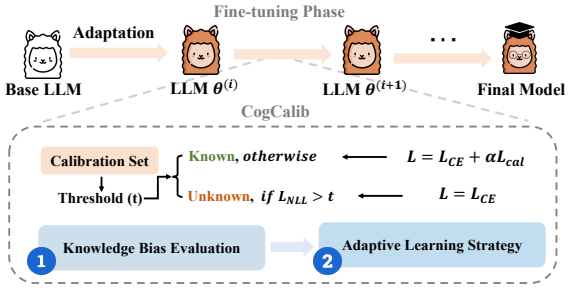


Figure 5: CogCalib's framework. CogCalib dynamically assesses knowledge bias during training through NLL, employing customized learning strategies with distinct loss functions to enhance calibration. Additionally, CogCalib incorporates a style adaptation process to improve the knowledge bias evaluation performance.

Algorithm 1: Adaptive Threshold Update

Data: Calibration set $S_c = \{(x_i, y_i)\}_{i=1}^{N}$, grid size M
Result: Updated threshold t^*

foreach *update step* (e.g., *every epoch*) **do**
  for $i \leftarrow 1$ **to** N **do**
    $(c_i, n_i) \leftarrow \text{Inference}(x_i)$;
    // c_i=1 if correct else 0
    // $n_i = -\log p_\theta(y_i|x_i)$
  end
  // Generate candidate thresholds
  $\mathcal{T} \leftarrow \text{linspace}\big(\min n_i,\ \max n_i,\ M\big)$;
  // Grid-search for optimal t
  $t^* \leftarrow \arg\max_{t\in\mathcal{T}}\big(\text{TPR}(t) + \text{TNR}(t)\big)$;
  // TPR: true positive rates
  // TNR: true negative rates
end

4.1 Knowledge Bias Evaluation

In section 3, we evaluate knowledge bias based on the correctness of the model's output through multiple inferences. However, during training, this method faces limitations due to the inability to perform multiple sampling iterations.

Therefore, we propose a more efficient method for knowledge bias assessment based on negative log-likelihood (NLL). Our hypothesis posits that known data aligns with the pre-trained model's prior knowledge bias, while unknown data represents novel information from the target distribution. From a distributional perspective, as shown in Equation (1), known data should exhibit lower NLL values compared to unknown data:

$$\mathcal{L}_{\text{NLL}} = \mathbb{E}_{\mathbf{q}}\left[\mathbf{p}\right] = -\sum_{k=1}^{C} q_k \log p_k, \qquad (1)$$

where $\mathbf{q}$ is the target one-hot distribution, $\mathbf{p}$ is the vocabulary distribution output by LLM, and p_k and q_k represent the probabilities for the k-th class in $\mathbf{p}$ and $\mathbf{q}$, respectively. Based on this, we evaluate knowledge bias during training, using Equation (2),

$$\mathbf{I}(\mathbf{p}, \mathbf{q}) = \begin{cases} 1, & \text{if } \mathcal{L}_{\text{NLL}} \leq t \\ 0, & \text{otherwise} \end{cases}, \qquad (2)$$

where t denotes the threshold, and $I = 1$ indicates the model has already mastered this knowledge. In addition, t requires adjustment to accommodate the evolving knowledge distribution during training. To solve this, we establish a calibration set to identify the optimal t according to Algorithm 1 (details of the calibration set are shown in Appendix A).

However, the discrepancy in LLMs' linguistic style and label formats prevents NLL from accurately assessing knowledge bias. Thus, the model requires a style adaptation process for calculating the initial threshold t_0, based on findings (Zhang and Wu, 2024; Mai et al., 2024) that LLMs rapidly adapt to downstream task syntax during early fine-tuning (details are shown in Appendix D).

4.2 Adaptive Learning Strategy

Our previous analysis reveals that during fine-tuning, model confidence increases rapidly for known data, while unknown data contributes positively to both calibration and downstream task performance. Therefore, we moderate confidence fitting for known data while preserving the learning dynamics for unknown data as Equation (3),

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathbf{I}(\mathbf{p}, \mathbf{q}) \cdot \alpha \mathcal{L}_{cal}, \qquad (3)$$

where α represents the regularization strength and $\mathcal{L}_{cal}$ denotes the calibration loss during training. The calibration term could be Label Smoothing (LS) (Szegedy et al., 2016), Margin-based Label Smoothing (MbLS) (Liu et al., 2022), or ECP (Pereyra et al., 2017), which have been proved to be helpful for confidence overfitting (details shown in Appendix C). In CogCalib, we call these methods CoLS, CoMbLS, and CoECP.

4.3 Integrated Framework

Building on our previous analysis, we propose an integrated framework aiming to achieve objective fine-tuning. Figure 5 illustrates the architecture of CogCalib. Initially, LLM undergoes style adaptation to align with the grammatical patterns of downstream tasks. Subsequently, CogCalib dynamically assesses knowledge bias using NLL and adaptive t. For low-bias data, we incorporate a calibration term to mitigate confidence overfitting, while cross-entropy loss is applied to new knowledge to maintain task alignment.

Type	Dataset	Accuracy	TPR	TNR
Multi-Choice	OBQA	99.44	99.44	99.52
	ARC-C	99.51	99.54	99.15
	WG-S	98.83	98.77	99.52
	WG-M	99.13	99.10	99.55
	BoolQ	98.69	98.63	98.21
Open-End	HotpotQA	83.64	79.43	90.59
	MedMCQA	83.69	79.77	87.78

Table 1: Accuracy, True Positive Rate (TPR), and True Negative Rate (TNR) for identifying known/unknown data using NLL in the fine-tuning process of Llama3-8B.

5 Experiments

In this section, we will evaluate the universality of CogCalib across 3 aspects: diverse datasets (from multiple-choice to open-ended), various LLM families and sizes, and different fine-tuning approaches (LoRA and FFT). Complete experimental results are presented in Appendix F, and hyperparameters settings are shown in Appendix G.3.

5.1 Experimental Setup

Datasets. To ensure the universality of CogCalib, we select a wide range of tasks, including HotpotQA (Yang et al., 2018) MedMCQA (Pal et al., 2022) for open-ended QA tasks, while utilizing OpenBookQA (OBQA) (Mihaylov et al., 2018), ARC-Challenge (ARC-C) (Clark et al., 2018), Winogrande-small (WG-S), Winogrande-medium (WG-M) (Sakaguchi et al., 2021) and BoolQ (Clark et al., 2019) for multiple-choice QA scenarios. Additionally, we extend our evaluation of CogCalib to various OOD tasks, including MMLU (Hendrycks et al., 2021) and ARC-E (Clark et al., 2018). See Appendix A for more details.

Models. We validate CogCalib across models of diverse families and scales, including Llama3-8B, Llama2-13B, Mistral-7B, and Qwen2.5-7B. All models employed in our experiments are instruction-tuned variants of their respective base models.

Evaluation Metrics. In addition to evaluating the accuracy of fine-tuned models, we also select ECE with a bin size of 10 to assess calibration. See Appendix E for more details of ECE.

Baselines. We consider 4 baseline methods: (1) *Vanilla SFT*: We use standard LoRA or FFT as a lower performance bound. (2) *MC-Dropout (MCD)* (Gal and Ghahramani, 2016): We use a dropout rate of 0.02 during fine-tuning and perform sampling 4 times. (3) *Deep Ensemble (Ensemble)* (Lakshminarayanan et al., 2016): We use 3 fine-tuned LLMs. (4) *Temperature Scaling (TS)* (Guo et al., 2017b): The optimal temperature is calculated on the ID validation set and applied to both the ID and OOD datasets. See Appendix G.2 for more details.

5.2 Main Results

In this section, we validate the effectiveness of CogCalib through comprehensive experiments. First, we demonstrate the validity of using NLL for assessing knowledge bias, which enhances the interpretability of our framework. Subsequently, we evaluate CogCalib's performance on both multiple-choice and open-ended tasks, showing that it not only maintains fine-tuning performance but also significantly improves calibration.

5.2.1 Effectiveness Evaluation of Knowledge Bias via NLL

In this section, we aim to validate the effectiveness of using NLL for evaluating knowledge bias. Table 1 presents accuracy in distinguishing between unknown/known data using NLL during training (average accuracy throughout the training process). The high accuracy proves NLL-based method aligns well with SliCK (Gekhman et al., 2024), validating NLL as an effective knowledge bias evaluation metric. Moreover, we demonstrate our threshold calculation method outperforms alternative approaches in Appendix F.8.

5.2.2 Calibration of Multi-Choice Task

To verify CogCalib's robustness and generalizability, our evaluation for CogCalib consists of two dimensions: in-domain performance assessment and out-of-domain evaluation.

Dataset	Metric	Vanilla SFT	MCD	Ensemble	TS	CoLS (ΔTS)	CoMbLS (ΔTS)	CoECP (ΔTS)
OBQA	**ACC↑**	84.80	83.60	**88.00**	84.80	85.60 (+0.80)	86.20 (+1.40)	86.20 (+1.40)
	ECE↓	11.20	9.10	6.02	9.90	**2.50** (-7.40)	3.70 (-6.20)	7.30 (-2.60)
ARC-C	**ACC↑**	81.40	80.70	81.14	81.40	81.60 (+0.20)	**81.70** (+0.30)	81.60 (+0.20)
	ECE↓	16.50	13.80	13.08	12.30	4.80 (-7.50)	**4.20** (-8.10)	7.40 (-4.90)
WG-S	**ACC↑**	78.00	78.21	80.27	78.00	80.10 (+2.10)	79.20 (+1.20)	**80.30** (+2.30)
	ECE↓	20.50	17.77	14.81	15.40	8.90 (-6.50)	9.40 (-6.00)	**7.00** (-8.40)
WG-M	**ACC↑**	84.50	84.37	**85.16**	84.50	84.50 (+0.0)	84.70 (+0.20)	84.70 (+0.20)
	ECE↓	14.80	13.51	10.09	11.50	4.10 (-7.40)	3.10 (-8.40)	**1.00** (-10.50)
BoolQ	**ACC↑**	90.09	90.15	**90.86**	90.09	90.15 (+0.06)	89.63 (-0.46)	89.54 (-0.55)
	ECE↓	9.54	8.95	6.69	7.70	**1.97** (-5.73)	2.36 (-5.34)	7.68 (-0.02)

Table 2: Comparison of our method's performance against baselines on in-domain (ID) datasets. Results are evaluated on Llama3-8B model fine-tuned by LoRA on 5 widely used domain-specific datasets. We integrate LS, MbLS, and ECP as calibration terms in CogCalib, resulting in 3 variants: CoLS, CoMbLS, and CoECP.

