CALIBRATING LLMS VIA REGION-SPECIFIC TEMPERATURE REFINEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) achieve strong performance across a wide range of tasks, yet post-training procedures such as instruction tuning and preference optimization often make their predicted probabilities substantially overconfident, undermining the reliability of their confidence estimates in downstream applications. We identify the reliability of the representation space as a key factor in sample-wise calibration: in reliable regions, correct and incorrect predictions are well separated, whereas in noisy regions they are mixed. Existing sample-wise methods predict temperatures uniformly across both, and therefore overfit in noisy regions and generalize poorly. To address this, we propose TS3, a theoretically grounded three-stage post-hoc framework that progressively refines temperatures from the global to the sample level. TS3 first applies global temperature scaling, then learns sample-specific temperature shifts in reliable regions while regularizing shifts toward zero in noisy regions, and finally learns a dedicated temperature for the noisy regions. Experiments on 50 tasks from three benchmarks with four instruction-tuned LLMs show that TS3 consistently achieves the lowest calibration error among a range of global, sample-wise, and prompt-based baselines. These results demonstrate the benefit of combining global, sample-level, and noise-aware calibration.

1 INTRODUCTION

Large language models (LLMs) achieve strong performance across a wide range of applications, including question answering Brown et al. (2020), reasoning Wei et al. (2022), and planning Yao et al. (2022). For reliable deployment, however, accuracy alone is insufficient: models must also provide confidence estimates that reflect the true likelihood of being correct. While pretrained LLMs are typically better calibrated than their post-trained counterparts, post-training techniques such as instruction tuning Wang et al. (2023) and preference optimization Christiano et al. (2017); Rafailov et al. (2023) make them substantially overconfident Tian et al. (2023); Shen et al. (2024); Luo et al. (2026). Such overconfidence is especially concerning in safety-critical domains such as healthcare Begoli et al. (2019); Pal et al. (2022), where confident but incorrect predictions can lead to costly decisions.

Calibrating LLMs through retraining Mukhoti et al. (2020); Kumar et al. (2018); Thulasidasan et al. (2019) or repeated inference Jiang et al. (2023b); Xiong et al. (2024) is expensive at scale, making post-hoc calibration particularly appealing. Among post-hoc methods, Temperature Scaling (TS) Guo et al. (2017) is widely used for its simplicity and its ability to preserve accuracy. However, TS learns a single global temperature, which cannot capture the heterogeneous confidence behavior of individual samples. We show that under such heterogeneity, no single temperature is optimal (Proposition 5.2).

Existing refinements of TS fall into two mechanisms. *Global methods*, such as TvA Le Coz et al. (2024) and unsupervised DACA Luo et al. (2026), still apply a single temperature to all samples and cannot adapt to sample-level heterogeneity. *Sample-wise methods* Tomani et al. (2022); Shen et al. (2024); Joy et al. (2023); Xie et al. (2024); Tomani et al. (2023) train a neural network to predict a temperature for each sample, and either apply these predictions directly or aggregate them into coarser temperatures, such as the task-level temperature in Thermometer Shen et al. (2024). However, the LLM representation space is not uniformly reliable: in some regions, correct and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

incorrect predictions are mixed, leaving no consistent signal for predicting a temperature. Because sample-wise predictors are trained on all samples, including those in these noisy regions, they overfit the calibration set and generalize poorly. Even DAC Tomani et al. (2023), which leverages feature-space density, does not distinguish reliable regions from noisy ones.

We propose **TS3**, a three-stage post-hoc framework consisting of **T**emperature **S**caling → **T**emperature **S**hifting → **T**emperature **S**caling. The key idea is to learn sample-level temperatures only in reliable regions of the representation space where the signal is consistent, and to assign a shared temperature to noisy regions. **Stage 1** learns a global base temperature. **Stage 2** learns a sample-specific shift from this base, guided by *agreement purity*, which measures whether a sample's neighbors in representation space share its correctness. An uncertainty regularizer drives the shift toward zero in low-purity regions i.e. noisy regions. **Stage 3** treats near-zero shifts as a signal of noisy regions and re-learns a dedicated temperature for them. Because every stage only rescales logits, TS3 preserves the model's predictions.

Our contributions: (1) We identify the reliability of the representation space as a key factor in sample-wise calibration and propose TS3, which combines global, sample-level, and noise-aware temperatures. (2) We provide a theoretical analysis showing that a global temperature is suboptimal under heterogeneity, that sample-specific shifts provably reduce risk on reliable regions, and that the ECE bin structure motivates the re-calibration in Stage 3. (3) Experiments on 50 tasks from BIG-Bench, MedMCQA, and GLUE with four instruction-tuned LLMs show that TS3 achieves the lowest ECE in all 12 (benchmark, model) settings, reducing the ECE of TS by 35.5% on average, and remains effective under multi-domain and out-of-distribution calibration. Ablations further show that the uncertainty regularizer is essential: without it, sample-wise shifts overfit noisy regions and perform worse than TS.

2 Related Work

Post-hoc calibration. Post-hoc methods adjust confidence estimates after training without modifying model parameters. Classical approaches include histogram binning Zadrozny & Elkan (2001); Naeini et al. (2015), isotonic regression Zadrozny & Elkan (2002), and Temperature Scaling (TS) Guo et al. (2017). TS is widely adopted for its simplicity and its ability to preserve accuracy, and has been extended with class-wise, bin-wise, and binary-reduction variants, including Dirichlet calibration Kull et al. (2019), Bin-wise TS Ji et al. (2019), and TvA Le Coz et al. (2024). For LLMs, DACA Luo et al. (2026) learns a temperature without labels by leveraging the pretrained model. These methods apply the same calibration map to all samples. However, a model can be calibrated overall while remaining miscalibrated within sub-populations Hébert-Johnson et al. (2018); Perez-Lebel et al. (2023), which cannot be corrected by any global recalibration.

Sample-wise temperature scaling. Adaptive methods instead predict a temperature for each input. PTS Tomani et al. (2022) maps logits to a sample-specific temperature, while other variants condition the temperature on latent representations Joy et al. (2023) or predictive entropy Balanya et al. (2024). DAC Tomani et al. (2023) uses k-NN density over intermediate-layer features, and ATS Xie et al. (2024) predicts per-token temperatures from LLM hidden states. Thermometer Shen et al. (2024) trains an auxiliary network across tasks and aggregates its predictions into task-level temperatures. Unlike these methods, which fit the temperature predictor on all samples, our proposed TS3 restricts sample-level estimation to reliable regions and assigns a shared temperature to noisy regions.

Calibration-oriented training. Other methods improve calibration during training, such as focal loss Mukhoti et al. (2020), mixup Thulasidasan et al. (2019), and MMCE Kumar et al. (2018). Liu et al. show that label smoothing Müller et al. (2019) and focal loss Lin et al. (2017) can be viewed as approximations of a linear penalty that mitigates the overconfidence induced by NLL training. For LLMs, EliCal Ni et al. (2026) fine-tunes models to align confidence with self-consistency, which serves as a proxy for correctness. These approaches require additional training, which is costly at LLM scale.

Prompting and data augmentation. Elicitation Tian et al. (2023) prompts LLMs to verbalize their confidence, while CAPE Jiang et al. (2023b) and Elicitation-Ensemble Xiong et al. (2024) aggregate predictions over multiple prompts or samples. These methods work well on very large models but

degrade on smaller ones Shen et al. (2024); Luo et al. (2026), and ensemble variants require multiple inference passes per input.

3 PRELIMINARIES

Problem formulation. We consider a C-class prediction task with data distribution $\mathcal{D}$. We are given a calibration set $D_{\text{cal}} = \{z_i = (x_i, y_i)\}_{i=1}^{N}$ and a disjoint test set D_{test} of N_{test} samples, both drawn i.i.d. from $\mathcal{D}$. For each input x_i, let $h_i \in \mathbb{R}^d$ denote the hidden-state representation of the last token produced by the LLM, and let $t_i \in \mathbb{R}^C$ denote its logit vector, with $t_i^{(j)}$ the logit for class j. The predicted class probabilities are $\sigma(t_i)$, where $\sigma(\cdot)$ denotes the softmax function and $\sigma(t_i)_j$ is the probability of class j. The predicted label and its confidence are $\hat{y}_i = \arg\max_{1\le j\le C} \sigma(t_i)_j$ and $\hat{p}_i = \max_{1\le j\le C} \sigma(t_i)_j$, respectively.

Objective. Let Y, $\hat{Y}$, and $\hat{P}$ denote the random variables for the true label, predicted label, and confidence, respectively. A model is perfectly calibrated if, among all predictions made with confidence γ, a fraction γ is correct, i.e., $\Pr(Y = \hat{Y} \mid \hat{P} = \gamma) = \gamma$ for all $\gamma \in [0, 1]$. Our goal is to learn a post-hoc mapping that brings $\hat{P}$ as close as possible to this condition.

A widely used metric for quantifying calibration quality is the Expected Calibration Error (ECE) Naeini et al. (2015), defined as $\mathbb{E}_{\hat{P}}\big[\,|\Pr(Y = \hat{Y} \mid \hat{P}) - \hat{P}|\,\big]$. In practice, ECE is estimated on D_{test} by partitioning $[0, 1]$ into M equal-width bins and assigning each sample to a bin according to its confidence $\hat{p}_i$. Letting B_m denote the set of samples whose confidence falls in $\big(\frac{m-1}{M}, \frac{m}{M}\big]$, the bin-wise accuracy and average confidence are $\mathrm{acc}(B_m) = \frac{1}{|B_m|}\sum_{i\in B_m} \mathbf{1}[y_i = \hat{y}_i]$ and $\mathrm{conf}(B_m) = \frac{1}{|B_m|}\sum_{i\in B_m} \hat{p}_i$. The empirical ECE is the weighted average of their discrepancy across bins:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N_{\text{test}}} \big|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\big|, \tag{1}$$

where a lower ECE indicates better calibration.

Temperature scaling (TS) Guo et al. (2017). TS calibrates a model by dividing its logits by a single scalar temperature $\tau > 0$, learned by minimizing the negative log-likelihood (NLL) on a held-out calibration set D_{cal} and then applied to all test samples. Because dividing the logits by a positive scalar does not change their ordering, TS preserves the predicted label and hence the model's accuracy.

4 PROPOSED METHOD

Motivation. Our method is motivated by an idealized oracle setting in which the correctness of each sample is known. In this setting, assigning a temperature close to zero to correctly classified samples drives their confidence toward one, while assigning a large temperature to misclassified samples drives their confidence toward the floor of $1/C$ (Lemma A.2), yielding near-optimal calibration. Although such an oracle is unavailable in practice, it suggests that calibration can be improved by predicting, for each sample, a temperature that reflects its likelihood of being correct. However, this likelihood can be inferred reliably only when a sample's representation is close to those of samples with the same correctness. When correct and incorrect samples are mixed in the representation space, a per-sample temperature predictor has no consistent signal to learn from. Existing sample-wise methods predict a temperature for every sample regardless of this, so their estimates for such mixed samples are unreliable. TS3 approximates this oracle through three stages that progressively refine the temperature from the global to the sample level. **Stage 1** learns a global base temperature. **Stage 2** learns sample-specific shifts from this base in reliable regions, while keeping shifts near zero in noisy regions. **Stage 3** uses near-zero shifts to identify noisy regions and re-learns a dedicated temperature for them. A visualization of TS3 is shown in Figure 1.

