CITP: CONTEXT-INDUCED TILT POSTERIOR

Anonymous authors
Paper under double-blind review

ABSTRACT

Predictive distributions of large language models (LLMs) remain sensitive even to semantically equivalent prompts. However, the variation in predictions across these semantically equivalent prompts can also provide useful information for predictive refinement. We propose **CITP (Context-Induced Tilt Posterior)**, a Bayesian method for refining a frozen LLM's predictive distribution using its logit responses across semantically equivalent prompts. Using a small calibration set of labeled examples from the target task, CITP learns a posterior over how to refine the model's predictions using its responses to semantically equivalent prompts. It then averages predictive distributions over this posterior to account for uncertainty in the learned refinement. Across three model families and nine tasks spanning fixed-label classification and multiple-choice question answering (MCQA), CITP achieves the best macro-averaged normalized Brier score and normalized negative log-likelihood in the main comparison, with relative reductions of 9.5% and 20.2%, respectively, compared with the strongest baseline for each metric. Further analyses show that learning from how predictions change across semantically equivalent prompts can improve predictive performance, and that averaging over the learned posterior provides additional gains.

1 INTRODUCTION

LLMs have been used in a broad array of downstream applications based on their strong performance and versatility (Brown et al., 2020; Bommasani et al., 2021). However, the predictions of LLMs appear to be sensitive to the surface-level format of the prompt. In particular, changes to prompts that preserve task-relevant information, such as the order of the demonstrations and the tokens used for output labels, can also change the class probabilities the model outputs and sometimes even the argmax prediction (Zhao et al., 2021; Lu et al., 2022; Chen et al., 2023; Cox et al., 2025). We refer to prompt variants that preserve task-relevant content and candidate semantics while changing such surface properties as *task-equivalent prompt variants*.

Prompt-induced variation has also been exploited as a useful source of information rather than treated solely as a failure mode. Prior work has aggregated predictive distributions obtained from multiple prompt realizations to improve prediction (Jiang et al., 2023b; Tonolini et al., 2024; Li et al., 2025), while other studies have analyzed cross-prompt sensitivity or disagreement as signals of predictive reliability or uncertainty (Chen et al., 2023; Cox et al., 2025). In particular, aggregating predictions across semantically equivalent prompt realizations can substantially improve over a single-prompt prediction (Jiang et al., 2023b; Cox et al., 2025), and even outperform the best single prompt selected on a validation set (Tonolini et al., 2024). Motivated by these findings, we ask the following research question: can observed reference-relative responses to task-equivalent prompt variations serve as query-specific correction directions to improve a frozen LLM's predictive distribution, and can accounting for uncertainty in the refinement through Bayesian inference yield further improvements?

To this end, we propose **CITP** (Context-Induced Tilt Posterior). For each query, CITP represents responses to multiple task-equivalent prompts as reference-relative logit contrasts, allowing these response vectors to be linearly combined into an additive correction. Using a calibration set of labeled examples from the target task, CITP jointly learns a coefficient vector. This vector comprises a global scaling coefficient (Scale) for adjusting predictive confidence, additive bias parameters (Bias) for shifting candidate logits, and coefficients for prompt-contrast refinement. This coefficient vector is shared across queries within a task, while the prompt contrasts determine the query-specific adjustment. CITP then approximates the posterior over this vector using a Laplace approximation (Daxberger

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

et al., 2021) and averages the corresponding predictive distributions over this posterior, propagating refinement uncertainty into the final predictive distribution.

To examine the additional predictive value of refinement using prompt-contrast directions, we analyze the point-predictor family. We first theoretically characterize when refinement using prompt-contrast directions can reduce population cross-entropy risk beyond optimal scale-and-bias correction and derive a lower bound on the available improvement. Empirically, we compare the maximum a posteriori (MAP) predictor from this family against scale-and-bias correction to assess the additional value of using prompt-contrast directions. We then compare the MAP predictor with the posterior predictive from the same fit to assess the effect of marginalizing uncertainty over the refinement coefficients.

Overall, CITP significantly reduces macro-averaged normalized Brier score (nBrier) and normalized negative log-likelihood (nNLL), our primary evaluation metrics, by 0.0432 and 0.1145, respectively, relative to the strongest evaluated baseline for each metric in the main comparison. Controlled comparisons support the additional predictive value of query-matched prompt contrasts beyond the Scale+Bias control, with a point-prediction diagnostic linking stronger residual alignment to larger NLL gains over the same control, consistent with our theoretical analysis. Further, comparisons with point predictions from the same fit show that marginalizing refinement uncertainty improves nBrier and nNLL, with larger NLL gains at smaller calibration budgets. NLL gains tend to be larger for queries with greater disagreement among predictions from the coefficient posterior, a measure of uncertainty about the learned refinement.

Our contributions are as follows:

- We propose CITP, a Bayesian method for frozen LLMs that uses responses to task-equivalent prompts to define query-specific correction directions while accounting for uncertainty in the learned refinement.
- We derive a residual-alignment criterion and a lower bound on the population cross-entropy risk reduction attainable within the proposed refinement family beyond optimal scale-and-bias correction.
- Across three model families and nine tasks with predefined candidate sets, CITP shows improved aggregate predictive performance over calibration and prompt-aggregation baselines. We provide controlled analyses to assess the contributions of query-matched responses and posterior marginalization.

2 Related Work

Prompt Sensitivity LLM predictions are sensitive to prompt components such as prompt format and label words (Zhao et al., 2021; Sclar et al., 2024), demonstration order (Zhao et al., 2021; Lu et al., 2022), and option order in multiple-choice tasks (Zheng et al., 2024). Cox et al. (2025) argued that the same semantic concept can admit multiple semantically equivalent surface realizations, and that the response obtained from a single realization may not adequately reflect the model's uncertainty about the underlying meaning.

Prompt Optimization and Bayesian Prompt Inference Alongside studies of prompt dependence, prior work has treated the prompt itself as an object of search or inference. Zhou et al. (2022) and Yang et al. (2024) used LLMs to generate and improve prompts, while Ross et al. (2026) explored the prompt space using LLM proposals to construct a posterior over prompts and the corresponding posterior predictive distribution. Whereas these approaches search over or infer the prompts themselves, CITP keeps a fixed bank of task-equivalent prompts and instead learns how to use the resulting responses for predictive refinement.

Prompt Aggregation and Other Uses of Prompt Variants Prompt variants have been used for many purposes, like aggregation, uncertainty estimation, and debiasing. Cross-prompt sensitivity or disagreement has been studied as a signal of predictive reliability or uncertainty (Chen et al., 2023; Cox et al., 2025), while other work aggregates predictions across prompt or demonstration variations to improve calibration and task performance (Jiang et al., 2023b; Li et al., 2025). In particular, BayesPE (Tonolini et al., 2024) learns Bayesian weights over semantically equivalent prompts to

form a weighted probability ensemble. CITP instead refines the reference logits using global scale and additive bias together with a signed linear combination of reference-relative prompt contrasts.

Calibration Methods Other works estimate and correct contextual biases or label shifts observed in prompted predictions, using content-free inputs (Zhao et al., 2021), unlabeled batch statistics (Zhou et al., 2024), or model-generated samples (Jiang et al., 2023c). Temperature scaling (Guo et al., 2017) learns a single global temperature, while Dirichlet calibration (Kull et al., 2019) learns a regularized affine transformation for calibration. LinC (Abbas et al., 2024) learns an affine correction using a small labeled dataset. CITP shares a similar setting with supervised output calibration methods in that it uses a small labeled calibration set to refine the predictive distribution of a frozen LLM post hoc.

Logarithmic Pooling Family Logarithmic pooling has long been studied as a way to combine predictive distributions through weighted log probabilities (Heskes, 1997), and log-linear combinations of predictions from multiple prompts have also been explored for factual knowledge probing (Jiang et al., 2020). CITP averages predictive distributions that each admit an exact generalized signed log-pooling representation. This connection motivates our comparisons with simplex and signed affine log pooling.

3 METHOD

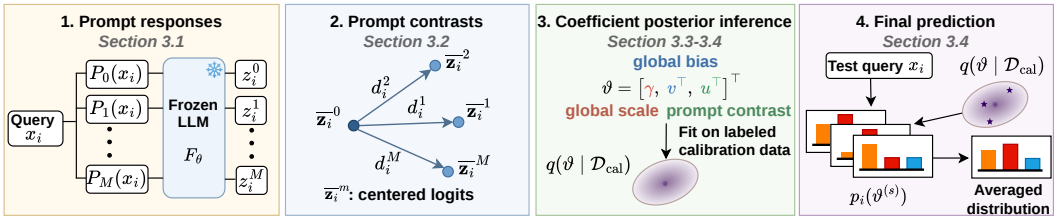


Figure 1: Overview of CITP. Task-equivalent prompt responses define query-specific logit contrasts, while a small labeled calibration set is used to infer task-shared refinement coefficients whose posterior predictive average forms the final prediction.

3.1 SETUP AND NOTATION

We consider a downstream task $\mathcal{T}$ with K candidate labels and a frozen language model F_θ. We implement CITP using candidate logits extracted from F_θ. Let P_0 denote a reference prompt. We construct M task-equivalent variants

$$\mathsf{P}_m = T_m(\mathsf{P}_0), \qquad m = 1, \ldots, M,$$

where each T_m preserves task-relevant content and candidate semantics. They only modify surface properties of the prompt, such as question format, demonstration order, or candidate presentation. Examples of transformations used in our experiments are provided in Appendix D.

For a query x_i, we denote the candidate logits and predictive distribution obtained from prompt realization P_m by

$$\mathbf{z}_i^m = F_\theta(\mathsf{P}_m(x_i)) \in \mathbb{R}^K, \qquad \mathbf{p}_i^m = \mathrm{softmax}(\mathbf{z}_i^m) \in \Delta^{K-1}.$$

For transformations that permute candidate positions, we map the resulting logits back to the canonical candidate order before any subsequent computation.

For each task, we use a small labeled calibration set $\mathcal{D}_{\text{cal}}$ to learn task-level refinement coefficients, while keeping F_θ frozen throughout.

3.2 PROMPT-INDUCED PREDICTIVE CONTRASTS

Softmax predictions are invariant to adding the same constant to all candidate logits. We therefore work with centered logits. Define

$$C = I_K - \frac{1}{K}\mathbf{1}\mathbf{1}^\top, \qquad \overline{\mathbf{z}}_i^m = C\mathbf{z}_i^m = C \log \mathbf{p}_i^m$$

For each query x_i, we define the reference-relative response induced by the m-th prompt realization as

$$\mathbf{d}_i^m = \overline{\mathbf{z}_i}^m - \overline{\mathbf{z}_i}^0, \qquad m = 1, \ldots, M.$$

We then collect these responses column-wise into the prompt-contrast matrix

$$D_i = \left[\mathbf{d}_i^1, \ldots, \mathbf{d}_i^M\right] \in \mathbb{R}^{K \times M}.$$

Thus, D_i describes how the frozen predictor responds to task-equivalent prompt realizations for the same query.

The prompt contrasts may contain a component that is shared across queries. Because such a component is indistinguishable from a query-independent class-logit shift, we separate it from the query-varying response. Using only the calibration set, we compute

$$\overline{D}_{\mathrm{cal}} = \frac{1}{|\mathcal{D}_{\mathrm{cal}}|} \sum_{j \in \mathcal{D}_{\mathrm{cal}}} D_j,$$

and define the query-centered prompt-contrast matrix as

$$\widetilde{J}_i = D_i - \overline{D}_{\mathrm{cal}}.$$

The contrast mean is estimated on the fitting partition and applied unchanged to held-out queries. The query-independent contribution removed by centering can be represented by the global bias term introduced below.

3.3 Global Calibration and Prompt-Contrast Refinement

CITP jointly models global calibration terms and query-varying responses induced by prompt variations. We define the refined centered logits as

$$\overline{\mathbf{z}_i}(\gamma, b, u) = (1 + \gamma)\overline{\mathbf{z}_i}^0 + b + \widetilde{J}_i u,$$

with the corresponding predictive distribution

$$\mathbf{p}_i(\gamma, b, u) = \mathrm{softmax}(\overline{\mathbf{z}_i}(\gamma, b, u)).$$

This affine logit construction permits signed additive corrections along the prompt-contrast directions while retaining valid predictive distributions and a convex objective for joint coefficient fitting. The three terms play distinct roles. The scalar γ adjusts the global sharpness of the reference prediction, b captures a query-independent additive shift over candidate logits, and $u \in \mathbb{R}^M$ contains the signed coefficients for the query-varying prompt contrasts. All coefficients are shared across queries within the same task, while the prompt-induced correction $\widetilde{J}_i u$ remains query-dependent through $\widetilde{J}_i$.

For an identifiable parameterization of the bias term, we write $b = Bv$, where $B \in \mathbb{R}^{K \times (K-1)}$ spans the centered class-logit subspace, and $v \in \mathbb{R}^{K-1}$.

We constrain $\gamma > -1$ so that the global scale $1 + \gamma$ remains positive.

3.4 Task-Level Bayesian Posterior Inference and Prediction

We jointly infer the task-shared coefficient vector

$$\vartheta = \left[\gamma,\ v^\top,\ u^\top\right]^\top.$$

For notational convenience, we write

$$\mathbf{p}_i(\vartheta) := \mathbf{p}_i(\gamma, Bv, u).$$

We use a zero-mean Gaussian prior on the task-level coefficients,

$$p(\vartheta) = \mathcal{N}\left(\vartheta; 0,\ \Lambda^{-1}\right).$$

The MAP estimate is obtained by minimizing the regularized cross-entropy objective

$$\widehat{\vartheta} = \underset{\gamma > -1,\, v,\, u}{\arg\min} \left[\sum_{(x_i, y_i) \in \mathcal{D}_{\mathrm{cal}}} \mathrm{CE}\left(y_i, \mathbf{p}_i(\vartheta)\right) + \frac{1}{2}\vartheta^\top \Lambda \vartheta \right].$$

Because the refined logits are affine in ϑ, the negative log-likelihood is convex in the task-level coefficients. Together with the quadratic prior, this yields a regularized convex MAP objective under the constraint $\gamma > -1$.

We approximate the posterior around the MAP estimate using a Laplace approximation (Daxberger et al., 2021). Let H_{post} denote the Hessian of the negative log posterior at $\widehat{\vartheta}$.

$$q(\vartheta \mid \mathcal{D}_{\text{cal}}) = \mathcal{N}\left(\widehat{\vartheta}, H_{\text{post}}^{-1}\right).$$

For a test query x_i, CITP marginalizes uncertainty in the task-level coefficients and forms the posterior-predictive distribution

$$\begin{aligned}\widehat{\mathbf{p}}_i &= \mathbb{E}_{q(\vartheta|\mathcal{D}_{\text{cal}})}\left[\mathbf{p}_i(\vartheta)\right] \\ &\approx \frac{1}{S}\sum_{s=1}^{S} \text{softmax}\left((1+\gamma^{(s)})\overline{\mathbf{z}}_i^{\,0} + Bv^{(s)} + \widetilde{J}_i u^{(s)}\right),\end{aligned}$$

where

$$\vartheta^{(s)} = \left[\gamma^{(s)}, {v^{(s)}}^\top, {u^{(s)}}^\top\right]^\top \sim q(\vartheta \mid \mathcal{D}_{\text{cal}}).$$

Once the prompt-conditioned logits are obtained, both MAP fitting and posterior-predictive marginalization require only low-dimensional operations and softmax evaluations, without additional LLM forward passes.

