A Domain-Specific Post-Hoc Approach to Address the Failure of Platt Scaling in LLM Calibration

Anonymous ACL submission

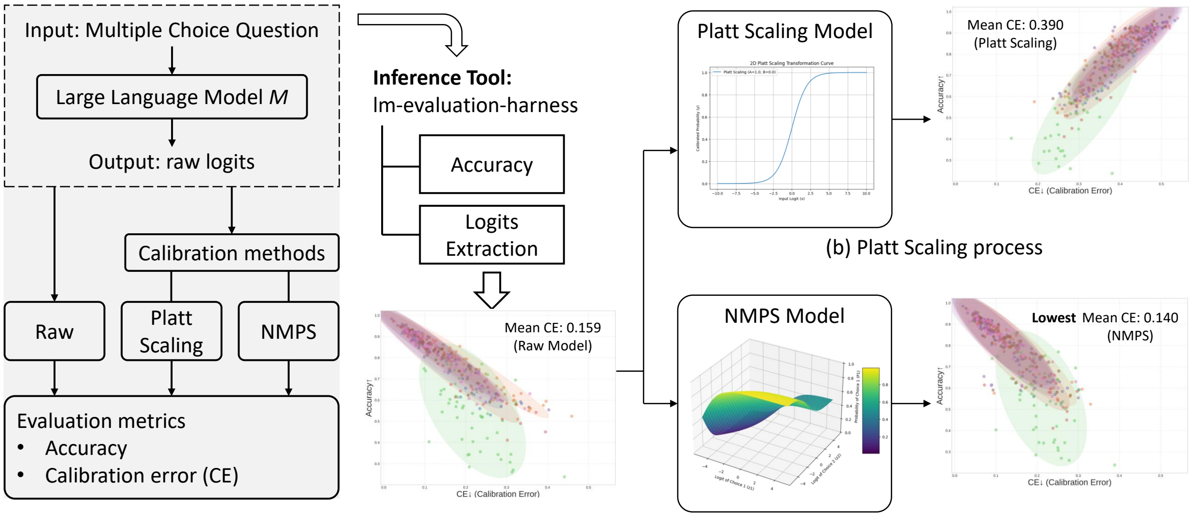
Abstract

The reliable deployment of trustworthy AI systems hinges upon precise model calibration. While LLM capabilities advance, a deeper empirical understanding of their calibration under diverse conditions and varying task demands, subjected to multiple choice questions, remains essential. This paper presents a comprehensive analysis of LLM calibration across multiple architectures and a spectrum of multiple choice questions in different domains. Our systematic investigation reveals that standard calibration techniques, including widely used temperature scaling and Platt Scaling, often show inconsistent efficacy across different models and different knowledge domains, underscoring the need for more adaptive calibration strategies. As part of this broad investigation, we introduce and evaluate Normalized Multiple Choice Platt Scaling (**NMPS**). This lightweight, post-processing technique is highly efficient, requiring no LLM fine-tuning and adding negligible computational overhead during inference. Our experiments demonstrate that this approach offers a substantial improvement over existing methods; it reduces the mean calibration error across our test suite by nearly 12%, whereas standard Platt Scaling shows detrimental, increasing the error to 145%. This work thus provides two key contributions: an effective, non-invasive calibration method and crucial insights into domain-dependent model reliability, offering a practical roadmap for developing more trustworthy AI systems.

1 Introduction

The rapid advancements and increasing scale of Large Language Models(LLMs) (Brown et al., 2020) have marked a significant leap in artificial intelligence, demonstrating remarkable capabilities across a multitude of tasks. As these models become increasingly integrated into real-world applications (Cheng et al., 2025), particularly in autonomous agent systems (Guo et al., 2024), their reliability and trustworthiness are paramount. While much of the recent research has centered on enhancing final accuracy, an often-overlooked aspect in this pursuit of performance is **calibration**—the alignment between a model's predicted confidence and its actual correctness (Dawid, 1982). Although early pre-trained models were found to be reasonably well-calibrated (Desai and Durrett, 2020), this property has degraded in modern, scaled-up LLMs, particularly after alignment tuning (Xie et al., 2024). Consequently, while simple post-hoc methods like temperature scaling (Guo et al., 2017) are common, achieving robust calibration on multiple-choice questions across diverse models at the same time presents significant ongoing challenges. We believe that robust calibration is the backbone of LLM reliability (Liu et al., 2025). It serves as a primary mechanism for identifying and mitigating unreliable outputs like hallucinations, where a model's confidence is a key signal of its potential factuality (Kuhn et al., 2023; Manakul et al., 2023). As such, it will inevitably become a primary target for ensuring trustworthy AI in future usage (Ali et al., 2024).

The need for better calibration methods is critical, as miscalibrated models can be deceptively confident in incorrect predictions, leading to unreliable behavior and exacerbating issues like hallucination, which undermines user trust and operational safety (Kalai and Vempala, 2024). The existing body of work (Proskurina et al., 2024) highlights these issues but often falls short of offering solutions that are robustly generalizable. Many methods focus on average calibration, which can conceal poor performance on specific tasks or subgroups of data—a phenomenon known as grouping loss (Chen et al., 2024). This underscores the need for solutions that are effective across the wide spectrum of tasks LLMs are expected to handle. Our initial investigations confirm this, revealing that the efficacy of standard calibration techniques varies significantly



(b) Platt Scaling process

(a) Our evaluation pipeline

(c) NMPS process

Figure 1: **Overview of the Calibration Framework and a Visual Summary of Results.** **(a)** Our evaluation pipeline uses the `lm-evaluation-harness` to extract raw logits from a base LLM. These logits are then grouped by their assigned category to train and evaluate the calibration scalers. **(b)** The standard Platt Scaling process fits a single, firm sigmoid function to the data. This approach lacks the flexibility for complex LLM outputs, resulting in a poor final Mean CE of 0.390. **(c)** In contrast, our NMPS method learns a more flexible, domain-specific transformation surface. This adaptability allows it to effectively calibrate the LLM's outputs, achieving a superior final Mean CE of 0.140.

with model architecture and knowledge domain, motivating the need for more adaptive approaches. This reveals a critical gap in existing evaluation methodologies: the heterogeneity of calibration performance across different domains has been largely overlooked. Our work is the first to systematically address this challenge, demonstrating that domain-specific analysis is essential for a true understanding of model reliability.