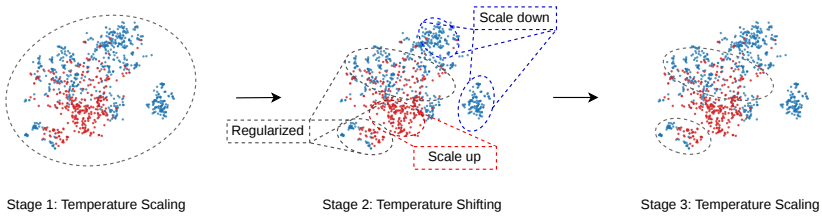


Figure 1: Visualization of TS3, which consists of three stages: temperature scaling on the entire calibration set, temperature shifting, and temperature scaling on uncertain samples. Blue dot represent the feature of correctly classified samples while red dot represent its counterpart.

4.1 Stage 1: Learning the Base Temperature

In Stage 1, we learn a global base temperature τ^* over the entire calibration set using standard TS Guo et al. (2017), i.e., by minimizing the negative log-likelihood (NLL):

$$\tau^* = \arg\min_{\tau>0} \frac{1}{N} \sum_{i=1}^{N} \ell\big(\sigma(t_i/\tau), y_i\big), \tag{2}$$

where ℓ denotes the NLL loss. τ^* serves as the initialization that Stage 2 refines at the sample level.

4.2 Stage 2: Learning the Shift Toward the Optimal Temperature

Let $\psi_\theta : \mathbb{R}^d \to \mathbb{R}$ denote a shift predictor parameterized by θ, which maps the representation h_i of each sample to a temperature shift, yielding the sample-specific temperature $\tau^* + \psi_\theta(h_i)$. To decide where such per-sample shifts can be learned reliably, we measure the local consistency of prediction correctness in the representation space. Let $c_i = \mathbf{1}[y_i = \hat{y}_i]$ denote the correctness of sample z_i, and let $\mathcal{N}_K(i) \subseteq D_{\text{cal}}$ denote the k nearest neighbors of h_i under the Minkowski distance, excluding z_i itself. We define the **agreement purity** of z_i as the fraction of its neighbors that share its correctness:

$$\text{Purity}_i = \frac{1}{K} \sum_{j \in \mathcal{N}_K(i)} \mathbf{1}[c_j = c_i]. \tag{3}$$

A high purity indicates that z_i lies in a **reliable region** where correct and incorrect predictions are well separated, whereas a low purity indicates a **noisy region** where they are mixed. Given a purity threshold β, we define the high-purity set $D_\beta = \{z_i \in D_{\text{cal}} : \text{Purity}_i > \beta\}$ and learn θ by minimizing:

$$\mathcal{L}^{(2)} = \underbrace{\frac{1}{|D_\beta|} \sum_{z_i \in D_\beta} \ell\left(\sigma\left(\frac{t_i}{\tau^* + \psi_\theta(h_i)}\right), y_i\right)}_{\text{temperature shift objective}} + \alpha \underbrace{\frac{1}{|D_{\text{cal}} \setminus D_\beta|} \sum_{z_i \in D_{\text{cal}} \setminus D_\beta} \psi_\theta(h_i)^2}_{\text{uncertainty regularizer}}, \tag{4}$$

where $\alpha, \beta \in [0, 1]$ are hyperparameters. β fits sample-specific temperatures in reliable regions, while α drives the predicted shift toward zero in noisy regions, where individual temperatures cannot be estimated reliably. We predict a shift relative to τ^* rather than the temperature itself, as we empirically find that this yields more stable training. A likely reason is that the optimal temperature varies substantially in scale across tasks and models; the residual form delegates this scale to Stage 1, so ψ_θ only needs to model the smaller deviations around τ^*. In practice, we clamp the predicted shift to ensure that $\tau^* + \psi_\theta(h_i) > 0$.

Importance of the uncertainty regularizer. Prior methods Shen et al. (2024); Tomani et al. (2023) train a neural network to predict a temperature for every sample. They do not distinguish samples in reliable regions from those in noisy regions. In noisy regions, correct and incorrect predictions are mixed, so the network fits noise and generalizes poorly. Our regularizer addresses this by using neighborhood purity during training. It lets ψ_θ learn sample-specific shifts only in reliable regions and keeps the shift near zero in noisy regions. Stage 3 then assigns these noisy samples a single shared temperature.

4.3 Stage 3: Re-calibrating the Noisy Region

After Stage 2, samples in noisy regions receive near-zero shifts and thus retain τ^*. Since this subset differs in distribution from D_{cal}, τ^* is no longer optimal for it. Stage 3 therefore re-learns a dedicated temperature $\bar{\tau}$ on this subset. Which samples to include depends on the number of ECE bins M relative to the number of classes C.

Dead bins when $M \geq C$. By Lemma A.2, a prediction's confidence cannot fall below $1/C$. When $M \geq C$, this floor lies in bin $b_0 = \lceil M/C \rceil$, so all bins below b_0 are unreachable (Proposition 5.5). Once a misclassified sample reaches b_0, a larger positive shift cannot move it to a lower-confidence bin and brings little further gain. We therefore also include positively shifted samples in Stage 3 and re-calibrate them jointly with the unshifted ones. When $M < C$, no such bin exists, so positive shifts remain effective and are kept.

Formally, we learn $\bar{\tau}$ on $D_{\text{sub}} = \{z_i \in D_{\text{cal}} : \psi_\theta(h_i) \in S\}$ by minimizing:

$$\mathcal{L}^{(3)} = \frac{1}{|D_{\text{sub}}|} \sum_{z_i \in D_{\text{sub}}} \ell\left(\sigma\left(\frac{t_i}{\bar{\tau}}\right), y_i\right), \quad \text{where} \quad S = \begin{cases} [-\epsilon, \infty), & \text{if } M \geq C, \\ (-\epsilon, \epsilon), & \text{otherwise,} \end{cases} \tag{5}$$

and ϵ is a small constant ($\epsilon = 10^{-3}$ in practice).

Inference. Given a test input x with representation h, TS3 assigns the temperature

$$\tau_x = \begin{cases} \bar{\tau}, & \text{if } \psi_\theta(h) \in S, \\ \tau^* + \psi_\theta(h), & \text{otherwise.} \end{cases} \tag{6}$$

The pseudo code of TS3 for calibration and inference is provided in Appendix A.2.

5 Theoretical Analysis

We analyze why a single global temperature (Stage 1) is insufficient under heterogeneity, why sample-specific shifts (Stage 2) reduce risk on reliable regions, and why the relation between M and C determines the subset used in Stage 3. Throughout, h denotes the representation of an input x, $\mathcal{R}$ denotes the reliable (high-purity) region, and all risks are measured by the NLL loss ℓ. Proofs and additional analysis are provided in Appendix A.1.

Assumption 5.1 (Regularity of the temperature-scaling risk)**.** *For every sub-population $\mathcal{D}_k \subseteq \mathcal{D}$ with $\pi_k = \Pr(z \in \mathcal{D}_k) > 0$, the risk $L_k(\tau) := \mathbb{E}[\ell(\sigma(t/\tau), y) \mid z \in \mathcal{D}_k]$ is differentiable on $(0, \infty)$ and has a unique minimizer $\tau_k^\star$, with $L_k'(\tau) < 0$ for $\tau < \tau_k^\star$ and $L_k'(\tau) > 0$ for $\tau > \tau_k^\star$.*

This holds whenever L_k has a finite minimizer, since the NLL is convex in $1/\tau$.

5.1 Why Stage 1 Alone Is Insufficient

Proposition 5.2 (Global temperature is a biased compromise)**.** *Let $\mathcal{D} = \bigcup_{k=1}^K \mathcal{D}_k$ be a partition into $K \geq 2$ sub-populations satisfying Assumption 5.1, with $\tau_1^\star, \ldots, \tau_K^\star$ not all equal, and let $\tau^\star = \arg\min_\tau \sum_k \pi_k L_k(\tau)$ be the Stage 1 optimum. Then $\min_k \tau_k^\star < \tau^\star < \max_k \tau_k^\star$. Hence $\tau^\star$ is strictly suboptimal on at least two sub-populations, even at the population level.*

Thus the limitation of Stage 1 is not estimation error: no single temperature is optimal once the population is heterogeneous.

5.2 Why Stage 2 Reduces Risk on the Reliable Region

Assumption 5.3 (Realizability on $\mathcal{R}$)**.** *There exists $\psi^\star : \mathbb{R}^d \to \mathbb{R}$ such that, for every $x \in \mathcal{R}$, $\tau^\star + \psi^\star(h)$ minimizes the conditional risk $\tau \mapsto \mathbb{E}[\ell(\sigma(t/\tau), y) \mid h]$.*

Proposition 5.4 (Global temperature is suboptimal on $\mathcal{R}$)**.** *Suppose Assumption 5.3 holds and, for each $x \in \mathcal{R}$, the conditional risk is m-strongly convex in τ between $\tau^\star$ and $\tau^\star + \psi^\star(h)$. Then*

$$\mathbb{E}_{\mathcal{R}}\left[\ell(\sigma(t/\tau^\star), y)\right] - \mathbb{E}_{\mathcal{R}}\left[\ell(\sigma(t/(\tau^\star + \psi^\star(h))), y)\right] \;\geq\; \frac{m}{2} \operatorname{Var}_{\mathcal{R}}\left[\psi^\star(h)\right].$$

The gain of Stage 2 over Stage 1 therefore scales with the dispersion of the optimal shift across $\mathcal{R}$. It is strictly positive when $\mathcal{R}$ is heterogeneous and vanishes when it is not. If the conditional risk is also smooth, any ψ_θ with $\mathbb{E}_{\mathcal{R}}[(\psi_\theta(h) - \psi^\star(h))^2] \to 0$ attains this gain. Outside $\mathcal{R}$, Assumption 5.3 need not hold, which motivates regularizing the shift toward zero there.

5.3 Why Stage 3 Depends on M and C

Let $c(x) = \max_j \sigma(t/\tau)_j$ denote the confidence under temperature τ.

Proposition 5.5 (Unreachable bins)**.** *Partition $[0, 1]$ into M equal-width ECE bins $\left(\frac{b-1}{M}, \frac{b}{M}\right]$. For any finite τ, $c(x) > 1/C$ (Lemma A.2), so the lowest reachable bin is $b_0 = \lfloor M/C \rfloor + 1$. If $M \geq C$, then $b_0 \geq 2$ and bins $1, \ldots, b_0 - 1$ are unreachable: once a misclassified sample reaches b_0, further increasing τ cannot move it to a lower-confidence bin. If $M < C$, then $b_0 = 1$ and every bin is reachable.*

When $M \geq C$, large positive shifts bring little benefit once a misclassified sample reaches b_0. Instead of relying on individually predicted large shifts, Stage 3 re-calibrates positively shifted samples jointly with the unshifted ones through the single temperature $\bar{\tau}$. When $M < C$, positive shifts can still move misclassified samples to the lowest bin, so they are kept.