3.5 Connections to Calibration and Logarithmic Pooling

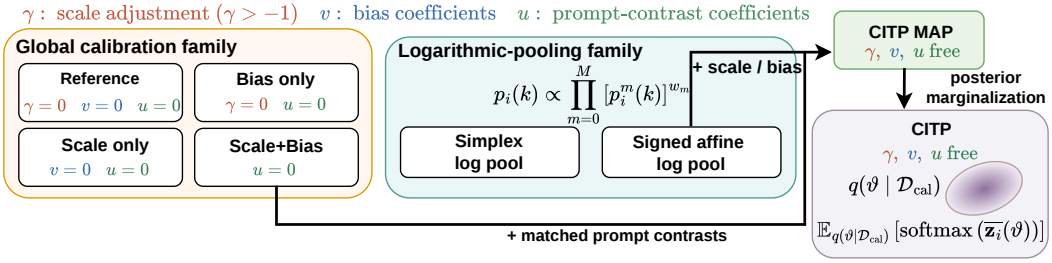


Figure 2: Calibration controls and pooling connections for analyzing CITP's refinement components and posterior marginalization.

The point-predictor family defined in Section 3.3 contains several calibration and prompt-pooling models as special cases. We use these restricted models as calibration and prompt-aggregation baselines, with selected controlled comparisons assessing the additional predictive value of prompt contrasts. Setting $b = 0$ and $u = 0$ yields temperature scaling (Guo et al., 2017) with inverse temperature $1 + \gamma$, while setting $\gamma = 0$ and $u = 0$ yields additive bias calibration. Setting $u = 0$ gives a joint scale-and-bias calibrator, referred to as Scale+Bias in our experiments.

The predictive distribution $\mathbf{p}_i(\gamma, b, u)$ admits an exact generalized logarithmic-pooling representation (Heskes, 1997). Since $\widetilde{J}_i = D_i - \overline{D}_{\text{cal}}$, an equivalent raw-contrast parameterization can be written using the reparameterized bias

$$b' = b - \overline{D}_{\text{cal}} u.$$

For candidate $k \in \{1, \ldots, K\}$, the corresponding probability is

$$p_i(k; \gamma, b, u) \propto \exp(b'_k) \prod_{m=0}^{M} \left[p_i^m(k)\right]^{w_m}, \qquad \sum_{m=0}^{M} w_m = 1 + \gamma.$$

Here,

$$w_0 = (1+\gamma) - \mathbf{1}^\top u, \qquad w_m = u_m, \quad m = 1, \ldots, M.$$

In the equivalent raw-contrast parameterization, setting $\gamma = 0$ and $b' = 0$ and constraining the weights to the simplex recovers simplex logarithmic pooling, while relaxing the nonnegativity constraint gives the signed affine log pool.

CITP can also be implemented using the corresponding candidate probabilities. We use the logit parameterization to express global calibration and prompt-contrast refinement in an additive form.

3.6 When Do Prompt Contrasts Provide Additional Predictive Value?

We analyze the population cross-entropy risk of the point-predictor family to characterize the additional value of prompt contrasts. Global scale and bias parameters are shared across queries, but the resulting predictions can still have query-dependent residual error. To characterize this, let $(\gamma^\star, v^\star)$ be an interior minimizer of the population cross-entropy risk within the scale-and-bias family, and define

$$\mathbf{p}_i^{\mathrm{SB}} = \mathrm{softmax}\left((1+\gamma^\star)\overline{\mathbf{z}}_i^{\,0} + Bv^\star\right).$$

We then define the residual-alignment vector between the remaining predictive residual and the query-dependent prompt contrasts as

$$a = \mathbb{E}\left[\widetilde{J}_i^\top \left(e_{Y_i} - \mathbf{p}_i^{\mathrm{SB}}\right)\right] \in \mathbb{R}^M,$$

where e_{Y_i} is the one-hot vector corresponding to the true label.

Proposition 1. Prompt-contrast improvement criterion. *The scale-and-bias optimum remains optimal after introducing the prompt-contrast coefficients u if and only if $a = 0$. Therefore, if $a \neq 0$, there exists u that strictly decreases the population cross-entropy risk.*

To quantify the available improvement, define

$$G = \mathbb{E}\Big[\widetilde{J}_i^\top \widetilde{J}_i\Big].$$

Proposition 2. Prompt-contrast improvement bound. *Let R_{SB} denote the optimal population cross-entropy risk within the scale-and-bias family and let R_{full} denote the infimum of the population cross-entropy risk within the CITP point-predictor family. Then*

$$R_{\mathrm{SB}} - R_{\mathrm{full}} \geq a^\top G^\dagger a,$$

where $G^\dagger$ is the Moore–Penrose pseudoinverse.

This bound relates the potential gain from prompt-dependent refinement to both the residual-alignment signal a and the geometry of the observable prompt-contrast features. The proofs of Propositions 1 and 2 are provided in Appendix A.1 and Appendix A.2, respectively.

The analysis above is stated entirely in terms of observable logits and prompt contrasts. For additional geometric intuition, we give a conditional local prompt-space interpretation of these contrasts: under smoothness conditions, linear combinations of the observed logit contrasts approximate the model's logit change from the reference to a hypothetical local prompt location. Appendix A.3 formalizes this approximation and shows that the resulting error also bounds the gap in optimal population risk between the observable and corresponding local prompt families.

4 Experiments

4.1 Experimental Setup

Datasets and Models We evaluated three pre-selected frozen 7–8B language models: Qwen3-8B (Yang et al., 2025), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023a). We also pre-selected four MCQA benchmarks (ARC-Challenge, CommonsenseQA, HellaSwag, and MMLU) (Clark et al., 2018; Talmor et al., 2019; Zellers et al., 2019; Hendrycks et al., 2021) and five fixed-label classification tasks (SST-2, AG News, DBpedia-14, TREC, BoolQ) (Socher et al., 2013; Zhang et al., 2015; Li & Roth, 2002; Clark et al., 2019).

Evaluation Protocol Unless otherwise stated, we use four-shot prompts, 50 labeled calibration examples per independently fitted group, and 200 test examples, and repeat each model–benchmark evaluation over ten outer seeds. The default inference budget is eight frozen-model forward passes: one reference prompt and seven task-equivalent variants. All supervised methods use the same labeled calibration examples, while all multi-prompt methods in the main comparison use the same eight-view prompt bank.

We use normalized Brier score (Brier, 1950) and normalized negative log-likelihood as our primary predictive-distribution metrics to aggregate performance across tasks with different numbers of candidates. We also report accuracy. Full data construction, fitting, aggregation, and statistical procedures are given in Appendix E.

4.2 Main Results

4.2.1 Main Comparison

Table 1: Overall predictive performance macro-averaged over 27 model–task pairs. The final row reports paired CITP-minus-best-baseline differences. Bold and underlined values indicate the best and second-best results, respectively. * denotes a statistically significant paired difference (95% bootstrap CI excludes zero).

Method	nBrier ↓	nNLL ↓	Accuracy (%) ↑
Reference and prompt selection			
Reference	0.6464	0.9160	70.08
Calibration-selected best prompt	0.5702	0.7742	73.05
Output calibration			
Temperature scaling (Guo et al., 2017)	0.5499	<u>0.5676</u>	70.08
Dirichlet-ODIR (Kull et al., 2019)	0.5076	1.1721	74.97
LinC (Abbas et al., 2024)	0.5456	0.6251	73.18
Prompt aggregation			
Mean probability ensemble (Jiang et al., 2023b)	0.5262	0.6179	72.82
Weighted probability ensemble	0.4740	0.5774	75.69
BayesPE (public impl.) (Tonolini et al., 2024)	0.4794	0.5707	75.39
Simplex log pool	0.4671	0.6023	77.44
Signed affine log pool	<u>0.4552</u>	0.5842	<u>78.03</u>
CITP (Ours)	**0.4120**	**0.4531**	**79.28**
Δ: CITP − best baseline	-0.0432^*	-0.1145^*	$+1.24^*$

CITP outperformed the output-calibration and prompt-aggregation baselines in Table 1 on macro-averaged nBrier, nNLL, and accuracy. Compared with the best-performing baseline for each metric, CITP reduced nBrier by 0.0432 and nNLL by 0.1145, while improving accuracy by 1.24 percentage points. The paired 95% confidence intervals excluded zero for all three improvements. Detailed results are provided in Appendix C.1.

4.2.2 Mechanism Decomposition

Table 2 compares restricted model families and controlled predictors to examine the roles of global calibration and query-matched prompt contrasts in CITP's predictive gains.

First, the Scale+Bias results indicate that global calibration can reduce a substantial portion of the residual error in the reference prediction, yet CITP significantly improves all three metrics over this control. To assess the additional value of prompt-induced contrasts at the point-prediction level, we compare CITP MAP with the Scale+Bias control. Both predictors include global scale and bias terms, and CITP MAP additionally uses query-matched prompt contrasts. CITP MAP reduces nBrier by 0.0534 and improves accuracy by 3.24 percentage points over Scale+Bias, with paired 95% confidence intervals excluding zero for both metrics. nNLL also decreases by 0.0170, although its confidence interval includes zero.

Table 2: Mechanism decomposition over all 270 model–task–seed cells. Shuffled-contrast results additionally average five label-independent shuffles within each cell. Paired comparisons evaluate CITP MAP and CITP against Scale+Bias. Accuracy differences are in percentage points. Bold and underlined values indicate the best and second-best results, respectively. * denotes a statistically significant paired difference (95% bootstrap CI excludes zero).

Method / restriction	nBrier ↓	nNLL ↓	Accuracy (%) ↑
Scale-and-Bias correction			
Reference	0.6464	0.9160	70.08
Scale only	0.5506	0.5743	70.08
Bias only	0.5024	0.6859	76.44
Scale+Bias	0.4679	0.5081	76.34
Prompt-contrast refinement			
CITP MAP — shuffled contrasts	0.5024	0.5813	75.01
CITP MAP — matched contrasts	<u>0.4145</u>	<u>0.4910</u>	**79.57**
CITP	**0.4120**	**0.4531**	<u>79.28</u>
Key paired comparisons (Δ)			
CITP MAP − Scale+Bias	-0.0534^*	-0.0170	$+3.24^*$
CITP − Scale+Bias	-0.0559^*	-0.0550^*	$+2.94^*$

Further, to test whether the correspondence between each query and its prompt-induced contrasts matters, we construct a shuffled control by permuting the prompt contrasts across queries while preserving the fitting procedure and coefficient dimension. The shuffled control performs substantially worse than the matched-contrast MAP predictor.

These comparisons provide empirical support for the additional predictive value of prompt contrasts beyond global scale-and-bias correction and for the importance of preserving the correspondence between queries and their prompt responses. Appendix C.2 reports the same decomposition separately for fixed-label classification and MCQA tasks.

4.2.3 Residual Alignment

Motivated by Propositions 1 and 2, we examine whether stronger alignment is associated with larger NLL improvements over the Scale+Bias control. Figure 3 shows a positive association between $\widehat{a}^\top \widehat{G}^\dagger \widehat{a}$ and NLL improvement of CITP MAP over Scale+Bias across 27 model–task pairs. The Spearman correlation is 0.562, with a 95% task-cluster bootstrap confidence interval of $[0.054, 0.843]$.

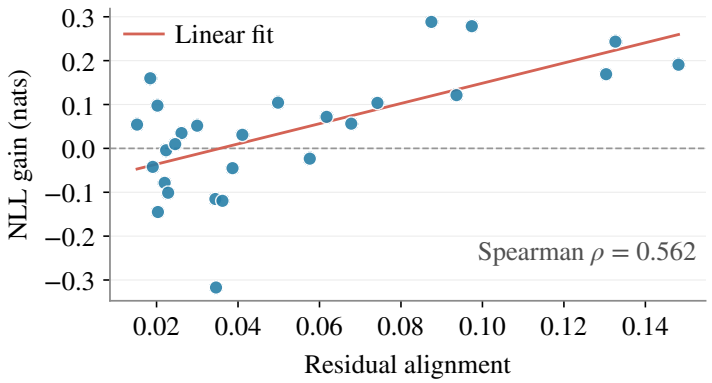


Figure 3: Empirical residual alignment $\widehat{a}^\top \widehat{G}^\dagger \widehat{a}$ versus NLL improvement of CITP MAP over Scale+Bias. Each point averages ten outer seeds for one model–task pair, covering three models and nine benchmarks. Alignment and NLL improvement are computed on disjoint halves of the test queries. Positive values favor CITP MAP; the line is a descriptive least-squares fit.

4.2.4 EFFECT OF POSTERIOR MARGINALIZATION

Table 3: Effect of posterior marginalization. λ_{MAP} and λ_{PP} are prior precisions selected using MAP and posterior-predictive validation NLL, respectively. CITP uses λ_{PP}, and Δ_{marg} is CITP minus MAP(λ_{PP}), isolating coefficient marginalization. The lower block reports the same difference across calibration-label budgets. For nBrier, nNLL, and accuracy in the upper block, * denotes a statistically significant paired difference (95% bootstrap CI excludes zero).

Metric	MAP (λ_{MAP})	MAP (λ_{PP})	CITP	Δ_{marg}
nBrier ↓	0.4145	0.4178	0.4120	-0.0058^*
nNLL ↓	0.4910	0.4816	0.4531	-0.0285^*
Accuracy (%) ↑	79.57	79.28	79.28	$+0.00$
NLL-CVaR$_{95}$ ↓	5.0052	4.5905	3.7735	-0.8170
Wrong conf. ↓	0.7175	0.7052	0.6626	-0.0426
Calibration labels	5	10	20	50
Δ_{marg} nBrier ↓	-0.0247	-0.0194	-0.0151	-0.0058
Δ_{marg} nNLL ↓	-0.1055	-0.0904	-0.0689	-0.0285

We compare CITP with the MAP predictor from the same fit using the posterior-selected prior precision. Marginalization leaves accuracy nearly unchanged while significantly improving both nBrier and nNLL. It also reduces tail loss, measured by NLL-CVaR$_{95}$, and average confidence on incorrect predictions. The improvements in nBrier and nNLL over the shared-prior MAP predictor are larger at smaller calibration budgets, consistent with the motivation for accounting for refinement uncertainty. Appendix C.3 further analyzes the effects of posterior marginalization. NLL gains tend to be larger for queries with greater disagreement among predictions from the coefficient posterior, reflecting greater epistemic uncertainty about the learned refinement, while the posterior prediction shows a lower incidence of high-confidence errors across confidence thresholds. We additionally report close predictive agreement between the Laplace approximation and NUTS.

5 CONCLUSION AND LIMITATIONS

We introduced CITP, a Bayesian method for refining the predictive distribution of a frozen LLM by exploiting query-specific responses to task-equivalent prompt variations. We theoretically characterized when prompt-contrast refinement can reduce population cross-entropy risk beyond optimal scale-and-bias correction and derived a lower bound on the attainable improvement. Across three model families and nine tasks, CITP outperformed the baselines in the main comparison on macro-averaged nBrier, nNLL, and accuracy. Our mechanism analysis further showed that matched query-specific prompt responses provide additional predictive value beyond the Scale+Bias control, while posterior marginalization improves proper scoring rules and predictive-loss tails over the corresponding MAP predictor from the same fit. Together, these results support the use of reference-relative responses to task-equivalent prompt variations as query-specific correction directions and the value of accounting for uncertainty in the learned refinement.

However, CITP requires multiple frozen-model forward passes compared to a single prompt prediction, although the prompt variants can be evaluated in parallel. CITP is also not uniformly preferable under extremely limited calibration data: in our calibration-budget analysis in Appendix C.4, BayesPE achieves lower nBrier and nNLL with five or ten labeled examples per fitted group, whereas CITP achieves lower mean values on both metrics at larger calibration budgets. Our experiments use a fixed, non-optimized bank of task-equivalent prompt variants and focus on tasks with predefined candidate sets. Adaptive construction of informative prompt variants and extension to free-form generation remain directions for future work.