Metric	Methods	In Domain	Smaller Distribution Shift		Larger Distribution Shift			
		OBQA	ARC-C	ARC-E	Business	Culture	History	Psychology
ECE↓	**Vanilla SFT**	11.20	18.00	13.50	18.40	17.61	19.22	23.38
	MCD	9.10	14.56	11.11	13.54	15.70	17.87	20.42
	Ensemble	6.02	14.59	8.92	14.09	15.33	15.99	18.76
	TS	9.90	15.90	10.40	16.10	16.70	17.40	21.30
	CoLS (ΔTS)	**2.50** (-7.4)	7.50 (-8.4)	2.40 (-8.0)	9.80 (-6.3)	10.30 (-6.4)	12.07 (-5.3)	14.75 (-6.6)
	CoMbLS (ΔTS)	3.70 (-6.2)	5.80 (-10.1)	**1.40** (-9.0)	8.20 (-7.9)	9.48 (-7.2)	9.83 (-7.6)	14.41 (-6.9)
	CoECP (ΔTS)	7.30 (-2.6)	**2.80** (-13.1)	4.90 (-5.5)	**3.80** (-12.3)	**3.46** (-13.2)	**6.27** (-11.1)	**9.51** (-11.8)
ACC↑	**Vanilla SFT**	84.80	79.10	84.10	79.20	79.52	77.63	73.47
	MCD	83.60	78.92	84.22	80.78	79.22	76.24	73.38
	Ensemble	**88.00**	79.35	**87.37**	80.32	79.52	78.39	**75.11**
	TS	84.80	79.10	84.10	79.20	79.52	77.63	73.47
	CoLS (ΔTS)	85.60 (+1.8)	79.30 (+0.2)	86.30 (+2.2)	79.40 (+0.2)	78.92 (-0.6)	76.24 (-1.4)	73.64 (+0.2)
	CoMbLS (ΔTS)	86.20 (+2.4)	**80.00** (+0.9)	86.70 (+2.6)	**81.70** (+2.5)	80.12 (+0.6)	**78.92** (+1.3)	74.50 (+1.0)
	CoECP (ΔTS)	86.20 (+2.4)	79.00 (-0.1)	84.60 (+0.5)	80.80 (+1.6)	**81.02** (+1.5)	77.96 (+0.3)	74.50 (+1.0)

Table 3: Comparison of our method's performance against baselines on distribution shift datasets is presented. Results are evaluated on Llama3-8B model which is fine-tuned on the OBQA dataset.

In-Distribution Performance. We first conduct in-domain tests on CogCalib across 5 commonsense reasoning datasets. On the one hand, our cognitive methods maintain competitive accuracy compared to baselines as shown in Table 2, such as CoECP achieving 86.20% and 80.30% accuracy on OBQA and WG-S datasets, respectively. On the other hand, our cognitive methods achieve substantial calibration improvements across all datasets. These indicate that CogCalib achieves more objective fine-tuning on ID tasks.

Performance Under Distribution Shift. Real-world applications demand robust model performance across different scenarios. We evaluate CogCalib under various distribution shifts, including ARC-C and ARC-E datasets for smaller shifts, and 4 MMLU subjects (Business, Culture, History, Psychology) for larger domain shifts. Table 3 indicates that CogCalib maintains competitive accuracy compared to the baselines under distribution shifts and achieves overall superior ECE. These findings demonstrate the robustness of the CogCalib in tasks with distribution shifts.

More experiments based on other LLMs. To validate the generalizability of CogCalib, we also conduct experiments on Mistral-7B, Qwen2.5-7B, and Llama2-13B in Appendix F. The results demonstrate that CogCalib consistently achieves significant calibration improvements across these models while maintaining fine-tuning performance, proving the cross-model generalizability of CogCalib.

More experiments based on FFT. To explore the applicability of CogCalib to other fine-tuning methods, we validate CogCalib with FFT on Llama3-8B, demonstrating its effectiveness beyond LoRA-based approaches (see Appendix F.5).

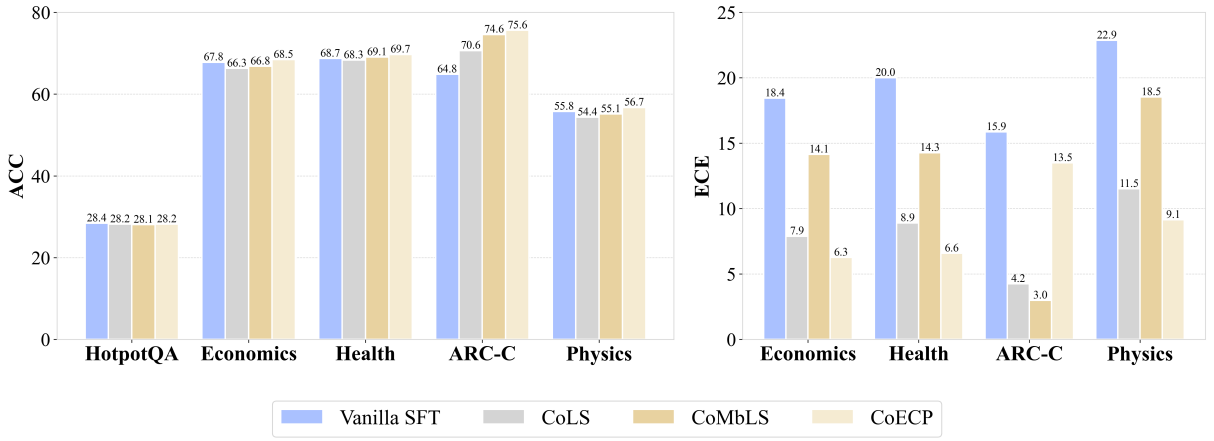


Figure 6: Comparison of our method's performance against baseline approaches on OOD datasets is presented. The results are evaluated on the Llama3-8B model, which is fine-tuned on the open-ended HotpotQA dataset.

5.2.3 Calibration of Open-End Task

In addition, we evaluate CogCalib on open-ended datasets HotpotQA (experiments on MedMCQA are presented in Appendix F.1). As shown in Figure 6, CogCalib maintains accuracy on ID and OOD tasks while providing larger accuracy gains on some datasets, e.g., the CoECP shows a 10.8% ACC gain over Vanilla SFT on ARC-C. Regarding calibration, our cognitive methods exhibit comprehensive improvement. These further demonstrate the task-agnostic nature of CogCalib.

5.3 Ablation Study

Comparision to Vanilla and Random Calibration. In this section, we validate the necessity of employing different learning strategies for known and unknown data within CogCalib. As baseline methods, we select (1) Vanilla calibration, which uniformly applies calibration loss to all data. (2) Random calibration, which randomly distinguishes between known and unknown data while maintaining a consistent number of known samples.

As shown in Figure 7, in these tasks, our cognitive methods achieved optimal results in both fine-tuning performance and calibration, thereby validating the necessity of employing distinct learning strategies within CogCalib. Whether using Vanilla Calibration or Random Calibration, the accuracy of downstream tasks declined (see more results in Appendix F.7). Further research revealed that applying calibration loss to unknown data impairs the model's performance on downstream tasks (detailed analysis is presented in the Appendix F.6), namely that unknown data are critical for aligning the model with downstream tasks.

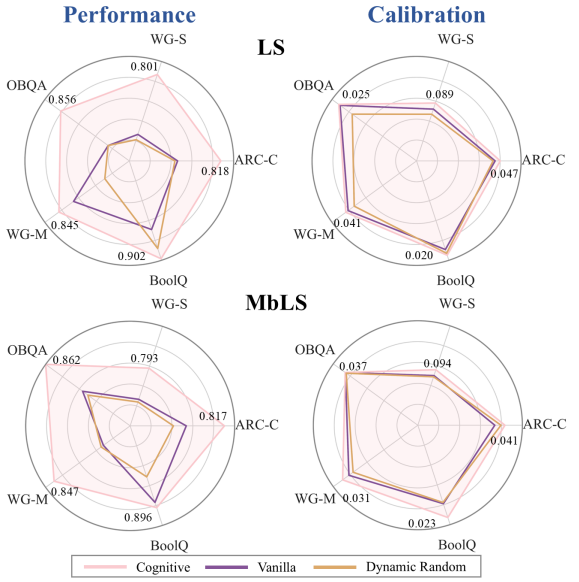


Figure 7: Comparison of CogCalib against baselines (Vanilla and Dynamic random) in terms of fine-tuning performance and calibration. Since a lower ECE is better, we normalize ECE to $[0, 1]$ using $\frac{ECE_{max}-ECE}{ECE_{max}-ECE_{min}}$ ($ECE_{max} = 0.2$, $ECE_{min} = 0.01$). ECP's results are shown in Figure 19 of Appendix F.7.

Sensitivity to Hyperparameters We investigate the impact of hyperparameter choices for CogCalib on performance. As illustrated in Figure 8, both LS and MbLS consistently demonstrate lower ECE across various hyperparameter configurations compared to the temperature scaling baseline, while maintaining comparable accuracy. More results regarding ECP are provided in the Appendix F.9. These findings demonstrate the robustness of CogCalib, with multiple hyperparameter configurations yielding calibration improvements.

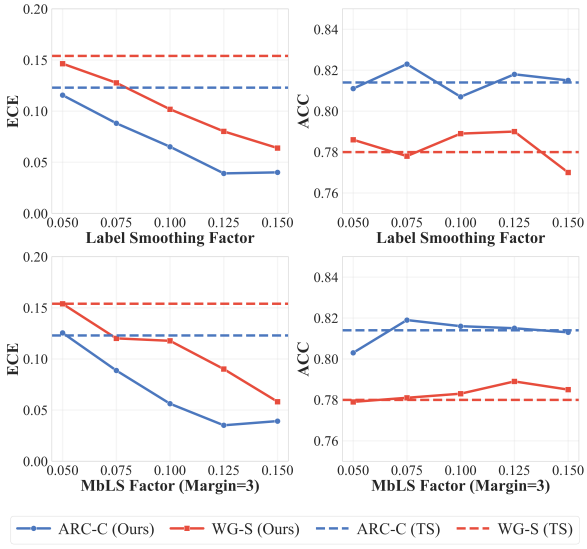


Figure 8: Sensitivity to Hyperparameters. We adjust the hyperparameters of CogCalib and compare its performance with the temperature scaling baseline on both ARC-C and WG-S datasets. The experimental results demonstrate the robustness of our method, showing consistent gains across various configurations.