6 Experiments

6.1 Experimental Setup

Models: We evaluate TS3 on four instruction-tuned LLMs from two model families, ranging from 7B to 32B parameters: `Qwen2.5-7B-Instruct` and `Qwen2.5-32B-Instruct` Qwen et al. (2025), and `Mistral-7B-Instruct-v0.3` and `Mistral-Small-24B-Instruct-2501` Jiang et al. (2023a); Liu et al. (2026). We focus on post-trained models because post-training substantially worsens calibration Luo et al. (2026) and these models are the ones typically deployed; results on pretrained models are reported in Appendix A.5 Table 5. **Datasets:** We use 50 multiple-choice tasks from three benchmarks: 21 from BIG-Bench Srivastava et al. (2022), 21 from MedMCQA Pal et al. (2022), and 8 from GLUE Wang et al. (2018). For each task, we cap the total number of samples at 5,000 and split them into calibration and test sets at a 1:4 ratio, following Shen et al. (2024). Details are provided in Appendix A.4. **Metric:** We measure calibration using ECE with $M = 10$ equal-width bins.

Baselines: Since evaluating all existing calibration methods is infeasible, we select representative methods from four categories, each shown to achieve strong calibration performance in prior work. *(i) Uncalibrated:* **Vanilla**, the original LLM confidence. *(ii) Prompt-based:* **Elicitation** Tian et al. (2023), which prompts the model to verbalize its confidence. *(iii) Global post-hoc:* Temperature Scaling (**TS**) Guo et al. (2017); Top-versus-All (**TvA**) Le Coz et al. (2024), which reformulates multi-class calibration as a binary problem; and **DACA** Luo et al. (2026), which learns a temperature without labels by aligning the post-trained model with its better-calibrated pretrained counterpart. *(iv) Sample-dependent post-hoc:* Density-Aware Calibration (**DAC**) Tomani et al. (2023), which sets a per-sample temperature from k-NN density over intermediate-layer features; and Thermometer (**Thermo**) Shen et al. (2024), which predicts temperatures with an auxiliary network.

6.2 Main results

Overall performance. As shown in Table 1, TS3 achieves the lowest ECE in all 12 (benchmark, model) settings. The strongest baseline varies across settings: TS in 7 settings, DAC and TvA in 2 each, and Thermometer in 1. TS and DAC represent complementary mechanisms: TS learns a single global temperature, whereas DAC adapts the temperature to each sample. Each is effective in some settings but not uniformly. TS3 combines the two mechanisms, a global base temperature with sample-specific shifts in reliable regions, and outperforms both. Relative to the second-best baseline in each setting, TS3 reduces ECE by 11.9% to 48.5% on BIG-Bench, 14.3% to 35.5% on MedMCQA, and 28.5% to 39.9% on GLUE, with consistent gains across both model families and all model sizes. The improvement persists at the task level (Tables 13–15). With `Qwen-7B-I`, TS3 achieves the best ECE among all methods on 13 of 21 BIG-Bench tasks, 16 of 21 MedMCQA tasks,

Table 1: ECE of calibration methods (lower is better), averaged over 21 BIG-Bench, 21 MedMCQA, and 8 GLUE tasks. Best results are in **bold** and second best are underlined.

		Uncalibrated	Prompt-based	Global post-hoc			Sample-dependent		Ours
		Vanilla	Elicitation	TS	TvA	DACA	DAC	Thermo	TS3
BIG-Bench	Qwen-7B-I	0.3995	0.3299	0.0443	0.0895	0.1487	0.0597	0.0863	**0.0304**
	Qwen-32B-I	0.2874	0.2634	0.0524	0.0776	0.1439	0.0667	0.0628	**0.0270**
	Mistral-7B-I	0.5079	0.5647	0.0698	0.1461	0.1335	0.0846	0.0927	**0.0615**
	Mistral-24B-I	0.2092	0.3104	0.0448	0.0894	0.1740	0.0412	0.0922	**0.0254**
MedMCQA	Qwen-7B-I	0.3214	0.2383	0.0622	0.0914	0.0659	0.0791	0.2182	**0.0401**
	Qwen-32B-I	0.2350	0.1535	0.0403	0.0764	0.0411	0.1010	0.1321	**0.0312**
	Mistral-7B-I	0.4194	0.4579	0.0563	0.0950	0.0591	0.0844	0.2175	**0.0431**
	Mistral-24B-I	0.1389	0.1973	0.0295	0.0600	0.0695	0.0251	0.1412	**0.0215**
GLUE	Qwen-7B-I	0.2097	0.1569	0.0864	0.0712	0.1567	0.0917	0.0969	**0.0428**
	Qwen-32B-I	0.1582	0.1183	0.0498	0.0811	0.1023	0.1238	0.1619	**0.0356**
	Mistral-7B-I	0.3044	0.2836	0.0821	0.0708	0.1668	0.1276	0.0929	**0.0440**
	Mistral-24B-I	0.1406	0.1355	0.0572	0.0559	0.1167	0.0635	0.0488	**0.0325**

and 2 of 8 GLUE tasks, and it ranks in the top two on 49 of the 50 tasks. In contrast, no single baseline is consistently competitive. The best baseline is typically a post-hoc method and varies from task to task. In addition, TS3 adds modest training overhead (under one minute) and negligible inference cost (Appendix A.5.2 7).

These results provide empirical support for our analysis in Section 5, which shows that a single global temperature is insufficient under heterogeneity, and for the region-aware design of TS3 in Section 4. We analyze the contribution of each stage in Section 6.3. The sample-dependent methods DAC and Thermometer are worse than the global TS in 10 and 11 of the 12 settings, respectively. Thermometer is particularly unstable on MedMCQA, with ECE between 0.13 and 0.22. This supports our motivation (Section 4): fitting a temperature predictor on all samples, including those in noisy regions, generalizes poorly. The label-free DACA is also worse than TS in all 12 settings, which suggests that alignment with the pretrained model cannot replace labeled calibration data. Consistent with prior findings Shen et al. (2024); Luo et al. (2026), prompt-based Elicitation performs substantially worse than logit-based post-hoc methods on all three benchmarks, with ECE between 0.12 and 0.56. For Mistral models on BIG-Bench and MedMCQA, it is even worse than the uncalibrated logits.

6.3 Ablation Study

Contribution of each stage. Table 2 reports the ECE after each stage of TS3. Per-task results are provided in Appendix A.5 (Tables 8–12). Stage 3 reduces ECE relative to Stage 1 (i.e., TS) in all settings, by 12.0% to 48.4%. Stage 2 alone does not consistently improve on Stage 1 and increases ECE in 4 of the 8 settings. This is expected: by design, Stage 2 keeps near-zero shifts in noisy regions, so these samples retain τ^*, which is suboptimal for this subset (Section 4). Stage 3 then re-learns a dedicated temperature for them. The gain of TS3 therefore comes from Stages 2 and 3 jointly: Stage 2 calibrates reliable regions and identifies noisy ones, and Stage 3 re-calibrates the noisy regions.

Table 2: ECE after each stage of TS3 and of the variant without the uncertainty regularizer (w/o reg.). Stage 1 is TS. For each benchmark and model: best in **bold**, second best underlined, worst shaded.

		Qwen		Mistral	
		7B	32B	7B	24B
BIG	Stage 1	0.0444	0.0525	0.0699	0.0448
	Stage 2	0.0445	0.0273	0.0616	0.0783
	Stage 3	**0.0305**	**0.0271**	**0.0615**	**0.0255**
	w/o reg.	0.0795	0.0686	0.1141	0.0549
Med	Stage 1	0.0623	0.0403	0.0563	0.0295
	Stage 2	0.0682	0.0455	0.0535	0.0276
	Stage 3	**0.0401**	**0.0313**	**0.0432**	**0.0216**
	w/o reg.	0.0628	0.0434	0.0599	0.0298
GLUE	Stage 1	0.0864	0.0498	0.0821	0.0572
	Stage 2	0.0653	0.0485	0.0791	0.0596
	Stage 3	**0.0428**	**0.0356**	**0.0440**	**0.0325**
	w/o reg.	0.0840	0.0570	0.1080	0.0686

Uncertainty regularizer. We further evaluate a variant of TS3 that omits the uncertainty regularizer in Stage 2 and keeps all other components unchanged. This variant is worse than the full TS3 in all settings, and even worse than Stage 1 (TS). Without the regularizer, Stage 2 fits shifts on all samples, including those in noisy regions. In our experiment, no sample receives a near-zero shift, Stage 3 therefore finds no noisy region to re-calibrate and does nothing. Its Stage 2 alone can occasionally outperform the regularized Stage 2 (e.g., on MedMCQA with Qwen-7B-I, 0.0628 vs. 0.0682). However, it loses the noisy-

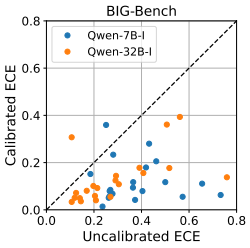


Figure 2: OOD results.

Table 3: ECE under **multi-domain calibration**. Calibration and test data are pooled across all tasks of each benchmark. Best results are in **bold**; second best underlined.

BIG-Bench	TS	DACA	TvA	DAC	Thermo	TS3
`Qwen-7B-I`	0.0748	0.0799	0.0838	<u>0.0732</u>	0.1021	**0.0702**
`Qwen-32B-I`	0.0537	0.0956	0.0782	0.0789	<u>0.0525</u>	**0.0506**
MedMCQA	**TS**	**DACA**	**TvA**	**DAC**	**Thermo**	**TS3**
`Qwen-7B-I`	0.0679	<u>0.0619</u>	0.0864	0.0974	0.0723	**0.0339**
`Qwen-32B-I`	<u>0.0374</u>	0.0453	0.0875	0.1466	0.1019	**0.0180**

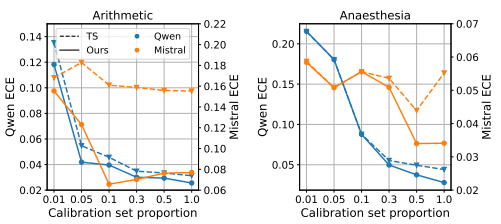


Figure 3: ECE on two sub-tasks: *Arithmetic*–BIG-Bench and *Anaesthesia*–MedMCQA using `Qwen-7B-I` and `Mistral-7B-I` backbones with various calibration set proportion–$\rho \in \{0.01, 0.05, 0.1, 0.3, 0.5, 1.0\}$.