REPRODUCIBILITY STATEMENT

The proposed method is specified in Section 3, and the main experimental protocol is summarized in Section 4.1. Appendix B provides inference and optimization details. Appendix D gives the examples of the prompt perturbations used in our experiments. Appendix E documents the model and data splits, prompt banks, baseline implementations, model selection, posterior computation, metrics, and statistical procedures.

AI USAGE STATEMENT

ChatGPT and Codex (OpenAI) were used for coding assistance, translation and improvements to the organization and clarity of the manuscript. All code, mathematical claims, citations, experimental settings, and reported results were reviewed and verified by the authors.

REFERENCES

Momin Abbas, Yi Zhou, Parikshit Ram, Nathalie Baracaldo, Horst Samulowitz, Theodoros Salonidis, and Tianyi Chen. Enhancing in-context learning via linear probe calibration. In *International conference on artificial intelligence and statistics*, pp. 307–315. PMLR, 2024.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

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

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.

Yanda Chen, Chen Zhao, Zhou Yu, Kathleen McKeown, and He He. On the relation between sensitivity and accuracy in in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 155–167, 2023.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, 2019.

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

Kyle Cox, Jiawei Xu, Yikun Han, Rong Xu, Tianhao Li, Chi-Yang Hsu, Tianlong Chen, Walter Gerych, and Ying Ding. Mapping from meaning: Addressing the miscalibration of prompt-sensitive language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23696–23703, 2025.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux-effortless bayesian deep learning. *Advances in neural information processing systems*, 34:20089–20103, 2021.

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.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Tom Heskes. Selecting weighting factors in logarithmic opinion pools. In M. Jordan, M. Kearns, and S. Solla (eds.), *Advances in Neural Information Processing Systems*, volume 10. MIT Press, 1997. URL https://proceedings.neurips.cc/paper_files/paper/1997/file/59f51fd6937412b7e56ded1ea2470c25-Paper.pdf.

Matthew D Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

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

Mingjian Jiang, Yangjun Ruan, Sicong Huang, Saifei Liao, Silviu Pitis, Roger Baker Grosse, and Jimmy Ba. Calibrating language models via augmented prompt ensembles, 2023b. URL https://openreview.net/forum?id=L0dc4wqbNs. Accepted at the ICML 2023 Workshop on Deployable Generative AI.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020. doi: 10.1162/tacl_a_00324. URL https://aclanthology.org/2020.tacl-1.28/.

Zhongtao Jiang, Yuanzhe Zhang, Cao Liu, Jun Zhao, and Kang Liu. Generative calibration for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2312–2333, 2023c.

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

Chengzu Li, Han Zhou, Goran Glavaš, Anna Korhonen, and Ivan Vulić. Large language models are miscalibrated in-context learners. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11575–11596, 2025.

Xin Li and Dan Roth. Learning question classifiers. In *Coling 2002: The 19th international conference on computational linguistics*, 2002.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098, 2022.

Brendan Ross, Noël Vouitsis, Atiyeh Ashari Ghomi, Rasa Hosseinzadeh, Ji Xin, Zhaoyan Liu, Yi Sui, Shiyi Hou, Kin Kwan Leung, Gabriel Loaiza-Ganem, et al. Textual bayes: Quantifying prompt uncertainty in llm-based systems. In *International Conference on Learning Representations*, volume 2026, pp. 133607–133630, 2026.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, volume 2024, pp. 25055–25083, 2024.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, 2019.

Francesco Tonolini, Nikolaos Aletras, Jordan Massiah, and Gabriella Kazai. Bayesian prompt ensembles: Model uncertainty estimation for black-box large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12229–12272, 2024.

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.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *International Conference on Learning Representations*, volume 2024, pp. 12028–12068, 2024.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4791–4800, 2019.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, volume 28, 2015.

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*, pp. 12697–12706. Pmlr, 2021.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations*, volume 2024, pp. 19426–19454, 2024.

Han Zhou, Xingchen Wan, Lev Proleev, Diana Mincu, Jilin Chen, Katherine Heller, and Subhrajit Roy. Batch calibration: Rethinking calibration for in-context learning and prompt engineering. In *International Conference on Learning Representations*, 2024.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*, 2022.

A Additional Theoretical Analysis

A.1 Proof of the Prompt-Contrast Improvement Criterion

We provide the proof of Proposition 1. Let $R(\gamma, v, u)$ denote the population cross-entropy risk of the full point-predictor family.

Proof. Cross-entropy is convex in the affine logit coefficients. Since $(\gamma^\star, v^\star)$ is an interior minimizer of the scale-and-bias family, the population gradients with respect to γ and v vanish at $(\gamma^\star, v^\star, 0)$. By the definition of a, the gradient with respect to u is

$$\nabla_u R(\gamma^\star, v^\star, 0) = -a.$$

If $a = 0$, all coordinates satisfy the first-order optimality condition at $(\gamma^\star, v^\star, 0)$, and convexity implies that the scale-and-bias solution remains optimal in the enlarged prompt-contrast family. If $a \neq 0$, $u = 0$ is not stationary in the enlarged family, and a sufficiently small step along a descent direction strictly decreases the population cross-entropy risk. □

A.2 Proof of the Prompt-Contrast Improvement Bound

We provide the proof of Proposition 2. Let

$$R_0(u) = R(\gamma^\star, v^\star, u),$$

so that

$$R_0(0) = R_{\mathrm{SB}}.$$

Proof. For multinomial cross-entropy, the Hessian of R_0 with respect to u is

$$\nabla_u^2 R_0(u) = \mathbb{E}\left[\widetilde{J}_i^\top \left(\mathrm{Diag}(\mathbf{p}_i(u)) - \mathbf{p}_i(u)\mathbf{p}_i(u)^\top\right) \widetilde{J}_i\right],$$

where $\mathbf{p}_i(u)$ denotes the predictive distribution at $(\gamma^\star, v^\star, u)$.

The softmax covariance matrix has operator norm at most $1/2$, and therefore

$$\nabla_u^2 R_0(u) \preceq \frac{1}{2} G.$$

Since

$$\nabla_u R_0(0) = -a,$$

it follows that

$$R_0(u) \leq R_{\mathrm{SB}} - a^\top u + \frac{1}{4} u^\top G u.$$

We also have $a \in \mathrm{Range}(G)$. Indeed, if $h \in \ker(G)$, then

$$0 = h^\top G h = \mathbb{E}\left[\|\widetilde{J}_i h\|_2^2\right],$$

so $\widetilde{J}_i h = 0$ almost surely and hence $h^\top a = 0$. Therefore a is orthogonal to $\ker(G)$ and lies in $\mathrm{Range}(G)$.

Choosing

$$u = 2G^\dagger a$$

then gives

$$R_0(u) \leq R_{\mathrm{SB}} - a^\top G^\dagger a.$$

Since R_{full} is the infimum of the population cross-entropy risk over the CITP point-predictor family,

$$R_{\mathrm{full}} \leq R_0(u),$$

and therefore

$$R_{\mathrm{SB}} - R_{\mathrm{full}} \geq a^\top G^\dagger a.$$

□

A.3 Conditional Local Prompt-Space Interpretation

The preceding results use only observable logits. Here, we provide a conditional local prompt-space interpretation of the prompt-induced response directions. This interpretation does not assume that a latent prompt representation is unique or identifiable.

Suppose that, locally around the observed prompt realizations, there exist coordinates $c_0, \dots, c_M$ corresponding to $\mathsf{P}_0, \dots, \mathsf{P}_M$. Define

$$\delta_m = c_m - c_0, \qquad D_c = [\delta_1, \dots, \delta_M], \qquad c(u) = c_0 + D_c u.$$

For query x_i, let $\overline{\mathbf{z}_i}(c)$ denote the centered-logit response at coordinate c. The observed prompt-response differences are

$$d_i^m = \overline{\mathbf{z}_i}(c_m) - \overline{\mathbf{z}_i}(c_0), \qquad D_i = [d_i^1, \dots, d_i^M].$$

For the point-predictor family, we use the raw-contrast parameterization in this section. When the additive bias is free, the raw contrasts D_i and the query-centered contrasts define the same point-predictor family up to a reparameterization of the bias. For fixed (γ, b, u), define the observable response as

$$s_i^{\mathrm{sec}}(\gamma, b, u) = (1+\gamma)\overline{\mathbf{z}_i}(c_0) + b + D_i u,$$

and the corresponding hypothetical local response as

$$s_i^{\mathrm{lat}}(\gamma, b, u) = (1+\gamma)\overline{\mathbf{z}_i}(c_0) + b + \overline{\mathbf{z}_i}(c_0 + D_c u) - \overline{\mathbf{z}_i}(c_0).$$

The scale and bias terms are shared by the two responses, so their difference is determined by the finite-response approximation error.

Proposition 3. Conditional local secant approximation. *Assume that $\overline{\mathbf{z}_i}$ has an L_i-Lipschitz Jacobian in a local region containing the line segments from c_0 to each c_m and to $c_0 + D_c u$. Then*

$$\left\| s_i^{\mathrm{lat}}(\gamma, b, u) - s_i^{\mathrm{sec}}(\gamma, b, u) \right\|_2 \le \frac{L_i}{2} \left(\|D_c u\|_2^2 + \sum_{m=1}^{M} |u_m| \|\delta_m\|_2^2 \right).$$

Proof. Let J_i^0 denote the Jacobian of $\overline{\mathbf{z}_i}$ at c_0 and write

$$\overline{\mathbf{z}_i}(c_0 + v) = \overline{\mathbf{z}_i}(c_0) + J_i^0 v + r_i(v).$$

The Lipschitz-Jacobian assumption gives

$$\|r_i(v)\|_2 \le \frac{L_i}{2} \|v\|_2^2.$$

Since

$$d_i^m = J_i^0 \delta_m + r_i(\delta_m),$$

we have

$$\overline{\mathbf{z}_i}(c_0 + D_c u) - \overline{\mathbf{z}_i}(c_0) - D_i u = r_i(D_c u) - \sum_{m=1}^{M} u_m r_i(\delta_m).$$

Applying the triangle inequality gives the result. □

Therefore, the observed prompt-induced response differences can be interpreted as finite-difference directions that approximate the frozen predictor's response variation within a restricted local prompt family. This result holds conditionally for any local prompt space satisfying the smoothness condition and does not require the latent space to be unique or identifiable.

The response approximation also transfers to predictive risk. The multiclass cross-entropy

$$\ell_y(s) = -\log \mathrm{softmax}(s)_y$$

is $\sqrt{2}$-Lipschitz with respect to the logits. Therefore, let $\mathcal{U}$ be a restricted coefficient set such that $c_0 + D_c u$ remains in the local region above, and let $R^*_{\mathrm{lat},\mathcal{U}}$ and $R^*_{\mathrm{sec},\mathcal{U}}$ denote the optimal population risks of the hypothetical local family and the observable secant family, respectively. Then

$$\left| R^*_{\mathrm{lat},\mathcal{U}} - R^*_{\mathrm{sec},\mathcal{U}} \right| \le \sqrt{2} \sup_{u \in \mathcal{U}} \mathbb{E}_i \left[\frac{L_i}{2} \left(\|D_c u\|_2^2 + \sum_{m=1}^{M} |u_m| \|\delta_m\|_2^2 \right) \right].$$

Thus, when the local approximation error is small, optimizing the observable response family can approximate the optimal predictive risk obtainable within the corresponding restricted local prompt family.

CITP does not estimate c_m or D_c and does not directly optimize latent coordinates. This result only gives a conditional interpretation of the observable prompt-response family.

B Additional Inference and Optimization Details

B.1 Query Centering and Reparameterization

Recall that

$$D_i = \overline{D}_{\mathrm{cal}} + \widetilde{J}_i.$$

For any coefficients u,

$$b + D_i u = \left(b + \overline{D}_{\mathrm{cal}} u\right) + \widetilde{J}_i u.$$

Hence, when a free query-independent bias is included, using raw prompt contrasts D_i or query-centered contrasts $\widetilde{J}_i$ defines the same point-predictor family up to a reparameterization of the bias.

We use the centered form because it separates the average query-independent response to prompt variation from the query-varying component that is central to our analysis. This equivalence concerns the point-predictor family; once priors are placed directly on a particular parameterization, different parameterizations need not induce identical posterior distributions.

B.2 Bias Parameterization

Because adding a class-common constant to all logits does not affect the softmax distribution, the additive bias is identifiable only within the centered class-logit subspace. We therefore parameterize

$$b = Bv,$$

where the columns of $B \in \mathbb{R}^{K\times(K-1)}$ form a fixed basis for

$$\left\{b \in \mathbb{R}^K : \mathbf{1}^\top b = 0\right\}.$$

Only v is inferred; B is a fixed orthonormal Helmert basis, so $B^\top B = I$.

B.3 Block Normalization and Prior

The scale, bias, and prompt-contrast blocks have different numerical scales. We therefore RMS-normalize the corresponding design blocks using a single RMS over all entries of each block on the fitting partition, covering its query, class, and coordinate axes. The same scale is applied to held-out queries; RMS values below 10^{-12} are replaced by one. We apply a common scalar Gaussian prior precision in these normalized coordinates. In the original coordinates, this is equivalent to the block-diagonal precision matrix Λ used in Section 3.4.

This normalization is used only to make the common prior precision comparable across parameter blocks; it does not introduce any additional query-dependent information.

B.4 Laplace Approximation

Let

$$A_i = \left[\overline{\mathbf{z}}_i^{\,0},\ B,\ \widetilde{J}_i\right],$$

so that

$$\overline{\mathbf{z}}_i(\vartheta) = \overline{\mathbf{z}}_i^{\,0} + A_i\vartheta.$$

At the MAP estimate $\widehat{\vartheta}$, the posterior Hessian is

$$H_{\mathrm{post}} = \Lambda + \sum_{(x_i,y_i)\in\mathcal{D}_{\mathrm{cal}}} A_i^\top \left[\mathrm{Diag}\left(\mathbf{p}_i(\widehat{\vartheta})\right) - \mathbf{p}_i(\widehat{\vartheta})\mathbf{p}_i(\widehat{\vartheta})^\top\right] A_i.$$

The Laplace approximation used for posterior prediction is

$$q(\vartheta \mid \mathcal{D}_{\mathrm{cal}}) = \mathcal{N}\left(\widehat{\vartheta}, H_{\mathrm{post}}^{-1}\right).$$

In normalized coordinates, we require the posterior Hessian to have a minimum eigenvalue greater than 10^{-10} and a condition number at most 10^{12}.

B.5 Boundary Safeguards

The global scale is constrained by $\gamma > -1$. To avoid using an unconstrained Gaussian approximation when the MAP solution lies too close to this boundary, we define the standardized boundary distance

$$z_{\partial} = \frac{\widehat{\gamma} + 1}{\sqrt{(H_{\mathrm{post}}^{-1})_{\gamma\gamma}}}.$$

We use the Gaussian Laplace posterior predictive only when

$$z_{\partial} \geq 4.$$

We reject a posterior candidate if any draw used for predictive integration violates $\gamma > -1$.

All 360 final CITP fits in the main study satisfied the boundary criterion. The minimum standardized distance was 4.2963, and no posterior draw crossed $\gamma = -1$.

C Additional Experimental Results and Ablations

C.1 Detailed Results by Model and Task

We report per-model and per-benchmark results corresponding to the aggregate results in Section 4. All values use the same outer seeds, calibration examples, test subsets, and aggregation rules as the main experiments.

Table 4: Detailed fixed-label results on AGNews, DBpedia14, and TREC (mean ± outer-seed SD over 10 runs). Brier and NLL are unnormalized. Arrows indicate the preferred direction. Bold and underlined values indicate the best and second-best results, respectively.