6 Conclusion

In this work, we reveal that LLMs' prior knowledge causes potential poor calibration due to the ubiquitous presence of known data in real-world fine-tuning, which we discover would induce overconfidence. To address this, we propose CogCalib, a real-time cognition-aware calibration, which could achieve more objective fine-tuning. Through extensive experiments, we demonstrate that CogCalib effectively improves calibration while maintaining model performance without additional computational overhead during deployment, enabling more objective and trustworthy fine-tuning in safety-critical applications.

Limitations

Our research focuses on how prior knowledge in large language models (LLMs) leads to poor calibration during the fine-tuning process, and we propose a real-time calibration framework to address this issue. However, our study has only investigated models up to 13B parameters, and larger-scale models remain unexplored. Given additional GPU resources, we can conduct more comprehensive experiments to validate our findings on larger models. Furthermore, our framework incorporates some calibration terms, and new calibration terms may potentially achieve better performance in the future. Nevertheless, our research provides a novel perspective on the problem of poor calibration during fine-tuning and offers a real-time solution.

Ethics Statement

In this work, ethical considerations have been carefully addressed, and all research activities were conducted in strict compliance with the ACL Ethics Guidelines. The primary focus of this study is to investigate the impact of LLM prior knowledge on calibration while proposing a real-time calibration framework. All models and datasets utilized in this research are publicly available and have been widely adopted by the research community. The experimental results presented herein have been rigorously validated for accuracy and reproducibility. Based on these considerations, we assert that this research does not raise any ethical concerns.

Acknowledgements

The work is supported by the grants from the Natural Science Foundation of China (62225202, 62202029), and Young Elite Scientists Sponsorship Program by CAST (No. 2023QNRC001). We owe sincere thanks to all authors for their valuable efforts and contributions. The corresponding author is Haoyi Zhou.

References

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

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, et al. 2024. A survey on data selection for language models. *arXiv preprint arXiv:2402.16827*.

Christoph Berger, Magdalini Paschali, Ben Glocker, and Konstantinos Kamnitsas. 2021. Confidence-based out-of-distribution detection: A comparative study and analysis. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis*, pages 122–132, Cham. Springer International Publishing.

Jochen Bröcker and Leonard A Smith. 2007. Increasing the reliability of reliability diagrams. *Weather and forecasting*, 22(3):651–661.

Dingshuo Chen, Zhixun Li, Yuyan Ni, Guibin Zhang, Ding Wang, Qiang Liu, Shu Wu, Jeffrey Xu Yu,

and Liang Wang. 2024. Beyond efficiency: Molecular data pruning for enhanced generalization. *arXiv preprint arXiv:2409.01081*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. 2023. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning llms on new knowledge encourage hallucinations? *arXiv preprint arXiv:2405.05904*.

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

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

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Tianxing He, Jun Liu, Kyunghyun Cho, Myle Ott, Bing Liu, James Glass, and Fuchun Peng. 2021. Analyzing the forgetting problem in pretrain-finetuning of open-domain dialogue response models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1121–1133, Online. Association for Computational Linguistics.

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

Jianheng Huang, Leyang Cui, Ante Wang, Chengyi Yang, Xinting Liao, Linfeng Song, Junfeng Yao, and Jinsong Su. 2024. Mitigating catastrophic forgetting in large language models with self-synthesized rehearsal. *arXiv preprint arXiv:2403.01244*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. Mathprompter: Mathematical reasoning using large language models. *arXiv preprint arXiv:2303.05398*.

Po-Nien Kung, Fan Yin, Di Wu, Kai-Wei Chang, and Nanyun Peng. 2023. Active instruction tuning: Improving cross-task generalization by training on prompt sensitive tasks. *arXiv preprint arXiv:2311.00288*.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2016. Simple and scalable predictive uncertainty estimation using deep ensembles. *Cornell University - arXiv,Cornell University - arXiv*.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Bingyuan Liu, Ismail Ben Ayed, Adrian Galdran, and Jose Dolz. 2022. The devil is in the margin: Margin-based label smoothing for network calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 80–88.

Xin Liu, Muhammad Khalifa, and Lu Wang. 2023. Litcab: Lightweight calibration of language models on outputs of varied lengths. *arXiv preprint arXiv:2310.19208*.

Zheda Mai, Arpita Chowdhury, Ping Zhang, Cheng-Hao Tu, Hong-You Chen, Vardaan Pahuja, Tanya Berger-Wolf, Song Gao, Charles Stewart, Yu Su, et al. 2024. Fine-tuning is fine, if calibrated. *arXiv preprint arXiv:2409.16223*.

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 *EMNLP*.

Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. 2020. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems*, 33:15288–15299.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference*

on Health, Inference, and Learning, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. 2017. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Sarah Sarabadani. 2019. Detection of adverse drug reaction mentions in tweets using elmo. In *Proceedings of the Fourth Social Media Mining for Health Applications (# SMM4H) Workshop & Shared Task*, pages 120–122.

Nabeel Seedat, Nicolas Huynh, Boris van Breugel, and Mihaela van der Schaar. 2023. Curated llm: Synergy of llms and data curation for tabular augmentation in ultra low-data regimes. *arXiv preprint arXiv:2312.12112*.

Maohao Shen, Subhro Das, Kristjan Greenewald, Prasanna Sattigeri, Gregory Wornell, and Soumya Ghosh. 2024. Thermometer: Towards universal calibration for large language models. *arXiv preprint arXiv:2403.08819*.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. 2024. Continual learning of large language models: A comprehensive survey. *arXiv preprint arXiv:2404.16789*.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. 2022. Mitigating neural network overconfidence with logit normalization. In *International conference on machine learning*, pages 23631–23644. PMLR.

Pijing Wei, Qianqian Wang, Zhen Gao, Ruifen Cao, and Chunhou Zheng. 2024. Dmfvae: mirna-disease associations prediction based on deep matrix factorization method with variational autoencoder. *Frontiers of Computer Science*, 18(6):186912.

Zishan Xu, Linlin Song, Shichao Liu, and Wen Zhang. 2024. Deepcrbp: improved predicting function of circrna-rbp binding sites with deep feature learning. *Frontiers of Computer Science*, 18(2):182907.

Adam X Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison. 2023. Bayesian low-rank adaptation for large language models. *arXiv preprint arXiv:2308.13111*.

Bufang Yang, Siyang Jiang, Lilin Xu, Kaiwei Liu, Hai Li, Guoliang Xing, Hongkai Chen, Xiaofan Jiang, and Zhenyu Yan. 2024a. Drhouse: An llm-empowered diagnostic reasoning system through harnessing outcomes from sensor data and expert knowledge. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(4):1–29.

Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. 2024b. Self-distillation bridges distribution gap in language model fine-tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1028–1043, Bangkok, Thailand. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Qiang Zhang, Juan Liu, Wen Zhang, Feng Yang, Zhihui Yang, and Xiaolei Zhang. 2024. A multi-stream network for retrosynthesis prediction. *Frontiers of Computer Science*, 18(2):182906.

Xiao Zhang and Ji Wu. 2024. Dissecting learning and forgetting in language model finetuning. In *The Twelfth International Conference on Learning Representations*.

Chiwei Zhu, Benfeng Xu, Quan Wang, Yongdong Zhang, and Zhendong Mao. 2023. On the calibration of large language models and alignment. *arXiv preprint arXiv:2311.13240*.

Yuchen Zhuang, Yue Yu, Kuan Wang, Haotian Sun, and Chao Zhang. 2023. Toolqa: A dataset for llm question answering with external tools. *Advances in Neural Information Processing Systems*, 36:50117–50143.

A Details of Datasets and Calibration Set

Dataset	Train	Test	Val	Calibration
HotpotQA	16k	2k	1k	1k
MedMCQA	10k	2k	1k	1k
OBQA	4452	500	500	500
BoolQ	8427	3270	1k	1k
ARC-C	1119	1172	299	200
WG-S	580	1267	80	80
WG-M	2258	1267	300	300

Table 4: Configuration of datasets for fine-tuning. The validation set is utilized for Temperature Scaling to search for optimal temperature for calibration, while the calibration set is employed for threshold updating during finetuning.

We present detailed statistics of the finetuning tasks in Table 4. For test-only tasks, including MMLU subtasks (Business, Culture, History, Psychology, Physics, Economics, Health, and Law) and the ARC-E task, we strictly adhered to their official dataset configurations. For datasets that originally lacked validation sets in Table 4, we partitioned a portion of their training data to create validation sets specifically for Temperature Scaling. The calibration set was designed to have a comparable size to the validation set, with samples randomly selected from the training set at fixed intervals for threshold updates. Notably, MedMCQA (Pal et al., 2022), a comprehensive medical multiple-choice dataset, was restructured into an open-ended format where option texts were directly used as answers, following the same question-answering format as HotpotQA.

B Addendum to Section 3

B.1 Construction of Datasets with Varying Knowledge Bias

In Section 3, we simulate varying knowledge bias by adjusting the ratio of unknown to known samples in the fine-tuning set. Specifically, N_k and N_{unk} denote the number of known and unknown samples in the original dataset, and $N = min\{N_k, N_{unk}\}$ represents the total data volume.

When $N_k \leq N_{unk}$, we first form $D_{0:r}$ by including all known samples (where the subscript indicates the ratio of unknown to known data, $r = 5$ in our experiments). We then construct the dataset $D_{i:(r-i)}{}_{i=1}^{r}$ according to these rules: randomly remove $\frac{N}{r}$ known samples from $D_{i:(r-i)}$ and add $\frac{N}{r}$ randomly selected unknown samples from the original dataset to form $D_{(i+1):(r-i-1)}$. For the case where $N_k > N_{unk}$, the process follows the same principle.

