BINARY AGREEMENT FOR EFFICIENT UNSUPERVISED CONFIDENCE CALIBRATION OF REASONING LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reasoning language models often lack the calibrated confidence needed for reliable deployment, particularly when correctness labels are unavailable and sampling budgets are limited. We introduce *Binary Agreement for Self-consistency Calibration* (BASC), which uses self-consistency as an unlabeled proxy for correctness without constructing costly self-consistency targets for each calibration query. Our key observation is that agreement between an anchor response and a single independent probe is a Bernoulli observation whose conditional mean is the anchor's latent self-consistency. Pooling these observations across unlabeled queries, BASC supports both recalibrating existing confidence scores and learning more discriminative scores before recalibration, using as few as two responses per calibration query and only one at deployment. The use of binary isotonic recalibration provides a distribution-free calibration guarantee with respect to latent self-consistency. Across 5 tasks and 9 reasoning models, BASC substantially improves calibration with respect to correctness.

1 INTRODUCTION

Reliable deployment of reasoning large language models (LLMs) requires confidence estimates that accurately reflect the probability that an answer is correct (Bommasani et al., 2022; Yang et al., 2026). This property, known as *confidence calibration* (Guo et al., 2017; Kadavath et al., 2022), is important for deciding when a model can be trusted to answer, when it should defer, or when additional intervention is needed. Modern LLMs nevertheless remain poorly calibrated (Leng et al., 2024; Damani et al., 2025), and existing confidence estimation approaches often rely either on correctness labels (Zadrozny & Elkan, 2001; Kumar et al., 2020; Band et al., 2024; Damani et al., 2025) or on repeated generations at inference time (Chen et al., 2024; Kuhn et al., 2023; Duan et al., 2024), both of which may be impractical in resource-constrained settings.

Self-consistency, the rate at which independently sampled responses to a prompt agree on a final answer, offers a useful unlabeled proxy for correctness (Wang et al., 2023; Lyu et al., 2025), but estimating it typically requires many generations per query. One approach is to amortize this computation by learning a predictor offline, allowing confidence to be produced from a single response at deployment. However, constructing accurate self-consistency targets for every calibration query still requires many generations, moving the sampling burden from deployment to offline calibration. These costs are especially consequential for on-device and edge systems, where memory, latency, and battery are constrained (Wang et al., 2025a); for black-box APIs, where every additional generation incurs monetary cost; and for agentic systems, where confidence may need to be evaluated repeatedly to determine when to act, defer, invoke tools, or escalate (Oh et al., 2026). Obtaining calibrated confidence therefore remains difficult in these important resource-constrained settings, where labels are scarce and repeated sampling is impractical.

In this work, we introduce ***Binary Agreement for Self-consistency Calibration*** **(BASC)**, an unsupervised framework that reduces sampling costs both offline and at deployment (see Figure 1). Our key observation is that accurately estimating self-consistency for every calibration query is unnecessary. Instead, we can learn to predict latent self-consistency from binary agreement between just two responses per query. BASC exploits this signal in two complementary ways. First, binary agreement can directly recalibrate an existing confidence score, such as token probabilities. Second, when stronger discrimination is desired or token probabilities are unavailable, these observations can train a score model before recalibration. In both cases, BASC applies binary isotonic regression

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.

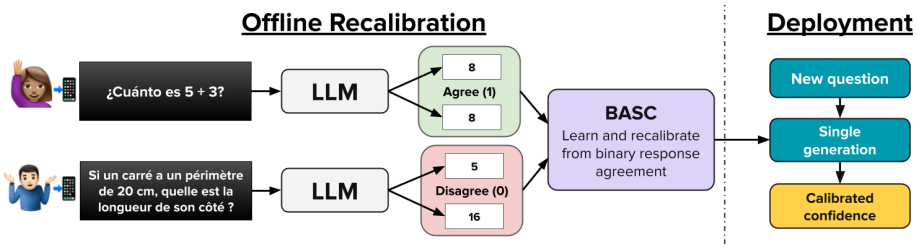

Figure 1: Binary Agreement for Self-consistency Calibration (BASC) learns a confidence calibrator offline by pooling binary agreement across unlabeled questions, enabling confidence-guided decisions from a single generation at deployment. It targets resource-constrained settings where correctness labels are scarce and latency, memory, or energy budgets make repeated test-time sampling impractical. Here, our motivating example is a multilingual on-device math tutor that must decide whether to answer directly or defer to a teacher. The figure illustrates $k = 2$ responses per question; larger k can be used for score learning.

to held-out agreement observations and requires only a single response at deployment, substantially reducing resource requirements. Building on Rossellini et al. (2026), this recalibration provides a finite-sample, distribution-free guarantee for calibration to latent self-consistency.

We evaluate BASC across 5 math and question-answering tasks using 9 reasoning models from the Qwen, DeepSeek, and Nemotron families. Our main contributions and findings are: **(1)** We introduce a binary-agreement framework that avoids costly per-query self-consistency estimation, supporting both label-free recalibration of existing scores and efficient learning of more informative scores. **(2)** We show that binary recalibration substantially reduces calibration error with respect to correctness, while agreement-based score learning improves probabilistic quality, discrimination, robustness, and selective prediction. **(3)** We justify the recalibration stage with a distribution-free guarantee for single-generation prediction of latent self-consistency, even though this target is never directly observed. Together, these results establish binary agreement as a practical, inexpensive signal for improving confidence estimates without labels or repeated test-time sampling.

Related Work Calibration has been extensively studied for classification (Naeini et al., 2015; Guo et al., 2017; Kumar et al., 2020; Minderer et al., 2021), but confidence estimation is more difficult for open-ended language generation, where outputs are multi-token and may admit many semantically equivalent forms (Kuhn et al., 2023; Yang et al., 2026). Common approaches include using token probabilities (Kadavath et al., 2022), verbalized confidence (Tian et al., 2023; Xiong et al., 2024; Yoon et al., 2025), and uncertainty estimated from repeated generations, including sample consistency (Si et al., 2023; Lyu et al., 2025) and semantic entropy (Kuhn et al., 2023). Because repeated sampling is expensive, prior work has explored amortizing multi-sample uncertainty signals into single-generation predictors (Kossen et al., 2024; Huang et al., 2025). These methods reduce inference-time sampling, but treat the uncertainty signal as a per-query target that must be estimated offline from many generations; moreover, Kossen et al. (2024) predicts entropy rather than calibrated confidence, while Huang et al. (2025) requires LLM tuning and eschews standard self-consistency in favor of a more expensive heuristic target.

Calibration is especially challenging for reasoning models, where post-training tends to induce overconfidence and poor abstention behavior (Leng et al., 2024; Kirichenko et al., 2025; Mei et al., 2025; Damani et al., 2025). Recent approaches address these issues through prompting or introspection (Kirichenko et al., 2025; Mei et al., 2025) or calibration-aware training (Band et al., 2024; Bani-Harouni et al., 2026; Damani et al., 2025). These methods generally require additional inference-time elicitation, supervised or reward-based training signals, or modification of the underlying model. Also related to our unsupervised setting are recent methods that derive confidence from unlabeled signals produced by a paired base model (Luo et al., 2025; Tan et al., 2026). However, we find that base-model-derived confidence is poorly calibrated in our setting (see Appendix C).

Finally, a separate line of work develops distribution-free methods that provide high-probability guarantees on LLM behavior, often by adapting conformal prediction or related risk-control ideas (Zollo et al., 2023; Quach et al., 2023; Campos et al., 2024; Dobriban, 2025; Ji et al., 2025). These methods share the broader goal of reliable behavior under uncertainty, but generally do not address answer-level confidence calibration.

2 METHOD

We study confidence estimation for reasoning LLMs when correctness labels are unavailable and both offline and inference-time sampling resources are limited. Let x denote an input prompt and $y = (r, a) \sim p_\theta(\cdot \mid x)$ a model response, where r is the reasoning trace and a the extracted final answer. Let $Z \in \{0, 1\}$ indicate whether a is correct. Given a single response y at deployment, our goal is a confidence estimate $\hat{c}(x, y) \in [0, 1]$ calibrated with respect to correctness: $\mathbb{E}[Z \mid C] = C$, where $C = \hat{c}(X, Y)$. Without correctness labels, we have only an unlabeled calibration set $\mathcal{D}_{\text{cal}} = \{x_i\}_{i=1}^n$ drawn i.i.d. from the target prompt distribution.

Self-consistency as an unlabeled target. Because Z is unobserved, we use *self-consistency* as an unlabeled proxy for answer confidence (Wang et al., 2023; Lyu et al., 2025). For an anchor response $y = (r, a)$ to prompt x, let $Y' = (R', A') \sim \tilde{p}_\theta(\cdot \mid x)$ be an independent response drawn from a fixed *probe* sampling distribution $\tilde{p}_\theta$. While the anchor and probe distributions may be the same, separating them allows low-entropy decoding at deployment while using more diverse probe samples to assess uncertainty.

We define the population self-consistency of a as

$$s^\star(x, a) = \mathbb{P}(A' \equiv a \mid x)\,, \tag{1}$$

where $\equiv$ denotes binary agreement under the task-specific answer-matching procedure (e.g., exact or numeric matching for math, and semantic matching via bidirectional entailment (Kuhn et al., 2023) for QA). Thus, $s^\star(x, a)$ is the probability that an independent response produces an answer equivalent to a. Self-consistency is not a correctness probability, but it measures how strongly the model favors its answer and empirically tracks correctness (Figure 5(a) previews this relationship across the 36 model/dataset combinations evaluated in Section 3).

A standard approach to estimating $s^\star(x, a)$ uses repeated sampling. Using k total responses, draw an *anchor* $Y^{(0)} = (R^{(0)}, A^{(0)}) \sim p_\theta(\cdot \mid x)$ and $k - 1$ probes $Y^{(1)}, \ldots, Y^{(k-1)}$ drawn i.i.d. from $\tilde{p}_\theta(\cdot \mid x)$. For each probe, define $B^{(j)} = \mathbf{1}[A^{(j)} \equiv A^{(0)}]$. The Monte Carlo estimate of the anchor's self-consistency is $\widehat{s}_k = (k-1)^{-1} \sum_{j=1}^{k-1} B^{(j)}$. Obtaining a low-variance $\widehat{s}_k$ for every calibration example can require many expensive generations. Our key observation is that this per-query precision is unnecessary. Let $S = s^\star(X, A^{(0)})$. Conditioned on the prompt and anchor response,

$$B^{(j)} \mid X, Y^{(0)} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(S), \qquad \mathbb{E}\Big[B^{(j)} \mid X, Y^{(0)}\Big] = S. \tag{2}$$

Each probe therefore provides an unbiased binary observation of latent self-consistency. Writing B for a single probe's agreement and $U = u(X, Y^{(0)})$ for any scalar score computed from the anchor by a fixed function u, the tower property gives

$$\mathbb{E}[B \mid U] = \mathbb{E}[S \mid U]. \tag{3}$$

Thus, pooling binary agreement observations across queries provides supervision for conditional latent self-consistency without accurately estimating S for any individual query.

