Test-Time User Alignment via Bayesian Population Guidance in Subjective Tasks

Hoe Sung Ryu[1] , **June Christoph Kang**[2] , **Christian Wallraven**[1,2,*]
[1]Department of Artificial Intelligence, Korea University
[2]Department of Brain and Cognitive Engineering, Korea University
wallraven@korea.ac.kr

Abstract

In subjective tasks, different individuals can have different correct answers for the same input—the ground truth is not fixed but rather determined by each user's personal perspective. Standard foundation models suppress this individual variation by producing population-averaged predictions; conversely, few-shot in-context learning (ICL) struggles to capture user-specific preferences from limited demonstrations. To address this limitation, we propose Population-Guided Bayesian Calibration (PGBC), a test-time personalization method that applies Bayesian inference to foundation model outputs for user-specific alignment. Unlike ICL methods that rely on the model to implicitly infer preferences from examples, PGBC explicitly quantifies each individual's deviation from the population distribution and incorporates this into prediction with minimal feedback ($K < 10$). Experiments on three subjective benchmark datasets across four foundation models demonstrate that PGBC significantly outperforms zero-shot baselines with a large effect size (Cohen's $d = 1.21$), and the improvement over ICL approaches is even more pronounced ($d = 1.40$). Furthermore, user-level distribution analysis reveals that PGBC aligns predictions with individual preferences when zero-shot baselines exhibit substantial bias, while ICL methods increase distributional mismatch. Our results show that PGBC enables practical deployment in cold-start scenarios where minimal user feedback must yield maximal personalization.

1 Introduction

Foundation model (FM) training pipelines operate under the assumption of a unique ground truth, treating inter-annotator disagreement as noise to be eliminated via alignment techniques like RLHF [Ouyang *et al.*, 2022]. While effective for objective tasks, this reductionist approach forces models to converge toward population averages, erasing individual nuances in subjective domains (such as affect and toxicity detection), where variability reflects meaningful heterogeneity in human cognition [Stanovich and West, 1998], personality [Jonas and Markon, 2019], and cultural background [He *et al.*, 2017] rather than error [Davani *et al.*, 2022; Basile *et al.*, 2021]. For users whose perspectives deviate from the population mean, such models fail to capture their true preferences [Santurkar *et al.*, 2023]; however, recovering this subjectivity is essential for true personalization.

Prior approaches, including PEFT [Hu *et al.*, 2021], retrieval augmentation [Salemi *et al.*, 2024], and ICL [Brown *et al.*, 2020], attempt to mitigate this issue, but suffer from prohibitive computational costs or reliance on extensive user history. Beyond such practical constraints, a fundamental issue remains. While evidence suggests that model uncertainty—quantified via sampling variance—might serve as a proxy for human disagreement to guide personalization [Uma *et al.*, 2021], our results show that this relationship does not hold in subjective tasks. Furthermore, a popular method like ICL relies on the assumption that models can implicitly infer user preferences from limited demonstrations, an expectation that fails when tasks diverge from pre-training distributions [Zhao *et al.*, 2021]. To enable practical personalization, a method must adapt from minimal examples without parameter updates—preferably at constant inference cost.

To address these limitations, we propose Population-Guided Bayesian Calibration (PGBC), a test-time method that explicitly models individual deviations from population consensus. PGBC treats foundation model outputs as semantic features and learns user-specific calibration via closed-form Bayesian updates, employing population disagreement to downweight inherently ambiguous items. Our contributions are threefold:

- We demonstrate that model uncertainty and human disagreement are independent phenomena in subjective tasks, establishing population disagreement as the principled signal for personalization.
- We empirically show that ICL-based personalization degrades performance even below zero-shot baselines when subjective tasks diverge from pre-training distributions, challenging the prevailing reliance on in-context examples.
- We propose PGBC, a training-free calibration method

*Corresponding author
Code: https://github.com/hoesungryu/Test-Time_User_Alignment

that achieves consistent improvement across diverse foundation models and subjective benchmarks while maintaining constant inference cost.

2 Related Works

2.1 Subjective Tasks and Annotation Ambiguity

Unlike objective classification tasks, subjective tasks—such as affect estimation, safety rating, or hate speech detection—inherently lack a single ground truth [Aroyo and Welty, 2015]. Variation in human judgment here reflect genuine differences in individual perspectives rather than annotation noise [Basile *et al.*, 2021; Larimore *et al.*, 2021]. Traditional approaches aggregate conflicting labels via majority voting [Snow *et al.*, 2008] or label smoothing [Szegedy *et al.*, 2016], suppressing minority perspectives. This limitation persists in foundation models, where alignment techniques such as RLHF homogenize outputs toward population-averaged preferences [Santurkar *et al.*, 2023].

Recent work advocates preserving disagreement as an intrinsic modeling signal [Basile *et al.*, 2021], yet practical mechanisms for aligning models with individual subjective standards remain underexplored. Unlike extensive research on factual confidence calibration [Desai and Durrett, 2020], personalized alignment typically requires costly parameter updates [Jang *et al.*, 2023], underscoring the need for efficient test-time frameworks.

2.2 Personalization in Foundation Models

Foundation models demonstrate impressive zero-shot performance across diverse tasks, yet their tendency to produce population-averaged predictions limits their ability to reflect individual user preferences. Parameter-Efficient Fine-Tuning (PEFT) methods such as LoRA [Hu *et al.*, 2021] and Prefix-Tuning [Li and Liang, 2021] have been employed to address this limitation [Houlsby *et al.*, 2019]. However, these approaches require maintaining user-specific parameters that scale linearly with the user base [Sheng *et al.*, 2023], suffer from cold-start issues [Lin and others, 2024], and necessitate auxiliary training for preference mapping [Geng *et al.*, 2022].