B.2 Additional Results of Section 3

In Section 3.1, we demonstrated that low-bias leads to overconfidence, while high-bias data contributes to better calibration. This section presents additional experimental evidence supporting this phenomenon across multiple datasets, following the experimental protocol established in Section 3.1. Figure 9 illustrates the ECE metrics and fine-tuning performance obtained from experiments on MathQA, where we constructed scenarios with varying degrees of knowledge bias. Figure 10 presents our test results on MedMCQA, a domain-specific open-ended dataset which is restructured by us as explained in Appendix A. Both figures clearly demonstrate that calibration performance deteriorates significantly as bias decreases. These findings further corroborate our conclusion from Section 3.1, supporting the principle of "minimal bias, maximal overconfidence". Furthermore, additional experiments are conducted on Llama2-13B (Figure 13), Qwen2.5-7B (Figure 14), and Mistral-7B (Figure 15), following the same setup described in Figure 2. The results substantiate that calibration degradation is a consistent phenomenon observed across various LLMs, regardless of their architecture or parameter scale.

To examine whether this phenomenon extends to Full Fine-Tuning (FFT) scenarios, we conducted additional experiments using Llama3-8B model on both MathQA and ARC-C datasets. The results, visualized in Figure 11 and Figure 12, reveal that the pattern persists in FFT settings. Low-bias data consistently leads to model overfitting, while the introduction of new knowledge helps mitigate this effect and improves calibration. The observation of this pattern in FFT scenarios further strengthens our findings, suggesting that the relationship between bias and calibration is a robust phenomenon that transcends specific training approaches.

As a supplement, to ensure the stability and consistency of the experiments in Figure 3b, standard deviations from three trials with different random seeds are presented in Table 5. These values demonstrate the consistency and reliability of the results.

Additionally, to investigate the potential interactions between different knowledge bias levels across datasets in realistic scenarios, we conducted a comparative analysis using the OBQA dataset. We randomly sampled equal portions of pure

Steps	50	100	150	200	250	300	350	400	450	500	550	600
Conf of CS (known)	0.34	1.16	0.71	1.34	0.66	1.05	0.71	0.88	0.45	0.65	0.36	0.66
Conf of IS (known)	1.76	3.24	0.17	3.15	3.47	3.11	1.33	0.58	1.10	0.88	1.73	1.76
Conf of CS (unknown)	5.48	6.67	5.26	2.89	5.01	0.51	3.01	4.97	2.88	4.24	2.87	1.82
Conf of IS (unknown)	3.66	5.61	3.03	2.29	3.76	0.63	3.57	4.24	4.65	4.51	3.72	1.19

Table 5: Standard deviation of confidence values at different steps. Standard deviations are recorded from three trials with different random seeds. These values demonstrate the consistency and reliability of the results.

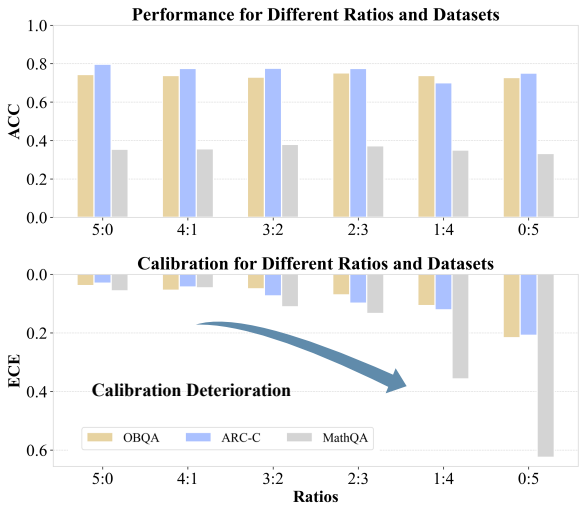


Figure 9: Accuracy and ECE of Llama3-8B fine-tuned with different knowledge biases in MathQA. The ratio varies from 5:0 to 0:5 (unknown data:known data), with equal dataset sizes. Calibration deteriorates as the knowledge bias lowers, while higher knowledge bias helps improve calibration aligning with findings in Section 3.1.

known data (low bias), pure unknown data (high bias), and mixed data. Figure 16 illustrates the calibration results after fine-tuning the model on these 3 dataset categories. The ECE curve for models fine-tuned on the mixed dataset consistently maintains an intermediate position between the other two curves throughout the fine-tuning process. This observation suggests that low-bias data effectively dilutes the calibration benefits achieved through high-bias fine-tuning.

C Details of LS, MbLS and ECP

C.1 Label Smoothing

Label Smoothing (LS) (Szegedy et al., 2016) replaces one-hot encoded labels with smoothed distributions by allocating small probabilities to non-target classes, effectively reducing model overconfidence and improving calibration performance in deep neural networks, as follows:

$$\mathcal{L}_{\mathrm{LS}} = -\sum_{k}((1-\epsilon)q_k + \frac{\epsilon}{K})\log p_k, \quad (4)$$

where ϵ is label smoothing factor, K denotes the numbers of total classes, $\mathbf{p}$ is the softmax probability predictions by model, which is computed as follows:

$$\mathbf{p} = (p_k)_{1\leq k\leq K} \in \mathbb{R}^K; \quad p_k = \frac{\exp^{l_k}}{\sum_j^K \exp^{l_j}}, \quad (5)$$

where $\mathbf{l} = (l_k)_{1\leq k\leq K} \in \mathbb{R}^K$ denotes logits vectors. Furthermore, the loss function of LS can be written as (Liu et al., 2022):

$$\mathcal{L}_{\mathrm{LS}} \triangleq \mathcal{L}_{\mathrm{CE}} + \frac{\epsilon}{1-\epsilon}\mathcal{D}_{\mathrm{KL}}\left(\mathbf{u}||\mathbf{p}\right), \quad (6)$$

where L_{CE} denotes Cross-Entropy loss, $\mathbf{u}$ denotes uniform distribution $\mathbf{u} = \frac{1}{K}$, $\triangleq$ stands for equality up to additive and/or nonnegative multiplicative constants.

C.2 Margin-based Label Smoothing

Margin-based Label Smoothing (MbLS) (Liu et al., 2022) addresses the calibration issue in deep neural networks by imposing inequality constraints on logit distances, unlike traditional methods that use equality constraints. This approach provides a better balance between model discrimination and calibration performance. MbLS introduces inequality constraints with controllable margins as follows:

$$\mathcal{L}_{MbLS} = \mathcal{L}_{\mathrm{CE}} + \gamma\sum_{k}\max(0, \max_{j}(l_j) - l_k - m), \quad (7)$$

where γ is the label smoothing factor, and m is the logits margin.

C.3 ECP

ECP (Pereyra et al., 2017) is a neural network regularization technique that works by penalizing low entropy output distributions as follows:

$$\mathcal{L}_{ECP} = \mathcal{L}_{\mathrm{CE}} - \beta\mathcal{H}(\mathbf{p}), \quad (8)$$

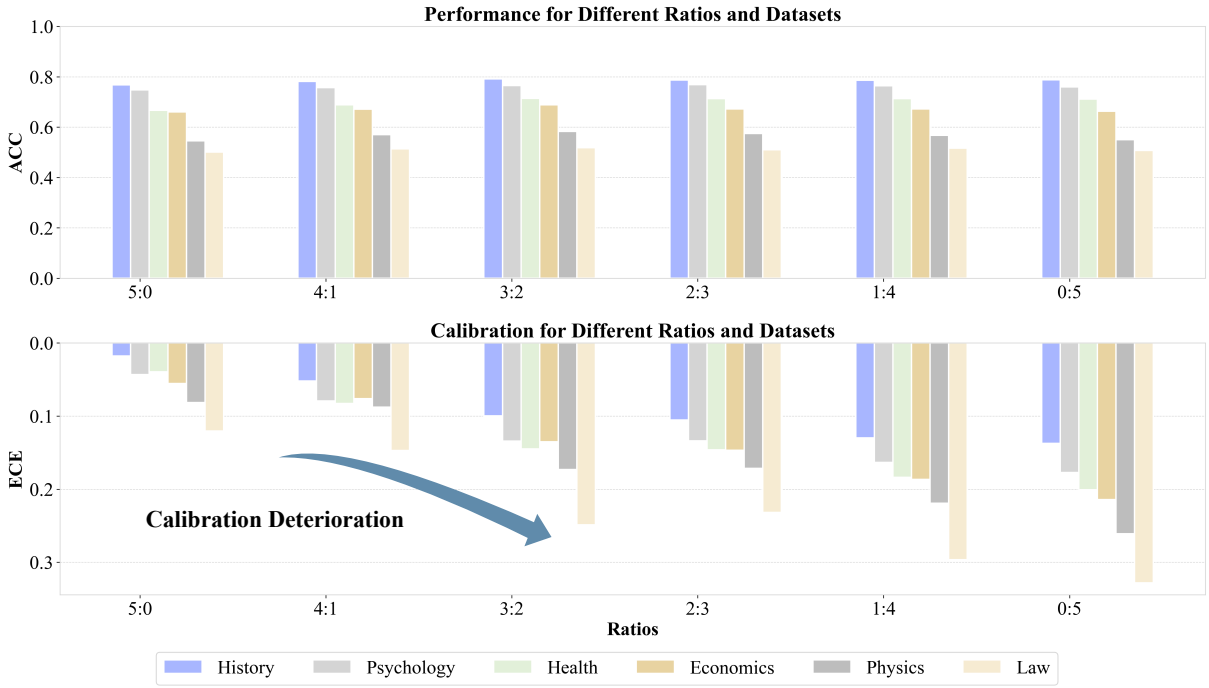


Figure 10: Accuracy and ECE of Llama3-8B fine-tuned with different knowledge biases in the open-ended dataset MedMCQA restructured by us. The ratio varies from 5:0 to 0:5 (unknown data:known data), with equal dataset sizes. Calibration also deteriorates as the knowledge bias lowers, while higher knowledge bias helps improve calibration in open-ended fine-tuning scenarios, aligning with findings in Section 3.1.

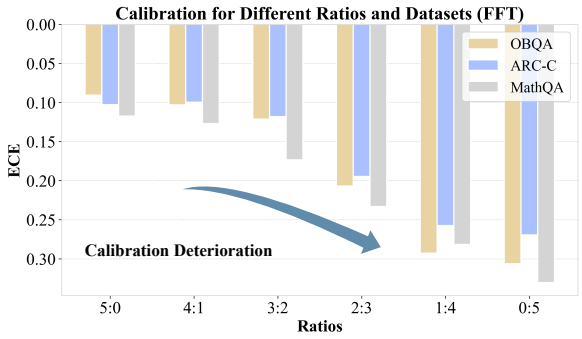


Figure 11: ECE of Llama3-8B fine-tuned with different knowledge biases in MathQA using Full Fine-Tuning (FFT). The ratio varies from 5:0 to 0:5 (unknown data:known data), with equal dataset sizes. Calibration deteriorates as the knowledge bias lowers, while higher knowledge bias helps improve calibration aligning with findings in Section 3.1.