To mitigate these challenges, we introduce **Normalized Multiple Choice Platt Scaling (NMPS)**, a lightweight post-processing strategy to calibrate LLM outputs without altering the base model. NMPS **obviates the need for costly re-training or fine-tuning** by training domain-specific scalers on the model's output logits. This process is highly efficient, requiring only seconds of CPU time per scaler. The NMPS framework offers significant practical benefits. At inference, applying the appropriate domain-specific scaler adds negligible latency. The scalers are both generalizable and robust: they can be transferred between models of varying sizes and can effectively calibrate outputs for unseen questions within a domain. By decoupling the calibration mechanism from the model itself, **NMPS provides a scalable, efficient, and non-invasive solution for improving the trustworthiness of deployed LLM systems**.

The contributions of our work are threefold:

1. We are the first to systematically analyze LLM calibration on a **domain-specific level**. This granular analysis reveals a critical finding: classic calibration methods like **Platt Scaling are fundamentally unsuitable** for LLMs. We find that the simple sigmoid function, effective for traditional binary classifiers, is too rigid to model the complex, high-dimensional logit distributions of LLMs. As shown by the degradation from Figure 2(a) to 2(b), this method systematically increases calibration error.

2. Our analysis further reveals that even simpler methods like Temperature Scaling are **brittle and high-risk**. We show that while the optimal temperature is consistently low, model-specific sensitivity creates "performance cliffs," making this approach unreliable for practical deployment.

3. To address the failure of existing calibration techniques, we propose **Normalized Multiple Choice Platt Scaling (NMPS)**. This novel, lightweight, post-hoc method uses domain-specific scalers to achieve robust calibration. As shown in Figure 2(c), NMPS successfully

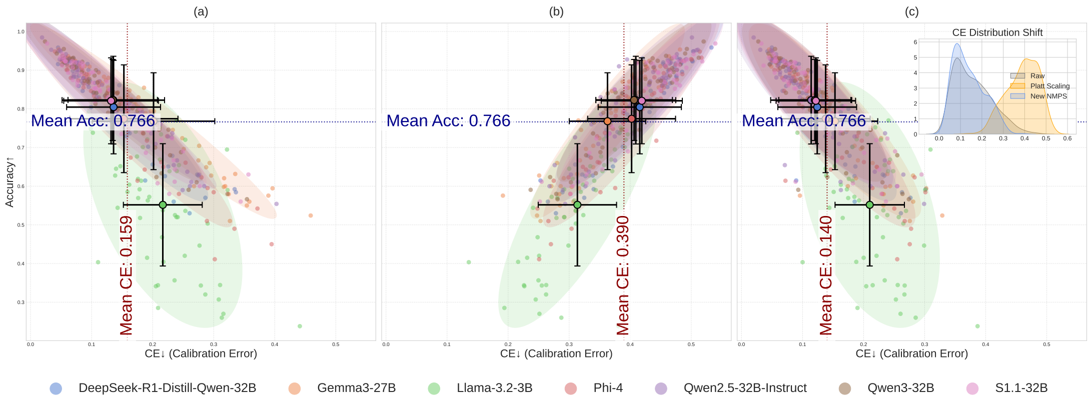


Figure 2: **NMPS Improves Model Calibration Without Affecting Accuracy.** Panels (a), (b), and (c) represent Raw, Platt Scaling, and NMPS calibration methods, respectively. Each filled circle indicates the mean performance across all data points for a given model, with error bars representing ±1 standard deviation (Table 4. The confidence ellipses illustrate the 2-sigma confidence regions for each model. **The inset in panel (c) shows the histogram of calibration errors, visually confirming the leftward shift of the error distribution achieved by NMPS.** Detailed performance metrics are presented in Table 1 and Table 2.

reduces the mean calibration error by nearly 12% where standard methods fail, all while adding negligible computational overhead.

This advancement contributes significantly to the development of more dependable LLMs, paving the way for more robust autonomous agents and safer AI applications.

2 Related Work

Platt Scaling (Platt, 1999), a common and effective post-processing technique for improving the calibration of probabilistic classifiers, including those used in LLMs, involves fitting a logistic regression model to the classifier's output scores (logits) to map them to more calibrated probability estimates. While originally proposed for Support Vector Machines (SVM), it has been widely adopted for various models to address issues of over- or under-confidence in their predictions (Gupta and Ramdas, 2023; Singh and Goshtasbpour, 2022). However, as our results show, its direct application to modern LLMs across diverse tasks can be detrimental. The adaptation of scaling methods to be sensitive to varying task demands is a key research direction. Recent work has explored this adaptation at different levels of granularity. For instance, Adaptive Temperature Scaling (ATS) learns to predict a unique temperature for each token based on model hidden states (Xie et al., 2024). Other work has focused on group-level adaptation to ensure fairness, calibrating models differently for distinct demographic subgroups to mitigate grouping loss (Chen et al., 2024). Our work contributes to this direction by proposing an adaptation at the level of semantic task categories, offering a balance between the flexibility of fine-grained methods and the robustness of group-level approaches.

The inherent challenge of hallucination in LLMs, as acknowledged by Kalai and Vempala (Kalai and Vempala, 2024), underscores the persistent need for robust calibration techniques. Their work suggests that hallucination may be an intrinsic property of these models, further emphasizing the importance of well-calibrated confidence scores as a means of identifying and potentially mitigating unreliable outputs.

Post-hoc adaptation of output probabilities, a broader body of research explores alternative paradigms for improving LLM reliability. One approach involves fine-tuning the model itself to better express confidence, for instance, by using a multi-agent speaker-listener framework to teach pragmatic confidence signaling (Stengel-Eskin et al., 2024). Another paradigm involves "white-box" probing, where lightweight classifiers are trained on the LLM's internal hidden states to directly predict the truthfulness of a statement (Azaria and Mitchell, 2023). Wei et al. (2024) focused on methods to measure and reduce hallucination without gold-standard answers, and Nguyen et al. (2025) explored distillation techniques to enhance factual consistency. While these approaches offer valuable tools, they often require costly model fine-tuning or direct access to internal model states. In contrast, our work provides a distinct yet complementary focus on a lightweight, post-hoc method that refines the confidence calibration of any model given only its output logits, making it a highly practical and scalable solution. The increasing attention to calibration and hallucination underscores the timeliness of our investigation.