In-Context Learning (ICL) offers an alternative by incorporating user history or personas directly into the prompt without parameter updates [Brown *et al.*, 2020]. However, ICL relies on the assumption that models can implicitly infer user-specific patterns from limited demonstrations—an assumption that fails when tasks diverge from pre-training distributions, where ICL can degrade performance below zero-shot baselines [Zhao *et al.*, 2021].

2.3 From Calibration to User Alignment

Standard calibration methods such as temperature scaling [Guo *et al.*, 2017] and deep ensembles [Lakshminarayanan *et al.*, 2017] aim to align model confidence with predictive accuracy, assuming a unique ground truth exists for each input. However, this assumption fails in subjective tasks where ground truth inherently varies across individuals. Calibrating toward aggregate accuracy suppresses user-specific nuances, forcing models toward population-averaged predictions [Davani *et al.*, 2022].

Dataset	Task	Users	Items	Labels
iNews	Affect Estimation	200	1,500	$[1, 7]$
DICES	Safety Rating	123	350	$[0, 2]$
MHS	Hate Speech	39,565	7,912	$[0, 4]$

Table 1: Datasets of highly-variable subjective used here, spanning affective and safety evaluation domains where individual perspectives lead to legitimate disagreement on identical items.

Our work addresses these limitations by shifting from population-level calibration to user-level alignment. Rather than assuming a single global truth, we explicitly model individual preferences through Bayesian inference, employing human disagreement rather than model uncertainty as the weighting signal for principled test-time personalization. This user-level alignment approach is closely related to Linear Mixed Models with fixed item effects [Krajewski and Matthews, 2010; Laird and Ware, 1982], which descend from Item Response Theory [Rasch, 1993].

3 Methods

3.1 Highly Subjective Datasets

We evaluate on three benchmarks—iNews [Hu and Collier, 2025], DICES [Aroyo *et al.*, 2023], and Measuring Hate Speech (MHS) [Kennedy *et al.*, 2020]—where annotations reflect personal beliefs and normative interpretations in the absence of objective ground truth. Such datasets exhibit substantial inter-user disagreement, yet relatively stable intra-user response patterns, making them ideal testbeds for test-time subjective alignment (see Table 1).

First, iNews contains news headlines annotated with valence and arousal scores on a 7-point Likert scale. The same headline frequently elicits divergent affective reactions across users due to differences in personal experience and emotional sensitivity. We predict the two dimensions independently and report averaged performance.

Second, DICES focuses on AI safety, providing rater judgments on model-generated responses across diverse conversational contexts. Labels (unsafe, ambiguous, safe) reflect individual interpretations of risk, harm, and policy boundaries rather than a single canonical notion of safety.

Third, MHS contains social media comments annotated across ten ordinal dimensions (e.g., sentiment, insult, dehumanization, violence) that are aggregated into a continuous hate speech score. Perceptions of hate speech vary systematically across individuals based on personal values and demographic background, making MHS a canonical benchmark for studying subjective disagreement rather than annotation noise.

3.2 Experimental Setup

Foundation Models

To demonstrate model-agnostic generalizability, we evaluate PGBC across four representative open-weights foundation models: (1) LLaMA-3.1-8B [Grattafiori *et al.*, 2024], (2) Mistral-7B [Jiang *et al.*, 2023], (3) Qwen3-8B [Yang *et al.*, 2025], and (4) Gemma-2-9B [Team *et al.*, 2024]. These models are selected from distinct sources (Meta,

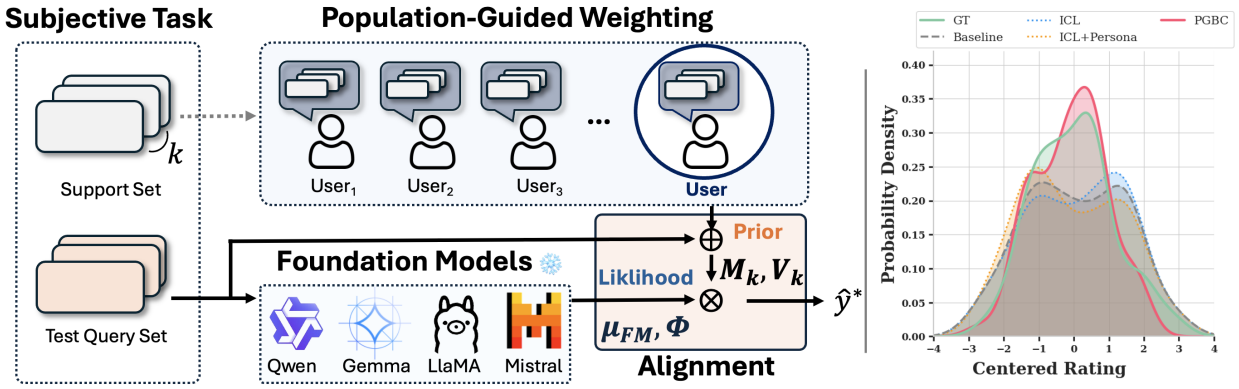

Figure 1: Overview of PGBC. On the left, given a subjective task, the support set (K items with user ratings) and test query are processed through FMs to obtain semantic features μ_{FM}. PGBC learns user-specific parameters via Bayesian inference, where population disagreement across users guides the likelihood weighting—items with high inter-user variance contribute less to parameter estimation. On the right, aggregated prediction distributions on DICES dataset. While Baseline (gray) exhibits bias from the ground-truth distribution (green), and ICL-based methods (blue, orange) fail to correct this misalignment, PGBC (red) achieves the closest alignment with individual user preferences.