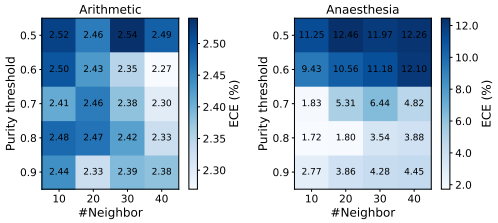


Figure 4: ECE sensitivity regarding the purity threshold–$\beta \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$ and the number of neighbor–$k \in \{10, 20, 30, 40\}$ on 2 tasks: *Arithmetic*–BIG-Bench and *Anaesthesia*–MedMCQA using `Qwen-7B-I` backbone.

region signal that Stage 3 relies on, so the full pipeline still achieves a much lower ECE (0.0401). This mirrors the behavior of existing sample-dependent methods: fitting per-sample temperatures in noisy regions degrades calibration below a single global temperature. The regularizer thus serves two purposes: it prevents overfitting in noisy regions, and it makes near-zero shifts a reliable indicator of these regions for Stage 3.

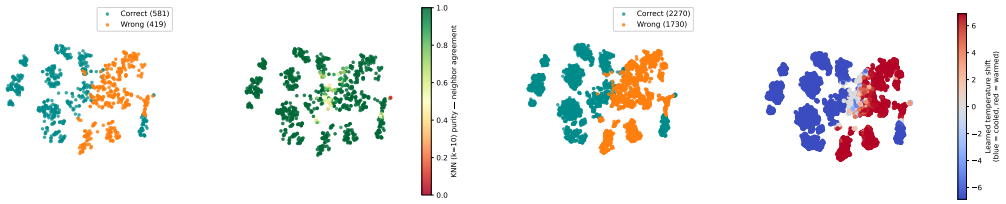
6.4 Multi-Domain & Out-of-Distribution Calibration

Multi-domain calibration. In many real-world applications, the task of each sample is unknown, and only pooled data without task labels is available. We evaluate TS3 in this setting by combining the calibration data from all tasks into a single calibration set and evaluating on the combined test set (Appendix A.4.5). As shown in Tables 3 and 6, TS3 achieves the lowest ECE in all settings. Pooling tasks increases heterogeneity, and a single global temperature becomes less adequate (Proposition 5.2). On BIG-Bench, which spans the most diverse tasks, the sample-dependent DAC or Thermometer outperforms TS with three of the four backbones. TS3 improves further on the best baseline by 3.6% to 9.1% on BIG-Bench and by 28.8% to 51.9% on MedMCQA. These results show that TS3 does not require task-specific calibration and remains effective when tasks are mixed.

Out-of-distribution (OOD) calibration. We evaluate TS3 when no calibration data from the target task is available. For each BIG-Bench task, we train TS3 on the pooled calibration sets of the remaining 20 tasks and evaluate it on the held-out task detailed experimental settings are provided in A.4.4. As shown in Figure 2, TS3 reduces ECE on 20 of the 21 tasks for both models, as indicated by the points below the diagonal in Figure 2. The average ECE drops from 0.3995 to 0.1205 for `Qwen-7B-I` and from 0.2874 to 0.1312 for `Qwen-32B-I`. We attribute this transferability to the systematic overconfidence of post-trained LLMs across tasks: because most tasks require a higher temperature, the adjustments learned on other tasks remain effective on unseen ones.

6.5 Sensitivity Analysis

Calibration set size sensitivity. Figure 3 shows how ECE changes as the calibration set grows from $\rho = 0.01$ to $\rho = 1.0$ of its original size. TS3 is more sensitive to the calibration set size than TS, but in a favorable direction: its ECE generally decreases as more calibration data becomes available, whereas the ECE of TS remains nearly unchanged in several settings. This is expected, since a



(a) Calibration set. (b) Calibration purity. (c) Test set. (d) Test temperature shift.

Figure 5: t-SNE visualizations of last-token representations for the Arithmetic task (BIG-Bench) with `Mistral-7B-I`. **(a)** Calibration set, colored by prediction correctness. **(b)** Calibration set, colored by agreement purity ($k = 10$). **(c)** Test set, colored by prediction correctness. **(d)** Test set, colored by the temperature shift predicted in Stage 2 (blue: negative shift; red: positive shift).

single global temperature has limited capacity to benefit from additional data, while Stage 2 of TS3 estimates purity from local neighborhoods and becomes more reliable as the calibration set grows.

Hyperparameter sensitivity. Figure 4 shows the ECE (%) of TS3 across purity thresholds $\beta \in \{0.5, \ldots, 0.9\}$ and neighborhood sizes $k \in \{10, 20, 30, 40\}$. The purity threshold is the dominant factor. On *Arithmetic*, ECE stays within a narrow range (2.27%–2.54%) over the entire grid. On *Anaesthesia*, ECE is low for $\beta \geq 0.7$ (1.72%–6.44%) but rises sharply for $\beta \leq 0.6$ (9.43%–12.46%). A low threshold admits samples from noisy regions into D_β, so Stage 2 fits shifts where correctness is mixed, which mirrors the effect of removing the uncertainty regularizer (Section 6.3). The neighborhood size k has a secondary effect: for a fixed $\beta \geq 0.7$, smaller neighborhoods generally yield lower ECE (e.g., 1.72% at $k = 10$ vs. 3.88% at $k = 40$ with $\beta = 0.8$), as purity estimated over fewer neighbors reflects more local structure. Overall, TS3 is robust as long as β is set sufficiently high ($\beta \geq 0.7$), and modest tuning within this range can further improve calibration.

6.6 Temperature Shifts

Figure 5 illustrates how TS3 adapts temperatures on the Arithmetic task. In the calibration set, correctly and incorrectly classified samples form largely separated clusters (Figure 5(a)). Purity is high within these clusters and drops in the boundary region where they overlap (Figure 5(b)). The test set exhibits a similar structure (Figure 5(c)), and the shifts predicted by Stage 2 closely follow it (Figure 5(d)). Correctly classified samples receive a mean shift of -6.60, which lowers the base temperature ($\tau^* = 6.87$) to below 1 and sharpens their predicted distributions, increasing confidence. Misclassified samples instead receive positive shifts, which flatten their distributions and reduce overconfidence. In the low-purity boundary region, the predicted shifts stay close to zero, as encouraged by the uncertainty regularizer. Stage 3 learns a dedicated temperature of $\bar{\tau} = 28.84$ for these samples, much higher than the base temperature, indicating that the noisy region is substantially more overconfident than the calibration set as a whole. Overall, TS3 reduces ECE on this task from 0.1553 (TS) to 0.0770. Additional visualizations on other BIG-Bench tasks and backbones are provided in Figure 8, 9. Results on the transferability of learned temperatures across model size are provided in Appendix A.5 Figure 6.

7 Conclusion

Reliable confidence estimates are essential for deploying LLMs, yet post-training often leaves them overconfident. A single global temperature cannot capture sample-level heterogeneity, while predicting a temperature for every sample overfits in regions of the representation space where correct and incorrect predictions are mixed. We introduced TS3, a three-stage post-hoc framework that learns sample-specific temperature shifts only in reliable regions and assigns a dedicated temperature to noisy ones. Our theoretical analysis shows why a global temperature is insufficient under heterogeneity and why region-aware shifts reduce risk. Experiments on 50 tasks with four instruction-tuned LLMs show that TS3 achieves the lowest ECE among all baselines in every setting, and remains effective under multi-domain and out-of-distribution calibration. **Limitations.** TS3 requires access to the model's hidden representations and logits, and therefore cannot be applied directly to closed-source models accessible only through APIs.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for formatting and polishing the paper draft. We did not use generative AI tools for idea generation or theoretical analysis and proof. The remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for debugging and fixing the implementation. We have carefully reviewed all AI-assisted work and revised all sentences and code modified or generated with the assistance of AI. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Sergio A Balanya, Juan Maronas, and Daniel Ramos. Adaptive temperature scaling for robust calibration of deep neural networks. *Neural Computing and Applications*, 36(14):8073–8095, 2024.

Edmon Begoli, Tanmoy Bhattacharya, and Dimitri Kusnezov. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1):20–23, 2019.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

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

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International conference on machine learning*, pp. 1939–1948. PMLR, 2018.

Byeongmoon Ji, Hyemin Jung, Jihyeun Yoon, Kyungyul Kim, et al. Bin-wise temperature scaling (bts): improvement in confidence calibration performance through simple scaling techniques. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 4190–4196. IEEE, 2019.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023a. URL `https://arxiv.org/abs/2310.06825`.

Mingjian Jiang, Yangjun Ruan, Sicong Huang, Saifei Liao, Silviu Pitis, Roger Baker Grosse, and Jimmy Ba. Calibrating language models via augmented prompt ensembles. 2023b.

Tom Joy, Francesco Pinto, Ser-Nam Lim, Philip HS Torr, and Puneet K Dokania. Sample-dependent adaptive temperature scaling for improved calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 14919–14926, 2023.

Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. *Advances in neural information processing systems*, 32, 2019.

Aviral Kumar, Sunita Sarawagi, and Ujjwal Jain. Trainable calibration measures for neural networks from kernel mean embeddings. In *International Conference on Machine Learning*, pp. 2805–2814. PMLR, 2018.

Adrien Le Coz, Stéphane Herbin, and Faouzi Adjed. Confidence calibration of classifiers with many classes. *Advances in Neural Information Processing Systems*, 37:77686–77725, 2024.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

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

Bingyuan Liu, Ismail Ben Ayed, Adrian Galdran, and Jose Dolz. 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*, pp. 80–88, 2022.

Beier Luo, Shuoyuan Wang, Sharon Li, and Hongxin Wei. Your pre-trained llm is secretly an unsupervised confidence calibrator. *Advances in Neural Information Processing Systems*, 38: 163514–163543, 2026.

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

Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Shiyu Ni, Keping Bi, Jiafeng Guo, Minghao Tang, Zengxin Han, Xueqi Cheng, et al. Annotation-efficient honesty alignment via confidence elicitation and calibration. In *International Conference on Learning Representations*, volume 2026, pp. 78415–78436, 2026.

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

Alexandre Perez-Lebel, Marine Le Morvan, and Gaël Varoquaux. Beyond calibration: Estimating the grouping loss of modern neural networks. In *International Conference on Learning Representations (ICLR)*, 2023.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL `https://arxiv.org/abs/2412.15115`.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

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

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in neural information processing systems*, 32, 2019.

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

Christian Tomani, Daniel Cremers, and Florian Buettner. Parameterized temperature scaling for boosting the expressive power in post-hoc uncertainty calibration. In *European conference on computer vision*, pp. 555–569. Springer, 2022.

Christian Tomani, Futa Kai Waseda, Yuesong Shen, and Daniel Cremers. Beyond in-domain scenarios: Robust density-aware calibration. In *International Conference on Machine Learning*, pp. 34344–34368. PMLR, 2023.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pp. 353–355, 2018.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 13484–13508, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Johnathan Xie, Annie S Chen, Yoonho Lee, Eric Mitchell, and Chelsea Finn. Calibrating language models with adaptive temperature scaling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18128–18138, 2024.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *International Conference on Learning Representations*, volume 2024, pp. 23650–23678, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, volume 1, pp. 2001, 2001.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 694–699, 2002.