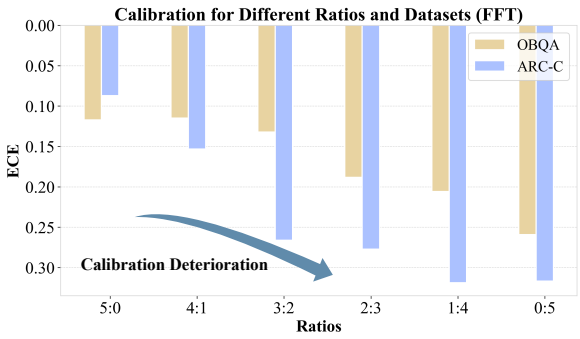


Figure 12: ECE of Llama3-8B fine-tuned with different knowledge biases in ARC-C using Full Fine-Tuning (FFT). The ratio varies from 5:0 to 0:5 (unknown data:known data), with equal dataset sizes. Calibration deteriorates as the knowledge bias lowers, while higher knowledge bias helps improve calibration aligning with findings in Section 3.1.

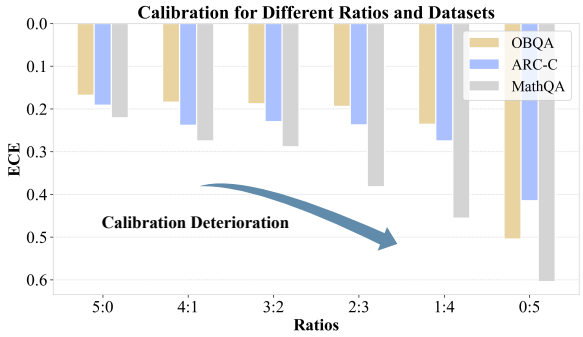


Figure 13: ECE of Llama2-13B fine-tuned with different knowledge biases following the same setup as Figure 2. The results confirm that calibration degradation is a consistent phenomenon across models.

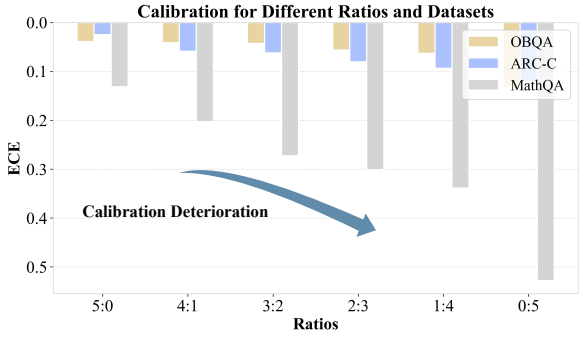


Figure 14: ECE of Qwen2.5-7B fine-tuned with different knowledge biases following the same setup as Figure 2. The results confirm that calibration degradation is a consistent phenomenon across models.

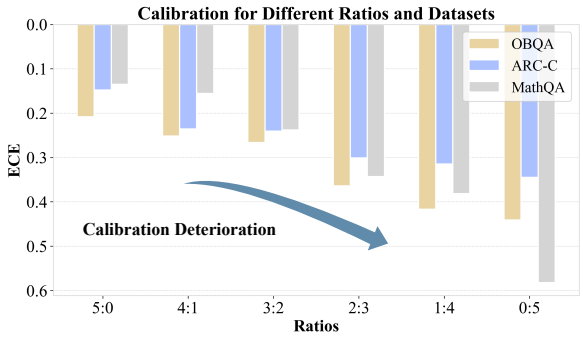


Figure 15: ECE of Mistral-7B fine-tuned with different knowledge biases following the same setup as Figure 2. The results confirm that calibration degradation is a consistent phenomenon across models.

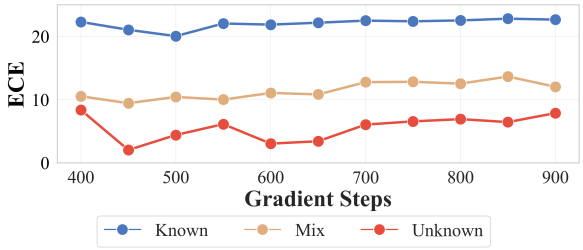


Figure 16: ECE of Llama3-8B after fine-tuning on unknown (high-bias), mixed, and known (low-bias) datasets, where the mixed dataset is randomly sampled from OBQA with an equal size. ECE curve for models fine-tuned on the mixed dataset maintains an intermediate position, indicating low-bias data would dilute the calibration benefits of high-bias data.

where β is ECP factor, and $\mathcal{H}$ denotes the Shannon entropy of the softmax prediction given by:

$$\mathcal{H}(\mathrm{p}) = -\sum_{k} p_k \log(p_k). \tag{9}$$

D Details of Style Adaptation

As shown in Table 6, the AUROC scores for t_0 discrimination demonstrate significant improvement with style adaptation. This improvement can be attributed to better alignment between the model's output style and fine-tuning data patterns after style adaptation, which enhances the model's capability to utilize NLL for distinguishing between known and unknown samples.

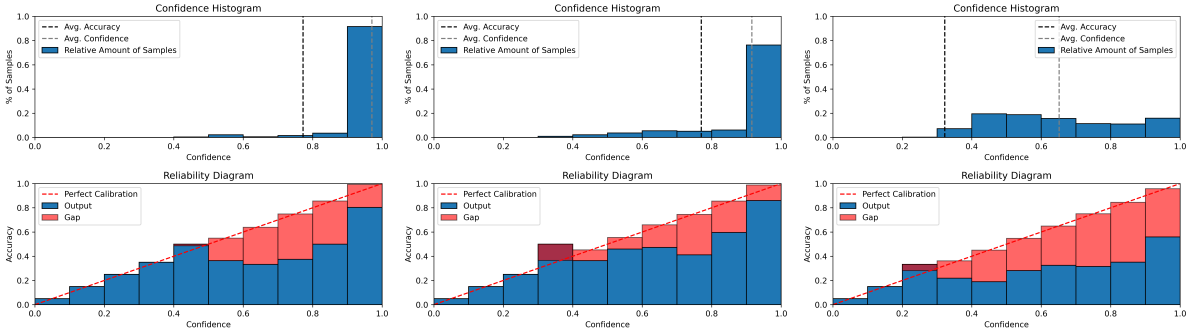
Dataset	t_0 w/o sa	t_0 w/ sa
HotpotQA	0.729	0.888
MedMCQA	0.712	0.908

Table 6: AUROC scores for t_0 discrimination in Llama3-8B with/without style adaptation (sa) on Open-End tasks.

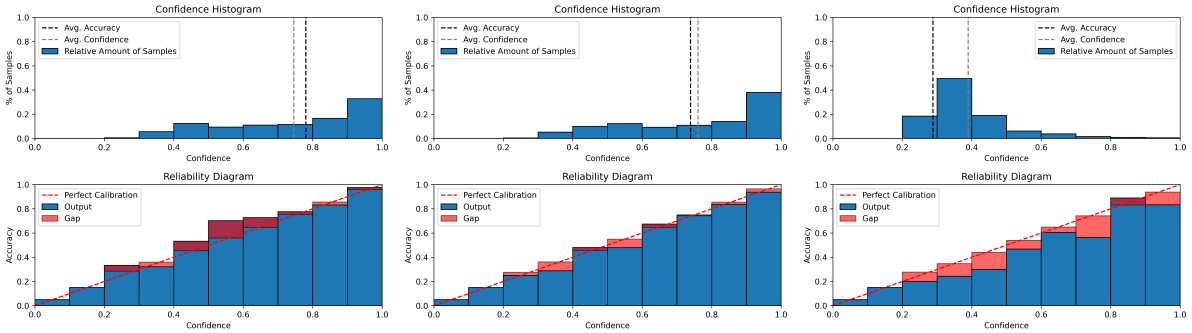
E Details of ECE and Reliability Diagram

Expected Calibration Error (ECE) serves as one of the primary metrics for assessing calibration, measuring the alignment between model confidence and accuracy. As demonstrated in Equation (10), ECE operates by partitioning model confidence (maximum output probabilities) into m bins, then computing a weighted sum of the discrepancies between accuracy and confidence across all bins.

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{N} |acc(B_m) - conf(B_m)|, \tag{10}$$



(a) known, OBQA, ECE=0.134 (b) known, ARC-C, ECE=0.148 (c) known, MathQA, ECE=0.330



(d) unknown, OBQA, ECE=0.047 (e) unknown, ARC-C, ECE=0.032 (f) unknown, MathQA, ECE=0.102

Figure 17: Reliability diagrams of models fine-tuned on OBQA known data or unknown data, evaluated on both ID test and OOD test (ARC-C, MathQA). Models trained on unknown data demonstrate better alignment between confidence and accuracy, further validating the conclusions drawn in Section 3.1.

where $|Bm|$ represents the number of samples in bin m, N denotes the total number of samples, while $acc(B_m)$ and $conf(Bm)$ are the average accuracy and average confidence in bin m, respectively.

In addition to ECE, calibration property can be visualized through Reliability Diagram (Bröcker and Smith, 2007). As illustrated in the Figure 17, there are significant differences in the reliability diagrams between models fine-tuned with unknown data versus known data. This distinction is evident in both ID and OOD scenarios, further corroborating the findings presented in Section 3.

F Additional Experimental Analysis

In this section, we demonstrate the versatility of the CogCalib through extensive experiments across a broader range of models and fine-tuning approaches. Our experiments encompass models of varying architectures and sizes, along with full-parameter fine-tuning methods (FFT). Additionally, we present supplementary experiments that evaluate the potential damage of calibrating unknown data on downstream tasks performance and assess the robustness of our method under different hyperparameter settings and threshold computation. Meanwhile, we provide complementary results comparing CoECP with Vanilla and Random Calibration approaches there.

F.1 Results of Llama3-8B on MedMCQA

We modified the question-answering format of MedMCQA and transformed it into an open-ended dataset for the medical domain, as explained in Appendix A. Table 7 presents the comprehensive experimental results of Llama3-8B on OOD datasets. The cognitive methods demonstrated superior calibration performance while maintaining comparable accuracy relative to baseline approaches. Different cognitive methods exhibited varying advantages across distinct datasets. Specifically, CoECP achieved optimal ECE on 4 datasets: Physics, Economics, Health, and Law. Meanwhile, CoMbLS showed superior ECE on both OBQA and ARC-C datasets. Notably, CoECP not only excelled in calibration but also maintained leading accuracy scores across almost all test sets.