Mistral AI, Alibaba, Google) to ensure diversity in training paradigms, and represent top-performing models in the 7B–9B regime [Myrzakhan *et al.*, 2024]—a practical balance of efficiency and capability for single-GPU deployment [Dettmers *et al.*, 2023].

Evaluation Protocol

In real-world personalization, new users typically possess limited interaction history (cold-start). Since the selection of informative context is critical when data is scarce [Rubin *et al.*, 2022; Zhang *et al.*, 2022], we focus our evaluation on the few-shot regime ($K < 10$) [Brown *et al.*, 2020; Finn *et al.*, 2017], covering scenarios from extreme cold-start to moderate data availability.

We employ a leave-one-out cross-validation protocol. For each user, every rated item is iteratively held out as the test instance. From the remaining history, we construct the support set by selecting exactly K examples with the highest population disagreement. This active sampling strategy prioritizes the most ambiguous—and thus informative—items to elicit the user's specific subjective boundary. For ordinal rating prediction, we report Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) averaged across all users and test instances. To ensure strict data hygiene, all population statistics ($\mu_{\text{pop}}, s_{\text{pop}}$) are computed via a closed-form leave-one-out update that excludes the target user's rating on each held-out item, preventing any target-side leakage into the support weighting.

Prompt Engineering

All foundation models are used in their instruction-tuned variants to ensure output format consistency. As shown in Figure 2, we employ a unified prompt structure with four modular components: (A) a task-specific system instruction that defines the evaluation domain, (B) an optional user profile schema for persona-based personalization, (C) K in-context exemplars as text-label pairs actively sampled from user history, and (D) an output format specification that enforces strict JSON formatting for reliable automated parsing. While the overall structure remains consistent across all experiments, task-specific elements such as evaluation criteria and label scales are adapted to each dataset.

Comparison Methods

To validate the effectiveness of our proposed PGBC, we construct a series of comparison methods that progressively incorporate personalization components (detailed prompt structure in Figure 2).

We begin with FM (Baseline), which directly uses the foundation model prediction μ_{FM} for each item without any calibration or personalization. We set this as our baseline to represent the default behavior of foundation models and to quantify the benefit of personalization.

For all subsequent methods, we employ active sampling to select K examples from the user's history. Following prior work on example selection for in-context learning [Rubin *et al.*, 2022; Zhang *et al.*, 2022], we employ active sampling to select informative examples. Unlike similarity-based approaches, we prioritize items with high population disagreement s_{pop}, as these items better reveal individual deviations from the population consensus.

Building upon this sampling strategy, ICL-Base provides the selected K examples as part of the input prompt, allowing the model to implicitly infer user preferences from demonstrated input-output patterns. To further enhance prediction accuracy for subjective tasks, ICL-Persona additionally injects explicit user profile information into the system context, guiding the model with persona-aware reasoning. However, these ICL-based approaches have notable limitations. There is no mechanism to verify the quality of selected examples, and the optimal number of examples remains underexplored. Moreover, it is difficult to guarantee that the model correctly utilizes the provided persona information. Both ICL variants also incur $\mathcal{O}(K \times L)$ token processing cost per prediction, where L denotes the average tokens per example.

In contrast, our proposed PGBC learns a user-specific affine transformation $\hat{y} = \alpha \cdot \mu_{\text{FM}} + \beta$ through Bayesian in-

[System Context]

(A) TASK: `You are evaluating the <<{emotional response | safety | dehumanize}>>.`

(B) USER: `The following user profile schema is {age, sex, personality}.`

[In-Context Examples]

(C) HISTORY: K `examples` $\{(x_j, y_j)\}_{j=1}^{K}$ `actively sampled from user history.`

[Query]

(D) PREDICT: `Rate the following. Respond ONLY with JSON <<{"V":int,"A":int} | {"safety score": int} | {"dehumanize score": int}>> .`

Figure 2: Prompt structure for comparison methods with four modular components. Each method uses a different combination of components for input prompt. Baseline and PGBC use only (A) (D), while ICL-Base and Ensemble add in-context examples using (A) (C) (D). ICL-Persona additionally incorporates user profile information with (A) (B) (C) (D). Task-specific alternatives within each template are separated by '|' symbols.

ference, where s_{pop} is used for likelihood weighting to assign lower confidence to highly subjective items. Once calibrated, PGBC achieves $\mathcal{O}(1)$ inference cost, offering significant efficiency over ICL's $\mathcal{O}(K \times L)$ token processing.

3.3 Population-Guided Bayesian Calibration

Foundation models typically produce generic scores reflecting population-averaged preferences, often failing to capture individual subjectivity. We propose Population-Guided Bayesian Calibration (PGBC) to personalize these outputs by treating Baseline scores not as direct predictions, but as semantic covariates within a Bayesian framework. Algorithm 1 summarizes the inference procedure.

Subjective Rating Model

We assume a user u's subjective rating $y_u(x)$ as a personalized affine transformation of the FM output $\mu_{\text{FM}}(x)$

$$y_u(x) = \alpha_u \cdot \mu_{\text{FM}}(x) + \beta_u + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

where α_u represents the user's sensitivity relative to the FM scale and β_u captures their systematic bias. Since $\mu_{\text{FM}}(x)$ provides an anchor reflecting population-level semantics, estimating $\boldsymbol{\theta}_u = [\alpha_u, \beta_u]^\top$ allows us to map the generic score onto the user's personalized preference.