A Appendix

A.1 Additional Theoretical Analysis and Proofs

Throughout, $t \in \mathbb{R}^C$ denotes the logit vector of a sample, $t_{(1)} \geq t_{(2)} \geq \cdots \geq t_{(C)}$ its sorted entries, and $c(x) = \max_j \sigma(t/\tau)_j$ its confidence under temperature τ. ECE bins are $B_b = \big(\frac{b-1}{M}, \frac{b}{M}\big]$, $b = 1, \ldots, M$.

Theorem A.1 (Temperature needed to reach the lowest bin)**.** *Assume $C > M \geq 2$ and let $\Delta = t_{(1)} - t_{(2)} > 0$. If $c(x) \in B_1$, i.e., $c(x) \leq 1/M$, then*

$$\tau \;\geq\; \tilde{\tau} := \frac{\Delta}{\ln\big(\frac{C-1}{M-1}\big)},$$

with equality attainable when $t_{(2)} = \cdots = t_{(C)}$.

Proof. Since $t_{(j)} \leq t_{(2)}$ for all $j \geq 2$,

$$c(x) = \frac{e^{t_{(1)}/\tau}}{\sum_{j=1}^{C} e^{t_{(j)}/\tau}} \;\geq\; \frac{e^{t_{(1)}/\tau}}{e^{t_{(1)}/\tau} + (C-1)e^{t_{(2)}/\tau}} = \frac{1}{1 + (C-1)e^{-\Delta/\tau}}.$$

Hence $c(x) \leq 1/M$ requires $1 + (C-1)e^{-\Delta/\tau} \geq M$, i.e., $e^{-\Delta/\tau} \geq \frac{M-1}{C-1}$. Since $C > M$, the right-hand side lies in $(0, 1)$, and taking logarithms gives $\tau \geq \Delta / \ln\big(\frac{C-1}{M-1}\big)$. The bound is tight when all non-top logits are equal, since the inequality above then holds with equality. □

Lemma A.2 (Confidence floor)**.** *Let t have finite entries that are not all equal. Then $c(x)$ is strictly decreasing in τ on $(0, \infty)$, $c(x) > 1/C$ for every finite $\tau > 0$, and $c(x) \to 1/C$ as $\tau \to \infty$.*

Proof. Dividing by $\tau > 0$ preserves the ordering of logits, so the top class is fixed and $\log c(x) = t_{(1)}/\tau - \log \sum_j e^{t_j/\tau}$. Differentiating,

$$\frac{d}{d\tau} \log c(x) = -\frac{1}{\tau^2}\Big(t_{(1)} - \textstyle\sum_j \sigma(t/\tau)_j \, t_j\Big) < 0,$$

since $t_{(1)}$ strictly exceeds the softmax-weighted mean of t when the logits are not all equal. Moreover, $c(x) = \max_j \sigma(t/\tau)_j > \frac{1}{C} \sum_j \sigma(t/\tau)_j = 1/C$, as the maximum of non-identical entries exceeds their average. Finally, $t_j/\tau \to 0$ for every j, so $\sigma(t/\tau)_j \to 1/C$. □

A.1.1 Proof of Proposition 5.2

Proof. By the first-order condition, $\sum_k \pi_k L_k'(\tau^\star) = 0$. Suppose for contradiction that $\tau^\star \leq \min_k \tau_k^\star$. By Assumption 5.1, $L_k'(\tau^\star) \leq 0$ for every k, and $L_k'(\tau^\star) < 0$ for any k with $\tau_k^\star > \tau^\star$. Such a k exists because the $\tau_k^\star$ are not all equal. Since $\pi_k > 0$, $\sum_k \pi_k L_k'(\tau^\star) < 0$, a contradiction. The symmetric argument rules out $\tau^\star \geq \max_k \tau_k^\star$. Hence $\tau^\star$ lies strictly between the smallest and largest $\tau_k^\star$, and so differs from both. □

A.1.2 Proof of Proposition 5.4

Proof. For $x \in \mathcal{R}$, let $G_h(\tau) = \mathbb{E}[\ell(\sigma(t/\tau), y) \mid h]$ denote the conditional risk. By Assumption 5.3, G_h is minimized at $\tau^\star + \psi^\star(h)$, so $G_h'(\tau^\star + \psi^\star(h)) = 0$. By m-strong convexity of G_h between $\tau^\star$ and $\tau^\star + \psi^\star(h)$,

$$G_h(\tau^\star) - G_h\big(\tau^\star + \psi^\star(h)\big) \;\geq\; \frac{m}{2}\, \psi^\star(h)^2.$$

Taking expectations over $x \in \mathcal{R}$ and applying the tower property,

$$\mathbb{E}_{\mathcal{R}}\big[\ell(\sigma(t/\tau^\star), y)\big] - \mathbb{E}_{\mathcal{R}}\big[\ell(\sigma(t/(\tau^\star + \psi^\star(h))), y)\big] \;\geq\; \frac{m}{2}\, \mathbb{E}_{\mathcal{R}}\big[\psi^\star(h)^2\big] \;\geq\; \frac{m}{2}\, \mathrm{Var}_{\mathcal{R}}\big[\psi^\star(h)\big],$$

where the last step uses $\mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \leq \mathbb{E}[X^2]$. □

A.1.3 PROOF OF PROPOSITION 5.5

Proof. A confidence $c \in (0, 1]$ falls in bin $\lceil Mc \rceil$. By Lemma A.2, $c(x) > 1/C$, so $Mc(x) > M/C$ and the bin index satisfies $\lceil Mc(x) \rceil \geq \lfloor M/C \rfloor + 1 = b_0$. This bound is attained: $c(x)$ decreases continuously to $1/C$ as $\tau \to \infty$, so for sufficiently large finite τ, $Mc(x) \in \big(M/C,\ \lfloor M/C \rfloor + 1\big]$, i.e., the sample lies in B_{b_0}. Hence b_0 is the lowest reachable bin, and bins $1, \ldots, b_0 - 1$ are never populated for any τ.

If $M \geq C$, then $M/C \geq 1$ and $b_0 \geq 2$, so these unreachable bins are non-empty. Since $c(x)$ is strictly decreasing in τ, its bin index is non-increasing, so once a sample reaches B_{b_0}, it remains there for all larger τ. If $M < C$, then $M/C < 1$ and $b_0 = 1$, so every bin is reachable. The minimum temperature required to reach B_1 is characterized by Theorem A.1. □

A.2 TS3 PSEUDO CODE

Algorithm 1 TS3 training

Require: Calibration dataset $D_{cal} = \{z_i = (x_i, y_i)\}_{i=1}^{N}$; hyper-parameters: ϵ, β, k; initialized: τ, θ, hidden feature h, temperature shift regressor ψ_θ parameterized by θ.

```
Learn τ* via Eq 2                                              ▷ Stage 1.
Initialize D_β = ∅
for i = 1 to N do
    Calculate Purity_i as Eq 3
    if Purity_i > β then
        D_β = D_β ∩ {z_i}
    end if
end for
Learn θ by optimizing the loss function 4                      ▷ Stage 2.
Learn τ̄ following Eq 5                                         ▷ Stage 3.
return τ*, θ, τ̄
```

Algorithm 2 TS3 inference ($M < C$)

Require: A single data point $x_i \sim \mathcal{D}$, learned $\tau^*, \theta, \bar{\tau}$ and a hyperparameter ϵ

```
Compute temperature shift: ψ_θ(h_i)
if |ψ_θ(h_i)| < ε then          ▷ Replace this with ψ_θ(h_i) > −ε whenever M ≥ C.
    τ_i = τ̄                    ▷ Small shift ⇒ noisy region data point return Stage 3's temperature.
else
    τ_i = τ* + ψ_θ(h_i)
end if
return τ_i                      ▷ Return specific temperature for each input sample.
```

Noted that purity is just used in training due to we do not see all test data points in inference time.

A.3 IMPLEMENTATION DETAILS

All methods use a MLP network of 2 hidden layers with 256 and 128 neurons with ReLU activation. For Temperature scaling we run for 500 epochs and for the temperature shift with learning rate in $\{5e-3, 1e-3, 5e-4, 1e-4\}$ and a weight decay of $1e-6$ we train the shift predictor until the mean shift of all (negative shift points) less than 1.

For **Thermometer** Shen et al. (2024) we set the test batch size equals to 1 for fair comparison with other baselines where additional test data points are sampled independently.

For **DAC** Tomani et al. (2023) we set the number of neighbor equal to 10.

For **DACA** Luo et al. (2026) we tune the learning rate $\in \{0.001, 0.01, 0.05\}$ and set the maximum epoch of 500.

A.4 DETAILED EXPERIMENTAL SETUPS

A.4.1 DATASET

BIG-Bench : We conduct various multiple choice question answering tasks in BIG-Bench benchmark including: [arithmetic, bbq_lite_json, color, elementary_math_qa, epistemic_reasoning, fact_checker, formal_fallacies_syllogisms_negation, goal_step_wikihow, hyperbaton, logical_fallacy_detection, mnist_ascii, movie_dialog_same_or_different, play_dialog_same_or_different, real_or_fake_text, social_iqa, strategyqa, timedial, tracking_shuffled_objects, vitaminc_fact_verification, unit_conversion, winowhy]. Calibtation and test set are partitioned from BIG-Bench **train** set.

MedMCQA: We evaluate our approach using all MedMCQA sub-tasks: [Anaesthesia Anatomy, Biochemistry, Dental, ENT, Forensic Medicine, Gynaecology & Obstetrics, Medicine, Microbiology, Ophthalmology, Orthopaedics, Pathology, Pediatrics, Pharmacology, Physiology, Psychiatry, Radiology, Skin, Social & Preventive Medicine , Surgery, Unknown]. Calibtation and test set are partitioned from MedMCQA **train** set.

GLUE We also include several multiple choice QA tasks including: [cola, sst2, mrpc, stsb, qqp, mnli, qnli, rte]. Calibtation and test set are partitioned from GLUE **validation** set.

A.4.2 EXAMPLE PROMPT FORMAT

MedMCQA prompt format: "You are a helpful assistant answering multiple-choice medical questions. Respond with only the letter of the correct answer. $\backslash n$Question: `<question>`. $\backslash n$A. `<opa>` $\backslash n$B. `<opb>` $\backslash n$C. `<opc>` $\backslash n$D. `<opd>` $\backslash n$Answer:"

BIG-Bench and GLUE prompt format: "You are a helpful assistant answering multiple-choice questions. Respond with only the letter of the correct answer. Task: `<task>`. $\backslash n$Question: `<question>` $\backslash n$A. `<opa>` $\backslash n$B. `<opb>` $\backslash n$C. `<opc>` $\backslash n$D. `<opd>` $\backslash n$Answer:"

Elicitation prompt format : "You are a helpful assistant answering multiple-choice questions. For each question you must: $\backslash n1$. State the letter of your chosen answer. $\backslash n2$. State your confidence as an integer percentage (0–100) that your chosen answer is correct. $\backslash n$Format your response EXACTLY as: Answer: `<LETTER>` Confidence: `<NUMBER>` $\backslash n$Do not include any other text. $\backslash n$Task: `<task>`. $\backslash n$Question: `<question>` $\backslash n$A. `<opa>` $\backslash n$B. `<opb>` $\backslash n$C. `<opc>` $\backslash n$D. `<opd>`"

A.4.3 DATA STATISTICS

Table 4: Statistics of the three benchmarks.