Model	Method	AGNews					DBpedia14					TREC				
		Brier↓	NLL↓	Acc. (%)↑	ECE↓	Macro-F1↑	Brier↓	NLL↓	Acc. (%)↑	ECE↓	Macro-F1↑	Brier↓	NLL↓	Acc. (%)↑	ECE↓	Macro-F1↑
Qwen3-8B	Reference	0.389 ± 0.051	1.868 ± 0.347	78.7 ± 3.2	0.192 ± 0.025	0.787 ± 0.033	0.230 ± 0.073	1.176 ± 0.423	87.6 ± 3.7	0.114 ± 0.035	0.866 ± 0.038	0.480 ± 0.059	1.598 ± 0.287	71.4 ± 3.8	0.213 ± 0.027	0.670 ± 0.030
	Temperature scaling	0.335 ± 0.038	**0.674 ± 0.067**	78.7 ± 3.2	**0.083 ± 0.017**	0.787 ± 0.033	0.208 ± 0.061	0.535 ± 0.169	87.6 ± 3.7	0.071 ± 0.029	0.866 ± 0.038	0.413 ± 0.045	0.891 ± 0.093	71.4 ± 3.8	0.103 ± 0.018	0.670 ± 0.030
	Cal.-selected best prompt	0.371 ± 0.043	1.233 ± 0.163	78.3 ± 3.6	0.168 ± 0.014	0.785 ± 0.035	0.221 ± 0.070	0.895 ± 0.300	87.2 ± 3.9	0.104 ± 0.030	0.864 ± 0.040	0.574 ± 0.169	1.395 ± 0.424	63.3 ± 12.6	0.245 ± 0.096	0.587 ± 0.144
	Mean prob. ensemble	0.340 ± 0.046	1.056 ± 0.235	79.1 ± 3.0	0.136 ± 0.024	0.791 ± 0.033	0.181 ± 0.053	0.524 ± 0.215	87.9 ± 3.3	**0.057 ± 0.026**	0.871 ± 0.034	0.391 ± 0.041	0.882 ± 0.206	73.9 ± 2.4	0.086 ± 0.016	0.677 ± 0.034
	Weighted prob. ensemble	0.330 ± 0.035	1.035 ± 0.242	79.8 ± 2.0	0.129 ± 0.024	0.799 ± 0.022	0.189 ± 0.047	0.640 ± 0.224	88.3 ± 3.0	0.075 ± 0.021	0.874 ± 0.032	0.392 ± 0.053	0.851 ± 0.210	73.2 ± 4.2	0.101 ± 0.026	0.684 ± 0.050
	BayesPE (public implementation)	0.334 ± 0.033	0.971 ± 0.182	79.1 ± 2.4	0.131 ± 0.022	0.792 ± 0.025	0.180 ± 0.047	0.504 ± 0.197	88.1 ± 3.1	0.064 ± 0.019	0.874 ± 0.032	0.396 ± 0.053	0.840 ± 0.224	72.3 ± 4.2	0.101 ± 0.027	0.679 ± 0.054
	Dirichlet-ODIR	0.381 ± 0.070	1.906 ± 1.889	78.0 ± 3.4	0.128 ± 0.046	0.783 ± 0.031	0.219 ± 0.043	1.135 ± 0.868	86.5 ± 4.1	0.093 ± 0.026	0.853 ± 0.045	0.408 ± 0.080	0.976 ± 0.395	71.0 ± 6.1	0.123 ± 0.058	0.677 ± 0.117
	LinC	0.359 ± 0.039	0.733 ± 0.068	78.8 ± 3.1	0.087 ± 0.037	0.789 ± 0.033	0.301 ± 0.076	0.844 ± 0.197	84.0 ± 6.2	0.194 ± 0.116	0.825 ± 0.071	0.423 ± 0.042	0.832 ± 0.097	68.5 ± 3.9	0.156 ± 0.042	0.613 ± 0.067
	Simplex log pool	0.352 ± 0.031	1.230 ± 0.298	**79.8 ± 2.6**	0.158 ± 0.009	**0.800 ± 0.028**	0.211 ± 0.056	0.855 ± 0.294	88.2 ± 3.2	0.106 ± 0.030	0.874 ± 0.031	0.384 ± 0.065	0.837 ± 0.181	74.6 ± 5.0	0.131 ± 0.033	0.699 ± 0.047
	Signed affine log pool	0.361 ± 0.046	1.164 ± 0.444	78.8 ± 1.7	0.154 ± 0.035	0.790 ± 0.018	0.218 ± 0.058	0.872 ± 0.367	87.6 ± 3.6	0.108 ± 0.029	0.869 ± 0.035	0.378 ± 0.074	0.817 ± 0.204	74.6 ± 5.5	0.110 ± 0.038	0.697 ± 0.050
	CITP	**0.327 ± 0.022**	0.712 ± 0.108	78.9 ± 1.9	0.099 ± 0.027	0.793 ± 0.018	**0.162 ± 0.034**	**0.437 ± 0.145**	**90.5 ± 2.5**	0.083 ± 0.032	**0.897 ± 0.025**	**0.303 ± 0.047**	**0.582 ± 0.104**	**78.0 ± 3.7**	**0.059 ± 0.015**	**0.761 ± 0.041**
Llama-3.1-8B	Reference	0.436 ± 0.046	0.972 ± 0.130	72.9 ± 2.9	0.162 ± 0.023	0.725 ± 0.030	0.294 ± 0.035	0.735 ± 0.101	81.8 ± 2.6	0.104 ± 0.013	0.812 ± 0.028	0.693 ± 0.128	1.630 ± 0.288	49.5 ± 10.8	0.185 ± 0.093	0.404 ± 0.123
	Temperature scaling	0.412 ± 0.041	0.781 ± 0.066	72.9 ± 2.9	0.096 ± 0.034	0.725 ± 0.030	0.294 ± 0.031	0.686 ± 0.062	81.8 ± 2.6	0.083 ± 0.017	0.812 ± 0.028	0.676 ± 0.093	1.459 ± 0.181	49.5 ± 10.8	0.151 ± 0.057	0.404 ± 0.123
	Cal.-selected best prompt	0.394 ± 0.046	0.826 ± 0.100	75.2 ± 3.4	0.133 ± 0.033	0.752 ± 0.035	0.217 ± 0.034	0.537 ± 0.087	86.4 ± 2.7	0.065 ± 0.013	0.848 ± 0.030	0.584 ± 0.092	1.228 ± 0.187	55.4 ± 8.8	0.125 ± 0.062	0.477 ± 0.077
	Mean prob. ensemble	0.394 ± 0.032	0.786 ± 0.080	74.6 ± 2.5	0.111 ± 0.027	0.743 ± 0.028	0.247 ± 0.036	0.500 ± 0.070	83.7 ± 2.0	0.067 ± 0.022	0.833 ± 0.020	0.571 ± 0.065	1.251 ± 0.181	58.4 ± 5.9	0.132 ± 0.052	0.488 ± 0.056
	Weighted prob. ensemble	0.380 ± 0.036	0.754 ± 0.077	75.0 ± 3.5	0.114 ± 0.024	0.748 ± 0.038	0.195 ± 0.034	0.437 ± 0.087	87.6 ± 2.6	0.057 ± 0.015	0.865 ± 0.028	0.529 ± 0.086	1.104 ± 0.178	60.5 ± 8.9	0.133 ± 0.048	0.534 ± 0.077
	BayesPE (public implementation)	0.377 ± 0.041	0.743 ± 0.079	75.4 ± 3.6	0.102 ± 0.024	0.752 ± 0.040	0.193 ± 0.035	0.417 ± 0.076	87.2 ± 2.7	0.064 ± 0.018	0.861 ± 0.028	0.548 ± 0.083	1.136 ± 0.169	59.3 ± 8.9	0.132 ± 0.054	0.529 ± 0.082
	Dirichlet-ODIR	0.373 ± 0.025	1.251 ± 1.111	75.9 ± 2.5	0.085 ± 0.016	0.762 ± 0.027	0.224 ± 0.055	1.237 ± 1.567	87.2 ± 4.3	0.110 ± 0.019	0.864 ± 0.048	0.466 ± 0.110	0.976 ± 0.228	65.1 ± 12.1	**0.106 ± 0.034**	0.588 ± 0.113
	LinC	0.383 ± 0.021	0.880 ± 0.250	75.4 ± 1.4	0.104 ± 0.015	0.754 ± 0.016	0.367 ± 0.096	0.946 ± 0.268	80.8 ± 6.3	0.226 ± 0.114	0.789 ± 0.075	0.547 ± 0.087	1.087 ± 0.163	62.1 ± 11.8	0.184 ± 0.064	0.525 ± 0.096
	Simplex log pool	0.386 ± 0.043	0.773 ± 0.078	75.2 ± 3.5	0.130 ± 0.023	0.751 ± 0.037	0.175 ± 0.036	0.392 ± 0.081	88.2 ± 2.2	0.057 ± 0.019	0.875 ± 0.019	0.533 ± 0.100	1.132 ± 0.216	60.7 ± 9.7	0.140 ± 0.049	0.536 ± 0.099
	Signed affine log pool	0.393 ± 0.049	0.805 ± 0.115	74.6 ± 3.8	0.133 ± 0.023	0.747 ± 0.035	0.161 ± 0.026	0.367 ± 0.067	89.0 ± 1.9	**0.052 ± 0.010**	0.883 ± 0.020	0.467 ± 0.102	0.961 ± 0.235	66.5 ± 8.2	0.108 ± 0.048	0.619 ± 0.071
	CITP	**0.332 ± 0.033**	**0.628 ± 0.064**	**77.5 ± 2.2**	**0.078 ± 0.019**	**0.777 ± 0.023**	**0.134 ± 0.036**	**0.320 ± 0.080**	**93.0 ± 2.9**	0.091 ± 0.010	**0.925 ± 0.030**	**0.351 ± 0.047**	**0.714 ± 0.099**	**76.8 ± 3.8**	0.134 ± 0.039	**0.726 ± 0.043**
Mistral-7B	Reference	0.434 ± 0.079	1.429 ± 0.349	75.6 ± 4.7	0.208 ± 0.039	0.754 ± 0.056	0.305 ± 0.057	1.044 ± 0.188	82.8 ± 3.0	0.144 ± 0.031	0.815 ± 0.034	0.588 ± 0.108	2.087 ± 0.425	66.4 ± 5.8	0.273 ± 0.052	0.598 ± 0.057
	Temperature scaling	0.363 ± 0.050	**0.699 ± 0.093**	75.6 ± 4.7	**0.086 ± 0.033**	0.754 ± 0.056	0.267 ± 0.041	0.607 ± 0.072	82.8 ± 3.0	0.076 ± 0.021	0.815 ± 0.034	0.465 ± 0.071	0.989 ± 0.122	66.4 ± 5.8	0.099 ± 0.019	0.598 ± 0.057
	Cal.-selected best prompt	0.438 ± 0.083	1.436 ± 0.308	75.2 ± 4.6	0.207 ± 0.042	0.753 ± 0.049	0.272 ± 0.080	0.874 ± 0.235	84.4 ± 4.3	0.133 ± 0.040	0.830 ± 0.049	0.597 ± 0.107	1.889 ± 0.345	64.9 ± 6.7	0.276 ± 0.050	0.600 ± 0.083
	Mean prob. ensemble	0.376 ± 0.057	0.978 ± 0.253	76.0 ± 4.5	0.147 ± 0.029	0.759 ± 0.053	0.245 ± 0.045	0.546 ± 0.108	82.8 ± 3.3	0.093 ± 0.029	0.814 ± 0.041	0.492 ± 0.071	1.229 ± 0.237	67.1 ± 5.8	0.169 ± 0.044	0.598 ± 0.062
	Weighted prob. ensemble	0.376 ± 0.056	1.034 ± 0.244	76.2 ± 4.0	0.140 ± 0.037	0.762 ± 0.046	0.212 ± 0.055	0.522 ± 0.144	84.9 ± 3.8	0.076 ± 0.026	0.838 ± 0.039	0.483 ± 0.062	1.195 ± 0.215	65.5 ± 4.9	0.159 ± 0.048	0.606 ± 0.064
	BayesPE (public implementation)	0.375 ± 0.061	0.970 ± 0.252	76.1 ± 4.2	0.144 ± 0.042	0.761 ± 0.047	0.210 ± 0.051	0.484 ± 0.112	84.7 ± 4.0	**0.074 ± 0.027**	0.837 ± 0.041	0.485 ± 0.074	1.168 ± 0.237	65.7 ± 5.7	0.164 ± 0.043	0.607 ± 0.066
	Dirichlet-ODIR	0.429 ± 0.082	17.230 ± 39.889	75.4 ± 3.4	0.161 ± 0.076	0.756 ± 0.036	0.247 ± 0.054	1.912 ± 2.249	85.2 ± 4.9	0.111 ± 0.025	0.841 ± 0.061	0.438 ± 0.091	2.136 ± 2.001	70.3 ± 5.0	0.142 ± 0.055	0.652 ± 0.091
	LinC	0.375 ± 0.046	0.835 ± 0.345	76.8 ± 3.9	0.089 ± 0.024	0.769 ± 0.045	0.368 ± 0.076	0.962 ± 0.149	81.0 ± 5.1	0.229 ± 0.097	0.786 ± 0.062	0.472 ± 0.058	0.933 ± 0.113	67.0 ± 5.8	0.145 ± 0.047	0.573 ± 0.069
	Simplex log pool	0.409 ± 0.068	1.197 ± 0.272	76.5 ± 4.4	0.190 ± 0.033	0.764 ± 0.052	0.243 ± 0.055	0.680 ± 0.173	85.6 ± 3.5	0.111 ± 0.029	0.846 ± 0.036	0.522 ± 0.075	1.495 ± 0.246	68.3 ± 5.6	0.229 ± 0.038	0.631 ± 0.071
	Signed affine log pool	0.409 ± 0.055	1.182 ± 0.221	76.2 ± 3.6	0.185 ± 0.025	0.763 ± 0.042	0.242 ± 0.056	0.716 ± 0.198	86.0 ± 3.2	0.115 ± 0.025	0.848 ± 0.033	0.491 ± 0.028	1.390 ± 0.166	70.1 ± 2.6	0.209 ± 0.015	0.657 ± 0.051
	CITP	**0.352 ± 0.033**	0.713 ± 0.108	**77.2 ± 2.5**	0.109 ± 0.027	**0.776 ± 0.027**	**0.171 ± 0.038**	**0.377 ± 0.089**	**88.8 ± 2.3**	0.076 ± 0.035	**0.879 ± 0.025**	**0.391 ± 0.051**	**0.762 ± 0.109**	**71.8 ± 3.5**	**0.090 ± 0.029**	**0.714 ± 0.034**

Table 5: Detailed fixed-label results on SST-2 and BoolQ (mean ± outer-seed SD over 10 runs). Brier and NLL are unnormalized. Arrows indicate the preferred direction. Bold and underlined values indicate the best and second-best results, respectively.