Prior

We construct a prior that defaults to the FM prediction when user evidence is absent (Alg. 1, Lines 2–3)

$$p(\boldsymbol{\theta}_u) = \mathcal{N}(\boldsymbol{\theta}_u \mid \mathbf{m}_0, \mathbf{V}_0), \quad \mathbf{m}_0 = [1, 0]^\top \tag{2}$$

Algorithm 1 Population-Guided Bayesian Calibration (PGBC)

Require: User history $\mathcal{D}_u = \{(x_i, y_i)\}_{i=1}^{K}$, query item x^*, FM outputs $\mu_{\text{FM}}(\cdot)$
Require: Population disagreement $\{s_i\}_{i=1}^{K}$, hyperparameters $\sigma_{\text{base}}, \sigma_\alpha, \sigma_\beta$
Ensure: Calibrated prediction $\hat{y}^*$

1: **Step 1. Prior and Design Matrix**
2: $\mathbf{m}_0 \leftarrow [1, 0]^\top$; $\mathbf{V}_0 \leftarrow \text{diag}(\sigma_\alpha^2, \sigma_\beta^2)$
3: $\boldsymbol{\Phi} \leftarrow [\boldsymbol{\mu}, \mathbf{1}] \in \mathbb{R}^{K \times 2}$ where $\boldsymbol{\mu} = [\mu_{\text{FM}}(x_1), \ldots, \mu_{\text{FM}}(x_K)]^\top$
4: **Step 2. Population-Guided Weighting**
5: $\mathbf{W} \leftarrow \text{diag}\left((\sigma_{\text{base}}^2 + s_1^2)^{-1}, \ldots, (\sigma_{\text{base}}^2 + s_K^2)^{-1}\right)$
6: **Step 3. Bayesian Update for Align**
7: $\mathbf{V}_K \leftarrow \left(\mathbf{V}_0^{-1} + \boldsymbol{\Phi}^\top \mathbf{W} \boldsymbol{\Phi}\right)^{-1}$
8: $\mathbf{m}_K \leftarrow \mathbf{V}_K \left(\mathbf{V}_0^{-1} \mathbf{m}_0 + \boldsymbol{\Phi}^\top \mathbf{W} \mathbf{y}\right)$
9: **Step 4. Prediction**
10: **return** $\hat{y}^* = \mathbf{m}_K^\top [\mu_{\text{FM}}(x^*), 1]^\top$

Setting the prior mean to $\alpha = 1$ and $\beta = 0$ implies that we trust the FM output as the best initial guess for any user. This prior acts as an identity anchor, ensuring that with insufficient data the posterior shrinks toward the original FM prediction and prevents overfitting to sparse observations.

Likelihood

Not all observations are equally reliable for estimating user parameters. Items with high population disagreement are inherently noisy targets, as they reflect items where subjective interpretation varies substantially across users. We incorporate this insight through a heteroscedastic likelihood (Alg. 1, Lines 4–5)

$$p(y_j \mid \boldsymbol{\theta}_u, x_j) = \mathcal{N}(\alpha_u \mu_j + \beta_u, \sigma_j^2), \quad \sigma_j^2 = \sigma_{\text{base}}^2 + s_j^2 \tag{3}$$

where $s_j = s_{\text{pop}}(x_j)$ denotes the standard deviation of ratings for item x_j across all users. High s_j increases the noise variance σ_j^2, thereby reducing the item's contribution to parameter estimation. Intuitively, when population opinions diverge on an item, any individual rating on that item carries less information about the user's underlying preference.

Posterior

Combining prior and likelihood via Bayes' rule yields a Gaussian posterior due to conjugacy (Alg. 1, Lines 6–8)

$$p(\boldsymbol{\theta}_u \mid \mathcal{D}_u) = \mathcal{N}(\mathbf{m}_K, \mathbf{V}_K) \tag{4}$$

with the closed-form posterior mean

$$\mathbf{m}_K = \mathbf{V}_K \left(\mathbf{V}_0^{-1} \mathbf{m}_0 + \boldsymbol{\Phi}^\top \mathbf{W} \mathbf{y}\right) \tag{5}$$

This formulation naturally handles varying levels of data availability. When $K = 0$, the posterior equals the prior $\mathbf{m}_0 = [1, 0]^\top$, and predictions default to the uncalibrated FM output. With small K, the posterior remains close to the prior due to shrinkage, providing robustness against overfitting. As K grows, the posterior shifts toward data-driven estimates, enabling full personalization.