Binary Agreement for Self-consistency Calibration (BASC)

Building on this observation, we propose ***Binary Agreement for Self-consistency Calibration*** **(BASC)**, a 2-stage algorithm using binary agreement to learn and recalibrate LLM confidence scores. Stage 1 supplies a scalar score $U = u(X, Y^{(0)})$, either pre-existing or learned from agreement observations; Stage 2 recalibrates it on held-out agreement observations. For a learned score, we split the calibration data into disjoint subsets $\mathcal{D}_{\text{cal}} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{iso}}$ for score learning and recalibration, respectively. For a pre-existing score, no training split is needed and $\mathcal{D}_{\text{cal}} = \mathcal{D}_{\text{iso}}$.

Stage 1: Scalar score. The simplest instantiation uses a pre-existing score $u(x, y)$, such as length-normalized token probability, and requires no score training. Alternatively, let $H = h(X, Y^{(0)})$ represent an anchor prompt-response pair (e.g., the reasoning model's last-token hidden state), and let $f_\phi(H) \in [0, 1]$ predict repeat agreement. For each $x_i \in \mathcal{D}_{\text{train}}$, draw one anchor $Y_i^{(0)} \sim p_\theta(\cdot \mid x_i)$ and $k - 1$ probes from $\tilde{p}_\theta(\cdot \mid x_i)$. Set $H_i = h(x_i, Y_i^{(0)})$ and $B_{ij} = \mathbf{1}[A_i^{(j)} \equiv A_i^{(0)}]$, $j = 1, \ldots, k-1$. For $k = 2$, each anchor contributes one Bernoulli realization, not the latent probability itself; larger k provides repeated conditionally i.i.d. observations for the same anchor. We

train f_ϕ using binary cross-entropy over these outcomes. This Bernoulli log-loss has Bayes-optimal prediction $\mathbb{P}(B = 1 \mid H) = \mathbb{E}[S \mid H]$, by the argument of Equation 3, making it a natural objective for predicting conditional latent self-consistency. The learned score is $u(x, y) = f_\phi(h(x, y))$.

Stage 2: Binary isotonic recalibration. We freeze u and recalibrate it using independent binary agreement observations. For each $x_i \in \mathcal{D}_{\text{iso}}$, draw one anchor $Y_i^{(0)} \sim p_\theta(\cdot \mid x_i)$ and one independent probe $Y_i^{(1)} \sim \tilde{p}_\theta(\cdot \mid x_i)$. Compute $U_i = u(x_i, Y_i^{(0)})$ and $B_i = \mathbf{1}[A_i^{(1)} \equiv A_i^{(0)}]$. We fit a one-dimensional isotonic regressor $\hat{g}$ to $\{(U_i, B_i) : x_i \in \mathcal{D}_{\text{iso}}\}$, assigning each unseen score the fitted value of its nearest observed calibration score (Rossellini et al., 2026) (see Appendix A for implementation details). By Equation 3, this fits a monotone approximation to $\mathbb{E}[S \mid U]$ at the population level (Figure 5(b–c)). The final confidence estimator is

$$\hat{c}(x, y) = \hat{g}(u(x, y)) \,. \tag{4}$$

At deployment, only a single response y is generated; no probe is needed.

Distribution-free calibration to latent self-consistency. Although latent self-consistency S is never observed directly, BASC inherits a calibration guarantee from binary isotonic regression. Let $n_{\text{iso}} = |\mathcal{D}_{\text{iso}}|$, and assume that the Stage 2 and test examples are i.i.d. and independent of any data used to construct u. For a fresh test example, write $U = u(X, Y^{(0)})$ and $C = \hat{g}(U)$. For readability, for the rest of this section all expectations are over fresh test data with the fitted functions u and $\hat{g}$ held fixed; Appendix A.1 gives the explicit conditional form. The binary-isotonic result of Rossellini et al. (2026) then gives

$$\mathbb{E}[|\mathbb{E}[B \mid C] - C|] = O_p\Big(n_{\text{iso}}^{-1/6}\sqrt{\log n_{\text{iso}}}\Big) \,, \tag{5}$$

where O_p is over the Stage 2 calibration sample. The left-hand side is the population calibration error of C with respect to binary agreement B. The tower property transfers this guarantee exactly to latent self-consistency: because C depends only on the anchor and fitted functions,

$$\mathbb{E}[B \mid C] = \mathbb{E}[S \mid C].$$

Thus, the population calibration error with respect to S is identical to that with respect to B, yielding

$$\mathbb{E}[|\mathbb{E}[S \mid C] - C|] = O_p\Big(n_{\text{iso}}^{-1/6}\sqrt{\log n_{\text{iso}}}\Big) \,. \tag{6}$$

The guarantee depends only on Stage 2 and applies to any frozen scalar score, regardless of its predictive quality, without requiring $\mathbb{E}[S \mid U]$ to be monotone. Appendix A.1 provides the full argument. To our knowledge, this is the first distribution-free calibration guarantee for single-generation prediction of LLM self-consistency.

Relationship to correctness calibration. The self-consistency guarantee also clarifies when BASC yields well-calibrated confidence with respect to correctness. Under the same convention of holding the fitted functions fixed, for correctness Z the triangle inequality gives

$$|\mathbb{E}[Z \mid C] - C| \le |\mathbb{E}[S \mid C] - C| + |\mathbb{E}[Z - S \mid C]| \,. \tag{7}$$

The left-hand side is the absolute correctness calibration gap at reported confidence C. Averaged over test examples, Equation 6 controls the first term, i.e., the calibration error with respect to self-consistency, while the second measures how closely self-consistency tracks correctness conditional on reported confidence. Thus, when this second term is small, BASC can achieve small correctness calibration error using only unlabeled agreement observations.

3 Experiments

We evaluate our proposed approach along four dimensions. **First,** we test whether self-consistency is in fact a useful unlabeled proxy for answer correctness in reasoning LLMs. **Second,** we ask whether binary agreement observations suffice to calibrate a single-generation score. **Third,** we examine whether BASC remains reliable under distribution shift between unsupervised training and deployment. **Finally,** we assess whether the confidence estimates are useful for downstream decision-making via selective prediction. Below we detail our experimental setup; see Appendix B for additional details. All parameters and prompts used to produce our data and results are contained in this section and Appendix B. Our study relies solely on open source models and datasets, and our code will be available to ensure reproducibility.

Reasoning Language Models Our models include: 5 Qwen3 models (Yang et al., 2025): 0.6B, 1.7B, 4B-Thinking-2507, 8B, 14B; 2 Nemotron models (NVIDIA et al., 2025; Moshkov et al., 2025; Wang et al., 2026): OpenReasoning-Nemotron-7B, Nemotron-Cascade-8B-Thinking; 2 DeepSeek models (Guo et al., 2025): DeepSeek-R1-Distill-Llama-8B, DeepSeek-R1-Distill-Qwen-7B.

Datasets Our tasks are split between math problem solving and question answering (QA) tasks, following previous work in LLM calibration (e.g., Leng et al. (2024); Luo et al. (2025); Damani et al. (2025)). For math, we use **GSM8K** (Cobbe et al., 2021) and **PolyMath** (Wang et al., 2025b). PolyMath is a multilingual dataset, and we use data from 8 languages: EN, ES, FR, PT, IT, RU, AR, ID. We perform science QA with **SciQ** (Welbl et al., 2017), and open-domain QA with **TriviaQA** (Joshi et al., 2017) and **WebQuestions (WebQ)** (Berant et al., 2013). All models are evaluated on the math datasets. For free-form QA, we include only the Qwen3 models and the Nemotron-Cascade-8B-Thinking model, as the other 3 models failed to reliably follow the formatting required for automatic extraction for these tasks. In total, we run 9 models on 2 math datasets and 6 models on 3 QA datasets for a total of 36 combinations (see Table 6 for model/dataset accuracy). We use 1000 examples per dataset, and average results over 10 random 40/60 calibration/test splits.

Our Method We evaluate BASC recalibration (Stage 2) applied to different Stage 1 scores. **BASC-L** includes score learning, where we split the calibration set 50/50, and fit an ℓ_2-regularized binomial-logistic regressor from the model's own last-token hidden state to the $k-1$ agreement outcomes per query, with $k \in \{2, 5\}$. Using the model's own embeddings integrates seamlessly with existing generation pipelines. A black box variant (**BASC-L-bb**) embeds the prompt and response with an external model, Gemma-3-4B-it; we use $k = 5$ for this variant. **BASC-TP** uses the pre-existing token probabilities, so all calibration data goes to Stage 2. Stage 2 recalibration always uses 2 decodes per query.

For each example we sample one anchor response at temperature 0.6 (TopP=0.95, TopK=20), and a pool of 100 probe responses at temperature 0.7 (100 probes allow us to run varied random trials, as well as calculate an accurate empirical self-consistency for each example). Stage 1 subsamples $k-1$ probes per anchor and Stage 2 uses one. Agreement uses exact matching for math and, for QA, the bidirectional-entailment rule of Kuhn et al. (2023) with a DeBERTa-v3 NLI cross-encoder. Sample code for our method can be seen in Code Box 1. Given our unsupervised setting, we perform no model- or dataset-specific tuning; all method and sampling hyperparameters are fixed across the 36 model/dataset combinations. We ablate probe sampling temperature in $[0.5, 0.7, 0.9]$ and QA answer matching (compared to string matching) in Appendix D.2; we find neither choice to make a significant difference to our results.

Evaluation We measure expected calibration error using ECE_1 with 12 equal-mass bins (ECE_2 is included in Appendix) (Guo et al., 2017). We also measure Brier score (BS) (Brier, 1950), a proper scoring rule for calibration that encourages more discriminative predictions, and AUROC. While AUROC does not measure calibration, it captures another important quality for useful confidence estimates, that they effectively separate correct and incorrect answers (Kadavath et al., 2022).