It's worth noting that for a fairer comparison of

Metric	Methods	Physics	Economics	Health	Law	OBQA	ARC-C
ECE↓	**Vanilla SFT**	22.87	18.44	20.00	28.82	12.71	15.87
	MCD	18.15	15.44	15.94	24.88	6.93	10.72
	Ensemble	20.43	15.52	17.68	25.45	7.49	10.11
	TS	25.22	28.53	26.17	21.39	40.71	37.89
	CoLS	11.51	7.88	8.89	15.14	5.39	4.24
	CoMbLS	18.52	14.14	14.26	22.28	**3.83**	**2.99**
	CoECP	**9.14**	**6.26**	**6.58**	**14.89**	11.50	13.49
ACC↑	**Vanilla SFT**	55.78	67.78	68.70	**52.30**	59.20	64.84
	MCD	55.46	65.22	67.43	49.63	59.00	64.50
	Ensemble	55.00	67.78	68.90	50.70	64.00	69.02
	TS	55.78	67.78	68.70	**52.30**	59.20	64.84
	CoLS	54.37	66.30	68.35	51.27	67.00	70.64
	CoMbLS	55.15	66.84	69.08	50.31	70.60	74.57
	CoECP	**56.71**	**68.46**	**69.69**	50.76	**71.00**	**75.60**

Table 7: Comparison of our method's performance against baseline approaches on out-of-domain (OOD) datasets is presented. The results are evaluated on the Llama3-8B model, which is fine-tuned on the open-ended HotpotQA dataset.

Metric	Methods	Physics	Economics	Health	Law	OBQA	ARC-C
ECE↓	**Vanilla SFT**	18.14	14.07	16.49	22.55	8.00	6.87
	MCD	15.35	11.16	13.61	19.82	5.81	5.19
	Ensemble	15.26	9.27	12.51	21.53	5.60	5.50
	TS	34.80	40.10	36.80	34.20	54.80	58.60
	CoLS	10.39	6.01	8.84	**14.08**	4.96	2.75
	CoMbLS	**9.81**	**5.28**	**7.90**	14.70	**3.11**	2.85
	CoECP	13.39	7.82	9.42	16.98	5.30	**2.44**
ACC↑	**Vanilla SFT**	51.72	59.16	59.21	48.21	65.20	73.55
	MCD	53.13	59.03	59.15	48.44	65.20	72.27
	Ensemble	**53.75**	**64.02**	**61.95**	48.67	**67.60**	**74.49**
	TS	51.72	59.16	59.21	48.21	65.20	73.55
	CoLS	53.13	59.03	59.15	48.44	65.20	72.27
	CoMbLS	51.72	58.63	57.20	47.65	65.00	72.01
	CoECP	51.72	60.51	60.49	**48.72**	65.40	73.46

Table 8: Comparison of our method's performance against baseline approaches on out-of-domain (OOD) datasets is presented. The results are evaluated on the Llama3-8B model, which is fine-tuned on the open-ended MedMCQA dataset.

all methods in the OOD scenario, the temperature scaling baseline uses the best temperature found on the ID data when applied to OOD data. For long texts, we used the geometric mean of probabilities as confidences (Liu et al., 2023) to find the optimal calibration temperature.

Table 8 presents the experimental results of Llama3-8B on another open-ended dataset, MedMCQA. Across all OOD test sets, CogCalib demonstrates superior calibration performance, achieving the ECE compared to all baseline methods. While Deep Ensemble maintains a slight lead in ACC, CogCalib achieves comparable accuracy metrics with other baseline approaches. These findings suggest that CogCalib exhibits universal applicability across open-ended datasets.

F.2 Results of Llama2-13B

Table 9 and Table 10 demonstrate the experimental results of CogCalib on the Llama2-13B model. For this model size, CogCalib achieves significant improvements in calibration performance while maintaining baseline accuracy. Furthermore, it demonstrates superior accuracy across multiple datasets. These experimental findings validate that CogCalib maintains its robustness when applied to larger-scale models, effectively preserving fine-tuning performance while enhancing calibration under both ID and distribution shift scenarios.

F.3 Results of Qwen2.5-7B

We conducted comprehensive evaluations on the Qwen2.5-7B model, with results presented in Ta-

Dataset	Metric	Vanilla SFT	MCD	Ensemble	TS	CoLS	CoMbLS	CoECP
OBQA	**ACC↑**	73.60	73.40	77.20	73.60	76.20	**78.60**	77.40
	ECE↓	20.91	16.70	10.10	18.90	7.55	6.10	**2.62**
ARC-C	**ACC↑**	70.90	70.73	**72.10**	70.90	71.16	71.50	70.56
	ECE↓	25.98	22.14	18.21	24.70	**13.34**	13.36	14.43
WG-S	**ACC↑**	74.59	74.35	**74.98**	74.59	73.56	72.45	72.63
	ECE↓	16.96	15.73	15.95	14.90	**7.99**	8.34	16.00
WG-M	**ACC↑**	82.08	81.61	**82.40**	82.08	80.82	81.22	81.06
	ECE↓	16.63	14.58	12.43	15.60	7.44	7.17	**6.71**
BoolQ	**ACC↑**	89.85	89.94	89.85	89.85	89.72	**90.06**	89.17
	ECE↓	9.59	8.69	7.58	9.40	**2.04**	2.47	4.49

Table 9: Comparison of our method's performance against baselines on in-domain (ID) datasets. Results are evaluated on Llama2-13B model fine-tuned by LoRA on 5 widely used domain-specific datasets.

Metric	Methods	ID	Smaller Distribution Shift		Larger Distribution Shift			
		OBQA	ARC-C	ARC-E	Business	Culture	History	Psychology
ECE↓	**Vanilla SFT**	20.91	26.93	21.12	20.86	22.41	22.41	30.40
	MCD	16.70	23.78	18.07	17.50	20.26	15.41	27.22
	Ensemble	10.10	18.87	14.27	17.76	18.26	14.07	25.80
	TS	18.90	25.00	19.70	20.90	21.00	22.70	28.80
	CoLS	7.55	13.48	8.51	11.65	**13.73**	12.45	20.78
	CoMbLS	6.10	12.10	7.00	**11.42**	13.90	11.72	19.98
	CoECP	**2.62**	**8.60**	**3.30**	12.49	14.63	**9.77**	**19.78**
ACC↑	**Vanilla SFT**	73.60	67.32	74.37	75.06	72.89	69.89	63.61
	MCD	73.40	66.72	74.03	**75.51**	72.89	68.49	63.70
	Ensemble	77.20	68.86	76.47	**75.51**	73.19	67.31	64.30
	TS	73.60	67.32	74.37	75.06	72.89	69.89	63.61
	CoLS	76.20	68.94	76.18	**75.51**	**74.10**	70.97	64.39
	CoMbLS	78.60	**69.20**	**76.98**	**75.51**	73.19	**71.61**	64.30
	CoECP	**77.40**	67.83	75.34	75.29	72.89	71.08	**65.25**

Table 10: Comparison of our method's performance against baselines on distribution shift datasets is presented. Results are evaluated on Llama2-13B model which is fine-tuned on the OBQA dataset.

Dataset	Metric	Vanilla SFT	MCD	Ensemble	TS	CoLS	CoMbLS	CoECP
OBQA	**ACC↑**	90.60	90.80	**91.80**	91.60	91.60	**91.80**	91.20
	ECE↓	8.48	8.09	5.05	7.50	5.65	**3.77**	7.51
ARC-C	**ACC↑**	87.46	**88.57**	87.54	88.48	87.63	87.97	88.48
	ECE↓	11.41	8.89	9.76	9.90	4.24	**3.22**	9.37
WG-S	**ACC↑**	78.37	78.93	**79.87**	79.40	76.95	78.37	78.45
	ECE↓	19.50	14.84	14.89	14.40	11.06	**10.41**	13.64
WG-M	**ACC↑**	83.90	83.50	**85.16**	83.90	84.53	83.74	84.61
	ECE↓	15.17	11.44	10.21	14.70	**4.20**	4.84	6.77
BoolQ	**ACC↑**	89.72	90.28	**90.61**	90.21	90.37	90.09	89.91
	ECE↓	9.71	7.91	7.10	8.80	**1.45**	4.22	7.19

Table 11: Comparison of our method's performance against baselines on in-domain (ID) datasets. Results are evaluated on Qwen2.5-7B model fine-tuned by LoRA on 5 widely used domain-specific datasets.

Metric	Methods	ID	Smaller Distribution Shift		Larger Distribution Shift			
		OBQA	ARC-C	ARC-E	Business	Culture	History	Psychology
ACC↑	**Vanilla SFT**	90.60	87.54	90.82	88.10	**85.24**	85.48	83.32
	MCD	90.80	87.12	90.49	86.96	84.04	84.85	82.89
	Ensemble	**91.80**	**88.05**	**91.04**	**88.79**	83.43	85.48	83.75
	TS	91.60	87.54	90.82	88.10	**85.24**	85.48	83.32
	CoLS	91.60	87.63	90.15	87.41	83.73	85.70	82.54
	CoMbLS	**91.80**	87.63	90.91	88.10	81.63	85.27	83.23
	CoECP	91.20	87.29	90.99	87.41	**85.24**	**85.91**	**83.75**
ECE↓	**Vanilla SFT**	8.48	11.51	8.60	10.78	13.66	12.65	15.09
	MCD	8.09	7.54	8.29	9.09	9.90	9.00	11.63
	Ensemble	5.05	9.27	7.38	8.51	12.17	10.51	13.08
	TS	7.50	11.20	8.30	9.90	13.00	12.10	14.60
	CoLS	5.05	**0.75**	2.43	**6.34**	**5.95**	**4.80**	**5.10**
	CoMbLS	**3.77**	2.56	**1.99**	**6.34**	8.17	4.90	6.90
	CoECP	7.51	8.36	8.33	7.66	6.18	6.75	6.66

Table 12: Comparison of our method's performance against baselines on distribution shift datasets is presented. Results are evaluated on Qwen2.5-7B model which is fine-tuned on the OBQA dataset.

ble 11 and Table 12. Although Qwen demonstrated superior accuracy across all datasets and inherently low ECE compared to other LLMs, our CogCalib framework still achieved significant calibration improvements over the baseline. In the in-distribution (ID) testing (Table 11), both CoLS and CoMbLS consistently outperformed other approaches, while maintaining accuracy comparable to the best-performing Ensemble methods. For out-of-distribution (OOD) scenarios (Table 12), CoLS and CoMbLS achieved optimal performance across all distribution shift conditions. Notably, CoECP exhibited competitive accuracy performance under multiple larger distribution shift scenarios.