Dataset	Signal	Independence $r(\tau, s)$	Predictive Power $r(\cdot, \lvert y - \mu_{\text{pop}} \rvert)$
iNews	τ: LLaMA3.1-8B	$-0.06^{**} \pm 0.04$	$-0.07^{***} \pm 0.02$
	τ: Mistral-7B	$+0.04 \pm 0.04$	$+0.00 \pm 0.02$
	τ: Qwen3-8B	$-0.06^{**} \pm 0.04$	$-0.07^{***} \pm 0.02$
	τ: Gemma2-9B	$+0.04 \pm 0.04$	-0.00 ± 0.02
	s_{pop}	—	$\mathbf{+0.65}^{***} \pm 0.01$
DICES	τ: LLaMA3.1-8B	-0.05 ± 0.10	$+0.15^{***} \pm 0.01$
	τ: Mistral-7B	$+0.05 \pm 0.10$	$+0.02^{***} \pm 0.01$
	τ: Qwen3-8B	$+0.01 \pm 0.10$	$-0.05^{***} \pm 0.01$
	τ: Gemma2-9B	-0.01 ± 0.10	-0.01 ± 0.01
	s_{pop}	—	$\mathbf{+0.44}^{***} \pm 0.01$
MHS	τ: LLaMA3.1-8B	$+0.00 \pm 0.01$	$+0.12^{***} \pm 0.01$
	τ: Mistral-7B	$+0.00 \pm 0.01$	$+0.03^{***} \pm 0.01$
	τ: Qwen3-8B	-0.01 ± 0.01	$+0.05^{***} \pm 0.01$
	τ: Gemma2-9B	-0.01 ± 0.01	$+0.02^{***} \pm 0.01$
	s_{pop}	—	$\mathbf{+0.67}^{***} \pm 0.00$

Table 2: Model uncertainty (τ_{FM}) versus population disagreement (s_{pop}). Left: Correlation between the two measures. Right: Correlation with individual deviations from population mean $|y - \mu_{\text{pop}}|$. Values shown as $r \pm \delta$ where δ is 95% CI half-width. $^{**}p<.01$, $^{***}p<.001$.

4 Experiment Results

4.1 Uncertainty in Subjective Tasks

Subjective tasks inherently involve uncertainty from multiple sources. We first verify the existence of model uncertainty and human disagreement in our data, then examine whether these two phenomena capture the same underlying signal.

Existence of Model Uncertainty and Human Disagreement

We measure model uncertainty via temperature sampling ($T = 0.7$) [Renze, 2024] with 10 stochastic forward passes per item. Across all 12 model-dataset combinations, sampling variance (τ_{FM}) is significantly greater than zero (one-sample t-test, $p < 0.001$; Cohen's d = 0.27–2.50). On the human side, population disagreement (s_{pop}), measured as the standard deviation of annotator ratings per item, is also significantly greater than zero across all three datasets (t-test, $p < 0.001$; Cohen's d = 1.01–6.48). Both sources of uncertainty are thus present and substantial in subjective tasks.

Independence and Predictive Power

Given that both model uncertainty and human disagreement exist, a natural question arises: do they capture the same underlying phenomenon? If so, model uncertainty could serve as a proxy for human disagreement in personalization.

Table 2 (left) shows the correlation between τ_{FM} and s_{pop}. The mean correlation is negligible ($r = -0.01$), with only 2 of 12 settings reaching significance. This near-zero correlation demonstrates that high sampling variance does not correspond to items where humans disagree—the two measures capture fundamentally different phenomena.

Since model uncertainty and human disagreement are independent, we must separately evaluate which better predicts individual deviations from the population mean. Table 2 (right) compares their predictive power for $|y - \mu_{\text{pop}}|$. Model uncertainty achieves negligible correlation (mean $r = +0.02$), while population disagreement yields strong effects (mean $r = +0.59$). Williams' test confirms s_{pop} significantly outperforms τ_{FM} in all 12 settings ($p < 0.001$); bootstrap 95% CIs exclude zero in all cases. These results establish s_{pop}—not τ_{FM}—as the appropriate weighting signal for personalization.

Comparison	MAE $\Delta \pm$ CI	MAE d	RMSE $\Delta \pm$ CI	RMSE d
B. vs ICL	$-.054 \pm .043^{*}$	$-.33$	$-.068 \pm .033^{***}$	$\mathbf{-.53}$
B. vs ICL+Per.	$-.053 \pm .042^{*}$	$-.33$	$-.058 \pm .034^{**}$	$-.44$
B. vs PGBC	$+.080 \pm .021^{***}$	$\mathbf{+1.01}$	$+.099 \pm .022^{***}$	$\mathbf{+1.21}$
ICL vs PGBC	$+.134 \pm .036^{***}$	$\mathbf{+0.97}$	$+.167 \pm .031^{***}$	$\mathbf{+1.40}$
ICL+Per. vs PGBC	$+.133 \pm .036^{***}$	$\mathbf{+0.96}$	$+.157 \pm .032^{***}$	$\mathbf{+1.28}$

Table 3: Pairwise comparisons via paired t-tests. Positive Δ indicates the second method achieves lower error. Values after $\pm$ denote half-widths of 95% CIs. Significance: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$. Bold d: large effect ($|d| \geq 0.5$). B.: Baseline; Per.: Persona.

4.2 Comparison Results

Quantitative Comparison

We conduct pairwise statistical comparisons across all 60 conditions (4 models $\times$ 3 datasets $\times$ 5 K values), summarized in Table 3. PGBC achieves significant improvements over all baselines with large effect sizes. Compared to the Zero-shot baseline, PGBC shows $\Delta = +0.099$ in normalized RMSE ($p < .001$, Cohen's $d = 1.21$), and the improvement is even more pronounced against ICL with $\Delta = +0.167$ ($p < .001$, $d = 1.40$). Notably, ICL-based methods show *negative* improvement over the baseline, with $\Delta = -0.068$ for ICL ($d = -0.53$) in normalized RMSE, confirming that in-context learning can harm rather than help personalization in low-data regimes. The comparison between ICL and ICL+Persona shows negligible difference ($p = 0.21$), indicating that persona augmentation provides no measurable benefit. To assess model-agnostic effectiveness, we conduct a Friedman test across the four foundation models, which yields $\chi^2 = 3.24$ and $p = 0.356$, suggesting that PGBC provides consistent improvements regardless of the underlying foundation model.