Baselines The most common baseline for producing LLM confidence estimates is the length-normalized token probability for the full model response (e.g., the geometric mean of the top token probabilities) (Kadavath et al., 2022; Luo et al., 2025). For reasoning models, to avoid scores being distorted by long thinking chains with many unimportant tokens, one might use the length-normalized token probability of only the tokens associated with the final answer (Damani et al., 2025). Given that both of these baselines are viable without labels or multiple samples, we adopt them (**Token Probs. (TP)** and **Answer Probs. (ATP)**) for comparison in our setting. Another popular approach to LLM calibration involves having the model directly verbalize its own confidence in its output (**Verbal Conf.** (VC)) (Tian et al., 2023). While this does require another full response (including reasoning), which may be prohibitive in resource-constrained settings, we include it because of its prevalence in the LLM calibration literature. Further, to ensure a strong verbalized confidence baseline, we search for our prompt among 5 prompts written by GPT-5.2 and 5 written by Gemini-3, and select the prompt with the best results across a selection of models and datasets (see Appendix B.3.2 for more details and the final prompt). Finally, we also include results for stronger baselines that violate the resource constraints of no labels and a single test-time generation. Namely, we include (1) a **Supervised** baseline where we perform our 2 stage algorithm with

	Average Across 36 Combinations			ECE_1 (↓) by Model and Task Family			
	ECE_1 (↓)	Brier (↓)	AUR (↑)	Qwen	non-Qwen	Math	QA
TP	0.197	0.222	0.655	0.198	0.195	0.102	0.293
ATP	0.296	0.297	0.622	0.314	0.254	0.147	0.445
VC	0.268	0.281	0.635	0.285	0.230	0.135	0.401
BASC-L (k=2)	**0.083**	**0.162**	0.685	**0.089**	**0.069**	**0.046**	**0.119**
BASC-L (k=5)	**0.084**	**0.159**	**0.708**	**0.091**	**0.068**	**0.045**	**0.123**
BASC-TP	**0.069**	**0.167**	0.647	**0.072**	**0.063**	**0.045**	**0.094**
BASC-L-bb (k=5)	**0.083**	**0.159**	**0.705**	**0.090**	**0.068**	**0.045**	**0.121**
TT-SC	0.090	0.125	0.836	0.099	0.071	0.037	0.144
Supervised	0.051	0.147	0.732	0.053	0.046	0.038	0.064

Table 1: Calibration results averaged over all 36 model/dataset combinations (left) and ECE_1 by model family and task type (right). Rows include the unsupervised baselines (token probs. (TP), answer token probs. (ATP), and verbalized confidence (VC)), our methods (BASC-L with k samples per training query, and BASC-TP), a black-box variant with embeddings from an external model (BASC-L-bb), and two strong baselines that break our constraints of no labels and one generation: test-time self-consistency with 100 samples (TT-SC) and supervised recalibration. **Bold results** are significantly better than each of TP, ATP, and VC (paired two-sided Wilcoxon test over the combinations in each column, $p < 0.05$). BASC-L improves ECE and Brier over all three unsupervised baselines (and AUROC with $k = 5$), with ECE_1 gains in every model and task family, including in the black-box setting; BASC-TP improves calibration but adds no discrimination to underlying token probabilities.

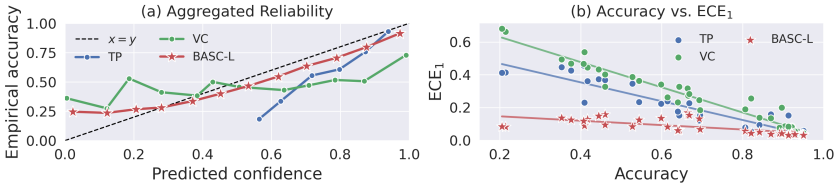

Figure 2: **(a)** Aggregated reliability diagram; perfect calibration lies on the $x = y$ line. **(b)** ECE_1 vs. accuracy across the 36 model/dataset combinations for our method and baselines.

ground truth correctness labels and (2) a multiple-sampling baseline measuring empirical **Test-Time Self-Consistency (TT-SC)** with 100 samples.

3.1 Main Results

Our first experiment addresses two questions: (1) does self-consistency provide a useful unlabeled signal of correctness in reasoning LLMs, and (2) can BASC learn this signal from binary agreement observations alone? We focus on the standard unsupervised setting, in which the unlabeled calibration data come from the same distribution as the test data (Ganin et al., 2016; Luo et al., 2025). Table 1 reports results averaged over all 36 combinations and ECE_1 by model and task family, and Appendix Figures 6, 7, and 8 show results by model and dataset (including standard error measurements).

With respect to question (1), in our results TT-SC provides a strong unlabeled confidence signal, echoing previous findings (Wang et al., 2023; Lyu et al., 2025). It matches the ECE of the supervised baseline on math and achieves a better overall Brier score and AUROC, indicating that agreement across sampled responses carries substantial information about correctness. This supports the central premise of our method: self-consistency is a high-quality target for learning answer-level confidence in reasoning LLMs. Further quantitative and qualitative measurements of the relationship between self-consistency and correctness for each model and dataset can be found in Appendix D.

For question (2), BASC-L significantly improves ECE and Brier score over token probabilities, answer probabilities, and verbalized confidence, with ECE_1 gains in every model and task family, and with $k = 5$ also significantly improves AUROC. With only two samples per training query, BASC-L ($k = 2$) approaches the ECE of test-time self-consistency on math and improves on it

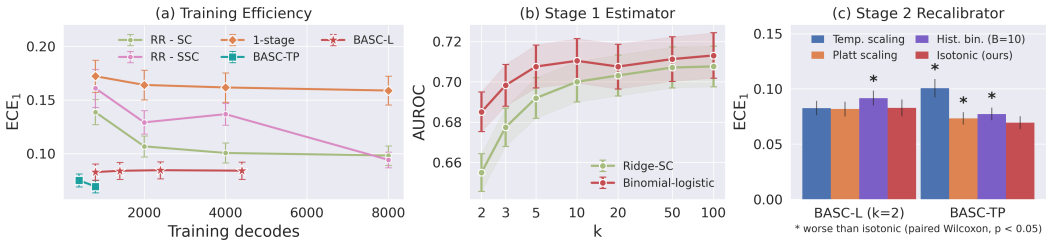

Figure 3: **(a)** ECE_1 vs. number of full-response training decodes for BASC-L, two single-stage regressors on continuous self-consistency targets (RR-SC, RR-SSC), and BASC-L's Stage-1 model without Stage-2 recalibration (1-stage), at $k \in \{2, 5, 10, 20\}$, and for BASC-TP fit on 200 or 400 queries (2 decodes each). **(b)** AUROC vs. k for BASC's binomial-logistic Stage 1 and ridge regression on empirical self-consistency (RR-SC), each followed by BASC's Stage-2 isotonic recalibration. **(c)** ECE_1 of BASC-L ($k = 2$) and BASC-TP with isotonic Stage 2 (ours) vs. Platt scalling, temperature scaling, and histogram binning (with 10 bins). Both stages of BASC matter: with only 800 training decodes, BASC-L has lower ECE_1 than the continuous-target baselines with $10\times$ as many decodes, and about half the ECE_1 of its Stage-1 model alone, while the binomial-logistic Stage 1 is more discriminative than the ridge-based RR-SC estimator at small k. As the Stage 2 recalibrator, isotonic regression is competitive with alternatives for BASC-L, and significantly outperforms for BASC-TP.

for QA, while requiring only a single generation at test time (although it clearly trails on Brier score and AUROC). BASC-TP, which applies Stage 2 directly to token probabilities, achieves even lower ECE than BASC-L. However, it inherits the weak discrimination of token probabilities, with lower AUROC and higher Brier score than BASC-L. This highlights that Stage 2 can substantially improve calibration, while the learned Stage 1 makes the score more informative. Finally, BASC only requires sampled outputs and can therefore operate without access to model internals. To test this setting, BASC-L-bb uses embeddings of the prompt and response from an external model, Gemma-3-4B-it, and otherwise runs the same method. As Table 1 shows, this black-box variant performs nearly identically to BASC-L on both task types, so our method extends naturally to API-only and proprietary models.

Reliability Diagram The reliability diagram in Figure 2(a), aggregated over all 36 combinations, offers a similar view. Reliability diagrams visualize calibration by plotting confidence vs. accuracy for each bin of data; a perfectly calibrated model produces a calibration curve that lies on the line $x = y$, where confidence matches accuracy. BASC closely tracks the diagonal, while token probabilities are consistently overconfident and verbalized confidence is only loosely related to correctness.

Calibration Error and Accuracy Previous work (Xiong et al., 2024) has shown that existing methods (e.g., verbalized confidence) can appear calibrated under ECE because they are usually highly confident and the underlying models are often highly accurate on the tasks under consideration; such calibration breaks down on harder tasks where the model is less accurate. To test for this effect, Figure 2(b) plots accuracy against ECE_1 across model/dataset combinations for our method and two unsupervised baselines. The baselines' calibration error increases substantially as accuracy decreases, whereas our method remains comparatively stable across accuracy levels. This suggests that its low ECE_1 is not merely due to matching average task accuracy with high confidence.

Design Decisions Figure 3(a) plots average ECE_1 against the total number of full-response decodes used for calibration. We compare BASC-L and BASC-TP to two single-stage ridge regressors on BASC-L's embedding features, fit with squared loss to continuous targets: the empirical self-consistency rate (**RR-SC**) and the confidence-weighted self-consistency of Huang et al. (2025) (**RR-SSC**). For $k \in \{2, 5, 10, 20\}$, BASC-L uses $200k + 400$ decodes (k per Stage-1 query, 2 per Stage-2 query), while the regressors use $400k$; BASC-TP, which requires no Stage-1 training, uses 2 decodes on each of 200 or 400 queries. With just 800 decodes, BASC-L is better calibrated than either ridge baseline at $10\times$ the budget, while BASC-TP performs even better with fewer decodes, demonstrating that BASC can exploit self-consistency without costly per-query estimates.

Panels 3(a)–(c) further examine the design choices underlying BASC. First and most importantly, a separate held-out recalibration stage is essential: the same binomial-logistic score model fit on

	Language		Math		QA		Average	
	ECE_1	Brier	ECE_1	Brier	ECE_1	Brier	ECE_1	Brier
TP	0.135	0.169	0.101	0.127	0.290	0.315	0.175	0.204
VC	0.175	0.195	0.135	0.150	0.398	0.409	0.236	0.251
RR-SC (k=10)	0.113	0.150	0.125	0.144	0.194	0.264	0.144	0.186
BASC-TP	0.081	0.153	0.140	0.153	0.180	0.257	0.134	0.188
BASC-L (k=2)	**0.076**	**0.139**	**0.095**	**0.126**	**0.157**	**0.244**	**0.110**	**0.170**
TT-SC	0.045	0.079	0.035	0.066	0.141	0.184	0.074	0.109

Table 2: Calibration results under distribution shift in language (using PolyMath) as well as across domains in math and QA; scores are averaged across models and shifts, lower is better for all metrics. Our method offers reliable performance even when calibration data does not match test data.

all 400 calibration queries without Stage 2 (**1-stage**, Figure 3(a)) has roughly twice the error of BASC-L at every k, across all 36 model/dataset combinations. RR-SC and RR-SSC likewise leave this calibration error unaddressed by fitting self-consistency in a single stage, even if the per-query self-consistency estimation may be more precise. Second, the binomial-logistic Stage 1 is more discriminative than RR-SC: with Stage 2 held fixed, it achieves significantly higher AUROC when few samples per query are available ($k \leq 10$; Figure 3(b)). Thus, BASC's advantage is not just post-hoc recalibration: its binomial-logistic Stage 1 also learns an informative ranking under limited sampling. Third, isotonic regression, which carries our distribution-free guarantee, is also empirically strong. As the Stage 2 recalibrator, isotonic regression performs similarly to the scaling variants and better than histogram binning for BASC-L, and significantly outperforms all three for BASC-TP (Figure 3(c)). Together, these results show that BASC's design cleanly separates and addresses the two key challenges of confidence estimation, learning an informative score and calibrating it reliably, both without correctness labels.

Ablations In Appendix D.2 we study the effects of two implementation choices: our probe sampling temperature of 0.7, and the use of semantic agreement instead of string matching for the QA tasks (we also discuss the computational costs of semantic matching). We find that these choices have no significant effects on our results (though a slightly higher sampling temperature seems to help, we do not report that as our main result given the unsupervised setting).