3 Preliminaries

3.1 Calibration in Large Language Models

In machine learning, a model is considered **calibrated** if its predicted probabilities accurately reflect the true likelihood of an event (Dawid, 1982). For instance, if a calibrated model assigns an 80% confidence to a set of predictions, then approximately 80% of those predictions should be correct. Conceptually, the goal of calibration is to minimize the difference between the model's confidence and its actual accuracy. This gap is often referred to as the **Calibration Error**. For a set of predictions, the error can be intuitively understood as:

$$\text{Calibration Error} = |\text{confidence} - \text{accuracy}|$$

While the **Calibration Error** is the absolute difference between confidence and accuracy. A perfectly calibrated model would have a calibration error of zero. The various metrics used to evaluate calibration, such as Adaptive Calibration Error (Pavlovic, 2025), are essentially sophisticated methods for averaging this fundamental error across different confidence levels and classes. To quantitatively assess LLM calibration, researchers often employ tasks with verifiable ground truth, such as multi-choice Question Answering (MCQA). In such setups, the model's confidence is typically derived from the probability it assigns to its chosen answer option. By comparing these confidence scores against the empirical accuracy of the predictions, we can evaluate calibration using metrics like the Adaptive Calibration Error (CE).

3.2 Standard Calibration Methods

Temperature Scaling is a simple post-hoc method that uses a single parameter, the temperature $T > 0$, to rescale a model's logits $\mathbf{z}$ before the softmax operation. Calibrated probabilities $\hat{\mathbf{q}}$ for each class c are given by:

$$\hat{q}_c = \frac{\exp(z_c/T)}{\sum_{j=1}^{C} \exp(z_j/T)}. \tag{1}$$

The parameter T is optimized on a validation set to minimize Negative Log-Likelihood (NLL).

Platt Scaling is another post-hoc method that learns a logistic regression model. For a binary problem with logit f, it computes a calibrated probability $P(y = 1|f) = \sigma(Af + B)$, where σ is the sigmoid function, f is the raw output before softmax from the last layer, and parameters A and B are optimized on a validation set. While effective for simple classifiers, our results show this standard approach is counterproductive for modern LLMs on diverse multi-choice questions.

4 Normalized Multiple Choice Platt Scaling

To address the inconsistent performance of standard calibration techniques, we introduce **Normalized Multiple Choice Platt Scaling (NMPS)**, a lightweight post-processing method designed to be both parameter-efficient and adaptable to the diverse demands of modern LLMs. This section first defines the core mathematical principles of NMPS and then details the adaptive, domain-specific process of training and inference.

4.1 Core Formulation

The formulation of NMPS is guided by a core insight: a calibration method for modern, multitalented LLMs must be flexible enough to handle varied multiple choice outputs but simple enough to be learned robustly from limited calibration data. Our design is based on two principles:

1. **Parameter Efficiency:** For multi-choice tasks where the number of options can vary, learning a separate parameter for each choice position as in methods like vector scaling is prone to overfitting. We enforce that our scaler uses only two global parameters, A and B, shared across all choices for a given instance. This simplicity is a key advantage to preventing overfitting, making the method robust and generalizable.

2. **Coherent Distribution:** The method must output a valid probability distribution that sums to one. Standard Platt Scaling, applied independently to each choice, does not guarantee this. Our method includes an explicit normalization step to ensure the final output is a coherent distribution, making it directly usable for downstream decision-making.

Based on these principles, the NMPS method maps a model's raw output logits for a given multi-choice instance to a calibrated probability distribution. Given an instance i with k_i choices and a vector of logits $\mathbf{l}_i = (l_{i,1}, \ldots, l_{i,k_i})$, NMPS computes the final calibrated probability vector $\hat{\mathbf{p}}_i$ as:

$$\hat{p}_{i,j} = \frac{\sigma(A \cdot l_{i,j} + B)}{\sum_{m=1}^{k_i} \sigma(A \cdot l_{i,m} + B)}, \qquad (2)$$

where $\sigma(z) = (1+e^{-z})^{-1}$ is the logistic (sigmoid) function, and $A, B \in \mathbb{R}$ are the learnable parameters. **This formulation allows NMPS to learn a flexible, two-dimensional transformation surface for the logits, as visualized in Figure 1(c), rather than the simple one-dimensional curve of standard Platt scaling.**

Parameter Learning. The optimal parameters (A, B) for a given domain category are learned by minimizing the Negative Log-Likelihood (NLL) on a dedicated calibration set, $\mathcal{D}_{\text{calib}}$:

$$\mathcal{L}(A, B) = -\frac{1}{|\mathcal{D}_{\text{calib}}|} \sum_{i \in \mathcal{D}_{\text{calib}}} \log(\hat{p}_{i,y_i}), \qquad (3)$$

where y_i is the index of the true class for instance i, and the optimization is performed using L-BFGS-B (Zhu et al., 1997).

4.2 Adaptive Calibration Process

Our key innovation is applying the NMPS formulation in a domain-category-dependent manner. Instead of relying on manual labels, our framework uses an LLM itself to categorize domains, allowing the calibration to adapt to the unique error profile an LLM may exhibit in different knowledge areas (e.g., mathematics vs. history). This adaptive process, formalized in Algorithm 1 and Algorithm 2, makes the system more autonomous and scalable.

The process operates in two phases. The first is a one-time, offline training phase (Algorithm 1). We begin with our calibration set, which consists of 80% of the MMLU benchmark data. To establish a robust set of domains, we first utilized **Google Gemini** to analyze this data and determine ten distinct semantic categories (e.g., Mathematics, History & Geography, etc.). Once these ten domains were defined, each question in the MMLU calibration set was automatically assigned its corresponding domain label.

With this fully categorized data, the algorithm then proceeds to train a specialized NMPS scaler for each of the ten categories by minimizing the NLL loss (Equation 3) on the subset of data belonging to that specific domain (D_c). This process results in a comprehensive toolkit of expert scalers, Θ, ready for the inference phase.

Algorithm 1 Training Domain-Category NMPS Scalers

```
Require: Base LLM $\mathcal{M}$, Categorizer LLM $\mathcal{M}_{cat}$,
    Training data $D_{\text{train}} = \{(q_i, o_i, y_i)\}_{i=1}^{N}$, Set of
    categories $\mathcal{C}$.
Ensure: A set of trained scalers $\Theta = \{\theta_c\}_{c \in \mathcal{C}}$.
    // Step 1: Auto-Categorize Training Data
$D_{\text{categorized}} \leftarrow []$
for each $(q_i, o_i, y_i)$ in $D_{\text{train}}$ do
    $c_i \leftarrow \mathcal{M}_{cat}(\text{prompt}, q_i)$ ▷ Assign category
    Append $(q_i, o_i, y_i, c_i)$ to $D_{\text{categorized}}$
end for
    // Step 2: Train Scalers on Categorized Data
$\Theta \leftarrow \{\}$
for each category $c$ in $\mathcal{C}$ do
    $D_c \leftarrow \text{FilterData}(D_{\text{categorized}}, c)$
    $logits, labels \leftarrow \text{ExtractLogits}(\mathcal{M}, D_c)$
    $\theta_c \leftarrow \text{TrainScaler}(logits, labels)$ ▷ Minimize Eq. 3
    $\Theta[c] \leftarrow \theta_c$
end for
return $\Theta$
```