Model	Method	SST-2					BoolQ				
		Brier↓	NLL↓	Acc. (%)↑	ECE↓	Macro-F1↑	Brier↓	NLL↓	Acc. (%)↑	ECE↓	Macro-F1↑
Qwen3-8B	Reference	0.401 ± 0.219	0.675 ± 0.392	73.5 ± 13.8	0.163 ± 0.146	0.697 ± 0.190	0.188 ± 0.015	0.513 ± 0.054	89.0 ± 0.9	0.091 ± 0.009	0.884 ± 0.010
	Temperature scaling	0.378 ± 0.102	0.554 ± 0.127	73.5 ± 13.8	0.162 ± 0.097	0.697 ± 0.190	**0.168 ± 0.009**	**0.296 ± 0.024**	89.0 ± 0.9	**0.047 ± 0.013**	0.884 ± 0.010
	Cal.-selected best prompt	0.228 ± 0.074	0.389 ± 0.112	84.4 ± 6.5	0.064 ± 0.022	0.840 ± 0.070	0.186 ± 0.018	0.503 ± 0.101	89.2 ± 1.3	0.089 ± 0.010	0.886 ± 0.014
	Mean prob. ensemble	0.274 ± 0.129	0.412 ± 0.149	77.5 ± 14.3	0.122 ± 0.084	0.746 ± 0.186	0.172 ± 0.010	0.414 ± 0.030	**89.5 ± 0.9**	0.074 ± 0.007	**0.890 ± 0.009**
	Weighted prob. ensemble	0.188 ± 0.044	0.315 ± 0.068	88.4 ± 3.6	0.071 ± 0.038	0.883 ± 0.037	0.175 ± 0.013	0.433 ± 0.056	89.2 ± 1.5	0.081 ± 0.013	0.886 ± 0.016
	BayesPE (public implementation)	0.193 ± 0.046	0.321 ± 0.066	88.2 ± 3.5	0.083 ± 0.040	0.882 ± 0.036	0.170 ± 0.010	0.403 ± 0.025	**89.5 ± 1.0**	0.074 ± 0.008	0.889 ± 0.011
	Dirichlet-ODIR	0.297 ± 0.074	0.457 ± 0.085	77.7 ± 11.0	0.111 ± 0.062	0.759 ± 0.158	0.188 ± 0.028	0.685 ± 0.796	88.8 ± 1.6	0.070 ± 0.025	0.882 ± 0.018
	LinC	0.306 ± 0.079	0.487 ± 0.105	76.2 ± 11.1	0.116 ± 0.053	0.745 ± 0.136	0.205 ± 0.026	0.489 ± 0.438	88.8 ± 1.5	0.089 ± 0.055	0.881 ± 0.017
	Simplex log pool	0.165 ± 0.037	0.277 ± 0.058	90.0 ± 2.8	0.056 ± 0.022	0.899 ± 0.029	0.183 ± 0.017	0.501 ± 0.077	**89.5 ± 1.3**	0.089 ± 0.010	0.889 ± 0.014
	Signed affine log pool	**0.148 ± 0.018**	**0.256 ± 0.041**	**90.7 ± 1.5**	**0.053 ± 0.029**	**0.907 ± 0.015**	0.186 ± 0.010	0.458 ± 0.087	88.8 ± 0.6	0.083 ± 0.011	0.882 ± 0.007
	CITP	0.195 ± 0.099	0.315 ± 0.126	86.8 ± 9.2	0.088 ± 0.048	0.861 ± 0.110	0.180 ± 0.018	0.368 ± 0.066	89.0 ± 1.6	0.074 ± 0.017	0.884 ± 0.016
Llama-3.1-8B	Reference	0.338 ± 0.237	0.581 ± 0.423	78.3 ± 13.7	0.142 ± 0.140	0.758 ± 0.180	0.197 ± 0.014	0.351 ± 0.015	87.3 ± 1.7	0.063 ± 0.013	0.870 ± 0.017
	Temperature scaling	0.284 ± 0.124	0.439 ± 0.149	78.3 ± 13.7	0.118 ± 0.079	0.758 ± 0.180	0.197 ± 0.012	0.336 ± 0.021	87.3 ± 1.7	**0.052 ± 0.012**	0.870 ± 0.017
	Cal.-selected best prompt	0.138 ± 0.018	0.248 ± 0.028	91.2 ± 2.1	0.045 ± 0.014	0.912 ± 0.021	0.191 ± 0.014	0.347 ± 0.026	87.7 ± 1.2	0.065 ± 0.015	0.873 ± 0.012
	Mean prob. ensemble	0.219 ± 0.092	0.352 ± 0.118	83.4 ± 10.8	0.085 ± 0.038	0.823 ± 0.124	0.175 ± 0.008	0.315 ± 0.012	**89.3 ± 1.0**	0.055 ± 0.006	**0.890 ± 0.011**
	Weighted prob. ensemble	0.130 ± 0.013	0.236 ± 0.021	91.6 ± 1.7	0.043 ± 0.017	0.915 ± 0.017	0.179 ± 0.014	0.323 ± 0.024	88.7 ± 1.8	0.062 ± 0.018	0.883 ± 0.018
	BayesPE (public implementation)	0.135 ± 0.020	0.244 ± 0.031	91.9 ± 1.3	0.056 ± 0.018	0.918 ± 0.013	**0.174 ± 0.008**	**0.313 ± 0.013**	89.2 ± 0.6	0.056 ± 0.012	0.889 ± 0.006
	Dirichlet-ODIR	0.223 ± 0.101	0.470 ± 0.271	84.7 ± 8.7	0.080 ± 0.042	0.841 ± 0.096	0.207 ± 0.021	0.443 ± 0.251	86.6 ± 1.3	0.067 ± 0.019	0.860 ± 0.016
	LinC	0.216 ± 0.092	0.405 ± 0.147	86.1 ± 8.3	0.088 ± 0.047	0.857 ± 0.094	0.200 ± 0.024	0.394 ± 0.200	87.4 ± 1.5	0.057 ± 0.020	0.868 ± 0.017
	Simplex log pool	**0.127 ± 0.012**	**0.229 ± 0.018**	**92.2 ± 1.1**	**0.035 ± 0.011**	**0.921 ± 0.011**	0.176 ± 0.011	0.320 ± 0.018	88.8 ± 1.6	0.061 ± 0.020	0.884 ± 0.016
	Signed affine log pool	0.130 ± 0.014	0.236 ± 0.029	91.4 ± 1.6	0.041 ± 0.013	0.914 ± 0.016	0.188 ± 0.032	0.338 ± 0.050	87.8 ± 2.6	0.063 ± 0.013	0.873 ± 0.029
	CITP	0.137 ± 0.021	0.248 ± 0.035	91.6 ± 1.5	0.051 ± 0.028	0.915 ± 0.015	0.184 ± 0.012	0.322 ± 0.019	88.6 ± 0.9	0.055 ± 0.014	0.881 ± 0.010
Mistral-7B	Reference	0.150 ± 0.045	0.412 ± 0.115	90.9 ± 2.8	0.074 ± 0.022	0.908 ± 0.029	0.240 ± 0.010	0.835 ± 0.047	87.0 ± 0.8	0.120 ± 0.005	0.864 ± 0.007
	Temperature scaling	0.137 ± 0.037	**0.250 ± 0.067**	90.9 ± 2.8	**0.054 ± 0.016**	0.908 ± 0.029	**0.209 ± 0.008**	**0.355 ± 0.014**	87.0 ± 0.8	**0.053 ± 0.012**	0.864 ± 0.007
	Cal.-selected best prompt	0.146 ± 0.035	0.387 ± 0.096	91.3 ± 1.8	0.072 ± 0.017	0.913 ± 0.019	0.225 ± 0.021	0.738 ± 0.127	**87.8 ± 1.3**	0.112 ± 0.012	**0.870 ± 0.015**
	Mean prob. ensemble	0.137 ± 0.032	0.338 ± 0.074	91.7 ± 2.5	0.064 ± 0.016	0.917 ± 0.025	0.222 ± 0.006	0.693 ± 0.066	87.2 ± 0.4	0.109 ± 0.008	0.867 ± 0.004
	Weighted prob. ensemble	0.133 ± 0.030	0.352 ± 0.066	**92.1 ± 2.1**	0.063 ± 0.017	**0.921 ± 0.021**	0.217 ± 0.014	0.681 ± 0.074	87.6 ± 1.1	0.107 ± 0.009	0.869 ± 0.013
	BayesPE (public implementation)	0.133 ± 0.027	0.329 ± 0.063	92.0 ± 1.7	0.060 ± 0.014	0.920 ± 0.017	0.219 ± 0.011	0.680 ± 0.073	87.2 ± 0.5	0.105 ± 0.011	0.865 ± 0.007
	Dirichlet-ODIR	0.201 ± 0.080	0.454 ± 0.240	91.5 ± 2.0	0.129 ± 0.083	0.914 ± 0.020	0.227 ± 0.033	0.377 ± 0.042	85.9 ± 2.3	0.061 ± 0.029	0.849 ± 0.031
	LinC	0.144 ± 0.023	0.277 ± 0.041	91.7 ± 1.8	0.073 ± 0.026	0.917 ± 0.018	0.239 ± 0.014	0.409 ± 0.022	87.0 ± 0.7	0.109 ± 0.044	0.864 ± 0.006
	Simplex log pool	0.142 ± 0.028	0.379 ± 0.071	91.8 ± 1.8	0.068 ± 0.014	0.917 ± 0.018	0.224 ± 0.017	0.745 ± 0.076	87.7 ± 1.4	0.113 ± 0.009	0.870 ± 0.016
	Signed affine log pool	0.145 ± 0.028	0.391 ± 0.075	91.6 ± 1.8	0.070 ± 0.013	0.915 ± 0.018	0.236 ± 0.019	0.712 ± 0.099	87.0 ± 1.5	0.113 ± 0.011	0.862 ± 0.016
	CITP	**0.130 ± 0.033**	0.301 ± 0.081	91.8 ± 2.3	0.057 ± 0.023	0.918 ± 0.024	0.219 ± 0.024	0.514 ± 0.115	87.0 ± 1.7	0.096 ± 0.012	0.863 ± 0.017

Table 6: Detailed multiple-choice results on ARC-Challenge and CommonsenseQA (mean ± outer-seed SD over 10 runs). Brier and NLL are unnormalized. Arrows indicate the preferred direction. Bold and underlined values indicate the best and second-best results, respectively.

Model	Method	ARC-C					CommonsenseQA				
		Brier↓	NLL↓	Acc. (%)↑	ECE↓	Macro-F1↑	Brier↓	NLL↓	Acc. (%)↑	ECE↓	Macro-F1↑
Qwen3-8B	Reference	0.485 ± 0.118	1.181 ± 0.244	68.3 ± 7.9	0.200 ± 0.057	0.671 ± 0.086	0.985 ± 0.107	2.938 ± 0.467	42.2 ± 4.7	0.450 ± 0.081	0.386 ± 0.057
	Temperature scaling	0.418 ± 0.079	0.804 ± 0.124	68.3 ± 7.9	0.093 ± 0.020	0.671 ± 0.086	0.685 ± 0.026	1.315 ± 0.059	42.2 ± 4.7	0.116 ± 0.035	0.386 ± 0.057
	Cal.-selected best prompt	0.452 ± 0.059	1.094 ± 0.181	71.2 ± 4.8	0.177 ± 0.031	0.702 ± 0.057	0.792 ± 0.120	2.059 ± 0.392	49.5 ± 10.5	0.351 ± 0.077	0.468 ± 0.126
	Mean prob. ensemble	0.362 ± 0.086	0.692 ± 0.163	73.8 ± 7.0	0.104 ± 0.024	0.735 ± 0.071	0.724 ± 0.072	1.482 ± 0.203	45.6 ± 5.6	0.245 ± 0.067	0.431 ± 0.061
	Weighted prob. ensemble	0.329 ± 0.051	0.648 ± 0.100	75.9 ± 4.0	0.107 ± 0.034	0.761 ± 0.040	0.568 ± 0.073	1.225 ± 0.202	57.7 ± 6.8	0.132 ± 0.033	0.563 ± 0.069
	BayesPE (public implementation)	0.326 ± 0.045	0.638 ± 0.108	76.8 ± 3.9	0.115 ± 0.026	0.768 ± 0.039	0.595 ± 0.081	1.261 ± 0.203	55.6 ± 9.2	0.153 ± 0.055	0.542 ± 0.097
	Dirichlet-ODIR	0.296 ± 0.125	0.657 ± 0.223	79.0 ± 16.3	0.091 ± 0.055	0.785 ± 0.181	0.487 ± 0.077	1.025 ± 0.171	64.2 ± 8.3	0.088 ± 0.015	0.628 ± 0.100
	LinC	0.394 ± 0.104	0.957 ± 0.513	74.6 ± 9.6	0.154 ± 0.040	0.744 ± 0.105	0.652 ± 0.045	1.840 ± 1.464	49.1 ± 7.7	0.143 ± 0.044	0.458 ± 0.099
	Simplex log pool	0.207 ± 0.030	0.428 ± 0.065	85.8 ± 2.6	**0.053 ± 0.010**	0.858 ± 0.028	0.510 ± 0.123	1.043 ± 0.289	63.5 ± 8.2	0.141 ± 0.087	0.620 ± 0.089
	Signed affine log pool	**0.197 ± 0.022**	0.432 ± 0.072	**86.7 ± 1.6**	0.059 ± 0.014	**0.868 ± 0.016**	0.404 ± 0.028	0.843 ± 0.117	71.2 ± 2.7	0.109 ± 0.034	0.702 ± 0.028
	CITP	0.203 ± 0.029	**0.411 ± 0.050**	86.2 ± 2.1	0.054 ± 0.011	0.863 ± 0.021	**0.379 ± 0.039**	**0.730 ± 0.076**	**71.8 ± 3.3**	**0.083 ± 0.014**	**0.715 ± 0.030**
Llama-3.1-8B	Reference	0.457 ± 0.049	0.868 ± 0.078	66.8 ± 5.2	0.084 ± 0.025	0.665 ± 0.057	0.504 ± 0.115	1.015 ± 0.170	63.8 ± 11.1	0.118 ± 0.090	0.626 ± 0.131
	Temperature scaling	0.454 ± 0.042	0.856 ± 0.067	66.8 ± 5.2	0.090 ± 0.022	0.665 ± 0.057	0.490 ± 0.069	0.981 ± 0.111	63.8 ± 11.1	0.098 ± 0.040	0.626 ± 0.131
	Cal.-selected best prompt	0.385 ± 0.069	0.767 ± 0.127	73.7 ± 6.4	0.074 ± 0.017	0.738 ± 0.063	0.450 ± 0.024	0.929 ± 0.049	69.2 ± 2.0	0.100 ± 0.021	0.687 ± 0.022
	Mean prob. ensemble	0.380 ± 0.029	0.738 ± 0.040	74.7 ± 3.5	0.118 ± 0.020	0.748 ± 0.035	0.445 ± 0.050	0.885 ± 0.091	68.0 ± 5.0	0.081 ± 0.024	0.674 ± 0.051
	Weighted prob. ensemble	0.339 ± 0.019	0.677 ± 0.031	77.8 ± 1.5	0.093 ± 0.034	0.778 ± 0.015	0.421 ± 0.029	0.840 ± 0.053	71.0 ± 3.3	0.087 ± 0.029	0.706 ± 0.033
	BayesPE (public implementation)	0.352 ± 0.027	0.697 ± 0.037	76.8 ± 2.9	0.112 ± 0.047	0.768 ± 0.027	0.426 ± 0.026	0.851 ± 0.051	70.6 ± 2.5	0.088 ± 0.038	0.699 ± 0.026
	Dirichlet-ODIR	0.354 ± 0.030	0.990 ± 0.363	79.1 ± 2.0	0.111 ± 0.059	0.790 ± 0.020	0.453 ± 0.045	0.972 ± 0.124	69.1 ± 3.1	0.125 ± 0.040	0.686 ± 0.028
	LinC	0.389 ± 0.032	0.896 ± 0.176	77.5 ± 1.8	0.161 ± 0.039	0.775 ± 0.018	0.482 ± 0.051	0.996 ± 0.087	69.8 ± 1.7	0.166 ± 0.060	0.691 ± 0.019
	Simplex log pool	0.302 ± 0.013	0.605 ± 0.026	**80.4 ± 0.9**	0.093 ± 0.023	**0.804 ± 0.009**	0.397 ± 0.031	0.792 ± 0.059	72.2 ± 3.0	**0.073 ± 0.025**	0.716 ± 0.032
	Signed affine log pool	0.302 ± 0.016	0.605 ± 0.044	79.0 ± 1.1	**0.068 ± 0.019**	0.790 ± 0.012	0.396 ± 0.027	0.790 ± 0.055	72.5 ± 2.5	0.076 ± 0.023	0.721 ± 0.026
	CITP	**0.292 ± 0.010**	**0.586 ± 0.015**	80.0 ± 1.8	0.073 ± 0.020	0.799 ± 0.018	**0.379 ± 0.019**	**0.756 ± 0.043**	**74.2 ± 2.0**	0.095 ± 0.022	**0.736 ± 0.020**
Mistral-7B	Reference	0.389 ± 0.018	1.123 ± 0.069	76.8 ± 1.6	0.177 ± 0.012	0.767 ± 0.016	0.509 ± 0.022	1.695 ± 0.104	71.4 ± 1.2	0.244 ± 0.011	0.705 ± 0.012
	Temperature scaling	0.326 ± 0.008	**0.623 ± 0.017**	76.8 ± 1.6	**0.077 ± 0.013**	0.767 ± 0.016	**0.392 ± 0.011**	**0.787 ± 0.025**	71.4 ± 1.2	**0.097 ± 0.016**	0.705 ± 0.012
	Cal.-selected best prompt	0.400 ± 0.046	1.204 ± 0.173	76.3 ± 2.7	0.176 ± 0.028	0.764 ± 0.027	0.493 ± 0.037	1.653 ± 0.133	72.2 ± 2.4	0.232 ± 0.024	0.714 ± 0.024
	Mean prob. ensemble	0.341 ± 0.009	0.843 ± 0.041	77.5 ± 1.2	0.134 ± 0.012	0.775 ± 0.012	0.441 ± 0.018	1.077 ± 0.085	71.9 ± 1.4	0.190 ± 0.017	0.709 ± 0.013
	Weighted prob. ensemble	0.324 ± 0.013	0.833 ± 0.052	78.6 ± 1.0	0.111 ± 0.023	0.786 ± 0.010	0.409 ± 0.017	1.015 ± 0.089	72.9 ± 1.2	0.145 ± 0.011	**0.721 ± 0.013**
	BayesPE (public implementation)	0.324 ± 0.010	0.811 ± 0.040	78.6 ± 1.2	0.109 ± 0.022	0.786 ± 0.012	0.411 ± 0.016	1.004 ± 0.080	72.9 ± 1.6	0.150 ± 0.017	0.721 ± 0.016
	Dirichlet-ODIR	0.389 ± 0.073	1.057 ± 0.511	77.3 ± 1.7	0.123 ± 0.075	0.772 ± 0.016	0.478 ± 0.068	1.295 ± 0.665	69.6 ± 4.0	0.146 ± 0.067	0.689 ± 0.039
	LinC	0.389 ± 0.030	1.035 ± 0.760	77.0 ± 1.1	0.116 ± 0.046	0.770 ± 0.011	0.488 ± 0.038	1.034 ± 0.069	71.5 ± 1.5	0.144 ± 0.070	0.706 ± 0.014
	Simplex log pool	0.330 ± 0.013	0.917 ± 0.052	**79.8 ± 1.0**	0.138 ± 0.011	**0.797 ± 0.010**	0.464 ± 0.023	1.377 ± 0.107	**73.0 ± 0.9**	0.215 ± 0.013	0.721 ± 0.011
	Signed affine log pool	0.338 ± 0.020	0.924 ± 0.046	79.0 ± 1.9	0.146 ± 0.016	0.790 ± 0.018	0.480 ± 0.045	1.420 ± 0.178	71.8 ± 2.3	0.220 ± 0.020	0.710 ± 0.025
	CITP	**0.309 ± 0.013**	0.653 ± 0.041	78.7 ± 2.0	0.083 ± 0.020	0.787 ± 0.020	0.405 ± 0.026	0.836 ± 0.117	72.0 ± 1.9	0.099 ± 0.035	0.714 ± 0.021