3.2 Distribution Shift

Next, we study whether our method is reliable when faced with shifts in the distribution between calibration and test data (noting that this shift nullifies the distribution-free guarantee). To do so, we evaluate two types of distribution shifts. First, using the PolyMath dataset, we study distribution shifts across groupings of the 8 languages in our data sample. We create 4 different splits based on various language groupings (for example, romance vs. non-romance), and use each split as both training and test data, for a total of 8 experiments (see Appendix D.3 for more details on the language splits). Second, we study distribution shifts across task domains. For math, we study transfer between PolyMath and GSM8K (2 combinations), and for QA we study transfer among SciQ, TriviaQA, and WebQ (6 combinations). Results for all models and shifts are summarized in Table 2, and presented in detail in Appendix D.3; average results with confidence intervals appear in Figure 13.

BASC-L remains robust under all three shifts and is the strongest single-generation method in every setting. Under language shift, it reduces ECE_1/Brier to 0.08/0.14, from 0.14/0.17 for token probabilities; it again performs best under math transfer, and yields the clearest advantage in the challenging QA setting. It also outperforms ridge regression on empirical self-consistency (RR-SC ($k = 10$)) on average within each shift family while using $5\times$ fewer training decodes. Despite strong in-distribution ECE, BASC-TP is worse than BASC-L on every shift and is even worse than raw token probabilities under math transfer, suggesting that the relationship between token probability and agreement changes across datasets. According to these results, a learned Stage 1 makes BASC meaningfully more robust to shift.

3.3 Selective Prediction

Selective prediction uses confidence estimates to decide which examples to answer and which to abstain on, with the goal of improving accuracy on the selected examples by abstaining on the most

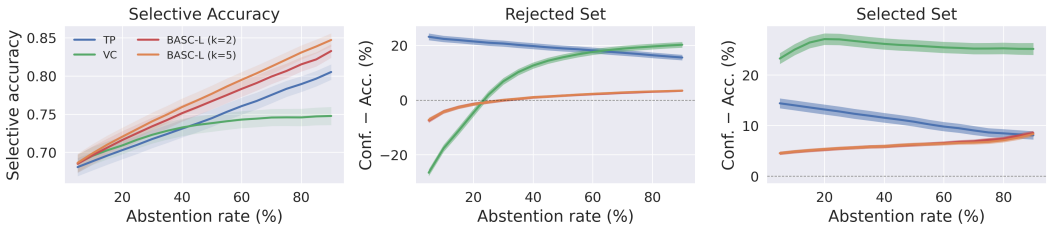

Figure 4: On the left, results for selective prediction using our method and baselines. The X-axis reflects the percent of examples on which the system abstains, and the y-axis reflects the corresponding accuracy for retained/selected examples. In the middle and right plots, we examine the difference between the average confidence and accuracy of the abstained and selected examples for each method at each level of abstention. Taken together, these plots highlight that our method enables selection that brings gains in accuracy while also allowing the decision-maker to understand the likely accuracy on both rejected and accepted predictions.

uncertain cases (Geifman & El-Yaniv, 2017; 2019; Yadkori et al., 2024). Although this is not strictly dependent on calibration (e.g., selective accuracy is invariant to strictly increasing transformations of the confidence scores), the ability to order examples from least to most likely to be correct is another desirable quality in a confidence estimator.

To study whether our method is useful in this way, we use the same experimental setup as Section 3.1 (40/60 data split), using the unlabeled calibration data to train our confidence estimation model. Using our method and baselines, we rank test examples by confidence and evaluate abstention rates from 5% to 90%, measuring the accuracy on the non-abstained (selected) set. In this experiment we focus on BASC-L, as BASC-TP does not learn a new ranking. Selective accuracy is averaged over 360 trials (36 combinations and 10 seeds), and we also analyze the rejected and selected sets by comparing their average confidence and realized accuracy; this lets us test not only whether a method selects good examples, but also whether it provides a calibrated picture of the quality of both the rejected and accepted examples.

Figure 4 (including standard error intervals) shows that our method yields the strongest overall selective prediction performance among the methods available in our setting. In Figure 4 (left), abstaining according to our confidence estimates produces substantially larger gains in retained-set accuracy across the full range of abstention rates, whereas the baselines improve more slowly, and VC plateaus around 50%. Moreover, the middle and right plots show that, for both the rejected and selected sets, our method's average confidence tracks realized accuracy much more closely than the alternatives, enabling a decision-maker to understand the likely accuracy of both rejected and accepted predictions. Appendix Figures 17 and 18 show that these trends hold across model families and datasets.

4 Conclusion

We find that calibrated confidence for reasoning LLMs need not require either labeled calibration data (and possibly compute-intensive RL training) or expensive repeated sampling. Importantly, this approach is not limited to open-weight models: our black-box results suggest that the same strategy can remain effective even when the underlying reasoning model is only accessible through a proprietary interface. Our results also highlight that verbalized confidence with reasoning LLMs remains a weak signal for calibration, at least without model training or extensive prompt optimization.

Limitations BASC is only guaranteed to offer calibration with respect to latent self-consistency, and not actual correctness. While our experiments show that self-consistency (and BASC-predicted confidence) do track correctness closely across a wide range of models and datasets, our method can only be as good as the self-consistency signal it learns: if agreement among samples is weakly related to correctness, or models are consistently wrong with high internal agreement, the learned predictor may inherit those failures (though one might argue that such a model should not be deployed anyway). Finally, our distribution-free bound is not tight at the calibration set sizes used in our experiments. Nevertheless, it establishes that population calibration error to latent self-consistency goes to zero as the number of calibration queries grows, even with only two responses per query.

AI USE STATEMENT

In this work, we have not used AI agents for autonomous research, but use generative AI tools regularly in an integrated research workflow. We used generative AI tools to draft (mainly in the Appendix) and polish some writing, to find related work, for research ideation, for coding implementation, and for proving mathematical claims. The generation, cleaning, or reformatting of datasets are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All models and datasets used in our experiments are publicly available for use. We have submitted our code as supplementary material, and will release it upon publication of this work.

REFERENCES

Neil Band, Xuechen Li, Tengyu Ma, and Tatsunori Hashimoto. Linguistic calibration of language models, 2024.

David Bani-Harouni, Chantal Pellegrini, Paul Stangel, Ege Özsoy, Kamilia Zaripova, Nassir Navab, and Matthias Keicher. Rewarding doubt: A reinforcement learning approach to calibrated confidence expression of large language models, 2026. URL https://arxiv.org/abs/2503.02623.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1533–1544, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D13-1160.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022. URL https://arxiv.org/abs/2108.07258.

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

Margarida M. Campos, António Farinhas, Chrysoula Zerva, Mário A. T. Figueiredo, and André F. T. Martins. Conformal prediction for natural language processing: A survey, 2024. URL https://arxiv.org/abs/2405.01976.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms' internal states retain the power of hallucination detection, 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. Beyond binary rewards: Training lms to reason about their uncertainty. *arXiv preprint arXiv:2507.16806*, 2025.

Edgar Dobriban. Statistical methods in generative ai, 2025. URL https://arxiv.org/abs/2509.07054.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models, 2024.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks, 2016. URL https://arxiv.org/abs/1505.07818.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks, 2017.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option, 2019.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks, 2017.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645 (8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Chengsong Huang, Langlin Huang, Jixuan Leng, Jiacheng Liu, and Jiaxin Huang. Efficient test-time scaling via self-calibration. *arXiv preprint arXiv:2503.00031*, 2025.

Wenlong Ji, Weizhe Yuan, Emily Getzen, Kyunghyun Cho, Michael I. Jordan, Song Mei, Jason E Weston, Weijie J. Su, Jing Xu, and Linjun Zhang. An overview of large language models for statisticians, 2025. URL https://arxiv.org/abs/2502.17814.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension, 2017. URL https://arxiv.org/abs/1705.03551.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J Bell. Abstentionbench: Reasoning llms fail on unanswerable questions. *arXiv preprint arXiv:2506.09038*, 2025.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in llms, 2024. URL https://arxiv.org/abs/2406.15927.

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

Ananya Kumar, Percy Liang, and Tengyu Ma. Verified uncertainty calibration, 2020. URL https://arxiv.org/abs/1909.10155.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. Taming overconfidence in llms: Reward calibration in rlhf. *arXiv preprint arXiv:2410.09724*, 2024.

Beier Luo, Shuoyuan Wang, Sharon Li, and Hongxin Wei. Your pre-trained llm is secretly an unsupervised confidence calibrator, 2025. URL https://arxiv.org/abs/2505.16690.

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. Calibrating large language models with sample consistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction, 2021.

Zhiting Mei, Christina Zhang, Tenny Yin, Justin Lidard, Ola Shorinwa, and Anirudha Majumdar. Reasoning about uncertainty: Do reasoning models know when they don't know?, 2025. URL https://arxiv.org/abs/2506.18183.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks, 2021. URL https://arxiv.org/abs/2106.07998.

Ivan Moshkov, Darragh Hanley, Ivan Sorokin, Shubham Toshniwal, Christof Henkel, Benedikt Schifferer, Wei Du, and Igor Gitman. Aimo-2 winning solution: Building state-of-the-art mathematical reasoning models with openmathreasoning dataset, 2025. URL https://arxiv.org/abs/2504.16891.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