BIG-Bench		MedMCQA		GLUE	
Tasks	**Total samples**	**Tasks**	**Total samples**	**Tasks**	**Total samples**
arithmetic	5000	Anaesthesia	3172	cola	1043
bbq_lite_json	5000	Anatomy	5000	sst2	872
color	3200	Biochemistry	5000	mrpc	408
elementary_math_qa	5000	Dental	5000	stsb	1500
epistemic_reasoning	1600	ENT	4919	qqp	5000
fact_checker	5000	Forensic Medicine	5000	mnli	5000
formal_fallacies_syllogisms_negation	5000	Gynaecology & Obstetrics	5000	qnli	5000
goal_step_wikihow	5000	Medicine	5000	rte	277
hyperbaton	5000	Microbiology	5000		
logical_fallacy_detection	2240	Ophthalmology	5000		
mnist_ascii	5000	Orthopaedics	2999		
movie_dialog_same_or_different	5000	Pathology	5000		
play_dialog_same_or_different	2612	Pediatrics	5000		
real_or_fake_text	5000	Pharmacology	5000		
social_iqa	1548	Physiology	5000		
strategyqa	1832	Psychiatry	4442		
timedial	2040	Radiology	4395		
tracking_shuffled_objects	3000	Skin	1771		
vitaminc_fact_verification	5000	Social & Preventive Medicine	5000		
unit_conversion	5000	Surgery	5000		
winowhy	2290	Unknown	3045		

Table 4 shows the statistics of the benchmarks. The calibration and test sets are randomly split at a 1:4 ratio.

A.4.4 Out-of-distribution setup

For each benchmark, we evaluate each sub-task as the target task by training TS3 exclusively on calibration data from all other sub-tasks in the benchmark (Leave-one-out). Specifically, for a given target sub-task, we combine the calibration sets from all remaining sub-tasks, apply the TS3 training procedure on this combined calibration set, and then evaluate the resulting model on the target task's test set. This setup evaluates the ability of TS3 to generalize its temperature prediction capability to an unseen task.

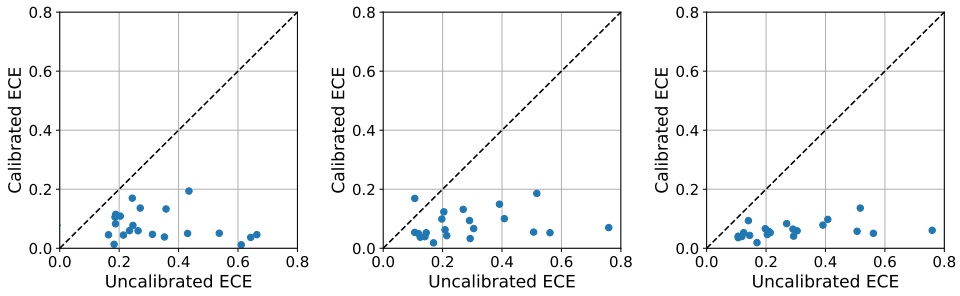
A.4.5 Multi-domain calibration setup

For each task in the benchmark, we uniformly sample up to 200 examples from its original extracted calibration set and combine the sampled data across all tasks into a single unified calibration set. The corresponding test sets are kept unchanged and combined in the same manner for evaluation. We exclude **Vanilla** and **Elicitation** from this experiment because their results are already reported in the main table (Table 1), and this unified training setup does not affect these baselines due to their calibration-set-independent properties.

A.5 Additional experiments

A.5.1 Temperature transferability on model size

We investigate whether temperatures learned from a smaller model can be transferred to larger models. To this end, we study temperature transferability within a model family, where models of different sizes are potentially trained using the same pipeline and may therefore share common representations and knowledge. Consequently, their predictive uncertainty may exhibit similar characteristics. Specifically, we first learn the optimal temperature on the test set using one model as the backbone. We then transfer the learned temperature to another model and evaluate its calibration error on the same test set without learning a new temperature. As shown in Figure 6 reusing the temperature substantially reduces the ECE compared with the uncalibrated models, suggesting that models of different sizes within the same family may exhibit similar predictive uncertainty. In particular, using the temperature learned from the 7B model results in an average ECE of 0.0711 for the 14B model and 0.0800 for the 32B model. Just 1 task is miscalibrated compare to the original model when transferring temperatures from 7B to 32B versions, all other tasks across different model size get improved.



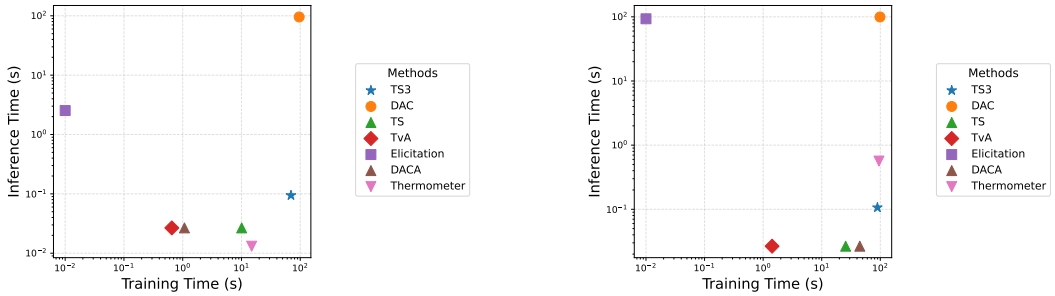
(a) `Qwen: 7B→14B`. (b) `Qwen: 7B→32B`. (c) `Qwen: 14B→32B`.

Figure 6: *Temperature transferability* across Qwen models on BIG-Bench. Temperatures learned from the source model are applied to the target model's output probabilities.

A.5.2 Computational cost

Computational cost. We compare the computational cost of TS3 with the baselines on the BIG-Bench Arithmetic task, using 1,000 calibration and 4,000 test samples and averaging over three runs

(Fig. 7). TS3 takes about 10× longer to train than TS. Because the calibration set is small, however, all methods finish training within two minutes. At inference, TS3 computes temperatures for all test samples in about 0.1 seconds, comparable to the other calibration methods.



(a) `Qwen-7B-I`. (b) `Qwen-32B-I`.

Figure 7: Computational cost of each calibration method, excluding the forward-pass time. Results are averaged over three runs on Arithmetic with a calibration set of 1,000 samples and a test set of 4,000 samples.

A.5.3 Additional Empirical Results

Table 5: ECE comparison across various calibration methods; lower ECE indicates better calibration. Results are averaged over 21 BIG-Bench tasks, 21 MedMCQA tasks, and 8 GLUE tasks. Best results are in **bold**, second best are underlined.

		Vanilla	TS	TvA	DAC	Thermo	TS3
BIG-Bench	`Qwen-7B`	0.3086	0.0764	0.1874	0.0816	0.0781	**0.0669**
	`Mistral-7B`	0.1695	0.0594	0.0838	0.0641	**0.0499**	0.0566
MedMCQA	`Qwen-7B`	0.1178	0.0277	0.1176	0.0287	0.1369	**0.0232**
	`Mistral-7B`	0.0300	0.0199	0.0318	0.0335	0.1619	**0.0191**

Table 6: ECE comparison across various calibration methods under **multi-domain calibration** settings.

		TS	DACA	TvA	DAC	Thermo	TS3
BIG-Bench	`Mistral-7B-I`	0.0745	0.0899	0.1264	0.0767	0.0701	0.0647
	`Mistral-24B-I`	0.0516	0.1408	0.0769	–	–	0.0469
MedMCQA	`Mistral-7B-I`	0.0594	0.0799	0.0905	0.0817	0.068	0.0413
	`Mistral-24B-I`	0.0260	0.0577	0.0597	–	–	0.0185

Table 7: ECE with varied calibration set proportions – ρ on 1 sub-tasks for each benchmarks.

		ρ = 0.1		ρ = 0.3		ρ = 0.5		ρ = 1.0	
		TS	TS3	TS	TS3	TS	TS3	TS	TS3
Arithmetic	Qwen-7B-I	0.0457	0.0396	0.0348	0.0301	0.0334	0.0294	0.0313	0.0256
	Mistral-7B-I	0.1610	0.0657	0.1588	0.0703	0.1559	0.0763	0.1553	0.0770
Anaesthesia	Qwen-7B-I	0.0878	0.0877	0.0553	0.0497	0.0494	0.0375	0.0441	0.0279
	Mistral-7B-I	0.0556	0.0556	0.0536	0.0509	0.0439	0.0340	0.0552	0.0341
QQP	Qwen-7B-I	0.1298	0.0602	0.1299	0.0573	0.1315	0.0446	0.1223	0.0328
	Mistral-7B-I	0.0603	0.0457	0.0623	0.0122	0.0623	0.0249	0.0603	0.0316

Table 8: TS3 ECE on BIG-Bench using Qwen family models as backbones.

ECE							
Qwen/Qwen2.5-7B-Instruct	TS	**stage 2**	**stage 3**	Qwen/Qwen2.5-32B-Instruct	TS	**stage 2**	**stage 3**
arithmetic	0.0313	0.0232	0.0239	arithmetic	0.0451	0.0588	0.0584
bbq_lite_json	0.0255	0.0320	0.0223	bbq_lite_json	0.0685	0.0039	0.0039
color	0.0404	0.0368	0.0366	color	0.0796	0.0642	0.0634
elementary_math_qa	0.0433	0.0417	0.0417	elementary_math_qa	0.0651	0.0184	0.0184
epistemic_reasoning	0.0278	0.0219	0.0219	epistemic_reasoning	0.0397	0.0004	0.0004
fact_checker	0.0643	0.0025	0.0025	fact_checker	0.0175	0.0016	0.0016
formal_fallacies_syllogisms_negation	0.0585	0.0564	0.0564	formal_fallacies_syllogisms_negation	0.0745	0.0025	0.0025
goal_step_wikihow	0.0470	0.0256	0.0256	goal_step_wikihow	0.0416	0.0030	0.0030
hyperbaton	0.0645	0.0022	0.0022	hyperbaton	0.0104	0.0027	0.0027
logical_fallacy_detection	0.0741	0.0740	0.0740	logical_fallacy_detection	0.0356	0.0036	0.0036
mnist_ascii	0.0453	0.0242	0.0242	mnist_ascii	0.0477	0.0686	0.0686
movie_dialog_same_or_different	0.0632	0.0690	0.0604	movie_dialog_same_or_different	0.0802	0.0015	0.0015
play_dialog_same_or_different	0.0094	0.0165	0.0054	play_dialog_same_or_different	0.0791	0.0539	0.0535
real_or_fake_text	0.0161	0.0141	0.0141	real_or_fake_text	0.0791	0.0781	0.0781
social_iqa	0.0238	0.0352	0.0242	social_iqa	0.0900	0.0052	0.0052
strategyqa	0.0548	0.1792	0.0532	strategyqa	0.0900	0.0576	0.0559
timedial	0.0581	0.0581	0.0581	timedial	0.0433	0.0448	0.0452
tracking_shuffled_objects	0.0435	0.0415	0.0415	tracking_shuffled_objects	0.0371	0.0506	0.0506
vitaminc_fact_verification	0.0852	0.1024	0.0137	vitaminc_fact_verification	0.0455	0.0401	0.0392
unit_conversion	0.0287	0.0437	0.0173	unit_conversion	0.0121	0.0076	0.0076
winowhy	0.0270	0.0335	0.0211	winowhy	0.0206	0.0052	0.0052
AVG	0.0444	0.0445	0.0305	AVG	0.0525	0.0273	0.0271

Table 9: TS3 ECE on BIG-Bench using Mistral family models as backbones.