Methodological Interpretation

Dataset-specific patterns reveal why PGBC succeeds where ICL fails (Figure 3). On iNews, where affect estimation aligns with foundation model pre-training objectives, all methods show gradual improvement as K increases, with ICL achieving 0.818 at K=9. However, on DICES and MHS, which involve safety judgments and dehumanization scoring that are underrepresented in pre-training, ICL-based methods fail to improve over the baseline. Most strikingly, on MHS, ICL achieves a normalized RMSE of 1.24 at K=1, *degrading* performance by 24%. Even at K=9, ICL remains above 1.0 on both datasets. PGBC, in contrast, maintains consistent improvement across all datasets, reaching normalized RMSE below 0.90 at K=9.

The contrasting performance stems from how each approach utilizes user information. ICL embeds user examples

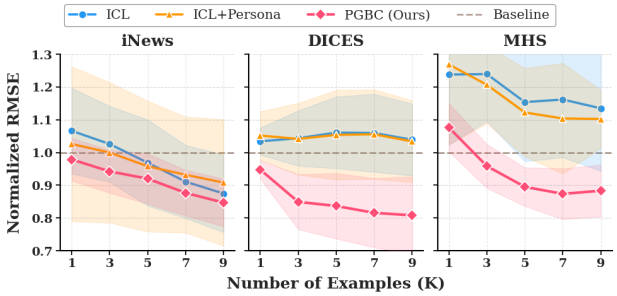

Figure 3: Normalized RMSE by dataset. Shaded regions indicate 95% confidence intervals. ICL effectiveness varies substantially across tasks, while PGBC maintains consistent improvement.

directly into the prompt, requiring the foundation model to infer user-specific patterns from limited demonstrations. This assumes the model can distinguish between general task patterns and individual user biases, an assumption that fails when the task is unfamiliar. On DICES and MHS, the model lacks the representational basis to interpret what user examples signify, leading to confusion rather than adaptation. The persona augmentation provides minimal benefit ($p = 0.21$), suggesting that verbal descriptions do not compensate for this fundamental limitation.

PGBC circumvents this by treating foundation model outputs as semantic features rather than direct predictions. The Bayesian calibration learns an explicit mapping from model outputs to user preferences without requiring in-context inference. This decoupling, separating what the model predicts from how the user differs, explains PGBC's robustness across tasks with varying pre-training alignment.

The divergent scaling behaviors reveal how each method leverages additional examples. For ICL, increasing K does not change the inference mechanism, and on unfamiliar tasks, additional examples amplify confusion rather than clarify preferences. For PGBC, each example directly refines the posterior over calibration parameters, with Bayesian updates reducing uncertainty monotonically. The steepest improvement occurs between K=1 and K=3 ($\Delta = 0.057$), demonstrating that minimal feedback yields substantial gains, a property particularly valuable in cold-start scenarios.

4.3 User-Level Distribution Alignment

To provide qualitative insights into how PGBC achieves personalization, we visualize the prediction distributions for representative users across all three datasets (Figure 4). For each user, we center all predictions at the user's ground-truth mean, enabling direct comparison of distribution shapes and alignment quality. We quantify distributional similarity using Wasserstein Distance (WD), where lower values indicate better alignment with individual user preferences.

Distributional Mismatch in Foundation Models

Across all datasets, Baseline predictions (gray dashed) consistently fail to align with the individual ground-truth distributions (green). The severity varies substantially across users: in iNews column 1, Baseline shows WD=1.30 with a clear rightward shift, while in column 3, Baseline achieves WD=0.10 with near-perfect alignment. This heterogeneity demonstrates that Baseline predictions work well for users whose preferences coincidentally align with population averages, but fail significantly for users with distinctive rating patterns.

A similar pattern appears in DICES, where the discrete 3-point scale creates multimodal ground-truth distributions. Baseline tends to produce smoother, less peaked distributions that fail to capture user-specific rating concentrations. In MHS column 1, Baseline (WD=0.70) shows both location shift and shape mismatch relative to the bimodal ground-truth.

Limitations of In-Context Learning

Providing user history as in-context examples does not reliably improve distribution alignment. Across all data, ICL-based methods frequently fail to improve upon Baseline, and in many cases perform worse.

In iNews, ICL (blue dotted) consistently achieves higher WD than Baseline across all four users (column 1: ICL=1.50 vs Baseline=1.30; column 4: ICL=1.70 vs Baseline=0.50). The ICL distributions appear more dispersed and shifted, suggesting that few-shot examples introduce noise rather than enabling calibration.

ICL+Persona (orange dotted) shows modest improvements over vanilla ICL but remains inconsistent. While it achieves the best result in DICES column 1 (WD=0.40, tied with PGBC's 0.20 for improvement over Baseline), it underperforms Baseline in other cases (DICES column 3: ICL+P=0.60 vs Baseline=0.40). This inconsistency limits the practical utility of persona-based approaches.

Per-User Alignment via PGBC

PGBC (red solid) achieves the best distribution alignment in cases where Baseline exhibits significant bias. The most pronounced improvements occur in iNews columns 1 and 2, where PGBC reduces WD from 1.30 to 0.10 and from 0.40 to 0.20, respectively. In these cases, the PGBC distribution closely overlaps with the ground-truth, capturing both location and spread. In DICES column 1, PGBC achieves WD=0.20, successfully aligning the primary mode while Baseline (WD=0.60) remains substantially misaligned. Across MHS, PGBC shows consistent improvements, demonstrating that the affine transformation effectively captures user-specific biases when the ground-truth distribution is approximately unimodal.