F.4 Results of Mistral-7B-v0.3

We evaluated CogCalib's performance on Mistral-7B-v0.3, with results presented in Table 13 and Table 14. For in-domain testing (Table 13), our approach demonstrated superior calibration metrics compared to all baselines. While the Ensemble method achieved optimal accuracy in most cases, our method maintained competitive accuracy scores. In out-of-distribution (OOD) scenarios (Table 14), our approach outperformed the baselines in both ECE and accuracy metrics.

F.5 Results of Llama3-8B Using FFT

In addition to LoRA, we validated CogCalib using Full-parameter Fine-Tuning (FFT) on the Llama3-8B model. Table 15 and Table 16 present the results for ID and OOD evaluations, respectively, demonstrating that CogCalib is effectively applicable to FFT. The method not only significantly enhances calibration performance but also improves model generalization. Specifically, the model trained on ARC-C achieved an average accuracy of 70.67% across other tasks, surpassing the conventional SFT baseline (67.13%). We hypothesize that the introduction of the calibration term mitigates overfitting

Dataset	Metric	Vanilla SFT	MCD	Ensemble	TS	CoLS	CoMbLS	CoECP
OBQA	**ACC↑**	68.40	67.60	65.40	68.40	**72.60**	70.20	69.00
	ECE↓	25.85	22.65	18.01	23.70	**9.61**	12.70	14.43
ARC-C	**ACC↑**	77.13	77.47	**79.18**	77.13	78.24	77.47	77.30
	ECE↓	20.99	17.86	13.16	20.10	**8.22**	9.39	8.90
WG-S	**ACC↑**	77.11	77.43	**80.19**	77.11	78.06	78.14	77.74
	ECE↓	21.34	19.41	13.36	20.40	11.17	**9.48**	10.84
WG-M	**ACC↑**	83.11	83.27	**83.90**	83.11	82.95	83.35	82.24
	ECE↓	15.47	14.35	10.47	14.70	6.27	**4.70**	6.10
BoolQ	**ACC↑**	89.57	89.60	**90.89**	89.57	89.94	89.48	90.21
	ECE↓	10.22	9.31	6.09	10.00	1.28	**0.56**	1.45

Table 13: Comparison of our method's performance against baselines on in-domain (ID) datasets. Results are evaluated on Mistral-7B model fine-tuned by LoRA on 5 widely used domain-specific datasets.

Metric	Methods	ID	Smaller Distribution Shift		Larger Distribution Shift			
		OBQA	ARC-C	ARC-E	Business	Culture	History	Psychology
ECE↓	**Vanilla SFT**	20.91	26.93	21.12	20.86	22.41	22.41	30.40
	MCD	16.70	23.78	18.07	17.50	20.26	15.41	27.22
	Ensemble	10.10	18.87	14.27	17.76	18.26	14.07	25.80
	TS	18.90	25.00	19.70	20.90	21.00	22.70	28.80
	CoLS	7.55	13.48	8.51	11.65	**13.73**	12.45	20.78
	CoMbLS	6.10	12.10	7.00	**11.42**	13.90	11.72	19.98
	CoECP	**2.62**	**8.60**	**3.30**	12.49	14.63	**9.77**	**19.78**
ACC↑	**Vanilla SFT**	73.60	67.32	74.37	75.06	72.89	69.89	63.61
	MCD	73.40	66.72	74.03	**75.51**	72.89	68.49	63.70
	Ensemble	77.20	68.86	76.47	**75.51**	73.19	67.31	64.30
	TS	73.60	67.32	74.37	75.06	72.89	69.89	63.61
	CoLS	76.20	68.94	76.18	**75.51**	**74.10**	70.97	64.39
	CoMbLS	**78.60**	**69.20**	**76.98**	**75.51**	73.19	**71.61**	64.30
	CoECP	77.40	67.83	75.34	75.29	72.89	71.08	**65.25**

Table 14: Comparison of our method's performance against baselines on distribution shift datasets is presented. Results are evaluated on Mistral-7B model which is fine-tuned on the OBQA dataset.

Dataset	Metric	Vanilla SFT	MCD	TS	CoLS	CoMbLS	CoECP
ARC-C	**ACC↑**	66.81	65.96	66.81	70.82	**72.18**	70.22
	ECE↓	29.84	28.31	28.61	13.09	14.19	**8.66**
ARC-E	**ACC↑**	**75.04**	74.71	**75.04**	72.85	73.99	74.92
	ECE↓	17.38	16.18	14.14	13.04	11.54	**4.28**

Table 15: Comparison of our method's performance against baselines on in-domain (ID) datasets. Results are evaluated on Llama3-8B model using full parameter fine-tuning (FFT) on 2 widely used domain-specific datasets.

Metric	Methods	ID		OOD			
		ARC-C	ARC-E	OBQA	Business	History	Psychology
ACC↑	**Vanilla SFT**	66.81	74.87	68.40	72.31	64.52	55.57
	MCD	65.96	74.71	69.40	72.31	63.23	55.14
	TS	66.81	74.87	68.40	72.31	64.52	55.57
	CoLS	70.82	72.85	**73.80**	71.85	70.97	**67.42**
	CoMbLS	**72.18**	73.99	67.60	72.31	**72.37**	64.82
	CoECP	70.22	**74.92**	65.60	**75.74**	69.14	66.64
ECE↓	**Vanilla SFT**	29.84	22.63	27.99	22.29	30.29	40.19
	MCD	28.31	20.87	23.59	20.10	28.67	37.72
	TS	28.61	21.60	26.20	20.90	28.60	38.40
	CoLS	13.09	13.04	**8.00**	9.60	12.73	16.14
	CoMbLS	14.19	11.54	15.72	11.45	13.32	20.69
	CoECP	**8.66**	**4.28**	9.47	**7.90**	**10.66**	**11.07**

Table 16: Comparison of our method's performance against baselines on OOD datasets. Results are evaluated on Llama3-8B model using full parameter fine-tuning (FFT) on OOD scenarios.

during fine-tuning, thereby enhancing the model's generalization capabilities.

F.6 Effects of Calibrating Unknown Data

We investigated the impact of calibrating unknown data on the model's downstream task performance. Beyond the adverse effects on OOD tasks observed with multiple-choice data in Section 5.3, we found that this negative impact was even more pronounced when using open-ended data as the fine-tuning dataset.

In our experimental setup, we utilized equal amounts (1k samples) of known or unknown data from MedMCQA as fine-tuning datasets, implementing various calibration enhancement methods. As illustrated in the Figure 18, while using calibration methods on known data did not significantly affect performance on OOD tasks, applying calibration methods to unknown data led to a consistent decline in OOD performance. These findings underscore the critical importance of thoroughly learning from unknown data during the fine-tuning process.

F.7 Comparision to Vanilla and Random Calibration.

We present a comparative analysis of CoECP against its corresponding variants: Vanilla calibration and Random Calibration. As illustrated in Figure 19, CoECP consistently outperforms both baseline methods across most datasets in terms of calibration and fine-tuning performance. These results further support our findings in Section 3, which emphasize that unknown data plays a crucial role in aligning the model with downstream tasks. The effective utilization of such data simultaneously enhances fine-tuning performance and improves calibration metrics.

F.8 Sensitivity to Threshold Computation.

We compare our threshold calculation called **Balanced**, with the **Accuracy-based** method seeking threshold t which achieves best known/unknown data classification accuracy using negative log-likelihood.

Figure 20 reveals that these 2 calculations have similar calibration effects, but our method attains a higher accuracy. This improvement can be attributed to the higher TNR achieved by our approach (Table 17). Accuracy-based calculation tends to misclassify unknown samples, consequently applying calibration to these samples as well, which prevents LLM from effectively learning critical knowledge.

In detail, Table 17 presents the classification results for both known and unknown data during the fine-tuning under different threshold calculation methods. The results demonstrate that when using our method, both TPR and TNR are well-balanced. In contrast, when employing the highest accuracy for threshold calculation, the TNR exhibits notably lower values. This discrepancy indicates that the latter method incorrectly classifies unknown data as known data and subsequently applies calibration methods, preventing the model from effectively learning crucial patterns in unknown data during the fine-tuning process. This limitation explains the consistently inferior performance of this method compared to the former approach, as illustrated in Figure 20.