Algorithm 2 Calibrated Inference with LLM-driven NMPS

```
Require: Base LLM $\mathcal{M}$, Categorizer LLM $\mathcal{M}_{cat}$,
    Trained scalers $\Theta = \{\theta_c\}_{c \in \mathcal{C}}$.
    Input: A new question $q_{\text{new}}$ with options $o_{\text{new}}$.
    Output: Calibrated probability distribution
    $\mathbf{p}_{calibrated}$.
$L_{\text{raw}} \leftarrow \mathcal{M}(q_{\text{new}}, o_{\text{new}})$
$c_{\text{new}} \leftarrow \mathcal{M}_{cat}(\text{prompt}, q_{\text{new}})$
$\theta_c \leftarrow \Theta[c_{\text{new}}]$
$\mathbf{p}_{calibrated} \leftarrow \text{ApplyScaler}(\theta_c, L_{\text{raw}})$ ▷ Uses Eq. 2
return $\mathbf{p}_{calibrated}$
```

The second phase is the lightweight online inference process (Algorithm 2). For any new input question q_{new}, the system first invokes the categorizer LLM to determine its category, c_{new}. It then retrieves the corresponding pre-trained scaler θ_c from the dictionary Θ and applies it to the base model's raw logits. This dynamic selection ensures that the most appropriate calibration is applied for every query with negligible latency, removing any need for manual intervention during deployment.

The inference process in Algorithm 2 is lightweight with negligible overhead. For any new input question q_{new}, the system first obtains the raw logits from the base LLM. It then identifies the question's domain category and retrieves the corresponding pre-trained scaler θ_c from the dictionary Θ. Finally, this specialized scaler is applied to the raw logits using the transformation in Equation 2 to produce the final, calibrated probability distribution. This dynamic selection ensures that the most appropriate calibration is applied for every query.

5 Experiment

5.1 Experimental Setup

We evaluate our method on a suite of contemporary LLMs, including models from the **Llama 3** series (Llama-3) (Grattafiori et al., 2024), a strong reasoning model **s1.1-32B** (Muennighoff et al., 2025), and other state-of-the-art models such as **Phi4-14B** (Abdin et al., 2024) and **Qwen3-32B**. These models were chosen for their strong performance and were accessed via the Huggingface Hub. To test for generalizability, we leveraged a diverse suite of MCQA benchmarks, including the 57 tasks of **MMLU** (Hendrycks et al., 2021) and a selection of tasks from **BigBench** (bench authors, 2023), accessed via the `lm-evaluation-harness` (Gao et al., 2023) tool from EleutherAI.

5.2 Baselines

We compare the performance of three approaches:

- **Raw (Uncalibrated):** The direct output of the base LLMs without any calibration.
- **Temperature Scaling:** The standard method where the logits is scaled with a single temperature T before softmax.
- **Platt Scaling:** The standard Platt transformation applies a sigmoid function to each logit independently, which does not produce a valid probability distribution that sums to one. For this baseline, we therefore implemented a multi-choice adaptation where the transformation is applied to each of the k_i choice logits, and the resulting scores are **then normalized** to form a coherent probability distribution. A single, global set of parameters (A, B) is trained on the entire calibration dataset.

5.3 Evaluation Metrics

To assess calibration, we use the Adaptive Calibration Error (CE). Unlike standard Expected Calibration Error (ECE) (Naeini et al., 2015), which uses equal-width bins, ACE uses quantile-based binning to ensure a stable error estimate even when high-confidence predictions are rare. For a given sample i, let $\hat{y}_i$ be its predicted class, y_i its true class, and conf_i its predicted confidence. The metric is computed class-wise: For each bin B_m and each true class k, we define the set of samples $S_{m,k} = \{i \mid \text{conf}_i \in B_m, y_i = k\}$. We then compute the accuracy, $\text{acc}(S_{m,k})$, and average confidence, $\text{conf}(S_{m,k})$, for this set. The final ACE is computed as the unweighted mean of the absolute differences over all $M \times C$ class-bin pairs:

$$\text{ACE} = \frac{1}{M \cdot C} \sum_{m=1}^{M} \sum_{k=1}^{C} |\text{acc}(S_{m,k}) - \text{conf}(S_{m,k})|, \tag{4}$$

where the error for an empty set $S_{m,k}$ is set to zero.

5.4 Implementation Details

Data Handling for Calibration and Testing. To ensure a fair and rigorous evaluation, we carefully partitioned our datasets. We designated an 80% split of the **MMLU** benchmark as our calibration set. This set was used exclusively for training the parameters of Platt Scaling, and our NMPS. The remaining 20% of MMLU was held out as a validation set. All other benchmarks were used purely as a test set to evaluate the generalization of the trained calibrators on entirely unseen data. This strict separation prevents any data leakage between the calibration training process and the final performance evaluation.

Logit Extraction. We used EleutherAI's `lm-evaluation-harness` tool throughout our experiments. For each multiple-choice question, we configured the tool to record the raw, pre-softmax logits for the tokens in every correct answer option.

Calibration Method Setup. All calibration methods were trained or optimized using the same 80% MMLU validation set for fair comparison. For the Temperature Scaling baseline, the optimal temperature T was determined via a **grid search over the range** $[0.01, 2.0]$, selecting the value that yielded the lowest CE on the validation set. For both Platt Scaling and our NMPS, the parameters were optimized using the L-BFGS-B algorithm to minimize the NLL, as described in Section 4.

Experimental Pipeline and Environment. Our entire experimental pipeline was automated with Python scripts using the **PyTorch** and **Hugging-Face Transformers** libraries. These scripts handled model loading, evaluation via the harness tool,

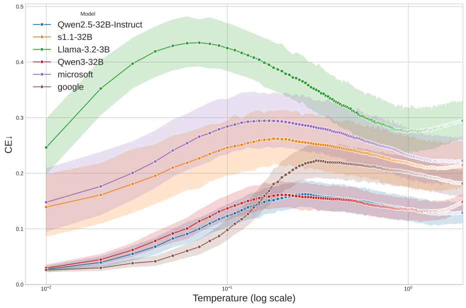


Figure 3: **Effect of Temperature on CE (↓).** The plot shows the relationship between CE and temperature for several foundation models, highlighting their drastically different sensitivity profiles.

and the extraction of raw logits to JSON files.