ECE							
Mistral-7B-Instruct	TS	**stage 2**	**stage 3**	Mistral-24B-Instruct	TS	**stage 2**	**stage 3**
arithmetic	0.1553	0.0783	0.0770	arithmetic	0.0248	0.0307	0.0307
bbq_lite_json	0.0228	0.0228	0.0228	bbq_lite_json	0.0414	0.0170	0.0170
color	0.0480	0.0310	0.0310	color	0.0810	0.3711	0.0489
elementary_math_qa	0.0377	0.0329	0.0329	elementary_math_qa	0.0242	0.0385	0.0239
epistemic_reasoning	0.0388	0.0364	0.0363	epistemic_reasoning	0.0216	0.0203	0.0203
fact_checker	0.0471	0.0471	0.0471	fact_checker	0.0376	0.0100	0.0100
formal_fallacies_syllogisms_negation	0.0593	0.0567	0.0567	formal_fallacies_syllogisms_negation	0.0458	0.1323	0.0310
goal_step_wikihow	0.0990	0.0878	0.0878	goal_step_wikihow	0.0221	0.0078	0.0078
hyperbaton	0.1648	0.1496	0.1496	hyperbaton	0.0158	0.0140	0.0140
logical_fallacy_detection	0.0683	0.0546	0.0546	logical_fallacy_detection	0.1492	0.0156	0.0156
mnist_ascii	0.0597	0.0563	0.0563	mnist_ascii	0.0107	0.0086	0.0060
movie_dialog_same_or_different	0.0558	0.0539	0.0539	movie_dialog_same_or_different	0.0930	0.3092	0.0169
play_dialog_same_or_different	0.1711	0.1689	0.1689	play_dialog_same_or_different	0.0106	0.0095	0.0106
real_or_fake_text	0.0384	0.0368	0.0368	real_or_fake_text	0.0313	0.0407	0.0407
social_iqa	0.0514	0.0494	0.0524	social_iqa	0.0411	0.0198	0.0198
strategyqa	0.0352	0.0272	0.0272	strategyqa	0.0756	0.0673	0.0744
timedial	0.0668	0.0668	0.0668	timedial	0.0337	0.0205	0.0205
tracking_shuffled_objects	0.0579	0.0558	0.0558	tracking_shuffled_objects	0.0458	0.1574	0.0425
vitaminc_fact_verification	0.0946	0.0911	0.0864	vitaminc_fact_verification	0.0322	0.0132	0.0132
unit_conversion	0.0139	0.0127	0.0127	unit_conversion	0.0529	0.3175	0.0480
winowhy	0.0814	0.0779	0.0779	winowhy	0.0509	0.0235	0.0235
AVG	0.0699	0.0616	0.0615	AVG	0.0448	0.0783	0.0255

Table 10: TS3 ECE on MedMCQA benchmark using Qwen family models as backbones

	ECE						
Qwen/Qwen2.5-7B-Instruct	TS	**stage 2**	**stage 3**	Qwen/Qwen2.5-32B-Instruct	TS	**stage 2**	**stage 3**
Anaesthesia	0.0441	0.0385	0.0177	Anaesthesia	0.0478	0.0501	0.0382
Anatomy	0.0692	0.0509	0.0256	Anatomy	0.0347	0.0462	0.0291
Biochemistry	0.0585	0.0775	0.0463	Biochemistry	0.0299	0.0423	0.0254
Dental	0.0593	0.05	0.0373	Dental	0.0381	0.0376	0.0275
ENT	0.0621	0.0677	0.047	ENT	0.047	0.0524	0.0267
Forensic Medicine	0.0798	0.0874	0.0486	Forensic Medicine	0.0594	0.0717	0.0505
Gynaecology & Obstetrics	0.051	0.0552	0.0403	Gynaecology & Obstetrics	0.0461	0.0553	0.0286
Medicine	0.0523	0.0459	0.0354	Medicine	0.0417	0.0459	0.0386
Microbiology	0.0714	0.068	0.0376	Microbiology	0.0391	0.0514	0.0337
Ophthalmology	0.0621	0.071	0.022	Ophthalmology	0.0395	0.042	0.0328
Orthopaedics	0.064	0.0726	0.0474	Orthopaedics	0.0484	0.0578	0.047
Pathology	0.0684	0.0702	0.0477	Pathology	0.0436	0.0478	0.0315
Pediatrics	0.0676	0.0727	0.0424	Pediatrics	0.0395	0.0535	0.0384
Pharmacology	0.0639	0.0685	0.0589	Pharmacology	0.0309	0.0084	0.0084
Physiology	0.0732	0.0911	0.0559	Physiology	0.0337	0.0406	0.0272
Psychiatry	0.0617	0.0792	0.0386	Psychiatry	0.0323	0.0545	0.04
Radiology	0.0665	0.0639	0.0253	Radiology	0.0398	0.0446	0.03
Skin	0.0635	0.0911	0.0535	Skin	0.0331	0.0092	0.0092
Social & Preventive Medicine	0.0619	0.086	0.0471	Social & Preventive Medicine	0.0303	0.0365	0.0217
Surgery	0.0529	0.0646	0.032	Surgery	0.0414	0.0404	0.0256
Unknown	0.0544	0.0612	0.0356	Unknown	0.0501	0.0667	0.0462
AVG	0.0623	0.0682	0.0401	AVG	0.0403	0.0455	0.0313

Table 11: TS3 ECE on MedMCQA benchmark using Mistral family models as backbones

	ECE						
Mistral-7B-Instruct	TS	**stage 2**	**stage 3**	Mistral-24B-Instruct	TS	**stage 2**	**stage 3**
Anaesthesia	0.0552	0.0481	0.0318	Anaesthesia	0.0301	0.0224	0.0201
Anatomy	0.0486	0.0484	0.0508	Anatomy	0.0308	0.0167	0.0132
Biochemistry	0.0639	0.0702	0.0588	Biochemistry	0.028	0.0449	0.0413
Dental	0.0457	0.0449	0.0427	Dental	0.0277	0.0228	0.0197
ENT	0.0586	0.0477	0.0381	ENT	0.0353	0.0235	0.0209
Forensic Medicine	0.0558	0.0554	0.0504	Forensic Medicine	0.0367	0.034	0.0226
Gynaecology & Obstetrics	0.0362	0.0368	0.0375	Gynaecology & Obstetrics	0.0294	0.0223	0.0123
Medicine	0.0669	0.063	0.0536	Medicine	0.032	0.0164	0.0204
Microbiology	0.0513	0.043	0.0197	Microbiology	0.0306	0.0427	0.0305
Ophthalmology	0.0459	0.0431	0.0318	Ophthalmology	0.0294	0.0321	0.0152
Orthopaedics	0.0705	0.0594	0.0506	Orthopaedics	0.0257	0.0172	0.0168
Pathology	0.0592	0.0545	0.0237	Pathology	0.0246	0.0292	0.0203
Pediatrics	0.0612	0.0559	0.0526	Pediatrics	0.0302	0.0342	0.0263
Pharmacology	0.0698	0.0643	0.0524	Pharmacology	0.0236	0.0349	0.0227
Physiology	0.0613	0.056	0.0477	Physiology	0.0258	0.0348	0.0291
Psychiatry	0.0699	0.0596	0.043	Psychiatry	0.0291	0.0153	0.0153
Radiology	0.0578	0.0645	0.0425	Radiology	0.0316	0.0295	0.0177
Skin	0.0549	0.0617	0.0547	Skin	0.0241	0.0236	0.0257
Social & Preventive Medicine	0.06	0.0636	0.0514	Social & Preventive Medicine	0.0308	0.0224	0.0138
Surgery	0.0454	0.0374	0.0325	Surgery	0.0296	0.0224	0.0217
Unknown	0.0451	0.0457	0.04	Unknown	0.0348	0.0389	0.0278

Table 12: TS3 ECE on GLUE benchmark.

ECE							
Qwen/Qwen2.5-7B-Instruct	TS	**stage 2**	**stage 3**	Qwen/Qwen2.5-32B-Instruct	TS	**stage 2**	**stage 3**
cola	0.0672	0.0667	0.0459	cola	0.088	0.0865	0.0647
sst2	0.0332	0.042	0.0213	sst2	0.0128	0.0095	0.0289
mrpc	0.0806	0.0825	0.0423	mrpc	0.0512	0.0476	0.0247
stsb	0.1139	0.1057	0.0853	stsb	0.063	0.0887	0.0598
qqp	0.1223	0.0858	0.0328	qqp	0.0557	0.053	0.0386
mnli	0.0494	0.0449	0.0194	mnli	0.0263	0.0318	0.0091
qnli	0.1064	0.051	0.051	qnli	0.0355	0.0225	0.0215
rte	0.1188	0.0445	0.0445	rte	0.0659	0.048	0.0377
AVG	0.086475	0.0653875	0.0428125	AVG	0.0498	0.0485	0.0356
ECE							
Mistral-7B-Instruct	TS	**stage 2**	**stage 3**	Mistral-24B-Instruct	TS	**stage 2**	**stage 3**
cola	0.0877	0.0867	0.0175	cola	0.0456	0.0348	0.0274
sst2	0.0433	0.0226	0.0271	sst2	0.0298	0.017	0.017
mrpc	0.0969	0.1294	0.0657	mrpc	0.0849	0.1499	0.0513
stsb	0.0902	0.0871	0.051	stsb	0.0504	0.0577	0.0421
qqp	0.0603	0.0566	0.0316	qqp	0.1364	0.1222	0.067
mnli	0.0425	0.05	0.036	mnli	0.0254	0.0346	0.0099
qnli	0.1297	0.1312	0.0538	qnli	0.0196	0.0255	0.0176
rte	0.1062	0.0699	0.0699	rte	0.0654	0.0352	0.0277
AVG	0.0821	0.0791875	0.044075	AVG	0.0572	0.0596	0.0325

Table 13: ECE of calibration methods (lower is better) on 21 BIG-Bench tasks using the `Qwen-7B-Instruct` backbone. Best results are in **bold** and second best are underlined.