NVIDIA, :, Aaron Blakeman, Aaron Grattafiori, Aarti Basant, Abhibha Gupta, Abhinav Khattar, Adi Renduchintala, Aditya Vavre, Akanksha Shukla, Akhiad Bercovich, Aleksander Ficek, Aleksandr Shaposhnikov, Alex Kondratenko, Alexander Bukharin, Alexandre Milesi, Ali Taghibakhshi, Alisa Liu, Amelia Barton, Ameya Sunil Mahabaleshwarkar, Amir Klein, Amit Zuker, Amnon Geifman, Amy Shen, Anahita Bhiwandiwalla, Andrew Tao, Anjulie Agrusa, Ankur Verma, Ann Guan, Anubhav Mandarwal, Arham Mehta, Ashwath Aithal, Ashwin Poojary, Asif Ahamed, Asit Mishra, Asma Kuriparambil Thekkumpate, Ayush Dattagupta, Banghua Zhu, Bardiya Sadeghi, Barnaby Simkin, Ben Lanir, Benedikt Schifferer, Besmira Nushi, Bilal Kartal, Bita Darvish Rouhani, Boris Ginsburg, Brandon Norick, Brandon Soubasis, Branislav Kisacanin, Brian Yu, Bryan Catanzaro, Carlo del Mundo, Chantal Hwang, Charles Wang, Cheng-Ping Hsieh, Chenghao Zhang, Chenhan Yu, Chetan Mungekar, Chintan Patel, Chris Alexiuk, Christopher Parisien, Collin Neale, Cyril Meurillon, Damon Mosk-Aoyama, Dan Su, Dane Corneil, Daniel Afrimi, Daniel Lo, Daniel Rohrer, Daniel Serebrenik, Daria Gitman, Daria Levy, Darko Stosic, David Mosallanezhad, Deepak Narayanan, Dhruv Nathawani, Dima Rekesh, Dina Yared, Divyanshu Kakwani, Dong Ahn, Duncan Riach, Dusan Stosic, Edgar Minasyan, Edward Lin, Eileen Long, Eileen Peters Long, Elad Segal, Elena Lantz, Ellie Evans, Elliott Ning, Eric Chung, Eric Harper, Eric Tramel, Erick Galinkin, Erik Pounds, Evan Briones, Evelina Bakhturina, Evgeny Tsykunov, Faisal Ladhak, Fay Wang, Fei Jia, Felipe Soares, Feng Chen, Ferenc Galko, Frank Sun, Frankie Siino, Gal Hubara Agam, Ganesh Ajjanagadde, Gantavya Bhatt, Gargi Prasad, George Armstrong, Gerald Shen, Gorkem Batmaz, Grigor Nalbandyan, Haifeng Qian, Harsh Sharma, Hayley Ross, Helen Ngo, Herbert Hum, Herman Sahota, Hexin Wang, Himanshu Soni, Hiren Upadhyay, Huizi Mao, Huy C Nguyen, Huy Q Nguyen, Iain Cunningham, Ido Galil, Ido Shahaf, Igor Gitman, Ilya Loshchilov, Itamar Schen, Itay Levy, Ivan Moshkov, Izik Golan, Izzy Putterman, Jan Kautz, Jane Polak Scowcroft, Jared Casper, Jatin Mitra, Jeffrey Glick, Jenny Chen, Jesse Oliver, Jian Zhang, Jiaqi Zeng, Jie Lou, Jimmy Zhang, Jinhang Choi, Jining Huang, Joey Conway, Joey Guman, John Kamalu, Johnny Greco, Jonathan Cohen, Joseph Jennings, Joyjit Daw, Julien Veron Vialard, Junkeun Yi, Jupinder Parmar, Kai Xu, Kan Zhu, Kari Briski, Katherine Cheung, Katherine Luna, Keith Wyss, Keshav Santhanam, Kevin Shih, Kezhi Kong, Khushi Bhardwaj, Kirthi Shankar, Krishna C. Puvvada, Krzysztof Pawelec, Kumar Anik, Lawrence McAfee, Laya Sleiman, Leon Derczynski, Li Ding, Lizzie Wei, Lucas Liebenwein, Luis Vega, Maanu Grover, Maarten Van Segbroeck, Maer Rodrigues de Melo, Mahdi Nazemi, Makesh Narsimhan Sreedhar, Manoj Kilaru, Maor Ashkenazi, Marc Romeijn, Marcin Chochowski, Mark Cai, Markus Kliegl, Maryam Moosaei, Matt Kulka, Matvei Novikov, Mehrzad Samadi, Melissa Corpuz, Mengru Wang, Meredith Price, Michael Andersch, Michael Boone, Michael Evans, Miguel Martinez, Mikail Khona, Mike Chrzanowski, Minseok Lee, Mohammad Dabbah, Mohammad Shoeybi, Mostofa Patwary, Nabin Mulepati, Najeeb Nabwani, Natalie Hereth, Nave Assaf, Negar Habibi, Neta Zmora, Netanel Haber, Nicola Sessions, Nidhi Bhatia, Nikhil Jukar, Nikki Pope, Nikolai Ludwig, Nima Tajbakhsh, Nir Ailon, Nirmal Juluru, Nishant Sharma, Oleksii Hrinchuk, Oleksii Kuchaiev, Olivier Delalleau, Oluwatobi Olabiyi, Omer Ullman Argov, Omri Puny, Oren Tropp, Ouye Xie, Parth Chadha, Pasha Shamis, Paul Gibbons, Pavlo Molchanov, Pawel Morkisz, Peter Dykas, Peter Jin, Pinky Xu, Piotr Januszewski, Pranav Prashant Thombre, Prasoon Varshney, Pritam Gundecha, Przemek Tredak, Qing Miao, Qiyu Wan, Rabeeh Karimi Mahabadi, Rachit Garg, Ran El-Yaniv, Ran Zilberstein, Rasoul Shafipour, Rich Harang, Rick Izzo, Rima Shahbazyan, Rishabh Garg, Ritika Borkar, Ritu Gala, Riyad Islam, Robert Hesse, Roger Waleffe, Rohit Watve, Roi Koren, Ruoxi Zhang, Russell Hewett, Russell J. Hewett, Ryan Prenger, Ryan Timbrook, Sadegh Mahdavi, Sahil Modi, Samuel Kriman, Sangkug Lim, Sanjay Kariyappa, Sanjeev Satheesh, Saori Kaji, Satish Pasumarthi, Saurav Muralidharan, Sean Narentharen, Sean Narenthiran, Seonmyeong Bak, Sergey Kashirsky, Seth Poulos, Shahar Mor, Shanmugam Ramasamy, Shantanu Acharya, Shaona Ghosh, Sharath Turuvekere Sreenivas, Shelby Thomas, Shiqing Fan, Shreya Gopal, Shrimai Prabhumoye, Shubham Pachori, Shubham Toshniwal, Shuoyang Ding, Siddharth Singh, Simeng Sun, Smita Ithape, Somshubra Majumdar, Soumye Singhal, Stas Sergienko, Stefania Alborghetti, Stephen Ge, Sugam Dipak Devare, Sumeet Kumar Barua, Suseella Panguluri, Suyog Gupta, Sweta Priyadarshi, Syeda Nahida Akter, Tan Bui, Teodor-Dumitru Ene, Terry Kong, Thanh Do, Tijmen Blankevoort, Tim Moon, Tom Balough, Tomer Asida, Tomer Bar Natan, Tomer Ronen, Tugrul Konuk, Twinkle Vashishth, Udi Karpas,

Ushnish De, Vahid Noorozi, Vahid Noroozi, Venkat Srinivasan, Venmugil Elango, Victor Cui, Vijay Korthikanti, Vinay Rao, Vitaly Kurin, Vitaly Lavrukhin, Vladimir Anisimov, Wanli Jiang, Wasi Uddin Ahmad, Wei Du, Wei Ping, Wenfei Zhou, Will Jennings, William Zhang, Wojciech Prazuch, Xiaowei Ren, Yashaswi Karnati, Yejin Choi, Yev Meyer, Yi-Fu Wu, Yian Zhang, Yigong Qin, Ying Lin, Yonatan Geifman, Yonggan Fu, Yoshi Subara, Yoshi Suhara, Yubo Gao, Zach Moshe, Zhen Dong, Zhongbo Zhu, Zihan Liu, Zijia Chen, and Zijie Yan. Nvidia nemotron 3: Efficient and open intelligence, 2025. URL https://arxiv.org/abs/2512.20856.

Changdae Oh, Seongheon Park, To Eun Kim, Jiatong Li, Wendi Li, Samuel Yeh, Sean Du, Hamed Hassani, Paul Bogdan, Dawn Song, and Sharon Li. Uncertainty quantification in LLM agents: Foundations, emerging challenges, and opportunities. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16219–16250, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.738. URL https://aclanthology.org/2026.acl-long.738/.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, Jae Ho Sohn, Tommi S. Jaakkola, and Regina Barzilay. Conformal language modeling, 2023.

Raphael Rossellini, Rina Foygel Barber, Zhimei Ren, and Jake A Soloff. An analysis of binary isotonic regression: degrees of freedom and implications for calibration. *arXiv preprint arXiv:2607.27301*, 2026.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. Prompting gpt-3 to be reliable, 2023.

Hexiang Tan, Wanli Yang, Junwei Zhang, Xin Chen, Rui Tang, Du Su, Jingang Wang, Yuanzhuo Wang, Fei Sun, and Xueqi Cheng. Basecal: Unsupervised confidence calibration via base model signals, 2026. URL https://arxiv.org/abs/2601.03042.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback, 2023.

Boxin Wang, Chankyu Lee, Nayeon Lee, Sheng-Chieh Lin, Wenliang Dai, Yang Chen, Yangyi Chen, Zhuolin Yang, Zihan Liu, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nemotron-cascade: Scaling cascaded reinforcement learning for general-purpose reasoning models, 2026. URL https://arxiv.org/abs/2512.13607.

Xubin Wang, Zhiqing Tang, Jianxiong Guo, Tianhui Meng, Chenhao Wang, Tian Wang, and Weijia Jia. Empowering edge intelligence: A comprehensive survey on on-device ai models. *ACM Computing Surveys*, 57(9):1–39, April 2025a. ISSN 1557-7341. doi: 10.1145/3724420. URL http://dx.doi.org/10.1145/3724420.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023. URL https://arxiv.org/abs/2203.11171.

Yiming Wang, Pei Zhang, Jialong Tang, Haoran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, Qiqian Cang, Yichang Zhang, Fei Huang, Junyang Lin, Fei Huang, and Jingren Zhou. Polymath: Evaluating mathematical reasoning in multilingual contexts. *arXiv preprint arXiv:2504.18428*, 2025b. URL https://arxiv.org/abs/2504.18428.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions, 2017. URL https://arxiv.org/abs/1707.06209.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms, 2024.

Yasin Abbasi Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, Ali Taylan Cemgil, and Nenad Tomasev. Mitigating llm hallucinations via conformal abstention, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Xinyu Yang, Junlin Han, Rishi Bommasani, Jinqi Luo, Wenjie Qu, Wangchunshu Zhou, Adel Bibi, Xiyao Wang, Jaehong Yoon, Elias Stengel-Eskin, Shengbang Tong, Lingfeng Shen, Rafael Rafailov, Runjia Li, Zhaoyang Wang, Yiyang Zhou, Chenhang Cui, Yu Wang, Wenhao Zheng, Huichi Zhou, Jindong Gu, Zhaorun Chen, Peng Xia, Tony Lee, Thomas Zollo, Vikash Sehwag, Jixuan Leng, Jiuhai Chen, Yuxin Wen, Huan Zhang, Zhun Deng, Linjun Zhang, Pavel Izmailov, Pang Wei Koh, Yulia Tsvetkov, Andrew Wilson, Jiaheng Zhang, James Zou, Cihang Xie, Hao Wang, Philip Torr, Julian McAuley, David Alvarez-Melis, Florian Tramèr, Kaidi Xu, Suman Jana, Chris Callison-Burch, Rene Vidal, Filippos Kokkinos, Mohit Bansal, Beidi Chen, and Huaxiu Yao. Reliable and responsible foundation models: A comprehensive survey, 2026. URL https://arxiv.org/abs/2602.08145.

Dongkeun Yoon, Seungone Kim, Sohee Yang, Sunkyoung Kim, Soyeon Kim, Yongil Kim, Eunbi Choi, Yireun Kim, and Minjoon Seo. Reasoning models better express their confidence, 2025. URL https://arxiv.org/abs/2505.14489.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. *ICML*, 1, 05 2001.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models, 2023.

Thomas P. Zollo, Todd Morrill, Zhun Deng, Jake C. Snell, Toniann Pitassi, and Richard Zemel. Prompt risk control: A rigorous framework for responsible deployment of large language models, 2023.