However, PGBC's affine transformation has inherent limitations—it preserves the relative structure of Baseline outputs while adjusting scale and shift, which means it cannot transform a unimodal Baseline prediction into a multimodal distribution. This limitation is visible in DICES columns 2 to 4, where both Baseline and PGBC show similar moderate performance on users with complex multimodal preferences.

There are also cases where PGBC improvement is minimal or where Baseline already performs well. In iNews column 3, Baseline achieves WD=0.10 while PGBC shows WD=0.30—here the user's preferences already align with population averages, making calibration unnecessary. Importantly, even in these cases, PGBC does not catastrophically degrade; the Bayesian prior anchors predictions to the Baseline when user-specific evidence provides limited signal for improvement.

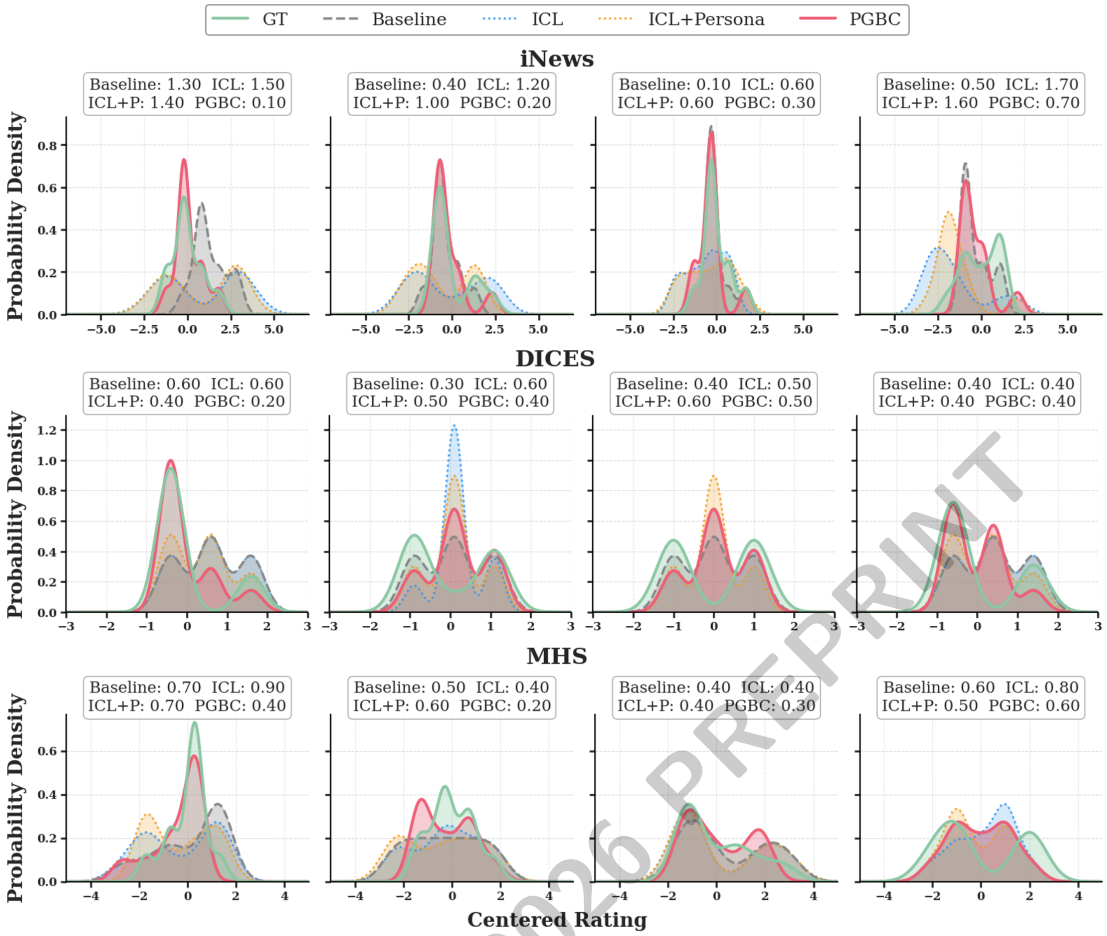

Figure 4: Per-user KDE distribution alignment across three datasets with $K = 9$ examples. Each row corresponds to a dataset (iNews, DICES, MHS), and each column shows a representative user selected based on PGBC improvement (from best (left) to bad (right) cases). All distributions are centered at each user's ground-truth (GT) mean for fair comparison. Inset boxes report Wasserstein Distance (WD) between each method's predictions and the ground-truth distribution (lower is better).

5 Conclusion

We introduced Population-Guided Bayesian Calibration (PGBC), a test-time user alignment method for subjective tasks. Using foundation model zero-shot prediction as the baseline, we compared against ICL-based methods including persona-augmented variants. Our empirical analysis reveals that model uncertainty and human disagreement capture fundamentally different phenomena, with only the latter predicting individual deviations from the population mean. We further demonstrate that existing ICL-based methods not only fail to improve personalization but can degrade performance below zero-shot baselines on subjective tasks that diverge from pre-training distributions. This occurs because few-shot examples introduce confusion rather than enabling adaptation when the model lacks the representational basis to interpret user-specific patterns.