A key step in our automated workflow is **LLM-driven domain categorization**. We established ten primary domains based on the MMLU benchmark. For any given question in our datasets, a dedicated instruction-tuned model (**Qwen1.5B-Instruct**) classifies it into one of these ten domains by processing it through a zero-shot classification prompt. This self-categorization step removes the need for manual labeling and makes our approach more scalable and generalized. Subsequent scripts then use these categorized logits for the training and evaluation of all calibration methods.

In addition to our primary analysis, we conducted a secondary set of experiments to evaluate how calibration is affected by **quantization**, a common technique for model compression. The detailed results for all experiments are provided in Appendix 2. All experiments were conducted on a system equipped with 2x NVIDIA A100 GPUs.

6 Results and Analysis

Our experiments are designed to rigorously evaluate the effectiveness of our proposed NMPS method against standard calibration methods. We analyze the results through 3 primary lenses: (1) the aggregate performance across all models and tasks; (2) the specific failure modes of common baselines; (3) a detailed, per-model, per-category breakdown that reveals the nuances of domain-specific multiple choice questions calibration.

6.1 Quantitative Performance: NMPS Achieves Superior Calibration

Our primary finding is that NMPS provides a significant and consistent improvement over both the uncalibrated baseline and standard Platt scaling, all without affecting model accuracy. Figure 2 provides a powerful visualization of the result across more than 300+ data points. The left panel shows the wide distribution of performance for the raw, uncalibrated models, which achieve a mean accuracy of 0.766 but exhibit a substantial mean CE of 0.159, showing that the modern SoTA LLMs perform a significantly high accuracy which lowers the issue of over-confidence.

The center panel reveals the catastrophic failure of the standard Platt Scaling baseline. Not only does it fail to improve calibration, it is actively detrimental, more than doubling the mean CE to a staggering 0.390. This demonstrates even with categorized applying calibration methods the standard calibration method designed for simple classifiers to the complex, multi-choice outputs of modern LLMs can systematically induce overconfidence and degrade reliability (Xiao et al., 2025).

In stark contrast, NMPS shows the clear success of our NMPS method. It reduces the mean CE to 0.140, a relative improvement of nearly 12% over the raw baseline, while leaving the mean accuracy entirely unchanged. This improvement is visualized by the decisive leftward shift of the entire data distribution, highlighted in the inset histogram. This result confirms that a method designed with the principles of parameter efficiency and coherent distribution for multi-choice outputs is essential for effective calibration.

6.2 The Brittle Nature of Temperature Scaling

A key motivation for our work is the unreliable and model-specific behavior of temperature scaling. While a visual inspection of Figure 3 suggests complex behavior, a deeper quantitative analysis in Table 5 reveals a more nuanced and critical problem: extreme sensitivity.

The data show that while all models achieve their theoretical minimum CE at the lowest tested temperature ($T = 0.01$), their performance landscapes dramatically. For example, Qwen3-32B is relatively stable across the temperature range. In stark contrast, Llama-3.2-3B is extremely brittle; its CE increases to a peak of 0.435 at a low temperature of just $T = 0.07$. This creates a **performance cliff** where a seemingly reasonable temperature setting could result in catastrophic miscalibration. This extreme, model-dependent sensitivity makes any fixed temperature scaling strategy a high-risk, unreliable solution, providing strong evidence that more adaptive, principled methods are required.

Table 1: **Quantitative Model Performance.** Average performance across all ten subject categories. The final row shows the mean performance across all models, highlighting the aggregate improvement of our method (NMPS). Full results are in Table 2.

Model	Accuracy	Calibration Error ↓		
		Raw	Platt	NMPS
DeepSeek-R1-Distill-Qwen-32B	0.8045	0.1360	0.4156	**0.1241**
Gemma3-27B	0.7680	0.2015	0.3628	**0.1531**
Llama-3.2-3B	0.5518	0.2167	0.3135	**0.2101**
Phi-4	0.7744	0.1529	0.4023	**0.1378**
Qwen2.5-32B-Instruct	0.8225	0.1356	0.4092	**0.1141**
Qwen3-32B	0.8229	0.1346	0.4064	**0.1201**
S1.1-32B	0.8208	0.1319	0.4190	**0.1218**
Mean	**0.7664**	**0.1585**	**0.3898**	**0.1402**

6.3 Detailed Analysis: A Deeper Look at Per-Category Performance

To systematically assess model calibration across diverse domains, we present a detailed breakdown of performance in Table 1 and Table 2. These tables report accuracy and CE across ten subject categories for each model under our three test conditions: Raw (uncalibrated), standard Platt Scaling, and our proposed NMPS. This granular view reveals several key trends.

First, the detailed data confirms that even strong foundation models suffer from significant miscalibration. While raw model outputs often yield high accuracy, they exhibit a consistent mismatch between confidence and correctness. For instance, Gemma3-27B shows a very high raw CE of 0.3020 on the challenging mathematics domain. Even in a comparatively strong area like psychology & sociology, its raw CE of 0.1308 indicates a notable level of miscalibration.

Second, the standard Platt Scaling baseline proves almost universally ineffective and frequently detrimental. Its global parameterization (a single set of A and B parameters applied to all choices) fails to adapt to domain-specific error profiles. With models like DeepSeek-R1, Platt Scaling degrades calibration in nearly every domain, causing the CE to balloon to over 0.40 in many cases, far worse than the uncalibrated baseline.

In contrast, our NMPS method demonstrates robust and consistent improvements across all models and domains. By incorporating domain-aware calibration, NMPS adapts flexibly to each category's unique data distribution. For Llama-3.2-3B in philosophy & ethics, NMPS lowers the CE from a detrimental 0.3310 (Platt) to 0.2219. A more striking example is seen with Qwen2.5-32B in psychology & sociology," where NMPS produces an exceptionally low CE of just 0.0587, a dramatic improvement over both the raw model (0.0679) and the failed Platt baseline (0.4682).

In summary, this detailed analysis confirms our central claims: 1) Modern LLMs exhibit significant and widespread miscalibration; 2) Generic, non-adaptive methods like standard Platt Scaling are insufficient and can actively harm performance; and 3) Our domain-aware NMPS method provides a consistent and significant reduction in CE, demonstrating that adaptive strategies are essential for building trustworthy LLM systems.