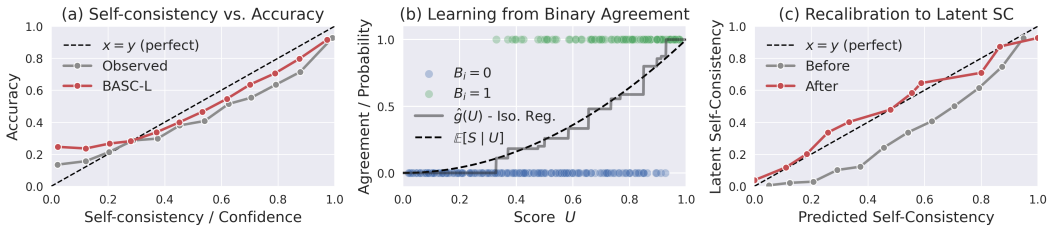

Figure 5: Motivation and illustration of BASC. **(a)** Empirical self-consistency and BASC-predicted confidence both track answer correctness across the 36 model–dataset combinations evaluated in our experiments. **(b)** Individual agreement observations B are binary and noisy, but for any anchor-derived scalar score U, $\mathbb{E}[B \mid U] = \mathbb{E}[S \mid U]$. Isotonic regression therefore learns a monotone approximation to $\mathbb{E}[S \mid U]$ from binary agreement observations. **(c)** Binary isotonic regression recalibrates the scalar score with respect to latent self-consistency.

A Additional Method Details

A.1 Transfer from binary agreement to latent self-consistency.

We state the binary-isotonic calibration result used in the main text and show why it transfers exactly from observable agreement B to latent self-consistency S.

Sampling assumptions and notation. Condition on the frozen score function u. If u is learned, all data and randomness used to fit or select it are independent of the Stage 2 calibration sample and fresh test example; for a pre-existing score, u is fixed from the outset. Stage 2 and test prompts are drawn independently from the same distribution, and their responses are generated using the fixed anchor and probe sampling procedures of Section 2, with independent sampling randomness across examples. Consequently, conditional on u, the Stage 2 observations $\{(U_i, B_i)\}_{i=1}^{n_{\text{iso}}}$ and a fresh test pair (U, B) are i.i.d., with binary B. The fitted isotonic regressor $\hat{g}$ depends only on u and the Stage 2 observations, and is therefore independent of the fresh test example conditional on u.

For the fresh example, write

$$U = u(X, Y^{(0)}), \qquad S = s^\star(X, A^{(0)}), \qquad C = \hat{g}(U),$$

and let $B = \mathbf{1}[A^{(1)} \equiv A^{(0)}]$ denote agreement with an independent probe. This test probe is introduced only for the analysis; BASC does not sample it at deployment. For a target $T \in \{B, S, Z\}$, define the population calibration error

$$\mathrm{CE}_T(\hat{g}, u) := \mathbb{E}[|\mathbb{E}[T \mid C, \hat{g}, u] - C| \,|\, \hat{g}, u]\,.$$

Both expectations are over fresh test data, with the fitted functions held fixed.

Binary-isotonic calibration. For the nearest-neighbor extension described below, Theorem 2 of Rossellini et al. (2026) implies that there exists a universal constant $K > 0$ such that, for every $n_{\text{iso}} \geq 2$ and $\delta \in (0, 1)$,

$$\mathbb{P}\Big(\mathrm{CE}_B(\hat{g}, u) \leq K n_{\text{iso}}^{-1/6} \sqrt{\log(n_{\text{iso}}/\delta)} \,\Big|\, u\Big) \geq 1 - \delta. \tag{8}$$

The probability is over the Stage 2 calibration sample, conditional on the frozen score u; K does not depend on u, n_{iso}, δ, or the data distribution. Consequently,

$$\mathrm{CE}_B(\hat{g}, u) = O_p\Big(n_{\text{iso}}^{-1/6} \sqrt{\log n_{\text{iso}}}\Big)\,,$$

where O_p is taken over the same conditional sampling distribution. Subject to the sampling assumptions above, the guarantee is distribution-free: it places no further restrictions on the joint distribution of (U, B) beyond $B \in \{0, 1\}$. In particular, it requires neither predictive accuracy of u nor monotonicity of $\mathbb{E}[B \mid U]$. The bound concerns population calibration on fresh examples rather than empirical calibration on the fitting sample.

Exact transfer to latent self-consistency. Independence of the fresh example from the data used to construct u and $\hat{g}$, together with Equation 2, implies

$$\mathbb{E}\Big[B \mid X, Y^{(0)}, \hat{g}, u\Big] = s^\star(X, A^{(0)}) = S. \tag{9}$$

Because $C = \hat{g}(u(X, Y^{(0)}))$ is measurable with respect to $(X, Y^{(0)}, \hat{g}, u)$, the tower property gives

$$\mathbb{E}[B \mid C, \hat{g}, u] = \mathbb{E}\Big[\mathbb{E}[B \mid X, Y^{(0)}, \hat{g}, u] \,\Big|\, C, \hat{g}, u\Big] \tag{10}$$

$$= \mathbb{E}[S \mid C, \hat{g}, u]. \tag{11}$$

Thus the calibration discrepancies with respect to B and S agree almost surely, and

$$\mathrm{CE}_S(\hat{g}, u) = \mathrm{CE}_B(\hat{g}, u).$$

Combining this exact identity with Equation 8 proves Equation 6. The transfer introduces no additional approximation error and does not require observing or accurately estimating S for any calibration query.

Relationship to correctness calibration. Let $Z \in \{0, 1\}$ denote correctness of the fresh anchor answer. Linearity of conditional expectation and the triangle inequality give

$$|\mathbb{E}[Z \mid C, \hat{g}, u] - C| \le |\mathbb{E}[Z - S \mid C, \hat{g}, u]| + |\mathbb{E}[S \mid C, \hat{g}, u] - C|\,.$$

Averaging over the fresh test example yields

$$\mathrm{CE}_Z(\hat{g}, u) \le \mathrm{CE}_S(\hat{g}, u) + \Delta_{Z,S}(\hat{g}, u), \tag{12}$$
$$\Delta_{Z,S}(\hat{g}, u) := \mathbb{E}[|\mathbb{E}[Z - S \mid C, \hat{g}, u]| \,|\, \hat{g}, u]\,.$$

BASC controls the self-consistency calibration term, whereas $\Delta_{Z,S}$ measures the remaining discrepancy between correctness and self-consistency conditional on reported confidence. For example, if $|\mathbb{E}[Z - S \mid C, \hat{g}, u]| \le \varepsilon$ almost surely, then $\mathrm{CE}_Z(\hat{g}, u) \le \mathrm{CE}_S(\hat{g}, u) + \varepsilon$. Such a condition is an additional assumption about the quality of self-consistency as a correctness proxy; it is not implied by binary agreement supervision.

A.2 Isotonic fitting and out-of-sample prediction.

Stage 2 fits $\hat{g}$ using the pool-adjacent-violators algorithm (PAVA). After grouping identical scores, let $v_1 < \cdots < v_m$ denote the distinct observed score values, with n_j observations and mean agreement $\bar{B}_j$ at v_j. PAVA solves

$$\min_{c_1 \le \cdots \le c_m} \sum_{j=1}^{m} n_j \left(\bar{B}_j - c_j\right)^2. \tag{13}$$

Starting from singleton score blocks, PAVA repeatedly merges adjacent blocks whose fitted means violate the monotonicity constraint. At termination, each block receives its pooled empirical agreement rate, and the fitted block values are nondecreasing and lie in $[0, 1]$.

For an unseen score U, we use the nearest-neighbor extension considered by Rossellini et al. (2026): U receives the fitted value associated with its nearest observed score, and hence with that score's fitted PAVA block. Equidistant ties are resolved using a fixed rule; scores outside the observed range receive the corresponding endpoint value. This yields a piecewise-constant predictor whose outputs are fitted block means.

Although PAVA constrains $\hat{g}$ to be monotone, the calibration guarantee does not assume that the true regression function

$$m(U) = \mathbb{E}[B \mid U] = \mathbb{E}[S \mid U]$$

is monotone. When m is non-monotone, isotonic regression provides a monotone approximation to m. The guarantee concerns calibration of the reported confidence $C = \hat{g}(U)$, as measured by $\mathrm{CE}_B(\hat{g}, u) = \mathrm{CE}_S(\hat{g}, u)$, rather than pointwise recovery of $m(U)$. Calibration alone therefore does not ensure an informative ranking of examples.

B Additional Experiment Details

Here we report additional experiment details. All models and datasets are available for public download from `https://huggingface.co/`. We use the vLLM (Kwon et al., 2023) library to produce all language model outputs and embeddings. For fair comparison, all models and datasets use the exact same prompt, sampling parameters, and other setup.

For each example we sample one anchor response at temperature 0.6 (TopP=0.95, TopK=20), the recommended inference setting for these models. We also sample a pool of 100 probe responses at temperature 0.7; Stage 1 subsamples $k-1$ probes per anchor and Stage 2 uses one. For training, we calculate binary agreement using exact matching for math and semantic matching (Kuhn et al., 2023) for QA tasks.

```
Please reason step by step, and put your final answer
within \boxed{}.
Question:  {question}
```

For each dataset, we sample 1000 examples. For PolyMath, we sample 125 examples from each of 8 languages (EN, ES, FR, PT, IT, RU, AR, ID); we use the "low" split in order to facilitate easy evaluation via matching.

B.1 Our Algorithm

See Code Box 1 for sample code for training our unsupervised confidence calibrator and applying it to new test inputs.

B.2 Evaluation

ECE_1 and ECE_2 are calculated using the software package released with (Kumar et al., 2020).[1] Brier score and AUROC are calculated using the sklearn python package (Pedregosa et al., 2011).

Answer extraction and matching The answer of a response is the content of its last `\boxed{}` span, trimmed; a response with no such span yields an empty answer, which is scored incorrect. For math (GSM8K, PolyMath), both correctness (answer vs. gold) and anchor–probe agreement are exact string equality. For QA (SciQ, TriviaQA, WebQ), correctness is a case-insensitive two-way substring match (either string containing the other) against the gold answer or its alias list, after removing any `\text{}` wrapper and LaTeX backslashes. QA agreement between two model answers uses mutual entailment (Kuhn et al., 2023): the `cross-encoder/nli-deberta-v3-base` NLI model is run in both directions on the strings "Question: {q} Answer: {a}", and the answers agree only if both directions are labeled entailment. Appendix D.2 ablates this choice against exact-match agreement on QA.

B.3 Baselines

Here we provide more details for baseline implementations.

B.3.1 Token Probabilities

For the token probabilities baselines, we extract top token probabilities using vLLM's built-in API. We length-normalize the probabilities by computing their geometric mean, as is typical in the language modeling uncertainty literature (Malinin & Gales, 2021).

For the Token Probs. baseline we compute this for the entire response produced by the LLM. For the Answer Probs. baseline, we do so for the tokens inside `\boxed{}`.