In contrast, PGBC does not require the model to implicitly infer preferences from in-context examples. Instead, it quantifies individual deviations from the population mean using population disagreement and combines this with foundation model outputs. This design ensures that predictions safely default to the foundation model when user evidence is sparse, while maintaining consistent effectiveness across diverse subjective tasks and foundation models. These results demonstrate practical applicability in cold-start scenarios where minimal user feedback must yield maximal personalization.

Future work includes extending PGBC to mixture models to relax the unimodality constraint of scale-and-shift calibration, and applying the framework to classification tasks through appropriate link functions.

Acknowledgments

This work was supported by the Ministry of Science and ICT (MSIT), Republic of Korea, through the AI Computing Infrastructure GPU Rental Support Program (RQT-25-110041); the National Research Foundation of Korea (BK21 FOUR; RS-2025-00555141); and IITP grants funded by MSIT (RS-2019-II190079; RS-2021-II212068; RS-2026-25522672).

References

[Aroyo and Welty, 2015] Lora Aroyo and Chris Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24, 2015.

[Aroyo *et al.*, 2023] Lora Aroyo, Alex Taylor, Mark Diaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. Dices dataset: Diversity in conversational ai evaluation for safety. *Advances in Neural Information Processing Systems*, 36:53330–53342, 2023.

[Basile *et al.*, 2021] Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, Alexandra Uma, et al. We need to consider disagreement in evaluation. In *Proceedings of the 1st workshop on benchmarking: past, present and future*, pages 15–21. Association for Computational Linguistics, 2021.

[Brown *et al.*, 2020] 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.

[Davani *et al.*, 2022] Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022.

[Desai and Durrett, 2020] Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. *arXiv preprint arXiv:2003.07892*, 2020.

[Dettmers *et al.*, 2023] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

[Finn *et al.*, 2017] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[Geng *et al.*, 2022] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *Proceedings of the 16th ACM Conference on Recommender Systems*, pages 299–315, 2022.

[Grattafiori *et al.*, 2024] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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

[He *et al.*, 2017] Jia He, Fons JR Van de Vijver, Velichko H Fetvadjiev, Alejandra de Carmen Dominguez Espinosa, Byron Adams, Itziar Alonso-Arbiol, Arzu Aydinli-Karakulak, Carmen Buzea, Radosveta Dimitrova, Alvaro Fortin, et al. On enhancing the cross–cultural comparability of likert–scale personality and value measures: A comparison of common procedures. *European Journal of Personality*, 31(6):642–657, 2017.

[Houlsby *et al.*, 2019] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[Hu and Collier, 2025] Tiancheng Hu and Nigel Collier. inews: A multimodal dataset for modeling personalized affective responses to news. *arXiv preprint arXiv:2503.03335*, 2025.

[Hu *et al.*, 2021] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[Jang *et al.*, 2023] Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*, 2023.

[Jiang *et al.*, 2023] 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, 2023.

[Jonas and Markon, 2019] Katherine G Jonas and Kristian E Markon. Modeling response style using vignettes and person-specific item response theory. *Applied Psychological Measurement*, 43(1):3–17, 2019.

[Kennedy *et al.*, 2020] Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*, 2020.

[Krajewski and Matthews, 2010] Grzegorz Krajewski and Danielle Matthews. Rh baayen, analyzing linguistic data: A practical introduction to statistics using r. cambridge: Cambridge university press, 2008. pp. 368. isbn-13: 978-0-521-70918-7. *Journal of Child Language*, 37(2):465–470, 2010.

[Laird and Ware, 1982] Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

[Lakshminarayanan *et al.*, 2017] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scal-

able predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[Larimore *et al.*, 2021] Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90, 2021.

[Li and Liang, 2021] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

[Lin and others, 2024] Zhaoxuan Lin et al. User-llm: Efficient llm contextualization with user embeddings. In *arXiv preprint arXiv:2402.13598*, 2024.

[Myrzakhan *et al.*, 2024] Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. Open-llm-leaderboard: From multi-choice to open-style questions for llms evaluation, benchmark, and arena. *arXiv preprint arXiv:2406.07545*, 2024.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[Rasch, 1993] Georg Rasch. *Probabilistic models for some intelligence and attainment tests.* ERIC, 1993.

[Renze, 2024] Matthew Renze. The effect of sampling temperature on problem solving in large language models. In *Findings of the association for computational linguistics: EMNLP 2024*, pages 7346–7356, 2024.

[Rubin *et al.*, 2022] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 2655–2671, 2022.

[Salemi *et al.*, 2024] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. Lamp: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392, 2024.

[Santurkar *et al.*, 2023] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR, 2023.

[Sheng *et al.*, 2023] Ying Sheng, Shiyi Cao, Dacheng Li, Coleman Hooper, Nicholas Lee, Shuo Yang, Christopher Chou, Banghua Zhu, Lianmin Zheng, Kurt Keutzer, et al. S-lora: Serving thousands of concurrent lora adapters. *arXiv preprint arXiv:2311.03285*, 2023.

[Snow *et al.*, 2008] Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263, 2008.

[Stanovich and West, 1998] Keith E Stanovich and Richard F West. Individual differences in rational thought. *Journal of experimental psychology: general*, 127(2):161, 1998.

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

[Team *et al.*, 2024] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[Uma *et al.*, 2021] Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470, 2021.

[Yang *et al.*, 2025] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[Zhang *et al.*, 2022] Yiming Zhang, Shi Feng, and Chenhao Tan. Active example selection for in-context learning. *arXiv preprint arXiv:2211.04486*, 2022.

[Zhao *et al.*, 2021] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR, 2021.