7 Discussion and Conclusion

Our investigation reveals that LLM calibration is a nuanced challenge where standard post-hoc methods are often inconsistent or even actively harmful. We find that Temperature Scaling is a brittle, high-risk strategy due to its extreme model-dependent sensitivity, while standard Platt Scaling catastrophically fails by inducing severe overconfidence. Our proposed NMPS method directly addresses these failures. By adapting its parameters to domain categories, NMPS achieves what simpler methods cannot: a consistent and significant improvement in calibration across a diverse suite of models without sacrificing accuracy. Its lightweight and non-invasive nature—requiring no LLM fine-tuning—makes it a practical and scalable tool.

In conclusion, by being the first to systematically analyze LLM calibration at a domain-specific level, this work uncovers the fundamental limitations of standard techniques. Our findings provide not only an effective tool for practitioners but also reinforce the critical insight that **domain-awareness** is an essential principle for building trustworthy and reliable AI systems.

Limitations

The primary limitation of our work is that the task categorization is manually predefined. While this provides a strong proof-of-concept for domain-specific multiple choice questions calibration, a natural next step is to explore methods for learning these categories automatically. To this end, a sensitivity analysis showing how performance changes with different category groupings would also strengthen the findings. Additionally, our study focused on multi-choice question answering; extending and evaluating NMPS for generative tasks and open-ended outputs remains an important avenue for future research (Liu et al., 2024).

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

S. Ali, A. S. Ryzhikov, D. A. Derkach, F. D. Ratnikov, and V. O. Bocharnikov. 2024. Calibrating for the future:enhancing calorimeter longevity with deep learning. *Preprint*, arXiv:2411.03891.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

BIG bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Lihu Chen, Alexandre Perez-Lebel, Fabian M. Suchanek, and Gaël Varoquaux. 2024. Reconfidencing LLMs from the grouping loss perspective. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1567–1581, Miami, Florida, USA. Association for Computational Linguistics.

Jingwen Cheng, Kshitish Ghate, Wenyue Hua, William Yang Wang, Hong Shen, and Fei Fang. 2025. Realm: A dataset of real-world llm use cases. *Preprint*, arXiv:2503.18792.

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

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. 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.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *Preprint*, arXiv:2402.01680.

Chirag Gupta and Aaditya Ramdas. 2023. Online platt scaling with calibeating. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Adam Tauman Kalai and Santosh S. Vempala. 2024. Calibrated language models must hallucinate. *Preprint*, arXiv:2311.14648.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *Preprint*, arXiv:2302.09664.

Han Liu, Yupeng Zhang, Bingning Wang, Weipeng Chen, and Xiaolin Hu. 2024. Full-ece: A metric for token-level calibration on large language models. *CoRR*, abs/2406.11345.

Hongfu Liu, Hengguan Huang, Xiangming Gu, Hao Wang, and Ye Wang. 2025. On calibration of llm-based guard models for reliable content moderation. *Preprint*, arXiv:2410.10414.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *Preprint*, arXiv:2501.19393.

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

Hieu Nguyen, Zihao He, Shoumik Atul Gandre, Ujjwal Pasupulety, Sharanya Kumari Shivakumar, and Kristina Lerman. 2025. Smoothing out hallucinations: Mitigating llm hallucination with smoothed knowledge distillation. *Preprint*, arXiv:2502.11306.

Maja Pavlovic. 2025. Understanding model calibration – a gentle introduction and visual exploration of calibration and the expected calibration error (ece). *Preprint*, arXiv:2501.19047.

John C. Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers*, pages 61–74. MIT Press.

Irina Proskurina, Luc Brun, Guillaume Metzler, and Julien Velcin. 2024. When quantization affects confidence of large language models? *Preprint*, arXiv:2405.00632.

Rishabh Singh and Shirin Goshtasbpour. 2022. Plattbin: Efficient posterior calibrated training for NLP classifiers. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3673–3684, Dublin, Ireland. Association for Computational Linguistics.

Elias Stengel-Eskin, Peter Hase, and Mohit Bansal. 2024. Lacie: Listener-aware finetuning for calibration in large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 43080–43106. Curran Associates, Inc.

Jiaheng Wei, Yuanshun Yao, Jean-Francois Ton, Hongyi Guo, Andrew Estornell, and Yang Liu. 2024. Measuring and reducing llm hallucination without gold-standard answers. *Preprint*, arXiv:2402.10412.

Jiancong Xiao, Bojian Hou, Zhanliang Wang, Ruochen Jin, Qi Long, Weijie J. Su, and Li Shen. 2025. Restoring calibration for aligned large language models: A calibration-aware fine-tuning approach. *Preprint*, arXiv:2505.01997.

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

Ciyou Zhu, Richard H. Byrd, Peihuang Lu, and Jorge Nocedal. 1997. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.*, 23(4):550–560.

A Acknowledgement of Artifacts

Our work prioritizes reproducibility. All models used are publicly available through the HuggingFace Hub, with specific model identifiers listed in the Experimental Setup. All datasets are standard public benchmarks accessed via the `lm-evaluation-harness` v0.4.1. Our experimental pipeline was fully automated using custom Python scripts to ensure consistency. These scripts handled model loading, execution of tasks via the harness, extraction of raw logits to CSV files for all predictions, and the subsequent application and evaluation of all calibration methods (Raw, Temperature Scaling, Platt Scaling, and NMPS). The code for our proposed NMPS method and the scripts to reproduce the main findings reported in this paper will be made publicly available upon publication.

A Additional Results and Analysis

This appendix provides supplementary data and analyses that support the main findings of our paper. We include detailed per-category results, a quantitative breakdown of our temperature scaling experiments, and a secondary analysis of quantization effects.

A.1 Detailed Per-Category Performance

Tables 2 and 3 contain the detailed numerical data for accuracy and Adaptive Calibration Error (CE) that are aggregated and visualized in Figure 2 and summarized in Table 1 of the main paper. These tables list the performance for each of the seven primary models on each of the ten subject categories under the three conditions: Raw (uncalibrated), standard Platt Scaling, and our proposed NMPS method. This granular data provides the full evidence for our claims regarding the widespread miscalibration of raw models, the detrimental effect of standard Platt Scaling, and the consistent, robust improvements provided by NMPS across diverse domains.

A.2 Temperature Scaling Analysis

Table 5 lists the maximum and minimum Calibration Error values and their corresponding temperatures for each model, obtained by sweeping the temperature parameter from 0.01 to 2.0. This table provides the quantitative data that supports the analysis in Section 6.2, particularly the finding that while all models achieve their minimum CE at a low temperature, their sensitivity profiles and peak error values (Max CE) differ dramatically, making temperature scaling a brittle strategy.