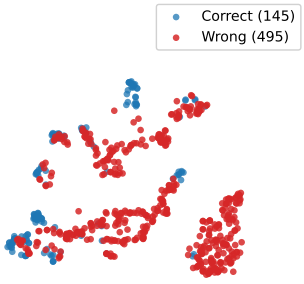
	Uncalibrated	Prompt-based	Global post-hoc			Sample-dependent		Ours
	Vanilla	Elicitation	TS	TvA	DACA	DAC	Thermo	TS3
arithmetic	0.2510	0.3836	0.0313	0.0511	0.0939	0.0520	0.0820	**0.0239**
bbq_lite_json	0.2690	0.0938	0.0255	0.0503	0.2349	0.0229	0.1400	**0.0223**
color	0.5000	0.4423	0.0404	0.1032	0.1286	0.0404	0.0710	**0.0366**
elementary_math_qa	0.5730	0.6481	0.0433	0.2299	0.1672	**0.0384**	0.0710	0.0417
epistemic_reasoning	0.2360	0.1568	0.0278	0.0445	0.1101	0.0519	0.0764	**0.0219**
fact_checker	0.2660	0.2404	0.0643	0.0087	0.0671	0.0795	0.0760	**0.0025**
formal_fallacies_syllogisms_negation	0.4620	0.3547	0.0585	**0.0531**	0.2746	0.1397	0.0808	0.0564
goal_step_wikihow	0.1850	0.0282	0.0470	**0.0176**	0.1159	0.0311	0.0420	0.0256
hyperbaton	0.2620	0.1313	0.0645	0.0158	0.0506	0.0473	0.0760	**0.0022**
logical_fallacy_detection	0.4320	0.3047	0.0741	**0.0590**	0.0634	0.1411	0.2240	0.0740
mnist_ascii	0.8180	0.7476	0.0453	0.4172	0.3843	0.0591	**0.0070**	0.0242
movie_dialog_same_or_different	0.4190	0.3012	0.0632	0.0611	0.3006	0.0811	0.1380	**0.0604**
play_dialog_same_or_different	0.3640	0.2328	0.0094	0.0097	0.1952	0.0163	0.0800	**0.0054**
real_or_fake_text	0.7330	0.7101	0.0161	0.3747	0.1586	0.0189	0.0470	**0.0141**
social_iqa	0.2650	0.0910	**0.0238**	0.0421	0.0907	0.0377	0.0600	0.0242
strategyqa	0.2780	0.1893	0.0548	0.0809	0.0860	0.0612	**0.0510**	0.0532
timedial	0.3730	0.2694	**0.0581**	0.0688	0.1240	0.1199	0.1560	**0.0581**
tracking_shuffled_objects	0.6550	0.7765	0.0435	0.0435	0.0435	0.0618	**0.0240**	0.0415
vitaminc_fact_verification	0.4040	0.2956	0.0852	0.0590	0.0863	0.0843	0.0600	**0.0137**
unit_conversion	0.2810	0.2705	0.0287	0.0616	0.0287	0.0410	0.1300	**0.0173**
winowhy	0.3650	0.2610	0.0270	0.0280	0.0270	0.0290	0.1430	**0.0211**

Table 14: ECE of calibration methods (lower is better) on 21 MedMCQA tasks using the `Qwen-7B-Instruct` backbone. Best results are in **bold** and second best are underlined.

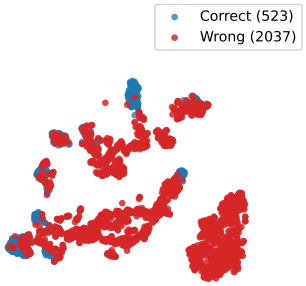
	Uncalibrated	Prompt-based	Global post-hoc			Sample-dependent		Ours
	Vanilla	Elicitation	TS	TvA	DACA	DAC	Thermo	TS3
Anaesthesia	0.3596	0.2781	0.0441	0.0634	0.0472	0.1199	0.2322	**0.0177**
Anatomy	0.3694	0.3034	0.0692	0.1245	0.0634	0.1038	0.2561	**0.0256**
Biochemistry	0.2229	0.1447	0.0585	0.0770	0.0819	**0.0137**	0.1266	0.0463
Dental	0.4065	0.3430	0.0593	0.0971	0.0760	0.0790	0.2199	**0.0373**
ENT	0.3417	0.2660	0.0621	0.0862	0.0600	0.1288	0.2585	**0.0470**
Forensic Medicine	0.4030	0.3344	0.0798	0.0668	0.0910	0.1081	0.2044	**0.0486**
Gynaecology & Obstetrics	0.3384	0.2532	0.0510	0.0747	0.0432	0.0937	0.2794	**0.0403**
Medicine	0.3053	0.2078	0.0523	0.0963	0.0535	0.1121	0.3032	**0.0354**
Microbiology	0.2814	0.2028	0.0714	0.0783	0.0659	0.1168	0.3201	**0.0376**
Ophthalmology	0.3742	0.2871	0.0621	0.1153	0.0667	0.1102	0.2231	**0.0220**
Orthopaedics	0.3521	0.2626	0.0640	0.1030	0.0635	0.0945	0.2359	**0.0474**
Pathology	0.2689	0.1738	0.0684	0.0996	0.0716	0.1080	0.1593	**0.0477**
Pediatrics	0.3415	0.2468	0.0676	0.0935	0.0561	0.0936	0.2787	**0.0424**
Pharmacology	0.2374	0.1381	0.0639	0.1010	0.0773	**0.0189**	0.1250	0.0589
Physiology	0.2807	0.2026	0.0732	0.1058	0.0772	**0.0304**	0.1589	0.0559
Psychiatry	0.2256	0.1308	0.0617	0.0724	0.0906	**0.0195**	0.1128	0.0386
Radiology	0.3598	0.2736	0.0665	0.1263	0.0636	0.1010	0.2602	**0.0253**
Skin	0.2749	0.1874	0.0635	0.0797	0.0566	0.0863	0.1235	**0.0535**
Social & Preventive Medicine	0.3307	0.2639	0.0619	0.0757	0.0567	**0.0259**	0.1892	0.0471
Surgery	0.3359	0.2499	0.0529	0.0802	0.0481	0.1267	0.2728	**0.0320**
Unknown	0.3401	0.2563	0.0544	0.1029	0.0511	0.1341	0.2460	**0.0356**

Table 15: ECE of calibration methods (lower is better) on 8 GLUE tasks using the `Qwen-7B-Instruct` backbone. Best results are in **bold** and second best are underlined.

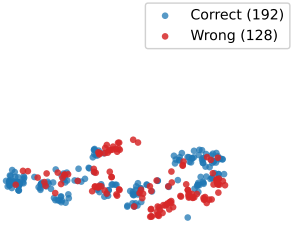
	Uncalibrated	Prompt-based	Global post-hoc			Sample-dependent		Ours
	Vanilla	Elicitation	TS	TvA	DACA	DAC	Thermo	TS3
cola	0.1621	0.1281	0.0672	0.0536	0.0821	0.1187	**0.0173**	0.0459
sst2	0.0673	0.0865	0.0332	0.0495	0.2054	0.0297	**0.0165**	0.0213
mrpc	0.3722	0.3072	0.0758	0.0769	0.0764	0.0579	**0.0416**	0.0423
stsb	0.3222	0.1276	0.1139	0.1021	0.1844	0.1097	0.1007	**0.0853**
qqp	0.2265	0.1642	0.1223	0.0613	0.1804	0.1560	0.1579	**0.0328**
mnli	0.1754	0.0678	0.0494	0.0364	0.2611	**0.0172**	0.0216	0.0194
qnli	0.2662	0.3459	0.1064	0.1157	0.1096	**0.0195**	0.1182	0.0510
rte	0.0860	0.0282	0.1188	0.0745	0.1542	0.2254	0.3017	**0.0445**



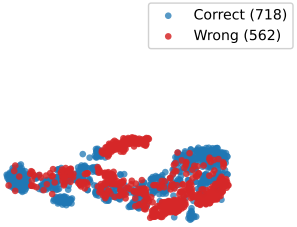
(a) Color–Calibration set.



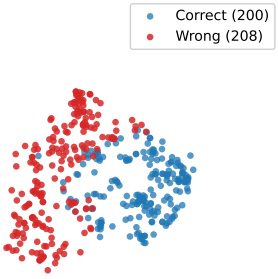
(b) Color–Test set.



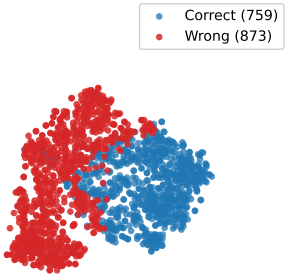
(c) Epistemic Reasoning–Calibration set.



(d) Epistemic Reasoning–Test set.

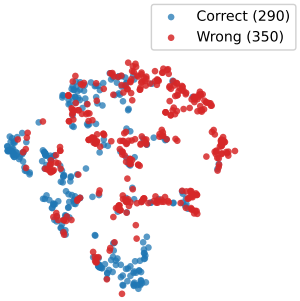


(e) Timedial–Calibration set.

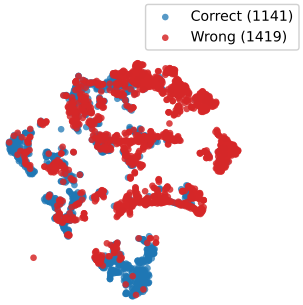


(f) Timedial–Test set.

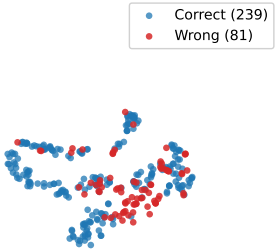
Figure 8: TSNE visualization of the validation and test sets on BIG-Bench using `Mistral-7B-I` backbone.



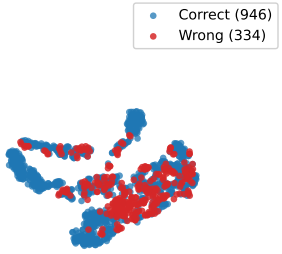
(a) Color – Calibration set.



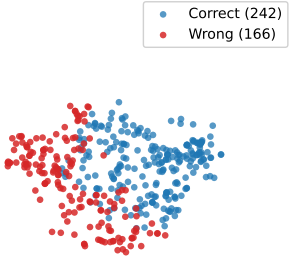
(b) Color – Test set.



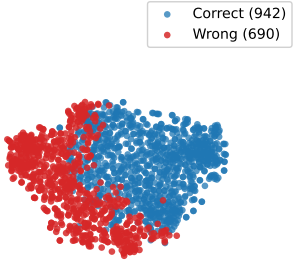
(c) Epistemic Reasoning–Calibration set.



(d) Epistemic Reasoning–Test set.



(e) Timedial – Calibration set.



(f) Timedial – Test set.

Figure 9: TSNE visualization of the validation and test sets on BIG-Bench using `Qwen-7B-I` backbone.