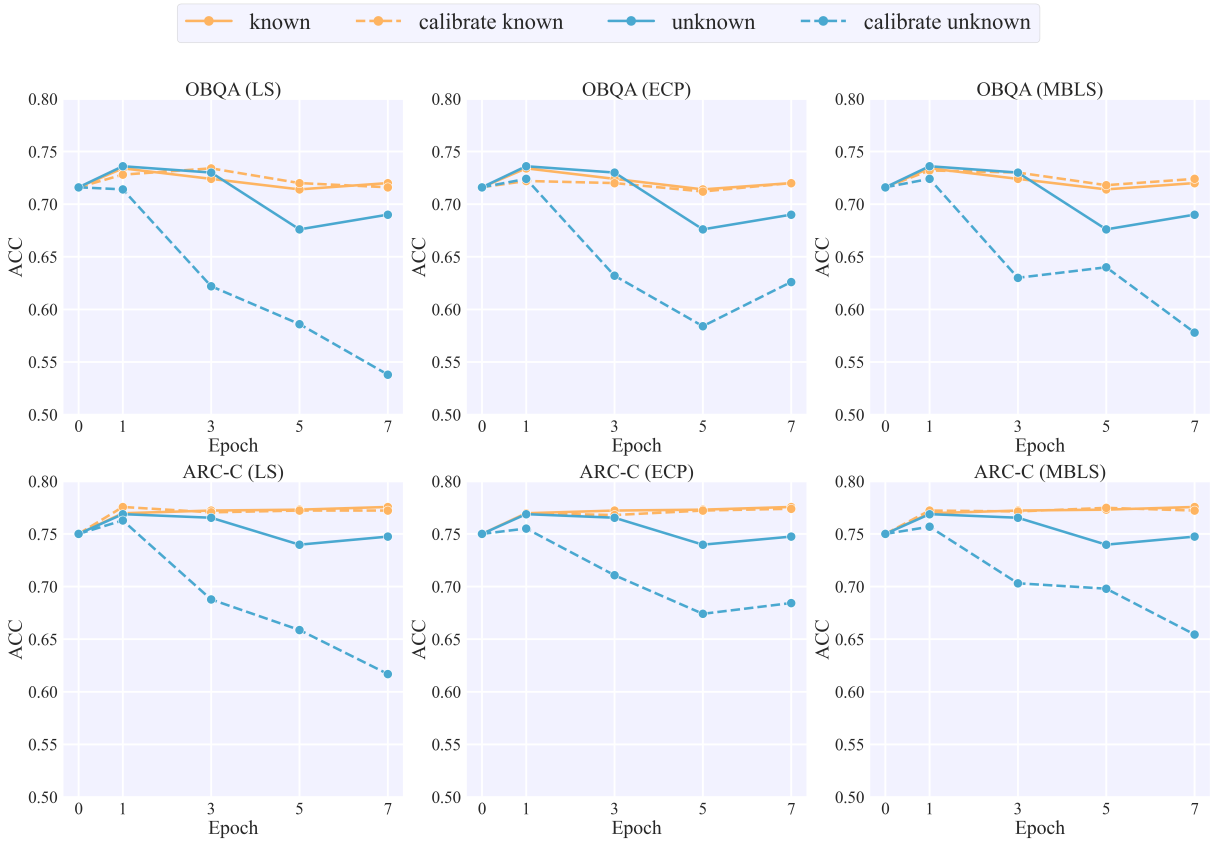


Figure 18: Comparison of accuracy on downstream tasks with and without calibration methods when fine-tuning on MedMCQA known or unknown data. It is observed that calibrating on unknown data significantly deteriorates the performance of other out-of-distribution (OOD) tasks.

Dataset	argmax(TNR+TPR)			argmax(Accuracy)		
	Accuracy	TPR	TNR	Accuracy	TPR	TNR
ARC-C	99.51	99.54	99.15	99.58	99.95	95.18
OBQA	99.44	99.44	99.52	99.78	99.91	96.82
WG-S	98.83	98.77	99.52	99.49	99.83	96.41

Table 17: Performance metrics (Accuracy, True Positive Rate (TPR), and True Negative Rate (TNR)) for different optimization criteria on the ARC-C, OBQA, and WG-S datasets.

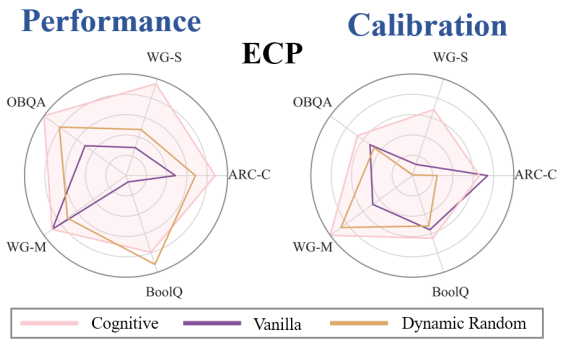


Figure 19: Comparsion between CogECP, vanilla ECP, and dynamic random ECP. Our proposed method, Co-ECP, consistently outperforms all baseline methods in terms of both calibration and accuracy across multiple datasets.

Dataset	Metrics	ECP factor 0.05	0.075	0.1	0.125
ARC-C	**ACC**	82.20	81.80	81.50	80.90
	ECE	15.38	15.66	**7.21**	15.85
wino-S	**ACC**	80.20	78.70	79.20	79.20
	ECE	17.05	18.72	**2.38**	6.28

Table 18: Performance metrics (ACC and ECE) for different ECP factors on the ARC-C and wino-S datasets.

F.9 Sensitivity to Hyperparameters.

We demonstrated additional robustness results regarding CogCalib hyperparameters in Table 18, Table 19 and Table 20. By varying the ECP Factor from 0.05 to 0.125 and MbLS Factor from 0.05

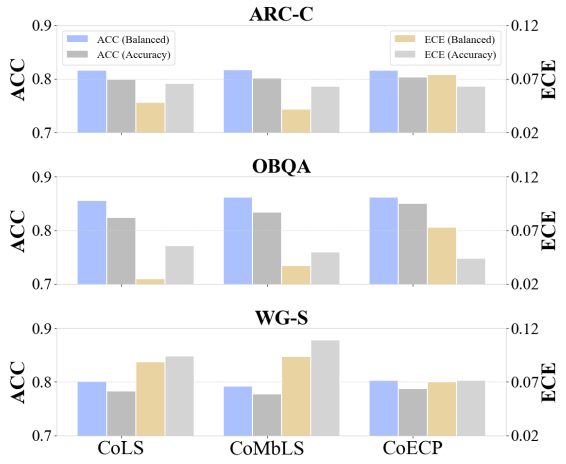


Figure 20: Sensitivity to Threshold Computation. Our method demonstrates robust performance across different threshold calculation approaches, while employing our proposed threshold computation methodology yields superior fine-tuning performance.

to 0.125 (with margins of 0 and 5), our method consistently achieved improved calibration performance compared to the Temperature Scaling baseline, which showed ECE values of 12.3 on ARC-C and 15.4 on Wino-S. These results across multiple parameter settings validate the robustness of our approach. However, it is worth noting that the ECP parameters may require more fine-grained tuning for optimal performance.

Dataset	Metrics	MbLS Factor (Margin=0) 0.05	0.075	0.1	0.125
ARC-C	**ACC**	80.30	81.90	81.60	82.30
	ECE	12.55	8.87	6.22	**2.37**
WG-S	**ACC**	77.30	78.10	79.60	78.60
	ECE	15.55	12.01	9.94	**8.06**

Table 19: Performance metrics (ACC and ECE) for different MbLS factors with Margin=0 on the ARC-C and WG-S datasets.

Dataset	Metrics	MbLS Factor (Margin=5) 0.05	0.075	0.1	0.125
ARC-C	**ACC**	81.20	81.10	81.40	82.20
	ECE	11.35	9.63	5.86	**2.53**
WG-S	**ACC**	78.10	78.20	78.80	79.30
	ECE	15.09	11.87	11.04	**6.19**

Table 20: Performance metrics (ACC and ECE) for different MbLS factors with Margin=5 on the ARC-C and WG-S datasets.

G Implementation Details

In this section, we present a detailed analysis of the SliCK method, the implementation of Temperature Scaling for both open-ended and multiple-choice data, along with our specific hyperparameter configurations.

G.1 Details of SliCK

In Section 3, we employed the SliCK (Gekhman et al., 2024) to classify known and unknown data. Specifically, the SliCK method concatenates 10 different randomly selected 4-shot prompts for each question-answer pair and performs 16 sampling iterations with temperature settings of either 0 or 0.5.

The prediction accuracy under greedy decoding is denoted as $P(T = 0)$, while $P(T > 0)$ represents the prediction accuracy when $T = 0.5$. Based on the accuracy calculations from multiple sampling iterations, the data is categorized into 4 classes: HighlyKnown, MaybeKnown, WeaklyKnown, and Unknown as shown in Table 21. For the experiments conducted in Section 3, we treated HighlyKnown samples as known data and maintained the Unknown classification as is.

Type	Definition
HighlyKnown	$P(T = 0) = 1$
MaybeKnown	$P(T = 0) \in (0, 1)$
WeaklyKnown	$P(T = 0) = 0 \wedge P(T > 0) > 0$
Unknown	$P(T \geq 0) = 0$

Table 21: SliCK's definition of different type of data.

G.2 Details of Temperature Scaling

For multiple-choice datasets, we employ the maximum probability of the first output token as the confidence. In contrast, for long-text datasets, we adopt the geometric mean probability as the confidence for long-text generation, following the approach proposed in LITCAB (Liu et al., 2023), as illustrated in the Equation (11),

$$p(y|x) = \sqrt[L]{\prod_{t=1}^{L} p(y_t|x, y_{<t})}. \quad (11)$$

To determine the optimal temperature for individual tokens, we implement the method developed by Guo et al.[1] , which utilizes gradient descent to minimize the ECE loss and identify the optimal

[1]https://github.com/gpleiss/temperature_scaling

temperature parameter. For long-text confidence, we employ a grid search strategy to determine the optimal temperature. The resulting optimal temperature is then applied to enhance calibration for both ID and OOD samples, ensuring a fair comparison.

Hyperparamter	value
LS ϵ	0.1
MbLS γ	0.1
MbLS Margin	0
ECP β	0.1

Table 22: Calibration term's hyperparameters for multi-choice QA task.

It is important to note that the optimal temperature discovered on the validation set typically fails to improve OOD calibration, which represents a significant limitation of Temperature Scaling. This limitation becomes particularly critical in the context of LLMs, which are required to handle diverse tasks effectively.

G.3 Hyperparameters

In CogCalib's framework, we experiment with 3 calibration losses: Label Smoothing (LS), Margin-based Label Smoothing (MbLS), and ECP. Considering the distinct nature of tasks we experimented on: multiple-choice QA with concentrated probability distributions and open-ended QA with inherently higher uncertainty in outputs, we adopted 2 sets of calibration loss hyperparameters. Specifically, the hyperparameter settings for multi-choice QA and open-ended QA are presented in Table 22 and Table 23, respectively. All experiments repeat three times, and the average results are recorded. Models smaller than 13B parameters are trained on an NVIDIA RTX-4090 GPU, while the 13B model is trained on an NVIDIA A100 GPU.

Hyperparamter	value
LS ϵ	0.15
MbLS γ	0.15
MbLS Margin	10
ECP β	0.15

Table 23: Calibration term's hyperparameters for open-ended QA task.

In our experimental setup, we fine-tune the LLM using both LoRA and FFT approaches. For LoRA implementation, we incorporate LoRA adapters into all linear layers of the LLM, maintaining the default PEFT configurations from Huggingface, as detailed in Table 24. The FFT parameters are specified in Table 25.

Hyperparamter	value
LoRA r	8
LoRA α	16
LoRA target	all
Learning Rate	6.0×10^{-5}
Batch size	2
Learning Rate scheduler	Linear
Max Sequence Length	1024

Table 24: Experimental hyperparameters used for LoRA fine-tuning.

Hyperparamter	value
Learning Rate	1.0×10^{-5}
Batch size	4
Learning Rate scheduler	Cosine
Warmup Ratio	0.1
Max Sequence Length	1024

Table 25: Experimental hyperparameters used for FFT fine-tuning.