A.3 Secondary Analysis: The Effect of Quantization

In addition to our primary analysis, we conducted a secondary set of experiments to evaluate how calibration is affected by quantization. Figure 4 shows the calibration performance for the Llama-3.2-1B model and its 8-bit and 4-bit quantized versions. The results demonstrate that while quantization can alter a model's calibration profile, our NMPS method remains an effective post-hoc solution for improving the reliability of these compressed models. Figure 5 further explores this by showing the temperature scaling curves for these quantized models. The results indicate that the general sensitivity profile to temperature remains largely consistent across different levels of model compression.

Table 2: **Comparative Analysis of Model Calibration Techniques.** A summary of performance metrics for various models across ten subject categories. We compare the raw model outputs (**Raw**), Platt Scaling, and our proposed method (**NMPS**). This comprehensive comparison demonstrates the efficacy of NMPS across diverse models and domains.

Model	**Category**	**Accuracy**	**Calibration Error ↓**		
			Raw	**Platt Scaling**	**NMPS**
DeepSeek-R1-Distill-Qwen-32B	Biological & Medical Sciences	0.8406	0.1210	0.4362	**0.1136**
	Computer Science & Engineering	0.7977	0.1457	0.3850	**0.1369**
	Economics & Business	0.8273	0.1167	0.4183	**0.0996**
	General Knowledge & Misc.	0.7595	**0.1424**	0.3907	0.1465
	History & Geography	0.8973	0.0793	0.4574	**0.0765**
	Law & Governance	0.8379	0.1145	0.4285	**0.0953**
	Mathematics	0.6931	0.2035	0.3703	**0.1846**
	Philosophy & Ethics	0.7839	0.1631	0.4248	**0.1492**
	Physical Sciences	0.7364	0.1700	0.3819	**0.1485**
	Psychology & Sociology	0.8838	0.0731	0.4643	**0.0650**
Gemma3-27B	Biological & Medical Sciences	0.7897	0.1893	0.3739	**0.1547**
	Computer Science & Engineering	0.7325	0.2288	0.3434	**0.1875**
	Economics & Business	0.7952	0.1826	0.3673	**0.1358**
	General Knowledge & Misc.	0.7438	0.1930	0.3572	**0.1519**
	History & Geography	0.8959	0.1051	0.4151	**0.0890**
	Law & Governance	0.8116	0.1706	0.3655	**0.0900**
	Mathematics	0.6368	0.3020	0.3159	**0.2070**
	Philosophy & Ethics	0.7523	0.2230	0.3671	**0.1673**
	Physical Sciences	0.7110	0.2386	0.3407	**0.1973**
	Psychology & Sociology	0.8538	0.1308	0.3896	**0.0928**
Llama-3.2-3B	Biological & Medical Sciences	0.6233	**0.2061**	0.3357	0.2088
	Computer Science & Engineering	0.5560	0.2764	0.3276	**0.2702**
	Economics & Business	0.5734	0.2100	0.3188	**0.2042**
	General Knowledge & Misc.	0.5333	0.1989	0.3087	**0.1749**
	History & Geography	0.6957	0.1807	0.3593	**0.1741**
	Law & Governance	0.6339	0.1804	0.3243	0.1900
	Mathematics	0.3121	0.2602	0.2440	**0.2408**
	Philosophy & Ethics	0.5281	0.2264	0.3310	**0.2219**
	Physical Sciences	0.4254	0.2450	0.2507	**0.2248**
	Psychology & Sociology	0.6879	0.1319	0.3413	**0.1333**
Phi-4	Biological & Medical Sciences	0.8187	0.1334	0.4239	**0.1214**
	Computer Science & Engineering	0.7497	**0.1739**	0.3651	**0.1739**
	Economics & Business	0.8313	0.1138	0.4284	**0.1085**
	General Knowledge & Misc.	0.7168	0.2026	0.3890	**0.1480**
	History & Geography	0.8883	0.0843	0.4494	**0.0823**
	Law & Governance	0.8241	0.0949	0.4116	**0.0743**
	Mathematics	0.5769	0.2587	0.3248	**0.2319**
	Philosophy & Ethics	0.7462	0.1805	0.4132	**0.1612**
	Physical Sciences	0.6959	0.2017	0.3522	**0.1826**
	Psychology & Sociology	0.8823	0.0809	0.4624	**0.0700**
Qwen2.5-32B-Instruct	Biological & Medical Sciences	0.8461	0.1228	0.4272	**0.1043**
	Computer Science & Engineering	0.8245	0.1498	0.3747	**0.1468**
	Economics & Business	0.8387	0.1184	0.4173	**0.0996**
	General Knowledge & Misc.	0.7897	0.1506	0.3933	**0.1265**
	History & Geography	0.9152	0.0696	0.4554	**0.0647**
	Law & Governance	0.8548	0.1040	0.4267	**0.0629**
	Mathematics	0.7084	0.2181	0.3668	**0.1818**
	Philosophy & Ethics	0.8199	0.1447	0.4111	**0.1187**
	Physical Sciences	0.7526	0.1750	0.3654	**0.1441**
	Psychology & Sociology	0.9082	0.0679	0.4682	**0.0587**

Table 3: **Comparative Analysis of Model Calibration Techniques (Part 2 of 2).** Performance metrics for the remaining models.

Model	**Category**	**Accuracy**	**Calibration Error ↓**		
			Raw	**Platt Scaling**	**NMPS**
Qwen3-32B	Biological & Medical Sciences	0.8490	0.1228	0.4221	**0.1148**
	Computer Science & Engineering	0.8347	0.1406	0.3903	**0.1364**
	Economics & Business	0.8285	0.1256	0.4082	**0.1079**
	General Knowledge & Misc.	0.7743	0.1690	0.3803	**0.1398**
	History & Geography	0.8981	0.0891	0.4287	**0.0886**
	Law & Governance	0.8126	0.1306	0.3943	**0.0862**
	Mathematics	0.7268	0.1879	0.3836	**0.1786**
	Philosophy & Ethics	0.7851	0.1566	0.3911	**0.1345**
	Physical Sciences	0.8209	0.1393	0.4052	**0.1249**
	Psychology & Sociology	0.9050	0.0730	0.4550	**0.0595**
S1.1-32B	Biological & Medical Sciences	0.8380	0.1226	0.4333	0.1082
	Computer Science & Engineering	0.8200	0.1562	0.3919	0.1561
	Economics & Business	0.8324	0.1190	0.4303	0.1026
	General Knowledge & Misc.	0.7856	0.1476	0.3965	0.1360
	History & Geography	0.9122	0.0709	0.4503	0.0687
	Law & Governance	0.8420	0.1138	0.4029	0.0809
	Mathematics	0.7376	0.1767	0.3990	0.1685
	Philosophy & Ethics	0.7956	0.1657	0.4220	0.1560
	Physical Sciences	0.7610	0.1514	0.3936	0.1511
	Psychology & Sociology	0.8987	0.0806	0.4588	0.0765

Table 4: **Quantitative Model Performance.** This variance and standard deviation is representing for Table 1.