Table 7: Detailed multiple-choice results on HellaSwag and MMLU (mean ± outer-seed SD over 10 runs). Brier and NLL are unnormalized. Arrows indicate the preferred direction. Bold and underlined values indicate the best and second-best results, respectively.

Model	Method	HellaSwag					MMLU				
		Brier↓	NLL↓	Acc. (%)↑	ECE↓	Macro-F1↑	Brier↓	NLL↓	Acc. (%)↑	ECE↓	Macro-F1↑
Qwen3-8B	Reference	0.681 ± 0.264	1.576 ± 0.757	54.8 ± 15.8	0.267 ± 0.155	0.502 ± 0.196	0.930 ± 0.025	2.134 ± 0.047	40.5 ± 2.4	0.420 ± 0.019	0.364 ± 0.022
	Temperature scaling	0.555 ± 0.124	1.037 ± 0.213	54.8 ± 15.8	0.120 ± 0.031	0.502 ± 0.196	0.670 ± 0.011	1.230 ± 0.019	40.5 ± 2.4	0.112 ± 0.027	0.364 ± 0.022
	Cal.-selected best prompt	0.533 ± 0.183	1.191 ± 0.468	64.3 ± 12.5	0.192 ± 0.112	0.613 ± 0.161	0.906 ± 0.055	2.109 ± 0.140	41.7 ± 3.6	0.413 ± 0.032	0.386 ± 0.027
	Mean prob. ensemble	0.535 ± 0.192	1.021 ± 0.370	60.3 ± 17.4	0.183 ± 0.104	0.566 ± 0.216	0.688 ± 0.018	1.319 ± 0.106	45.2 ± 2.0	0.197 ± 0.017	0.419 ± 0.019
	Weighted prob. ensemble	0.443 ± 0.173	0.857 ± 0.335	68.5 ± 13.4	0.141 ± 0.078	0.659 ± 0.167	0.617 ± 0.035	1.220 ± 0.129	51.2 ± 2.4	0.106 ± 0.037	0.486 ± 0.029
	BayesPE (public implementation)	0.452 ± 0.170	0.874 ± 0.335	67.1 ± 12.8	0.130 ± 0.088	0.648 ± 0.157	0.637 ± 0.040	1.255 ± 0.130	50.0 ± 2.8	0.109 ± 0.047	0.474 ± 0.032
	Dirichlet-ODIR	0.361 ± 0.072	0.776 ± 0.238	74.6 ± 6.2	0.082 ± 0.035	0.738 ± 0.067	0.635 ± 0.009	1.286 ± 0.113	47.0 ± 3.2	0.103 ± 0.025	0.460 ± 0.029
	LinC	0.429 ± 0.081	0.879 ± 0.124	69.5 ± 9.1	0.136 ± 0.040	0.686 ± 0.092	0.674 ± 0.017	1.291 ± 0.058	43.4 ± 3.0	0.113 ± 0.033	0.413 ± 0.028
	Simplex log pool	0.416 ± 0.208	0.832 ± 0.458	71.4 ± 13.8	0.127 ± 0.112	0.696 ± 0.164	0.547 ± 0.035	1.054 ± 0.105	57.6 ± 3.1	0.091 ± 0.032	0.560 ± 0.040
	Signed affine log pool	0.314 ± 0.114	0.630 ± 0.234	79.0 ± 8.1	0.090 ± 0.052	0.784 ± 0.085	0.545 ± 0.049	1.067 ± 0.145	58.2 ± 3.8	0.098 ± 0.035	0.565 ± 0.050
	CITP	**0.270 ± 0.052**	**0.536 ± 0.093**	**81.8 ± 4.3**	**0.082 ± 0.022**	**0.816 ± 0.043**	**0.516 ± 0.019**	**0.971 ± 0.039**	**59.5 ± 2.4**	**0.072 ± 0.027**	**0.583 ± 0.026**
Llama-3.1-8B	Reference	0.355 ± 0.086	0.704 ± 0.150	74.7 ± 7.0	0.087 ± 0.028	0.747 ± 0.066	0.883 ± 0.024	1.711 ± 0.071	36.5 ± 1.2	0.361 ± 0.019	0.363 ± 0.014
	Temperature scaling	0.352 ± 0.078	0.700 ± 0.128	74.7 ± 7.0	0.084 ± 0.029	0.747 ± 0.066	0.695 ± 0.012	1.265 ± 0.015	36.5 ± 1.2	0.127 ± 0.020	0.363 ± 0.014
	Cal.-selected best prompt	0.308 ± 0.044	0.628 ± 0.085	78.2 ± 4.0	**0.076 ± 0.023**	0.780 ± 0.040	0.721 ± 0.034	1.420 ± 0.071	48.8 ± 2.5	0.241 ± 0.028	0.483 ± 0.025
	Mean prob. ensemble	0.330 ± 0.054	0.652 ± 0.094	77.2 ± 4.7	0.120 ± 0.017	0.775 ± 0.043	0.682 ± 0.010	1.273 ± 0.031	43.8 ± 1.3	0.184 ± 0.015	0.445 ± 0.013
	Weighted prob. ensemble	0.292 ± 0.039	0.587 ± 0.077	80.2 ± 3.0	0.098 ± 0.028	0.801 ± 0.030	0.576 ± 0.022	1.126 ± 0.057	53.0 ± 2.3	0.113 ± 0.019	0.525 ± 0.025
	BayesPE (public implementation)	0.304 ± 0.042	0.608 ± 0.075	80.0 ± 3.2	0.139 ± 0.042	0.799 ± 0.030	0.593 ± 0.019	1.164 ± 0.053	52.9 ± 2.5	0.115 ± 0.017	0.524 ± 0.028
	Dirichlet-ODIR	0.303 ± 0.035	0.615 ± 0.055	79.6 ± 2.1	0.104 ± 0.048	0.797 ± 0.021	0.626 ± 0.008	1.183 ± 0.031	49.5 ± 3.7	**0.083 ± 0.021**	0.490 ± 0.039
	LinC	0.316 ± 0.036	0.684 ± 0.127	79.7 ± 2.8	0.135 ± 0.027	0.796 ± 0.028	0.645 ± 0.024	1.226 ± 0.058	46.3 ± 3.4	0.087 ± 0.024	0.465 ± 0.033
	Simplex log pool	0.276 ± 0.031	0.555 ± 0.063	**81.5 ± 2.8**	0.104 ± 0.035	0.813 ± 0.028	0.538 ± 0.020	1.033 ± 0.048	**58.7 ± 2.1**	0.087 ± 0.016	**0.584 ± 0.020**
	Signed affine log pool	0.278 ± 0.024	0.557 ± 0.049	**81.5 ± 2.7**	0.093 ± 0.033	**0.813 ± 0.028**	0.538 ± 0.022	1.032 ± 0.048	57.9 ± 2.8	0.106 ± 0.019	0.575 ± 0.028
	CITP	**0.271 ± 0.019**	**0.544 ± 0.040**	81.4 ± 2.5	0.102 ± 0.033	0.812 ± 0.026	**0.533 ± 0.010**	**1.017 ± 0.021**	58.1 ± 1.5	0.088 ± 0.014	0.578 ± 0.013
Mistral-7B	Reference	0.527 ± 0.056	1.364 ± 0.157	67.7 ± 4.3	0.215 ± 0.030	0.674 ± 0.044	0.726 ± 0.017	2.095 ± 0.060	55.9 ± 1.2	0.332 ± 0.014	0.555 ± 0.012
	Temperature scaling	0.462 ± 0.040	0.883 ± 0.079	67.7 ± 4.3	**0.095 ± 0.030**	0.674 ± 0.044	**0.545 ± 0.006**	**1.034 ± 0.013**	55.9 ± 1.2	0.092 ± 0.018	0.555 ± 0.012
	Cal.-selected best prompt	0.509 ± 0.041	1.266 ± 0.068	68.1 ± 3.7	0.213 ± 0.027	0.676 ± 0.042	0.680 ± 0.037	1.880 ± 0.141	57.4 ± 2.1	0.297 ± 0.022	0.574 ± 0.021
	Mean prob. ensemble	0.443 ± 0.032	0.912 ± 0.086	68.8 ± 3.0	0.130 ± 0.022	0.686 ± 0.032	0.624 ± 0.008	1.417 ± 0.043	57.0 ± 1.0	0.254 ± 0.013	0.566 ± 0.010
	Weighted prob. ensemble	0.431 ± 0.029	0.901 ± 0.103	70.2 ± 2.5	0.120 ± 0.027	0.697 ± 0.027	0.573 ± 0.012	1.329 ± 0.042	58.1 ± 1.8	0.199 ± 0.018	0.580 ± 0.019
	BayesPE (public implementation)	0.429 ± 0.031	0.881 ± 0.093	69.5 ± 3.1	0.111 ± 0.025	0.691 ± 0.035	0.569 ± 0.014	1.314 ± 0.044	**58.7 ± 1.7**	0.187 ± 0.024	**0.586 ± 0.018**
	Dirichlet-ODIR	0.477 ± 0.042	1.034 ± 0.388	69.2 ± 4.3	0.137 ± 0.039	0.689 ± 0.045	0.574 ± 0.018	1.865 ± 2.153	56.0 ± 1.4	**0.084 ± 0.018**	0.558 ± 0.014
	LinC	0.469 ± 0.049	0.915 ± 0.080	69.4 ± 3.8	0.100 ± 0.051	0.692 ± 0.040	0.608 ± 0.008	1.159 ± 0.030	56.2 ± 1.5	0.139 ± 0.017	0.558 ± 0.015
	Simplex log pool	0.447 ± 0.029	0.992 ± 0.075	**71.4 ± 2.1**	0.174 ± 0.019	**0.711 ± 0.024**	0.635 ± 0.021	1.564 ± 0.055	58.5 ± 1.8	0.267 ± 0.017	0.583 ± 0.018
	Signed affine log pool	0.449 ± 0.042	1.004 ± 0.086	71.2 ± 2.8	0.172 ± 0.024	0.709 ± 0.031	0.640 ± 0.023	1.575 ± 0.065	58.1 ± 2.0	0.271 ± 0.020	0.579 ± 0.020
	CITP	**0.397 ± 0.018**	**0.754 ± 0.055**	71.4 ± 1.8	0.097 ± 0.026	0.709 ± 0.017	0.563 ± 0.026	1.108 ± 0.060	58.0 ± 1.5	0.164 ± 0.037	0.577 ± 0.015

C.2 Gain Decomposition across Task Regimes

We report the model-family comparisons separately for fixed-label classification and MCQA.

Table 8: Regime-specific mechanism decomposition. Metrics are first averaged over the ten outer seeds within each model–task pair and then macro-averaged over the 15 fixed-label or 12 MCQA model–task pairs. Shuffled-contrast cells retain the five-shuffle average used in the main mechanism analysis.

Method / restriction	nBrier ↓	nNLL ↓	Accuracy (%) ↑
Fixed-label (150 cells)			
Global calibration			
Reference	0.5141	0.8015	78.18
Scale only	0.4620	0.4986	78.18
Bias only	0.4245	0.6210	81.65
Scale+Bias	0.3986	<u>0.4491</u>	81.74
Prompt-contrast refinement			
CITP MAP — shuffled contrasts	0.4323	0.5365	80.62
CITP MAP — matched contrasts	<u>0.3517</u>	0.4569	**84.62**
CITP	**0.3464**	**0.4028**	<u>84.49</u>
MCQA (120 cells)			
Global calibration			
Reference	0.8117	1.0592	59.95
Scale only	0.6613	0.6688	59.95
Bias only	0.5998	0.7670	69.91
Scale+Bias	0.5546	0.5817	69.58
Prompt-contrast refinement			
CITP MAP — shuffled contrasts	0.5899	0.6372	67.99
CITP MAP — matched contrasts	**0.4930**	<u>0.5337</u>	**73.26**
CITP	<u>0.4940</u>	**0.5158**	<u>72.76</u>

C.3 Posterior Marginalization

We compare posterior-predictive and MAP predictions from the same fit and posterior-selected prior.