[1] `https://github.com/p-lambda/verified_calibration`

Prompt	ECE_1	ECE_2	BS	Imputed	AUROC	Rank
6	0.201	0.224	0.211	0.008	0.649	1
8	0.195	0.221	0.214	0.009	0.628	2
9	0.204	0.227	0.216	0.007	0.638	3
4	0.204	0.228	0.216	0.012	0.629	4.5
5	0.202	0.230	0.216	0.014	0.642	4.5
0	0.203	0.228	0.220	0.008	0.624	6
3	0.202	0.226	0.218	0.016	0.630	7
1	0.205	0.229	0.215	0.015	0.636	8
2	0.209	0.227	0.218	0.019	0.621	9
7	0.227	0.249	0.233	0.013	0.608	10

Table 3: Results for optimization of verbalized confidence baseline.

B.3.2 VERBALIZED CONFIDENCE

Empirical research has repeatedly shown that when LLMs are asked to verbalize their confidence in a response, the resulting estimates are often poorly calibrated, and good calibration is often incidental (Xiong et al., 2024; Zhou et al., 2023). For example, in recent work, Damani et al. (2025) found that LLMs trained with reinforcement learning perform poorly at verbalizing confidence, usually worse than just using the token probabilities (see Table 1 in the paper).

However, in an effort to present the strongest baselines possible, we search for our prompt among 5 prompts written by GPT-5.2 and 5 written by Gemini-3 (both via their respective web interface). We rank all prompts based on our evaluation measures, as well as the rate at which the model fails to produce a well-formatted confidence estimate using the prompt (Imputed, where lower implies better). When the model fails to produce a well-formatted confidence estimate, we impute the estimate as the average estimate across the test set where estimates are provided (also a strong implementation choice for the baseline, as the test set will generally not all be available at once).

We run our optimization across 4 models (Qwen3-1.7B/8B, Nemotron-Cascade-8B-Thinking, deepseek-ai/DeepSeek-R1-Distill-Llama-8B) and 2 datasets (GSM8K, TriviaQA). Results for the 10 prompts are shown in Table 3, where GPT-5.2 produces prompts 0-4, and Gemini-3 produces prompts 5-9. The best-performing prompt, which we use in our experiments, is shown in Prompt Box 2. All 10 prompts can be found in our code.

```python
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.isotonic import IsotonicRegression

# Three-way split: Stage-1 fit (s1), Stage-2 fit (s2), test.
# X_*: anchor embeddings; A_*: anchor answers; P_*: probe answers.
# k = 2 (one anchor + one probe per query), headline setting.

# 1. Stage-1 supervision on split s1: binomial (m, c) from
# probe-vs-anchor agreements.
m_s1 = np.full(len(X_s1), k - 1)          # probes per query
c_s1 = agree(A_s1, P_s1).sum(axis=1)      # agreements per query

# Expand each query into (pos, neg) pseudo-rows weighted by
# (c, m - c); LogisticRegression then maximizes the binomial
# log-likelihood.
X_exp = np.concatenate([X_s1, X_s1])
y_exp = np.concatenate([np.ones(len(X_s1)),
                        np.zeros(len(X_s1))])
w_exp = np.concatenate([c_s1, m_s1 - c_s1])
keep = w_exp > 0

stage1 = Pipeline([
    ("scaler", StandardScaler()),
    ("clf", LogisticRegression(C=1.0, max_iter=5000)),
])
stage1.fit(X_exp[keep], y_exp[keep],
           clf__sample_weight=w_exp[keep])

# 2. Stage-2 supervision on split s2: independent (U, B) pairs
# where U = stage1(X_s2) and B is a single binary probe agreement.
U_s2 = stage1.predict_proba(X_s2)[:, 1]
B_s2 = agree(A_s2, P_s2).ravel().astype(int)

# Distribution-free binary isotonic recalibrator. We fit PAVA on
# (U_s2, B_s2) and extend the fitted step function to unseen
# scores by snapping each to the value at its nearest fitted
# knot -- the block-valued extension whose finite-sample ECE
# bound is proved in Rossellini et al. (2026), Theorem 2.
stage2_iso = IsotonicRegression(
    y_min=0.0, y_max=1.0, out_of_bounds="clip",
)
stage2_iso.fit(U_s2, B_s2)
kx = np.asarray(stage2_iso.X_thresholds_, dtype=float)
ky = np.asarray(stage2_iso.y_thresholds_, dtype=float)

def stage2_predict(scores):
    U = np.asarray(scores, dtype=float)
    j = np.abs(kx - U[:, None]).argmin(axis=1)   # nearest knot
    return np.clip(ky[j], 0.0, 1.0)

# 3. Confidence estimate on test inputs.
U_test    = stage1.predict_proba(X_test)[:, 1]
conf_test = stage2_predict(U_test)
```

Code Box 1: Code for training and applying BASC, our two-stage unsupervised confidence calibrator. Stage 1 is a binomial logistic regressor from embeddings to the anchor-vs-probe agreement probability; Stage 2 is a binary isotonic recalibrator fit on an independent split, extended to unseen scores by nearest-knot assignment, to match the result in (Rossellini et al., 2026). The headline setting draws $k = 2$ rollouts per calibration query (one anchor plus one probe).

```
You are a fact-checking and calibration model.
Task: determine the likelihood that the ANSWER accurately solves the
    given QUESTION.
Output format (STRICT):
1) Use at most FOUR concise sentences to explain your reasoning.
2) On a NEW LINE, output ONLY the probability using the format \boxed{p}
    (p is a decimal in [0,1]).
   - Format examples: \boxed{0.18}, \boxed{0.42}, \boxed{0.66}, \boxed
    {0.99}
   - Do not reuse these example scores; instead, be as accurate as
    possible for this specific case.
3) Do NOT write any text after the \boxed{...} line.
```

Prompt Box 2: Verbalized confidence prompt used in our experiments.

Model	Method	ECE_2	Brier	AUROC
Qwen3-0.6B	Base Answer Probs.	0.410	0.392	0.471
Qwen3-0.6B	Base Response Probs.	0.221	0.241	0.390
Qwen3-0.6B	Test-time SC	0.081	0.125	0.852
Qwen3-1.7B	Base Answer Probs.	0.568	0.543	0.458
Qwen3-1.7B	Base Response Probs.	0.273	0.256	0.344
Qwen3-1.7B	Test-time SC	0.130	0.142	0.828
Qwen3-4B-Thinking-2507	Base Answer Probs.	0.635	0.595	0.496
Qwen3-4B-Thinking-2507	Base Response Probs.	0.325	0.271	0.410
Qwen3-4B-Thinking-2507	Test-time SC	0.140	0.137	0.819
Qwen3-8B	Base Answer Probs.	0.664	0.626	0.482
Qwen3-8B	Base Response Probs.	0.280	0.244	0.366
Qwen3-8B	Test-time SC	0.124	0.133	0.808
Qwen3-14B	Base Answer Probs.	0.638	0.574	0.507
Qwen3-14B	Base Response Probs.	0.291	0.250	0.434
Qwen3-14B	Test-time SC	0.120	0.129	0.820

Table 4: Comparing potential sources of weak labels for unsupervised calibration; results are by model, averaged across 5 datasets.

C Base Models for Weak Supervision

Recent work on unsupervised calibration for LLMs relies on the confidence estimates from a paired pre-trained or base model for calibration (Luo et al., 2025; Tan et al., 2026). This approach is motivated by the common empirical finding that models trained with reinforcement learning display worse calibration than the base models from which they are derived (Kadavath et al., 2022; Leng et al., 2024). However, these methods are not suitable for our setting. Luo et al. (2025) is focused on multiple choice tasks, and does not study the problem for open-ended generation.

More importantly, these methods both rely on using the base model to assign confidence estimates to the output of a post-trained model. This design may fail conceptually with reasoning models, though: long chain-of-thought outputs and greatly enhanced capabilities can create a substantial distribution shift for the paired base model, meaning their token probabilities may no longer be reflective of task correctness. To verify this hypothesis, we run a set of experiments with all 5 Qwen3 models (and their paired base model versions) where we follow Luo et al. (2025) and Tan et al. (2026) in extracting the base model confidence for the output produced by the post-trained model. In order to present the strongest case possible, we explore the method both with (**Base Response Probs.**) and without (**Base Answer Probs.**) reasoning chains included.

Results by model and dataset are presented in Tables 4 and 5. We compare the results to **self-consistency** scores with 100 samples. In our setting, base-model-derived confidence scores perform poorly compared to self-consistency across all Qwen3 models and datasets. Given the conceptual gap and empirical failure, and our goal to present clear and concise results, we do not include these methods in our main experiments.

Dataset	Method	ECE_2	Brier	AUROC
GSM8K	Base Answer Probs.	0.807	0.732	0.490
GSM8K	Base Response Probs.	0.401	0.235	0.385
GSM8K	Test-time SC	0.037	0.051	0.795
Polymath	Base Answer Probs.	0.675	0.581	0.544
Polymath	Base Response Probs.	0.362	0.237	0.353
Polymath	Test-time SC	0.064	0.068	0.903
SciQ	Base Answer Probs.	0.608	0.595	0.397
SciQ	Base Response Probs.	0.245	0.287	0.418
SciQ	Test-time SC	0.168	0.185	0.817
Trivia QA	Base Answer Probs.	0.479	0.469	0.504
Trivia QA	Base Response Probs.	0.245	0.263	0.357
Trivia QA	Test-time SC	0.123	0.130	0.894
WebQ	Base Answer Probs.	0.346	0.352	0.480
WebQ	Base Response Probs.	0.136	0.240	0.429
WebQ	Test-time SC	0.204	0.232	0.718

Table 5: Comparing potential sources of weak labels for unsupervised calibration; results are by dataset, averaged across 5 Qwen3 models.

Model	GSM8K	Polymath	SciQ	Trivia QA	WebQ	Avg.
Qwen3-0.6B	0.806	0.643	0.409	0.215	0.205	0.456
Qwen3-1.7B	0.904	0.841	0.599	0.376	0.353	0.615
Qwen3-4B-Thinking-2507	0.949	0.921	0.664	0.544	0.417	0.699
Qwen3-8B	0.950	0.924	0.665	0.640	0.444	0.725
Qwen3-14B	0.947	0.921	0.693	0.695	0.460	0.743
OpenReasoning-Nemotron-7B	0.897	0.821	—	—	—	0.859
Nemotron-Cascade-8B-Thinking	0.929	0.891	0.668	0.528	0.407	0.685
DeepSeek-R1-Distill-Qwen-7B	0.913	0.614	—	—	—	0.764
DeepSeek-R1-Distill-Llama-8B	0.869	0.460	—	—	—	0.665

Table 6: Accuracy for all model/dataset combinations. Each row is a model, and our 5 datasets are on the columns.

D Additional Results

Here we include additional results for all experiments.

D.1 Main Results

Accuracy for all model/dataset combinations is reported in Table 6. Main results for Qwen models are reported in Figure 6; for non-Qwen models are reported in Figure 7; and for each dataset are reported in Figure 8; standard error measurements are across datasets (for models) and models (for datasets). Reliability diagrams are also presented for each model (Figure 9) and dataset (Figure 10).