Model	**Variance**			**Standard Deviation**		
	Raw	**Platt**	**NMPS**	**Raw**	**Platt**	**NMPS**
DeepSeek-R1-Distill-Qwen-32B	0.0059	0.0046	0.0042	0.0769	0.0681	0.0645
Gemma3-27B	0.0101	0.0039	0.0049	0.1003	0.0627	0.0700
Llama-3.2-3B	0.0042	0.0041	0.0032	0.0646	0.0642	0.0568
Phi-4	0.0078	0.0052	0.0053	0.0886	0.0720	0.0729
Qwen2.5-32B-Instruct	0.0070	0.0043	0.0044	0.0840	0.0654	0.0664
Qwen3-32B	0.0054	0.0039	0.0035	0.0733	0.0628	0.0590
S1.1-32B	0.0060	0.0044	0.0042	0.0775	0.0664	0.0651

Table 5: **Temperature Scaling Min Max Result.** Performance showing from sweeping through 0.01 to 2.0 per 0.01 temperature changing.

Model	Max CE ↓	Max Temp	Min CE ↓	Min Temp
Qwen2.5-32B-Instruct	0.162080	0.270000	0.027387	0.010000
s1.1-32B	0.262212	0.180000	0.139237	0.010000
Llama-3.2-3B	0.434978	0.070000	0.246408	0.010000
DeepSeek-R1-Distill-Qwen-32B	0.177493	0.170000	0.036425	0.010000
Qwen3-32B	0.161016	0.210000	0.029547	0.010000
microsoft	0.294479	0.170000	0.147911	0.010000
google	0.222729	0.310000	0.026073	0.010000

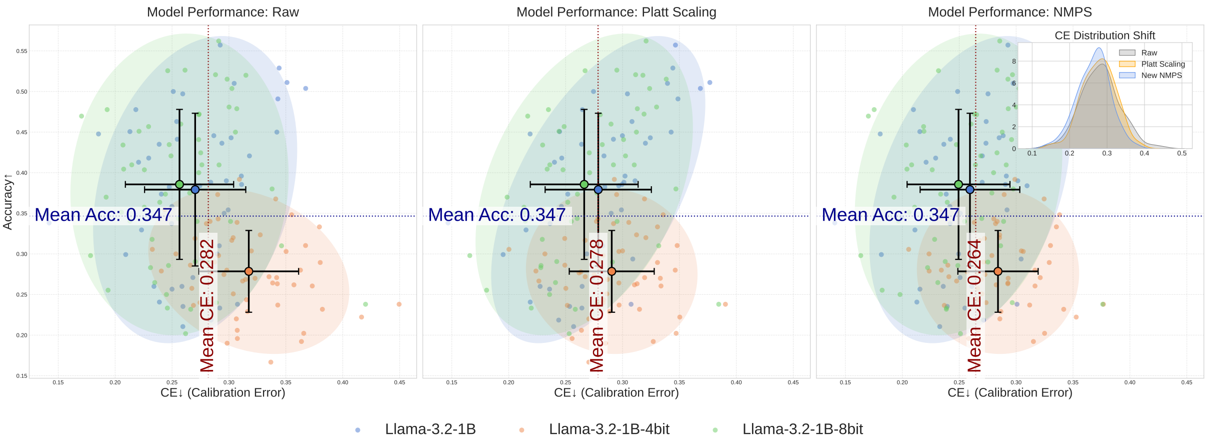


Figure 4: **Quantization Effect on Calibration.** This figure demonstrate our NMPS method on Llama 3.2-1B. This demonstrates that domain-agnostic can still work even on less parameter models, but not as useful.

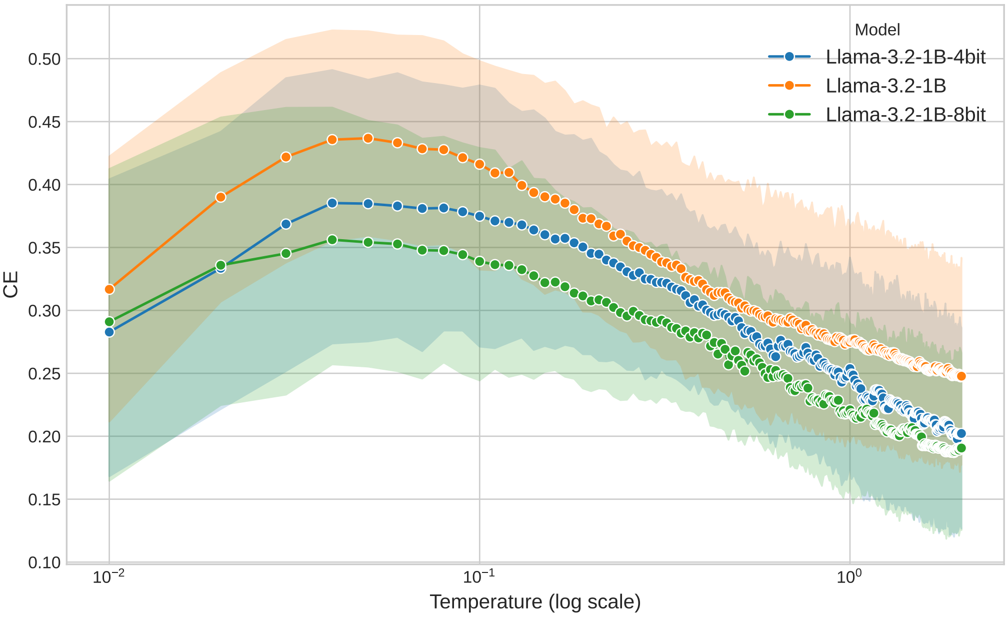


Figure 5: **The Influence of Quantization on the Efficacy of Temperature Scaling.** This plot illustrates the CE as a function of temperature, applied post-hoc to different model versions. The results indicate that across different levels of model compression through quantization, the optimal temperature for minimizing calibration error remains largely consistent. This suggests that the effectiveness of temperature scaling as a calibration method is not significantly impacted by the quantization of the model.