Figure 4 shows larger marginalization gains in nBrier and nNLL at smaller calibration budgets. Figure 5 shows that NLL gains tend to be larger for queries with greater coefficient-posterior disagreement, while the posterior predictive exhibits lower incidence of high-confidence errors across the examined thresholds.

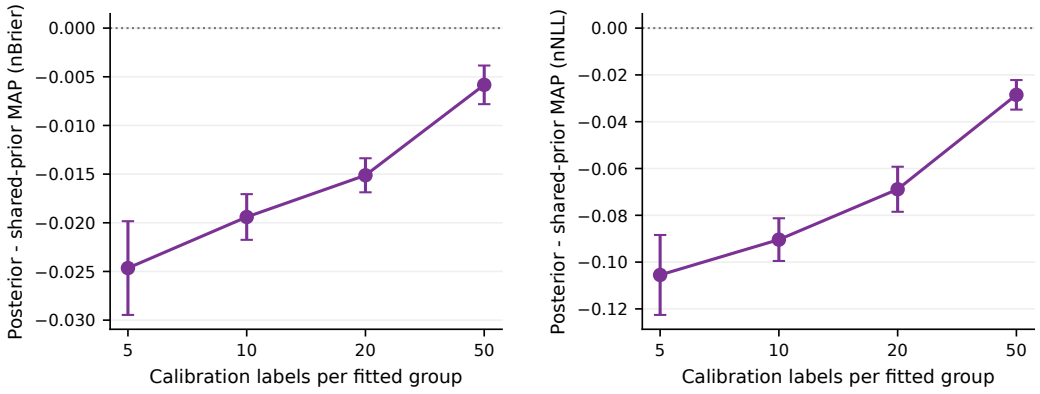


Figure 4: Differences in nBrier and nNLL (CITP minus shared-prior MAP) versus the number of calibration labels per fitted group. Within each seed, differences are averaged over 27 model–task pairs; error bars show paired 95% t confidence intervals over ten seeds. Negative values favor posterior marginalization.

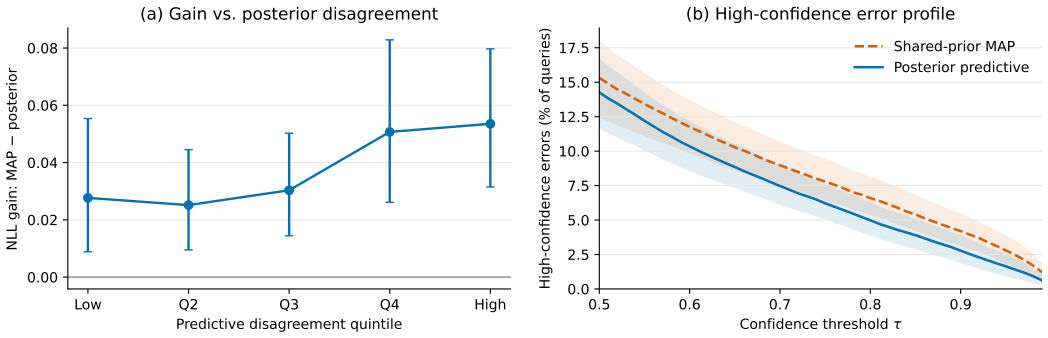


Figure 5: Posterior marginalization with the prior and MAP fit held fixed. (a) NLL gain (MAP minus posterior) by predictive-disagreement quintile. (b) Percentage of all queries that are incorrect and have confidence at least τ. Shading and error bars show pointwise 95% bootstrap intervals; definitions and resampling details are in Appendix E.

We further assess the Laplace approximation in CITP by comparing its posterior predictive probabilities with those obtained using the No-U-Turn Sampler (Hoffman & Gelman, 2014) (NUTS) in Table 9.

Table 9: Predictive agreement between Laplace and NUTS for CITP on three models and nine tasks at outer seed 42 (27 reporting cells, 36 fitted groups). For cell c, probability MAE is $\frac{1}{N_c K_c} \sum_{i=1}^{N_c} \sum_{k=1}^{K_c} |p^A_{cik} - p^B_{cik}|$. We report the unweighted mean over 27 cells; MMLU predictions are concatenated across its four groups before computing its cell mean. Metric differences are $A - B$. The larger Laplace sample is an integration diagnostic; the main predictor uses 200 draws.

Comparison ($A - B$)	Mean probability MAE	Δ nBrier	Δ nNLL
Laplace (200) – NUTS	0.0110	+0.0022	-0.0040
Laplace (8192) – NUTS	0.0110	+0.0020	-0.0045
Laplace (200) – Laplace (8192)	0.0019	+0.0002	+0.0005

Laplace and NUTS yield similar posterior predictive probabilities on average, with a mean probability MAE of 0.0110.

C.4 Budgets and Model Size Ablations

We study sensitivity to the number of labeled calibration examples, the number of available prompt realizations, and model size. These experiments evaluate how CITP performs across supervision and inference budgets and frozen-model sizes.

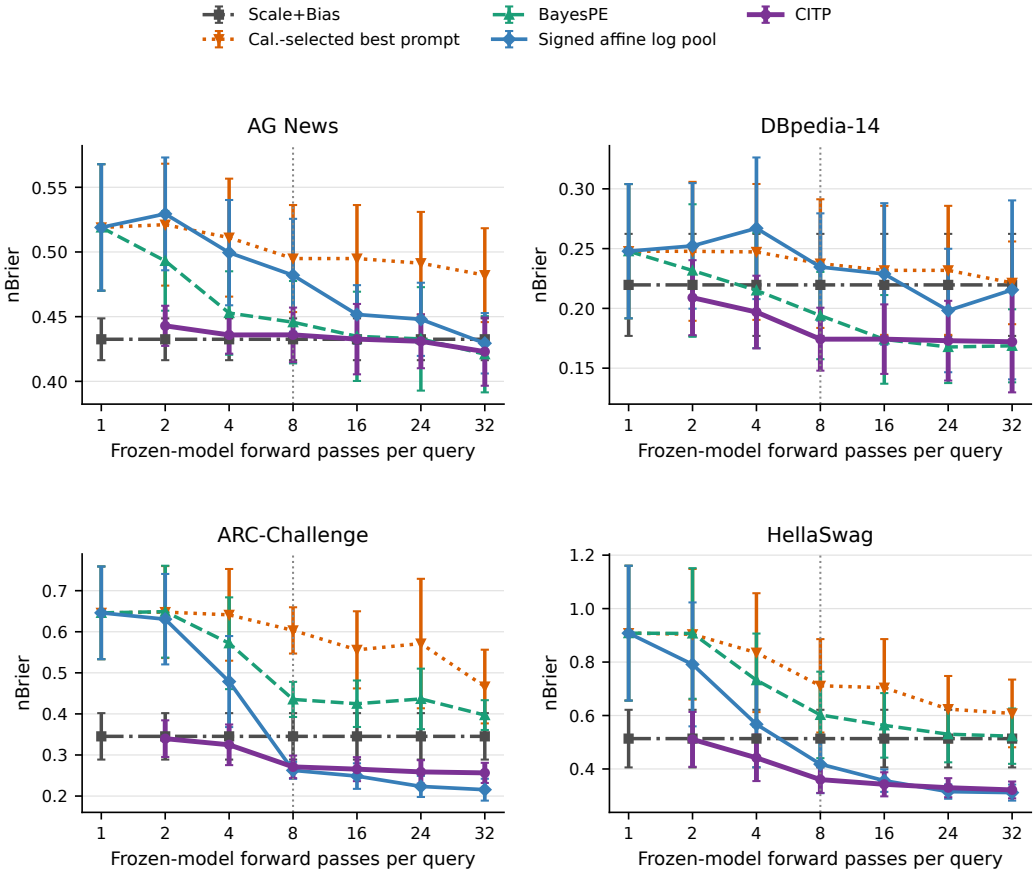


Figure 6: Normalized Brier score versus forward-pass budget for Qwen3-8B on AG News, DBpedia-14, ARC-Challenge, and HellaSwag, using 50 calibration examples. Prompt subsets are nested, and the dotted line marks the main eight-view bank. CITP starts at two views because it requires at least one prompt contrast. Curves show means and 95% t confidence intervals over ten outer seeds.

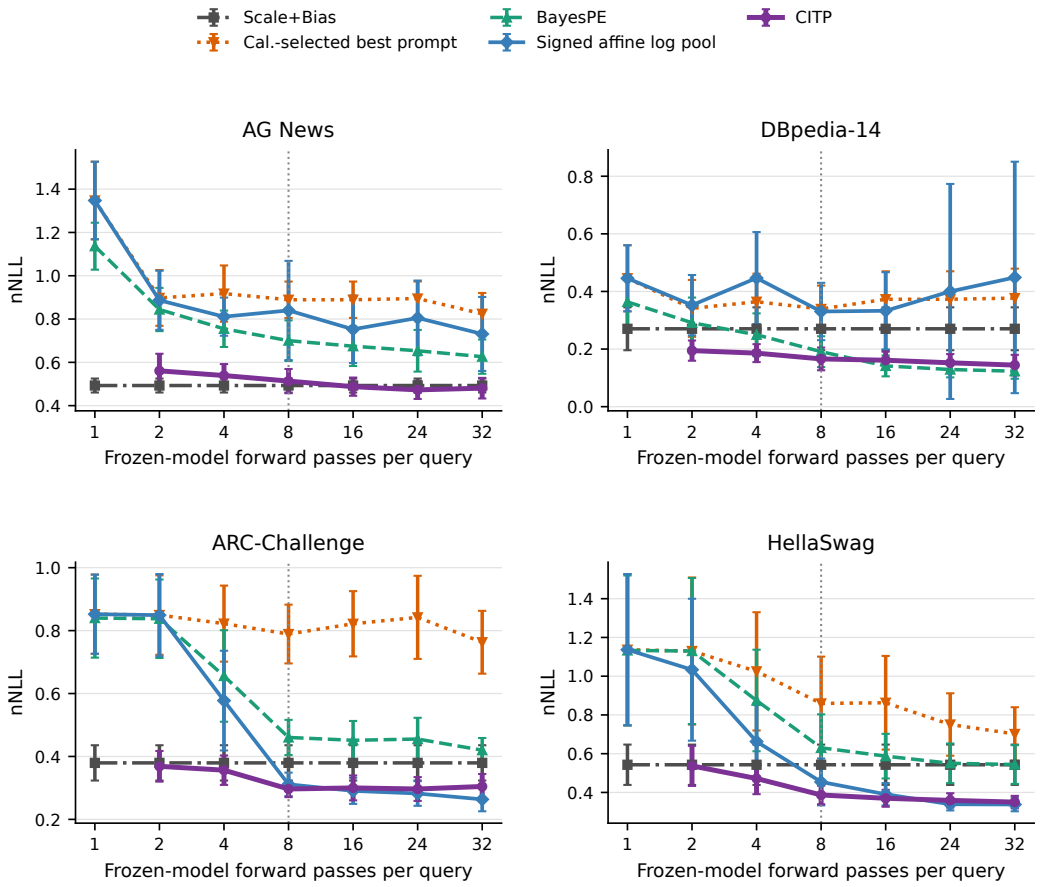


Figure 7: Normalized NLL versus forward-pass budget under the same settings as Figure 6. Curves show means and 95% t confidence intervals over ten outer seeds, with the dotted line marking the main eight-view bank.

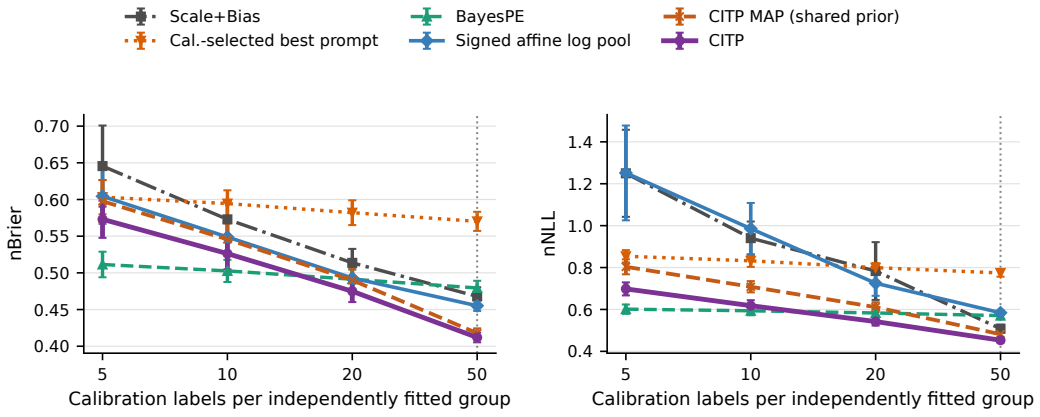


Figure 8: Calibration-label budget sensitivity with the eight-view prompt bank fixed. Metrics are macro-averaged over the 27 model–task pairs within each outer seed; curves show means and 95% t confidence intervals over ten seeds. Label counts are per independently fitted group, including each MMLU supercategory. The dotted line marks the main setting of 50 calibration labels.

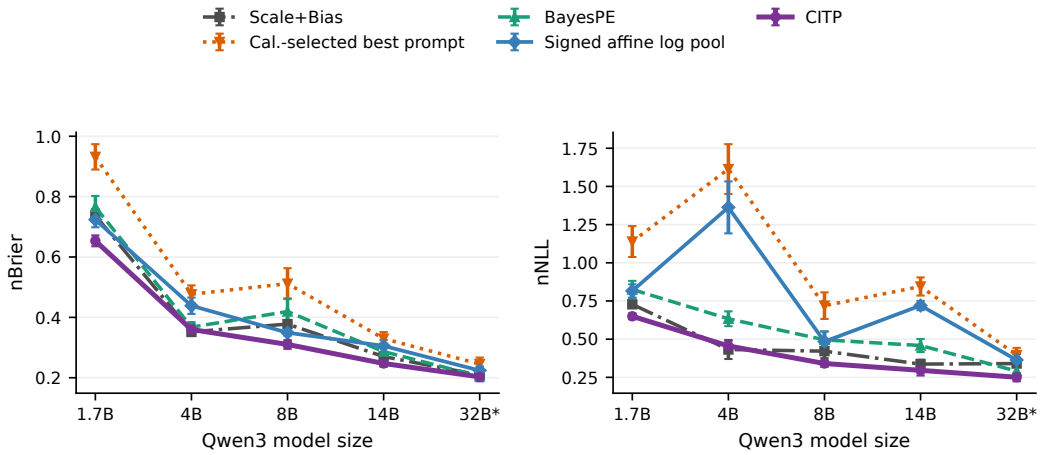


Figure 9: Model-size sensitivity for Qwen3 on AG News, DBpedia-14, ARC-Challenge, and HellaSwag. For each model size, metrics are averaged over the four tasks within each seed; curves show means and 95% t confidence intervals over ten seeds. *The 32B model was evaluated on an NVIDIA A100; smaller models were evaluated on an NVIDIA RTX 5090.

CITP remains competitive across the evaluated Qwen3 model sizes and forward-pass budgets. Its relative performance depends on the calibration budget: CITP achieves lower mean nBrier and nNLL than BayesPE with 20 or 50 labels per fitted group, whereas BayesPE performs better on both metrics with five or ten labels.

C.5 Option-Only Baseline

Option-order debiasing methods such as PriDE (Zheng et al., 2024) and cyclic permutation can also be interpreted as methods using task-equivalent prompt variants, since changing the option order does not change the semantics of the prompt. We therefore provide an additional comparison.

Table 10: Option-permutation comparison over 120 MCQA model–task–seed cells. Metrics are computed per cell, averaged over ten outer seeds within each of the 12 model–task pairs, and then macro-averaged over those pairs. PriDE follows its native $\alpha = 0.05$ protocol; its metrics are averaged over five unlabeled estimation-subset replicates. The lower block uses the same reference plus seven content-preserving option permutations. MMLU uses corrected subject-matched four-shot demonstrations. Supervised rows use only calibration labels. Native test-time frozen-model forwards per query are reported explicitly.

Method	Labels	Fwds./query	nBrier ↓	nNLL ↓	Accuracy (%) ↑	ECE ↓
Native option-order controls						
PriDE ($\alpha = 0.05$)	None	1.16	0.6295	0.7802	68.18	0.1651
Cyclic permutation average	None	4.25	0.5326	0.5716	75.08	0.1905
Matched eight-view option bank						
Matched-8 probability average	None	8.00	0.5247	0.5616	75.26	0.1840
Weighted probability ensemble	Calibration	8.00	0.5397	0.5933	71.45	0.1361
BayesPE (public impl.)	Calibration	8.00	0.5321	0.5703	71.53	0.1408
Simplex log pool	Calibration	8.00	<u>0.4632</u>	0.5432	<u>76.73</u>	0.1261
Signed affine log pool	Calibration	8.00	**0.4627**	0.5453	**76.77**	0.1258
CITP MAP	Calibration	8.00	0.4767	<u>0.5268</u>	74.63	<u>0.0961</u>
CITP	Calibration	8.00	0.4706	**0.4954**	74.55	**0.0881**

Table 10 compares native option-order controls and supervised refinement methods using an option-only bank. Within the matched eight-view bank, learned logarithmic pooling and CITP improve

nBrier, nNLL, and ECE over uniform and weighted probability averaging. CITP achieves the lowest nNLL and ECE, whereas the Signed affine log pool achieves the lowest nBrier and highest accuracy.

D Actual Prompt Perturbation Examples

We provide an actual four-shot prompt bank from ARC-Challenge (seed 42). The same four demonstrations are shared by all eight prompt members; only the designated prompt component is changed.

D.1 Demonstrations Used in the Example Run

Demonstration 1 (D1).

```
In early 2003, the Human Genome Project identified the
sequence of base pairs in the genes in human DNA. With all
of this information, many of the functions of the genes are
still unknown.  Currently, scientists are studying many of
these genes in order to learn more about them.  What is the
significance of this new genetic discovery?
Choices:
A. It can provide new methods for creating diseases.
B. It can lead to faster chromosome replication.
C. It can lead to a simpler structure for DNA.
D. It can provide new ways to treat diseases.
Answer:  D
```

Demonstration 2 (D2).

```
Ten grams of sugar is dissolved in 100 grams (g) of water.
How many grams is the sugar-and-water solution?
Choices:
A. 90 g
B. 100 g
C. 110 g
D. 1000 g
Answer:  C
```

Demonstration 3 (D3).

```
Which protist is a bi-flagellate autotroph?
Choices:
A. amoeba
B. euglena
C. paramecium
D. volvox
Answer:  D
```

Demonstration 4 (D4).

```
Which of these is one similarity of the elements in the
first group of the periodic table?
Choices:
A. The elements in the group have similar atomic masses.
B. The elements in the group have the same number of
electrons.
C. The elements in the group are very unreactive metals.
D. The elements in the group react with oxygen in similar
ways.
Answer:  D
```

D.2 TEMPLATE PERTURBATIONS

The reference prompt uses the following instruction:

```
Choose the best answer option.
```

The two template perturbations replace only this instruction. The four demonstrations and their order remain unchanged.

Template 0.

```
Select the most appropriate answer choice.
```

Template 1.

```
Read each problem and respond with the letter of its best
answer.
```

D.3 DEMONSTRATION-ORDER PERTURBATIONS

The actual demonstration sequences used in this run were:

Table 11: Actual demonstration-order perturbations. Each symbol refers to the complete demonstration shown above.

Prompt member	Position 1	Position 2	Position 3	Position 4
Clean	D1	D2	D3	D4
Order 0	D1	D3	D2	D4
Order 1	D1	D3	D4	D2

Thus, for example, an abbreviated illustration of the `Order 0` demonstration block is:

```
[D1] In early 2003, the Human Genome Project identified
the sequence of base pairs in the genes in human DNA. ...
Answer:  D

[D3] Which protist is a bi-flagellate autotroph?
A. amoeba
B. euglena
C. paramecium
D. volvox
Answer:  D

[D2] Ten grams of sugar is dissolved in 100 grams (g) of
water.  How many grams is the sugar-and-water solution?
A. 90 g
B. 100 g
C. 110 g
D. 1000 g
Answer:  C

[D4] Which of these is one similarity of the elements in the
first group of the periodic table?  ... Answer:  D
```

An abbreviated illustration of the `Order 1` demonstration block is:

```
[D1] In early 2003, the Human Genome Project identified
the sequence of base pairs in the genes in human DNA. ...
Answer:  D

[D3] Which protist is a bi-flagellate autotroph?
```

```
A. amoeba
B. euglena
C. paramecium
D. volvox
Answer:  D

[D4] Which of these is one similarity of the elements in the
first group of the periodic table?  ... Answer:  D

[D2] Ten grams of sugar is dissolved in 100 grams (g) of
water.  How many grams is the sugar-and-water solution?
A. 90 g
B. 100 g
C. 110 g
D. 1000 g
Answer:  C
```

D.4 Option-Presentation Perturbations

The option-presentation perturbations change both the displayed option letters and, for the second perturbation, their display order. The underlying candidate and the correct answer content remain unchanged. For D2, whose correct answer is *110 g*, the actual renderings are:

Reference.

```
Ten grams of sugar is dissolved in 100 grams (g) of water.
How many grams is the sugar-and-water solution?
Choices:
A. 90 g
B. 100 g
C. 110 g
D. 1000 g
Answer:  C
```

Option Permutation 0.

```
Ten grams of sugar is dissolved in 100 grams (g) of water.
How many grams is the sugar-and-water solution?
Choices:
B. 90 g
C. 100 g
D. 110 g
A. 1000 g
Answer:  D
```

Option Permutation 1.

```
Ten grams of sugar is dissolved in 100 grams (g) of water.
How many grams is the sugar-and-water solution?
Choices:
B. 1000 g
A. 110 g
D. 100 g
C. 90 g
Answer:  A
```

Before computing metrics or constructing logit contrasts, predictions from the permuted prompts are mapped back to the common canonical candidate order. Therefore, these perturbations do not change the semantic identity of the candidates.

D.5 Pure-Format Perturbation

The format perturbation preserves the reference instruction, demonstration contents, order, option labels, and answer labels. It inserts separators around the demonstration block:

```
Choose the best answer option.
-----
[D1]
[D2]
[D3]
[D4]
-----
[Test query]
Answer:
```

E Experimental and Implementation Details

E.1 Models and Data Splits

We evaluate Qwen3-8B, Llama-3.1-8B-Instruct, and Mistral-7B-Instruct-v0.3. Models use their native chat format, with Qwen3 thinking mode disabled, bfloat16 inference, batch size one, and no quantization. Candidate scores are next-token answer-token logits after the generation prompt, mapped to the canonical candidate order.

We use ten outer seeds, 42–51. Each unit contains four demonstrations, 50 calibration examples, and 200 test examples. For non-MMLU tasks, demonstrations and calibration examples are sampled without replacement from the training split, excluding demonstrations from the calibration pool. The test subset is fixed independently of the outer seed. SST-2, BoolQ, CommonsenseQA, and HellaSwag use their validation split for evaluation; AG News, TREC, DBpedia-14, and ARC-Challenge use their test split.

For MMLU, four subject-matched development examples serve as demonstrations. Calibration and evaluation examples come from the validation and test splits, respectively. Each official supercategory contributes 50 calibration and 50 test examples. The four supercategories are fitted independently, and their predictions are concatenated before computing one 200-example metric vector. The main study thus contains $3 \times 9 \times 10 = 270$ model–task–seed reporting cells and 360 fitted groups. Split identifiers and normalized semantic content are checked for overlap. Test labels are not used for fitting or selection.

Main inference uses an NVIDIA RTX 5090 with 32 GB memory, Python 3.12.13, PyTorch 2.12.0, CUDA 13.0, and Transformers 5.14.1. Coefficient fitting and post-processing use CPU float64 operations, except for the BayesPE optimizer, which uses CUDA float32 parameters.

E.2 Prompt Banks and Inference Budgets

In the main comparison, multi-prompt methods use the same eight-view bank: one reference prompt and seven perturbations comprising two template variants, two demonstration-order variants, two option permutations, and one formatting variant. Demonstration content is fixed within a unit. Option-permuted outputs are inverse-mapped before fitting or aggregation. Prompt contrasts and centering follow Sections 3.2 and 3.3.

The option-permutation controls use their separate matched bank. The calibration-budget analysis holds the eight-view bank fixed. Forward budgets of 1, 2, 4, 8, 16, 24, and 32 use one reference view plus (template, demonstration-order, option-permutation, format) counts of (0,0,0,0), (1,0,0,0), (1,1,1,0), (2,2,2,1), (4,4,4,3), (6,6,6,5), and (8,8,8,7), respectively. These nested subsets use fixed family-index prefixes without performance-based selection.

E.3 Baseline Implementations

Reference uses the uncalibrated reference-prompt distribution. Temperature scaling fits a positive temperature on the full calibration set without a prior. The Scale-only control instead uses CITP's normalization, regularization, and cross-validation protocol. Calibration-selected best prompt selects the bank member with the lowest full-calibration NLL. Mean probability ensemble averages the eight probability vectors. Weighted probability ensemble fits nonnegative, sum-to-one weights by full-calibration mixture NLL.

BayesPE uses the public implementation at revision `6a5e05a` on the same stored prompt probabilities, rather than its native prompt-generation procedure. We retain its probability smoothing and CUDA L-BFGS optimizer, with learning rate 10^{-5}, ten outer steps of up to 100 iterations, and the public code's regularization coefficient 0.1.

Dirichlet-ODIR takes reference-prompt log probabilities; its off-diagonal weight and bias penalties are selected from $\{10^{-5}, \ldots, 10^{3}\}$. LinC takes reference probabilities with an ℓ_2 grid $\{0, 10^{-6}, 10^{-4}, 10^{-2}, 1, 10^{2}, 10^{4}\}$. Simplex log pooling uses nonnegative, sum-to-one prompt weights, while signed affine log pooling permits negative weights but retains the sum constraint.

All supervised methods use the same calibration examples. Regularization selection uses deterministic three-fold calibration splits where applicable, with method-specific fixed fold seeds. LinC, the log pools, and CITP are refitted on the full calibration set after selection. Temperature scaling, the weighted ensemble, and BayesPE are fitted directly on that set. Dirichlet-ODIR instead averages the three fold-trained test predictions under its predefined protocol.

E.4 Model Selection and Posterior Computation

CITP RMS-normalizes the reference, bias, and query-centered prompt blocks. The two log pools instead normalize their sum-to-one weight-contrast block around uniform weights. All data-dependent centering and normalization are estimated on the training portion of each fold and recomputed on the full calibration set for the final fit. A common isotropic Gaussian precision in the corresponding normalized coordinates is selected from $\{10^{-6}, 10^{-5}, \ldots, 10^{6}\}$ by fold-size-weighted validation NLL. MAP rows use MAP validation predictions; CITP uses Laplace posterior-predictive validation predictions.

CITP MAP fitting uses L-BFGS-B with analytic gradients and positive global scale, $\gamma > -1$. The posterior predictive averages 200 antithetic Gaussian draws. The Hessian and boundary safeguards follow Appendix B. The marginalization diagnostic holds the posterior-selected precision and MAP fit fixed. The main CITP MAP and CITP rows instead use independently selected precisions.

Shuffled-Contrast Control. We reassign whole query-contrast matrices using five label-independent derangements, keeping reference logits and labels fixed. Shuffling occurs separately within each fold's training and validation partitions, and within final calibration and test partitions. Each shuffle repeats the matched MAP control's three-fold prior selection and calibration refit; reported cell metrics average the five runs.

E.5 Metrics and Statistical Analysis

For K classes and n examples, we compute

$$\mathrm{Brier} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}(p_{ik} - \mathbb{1}\{y_i = k\})^2, \qquad \mathrm{nBrier} = \frac{\mathrm{Brier}}{1 - 1/K}, \tag{1}$$

$$\mathrm{NLL} = -\frac{1}{n}\sum_{i=1}^{n}\log p_{i,y_i}, \qquad \mathrm{nNLL} = \frac{\mathrm{NLL}}{\log K}. \tag{2}$$

We also report accuracy, macro-F1 over all K classes, and 15-bin equal-width ECE. Metrics are computed per reporting cell, after concatenation for MMLU. Main aggregates first average ten seeds within each model–task pair and then macro-average the relevant pairs: 27 overall, 15 fixed-label, or 12 multiple-choice. Normalization and aggregation precede rounding. NLL-CVaR$_{95}$ averages

the largest 5% of query NLL losses within a cell; wrong-confidence averages maximum predicted probability over each method's incorrect predictions.

Main-table paired differences use identical test queries. The 95% intervals use 20,000 task-cluster bootstrap replicates: sample the relevant tasks with replacement, retain all three models, and resample the ten seed-level differences within each sampled model–task pair. Test queries remain fixed. Table 1 chooses its comparator separately by metric and holds it fixed during resampling; these intervals are conditional on that comparator choice. Budget and model-size curves instead use two-sided 95% Student t intervals over ten outer-seed values, with the task or model–task averaging specified in each caption.

For Figure 5, predictive disagreement is $H(\bar{p}) - S^{-1}\sum_s H(p^{(s)})$, with quintiles formed within each model–task–seed cell. Diagnostic statistics use the same seed and model–task averaging and task-cluster bootstrap described above; the intervals are pointwise.

Residual-Alignment Diagnostic. We use the saved Scale+Bias and matched CITP MAP fits from all 270 cells, without additional fitting or selection. Test queries are divided into two label-independent, equally sized halves within each fitted group. The deterministic split is fixed across outer seeds. On the first half A, we compute

$$\widehat{a} = \frac{1}{|A|}\sum_{i\in A}\widetilde{J}_i^\top(e_{y_i} - p_i^{\mathrm{SB}}), \qquad \widehat{G} = \frac{1}{|A|}\sum_{i\in A}\widetilde{J}_i^\top\widetilde{J}_i,$$

retaining the original calibration-set centering. The alignment score is $\widehat{a}^\top\widehat{G}^\dagger\widehat{a}$, retaining eigenvalues greater than 10^{-10} times the largest eigenvalue of $\widehat{G}$ in the pseudoinverse.

On the other half, we measure NLL(Scale+Bias)−NLL(CITP MAP). MMLU group statistics are weighted by query count. We average ten seeds within each model–task pair and compute correlation intervals using 2,000 task-cluster bootstrap replicates, retaining all three models. Test labels are used only for this post-hoc analysis; the fitted predictors and their selected priors remain unchanged.

E.6 NUTS Setup and Sampling Diagnostics

We use NumPyro's NUTS implementation (Hoffman & Gelman, 2014) on all 27 model–task pairs at outer seed 42, covering 36 fitted groups. It targets the same summed calibration log-likelihood and Gaussian prior as CITP under the positive-scale constraint, with the selected prior, feature normalization, and MAP fit held fixed. Sampling uses MAP-covariance whitening and a constrained-coordinate transformation with its log-Jacobian correction. We run four chains in CPU float64, with 1,000 warmup and 2,000 retained draws per chain, dense mass-matrix adaptation, target acceptance probability 0.9, and maximum tree depth 10.

Across all 36 fitted groups, the maximum rank-normalized split $\widehat{R}$ (including folded-rank checks) was 1.0030; the minimum bulk and tail effective sample sizes were 7342 and 4976. No divergences or maximum-tree-depth hits were observed.

The separate 8,192-draw Gaussian integration diagnostic retains all draws, including one that crosses the scale boundary; the main results use the 200-draw predictor.