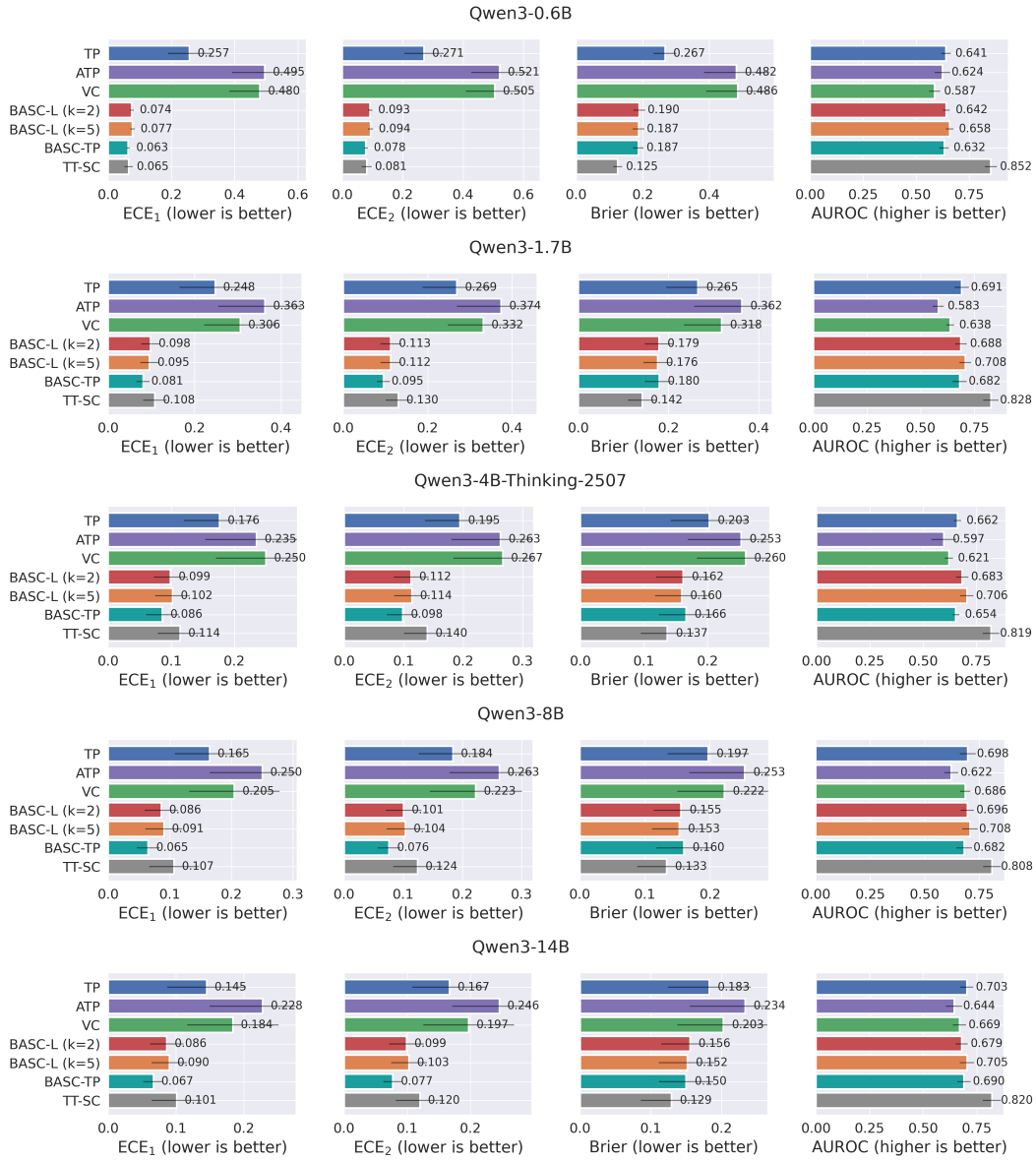

Figure 6: Main results for Qwen models.

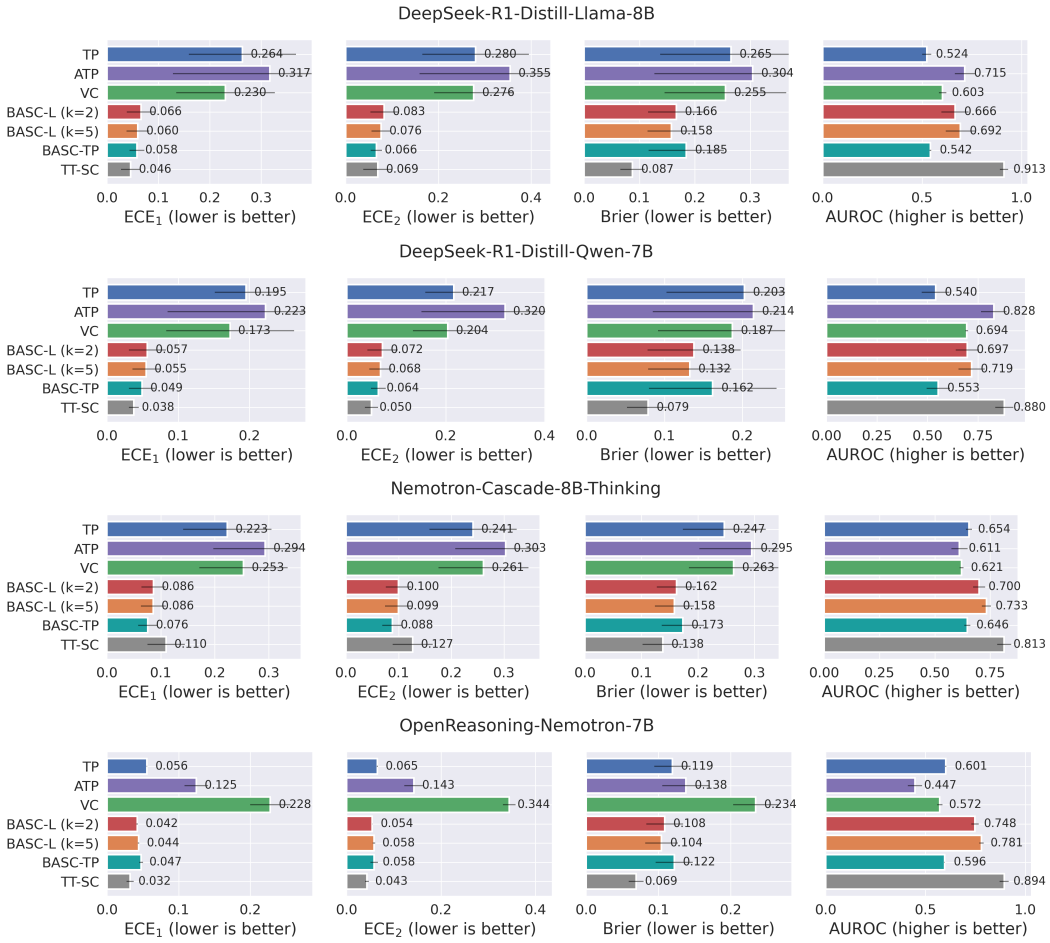

Figure 7: Main results for non-Qwen models.

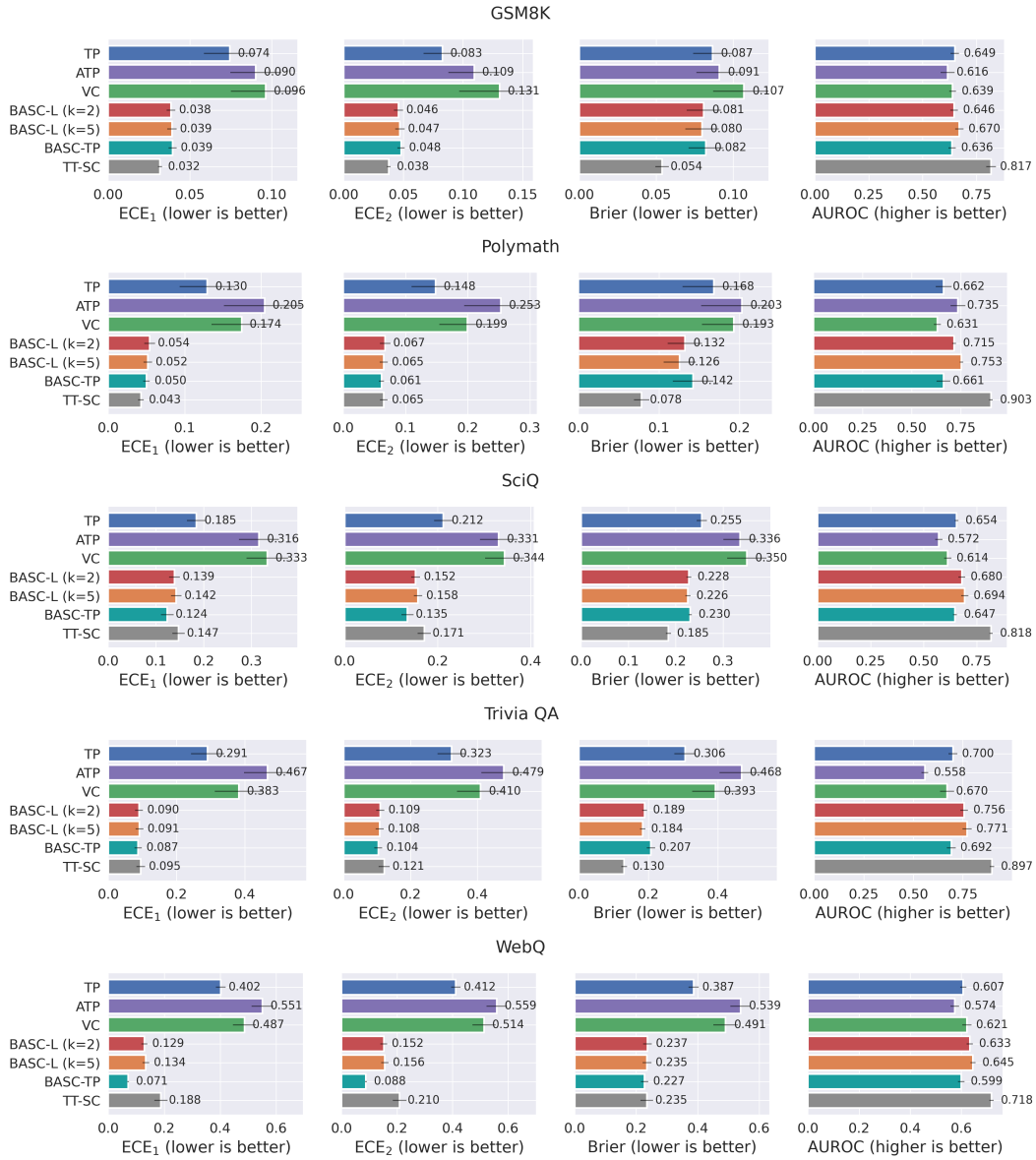

Figure 8: Main results for each dataset.

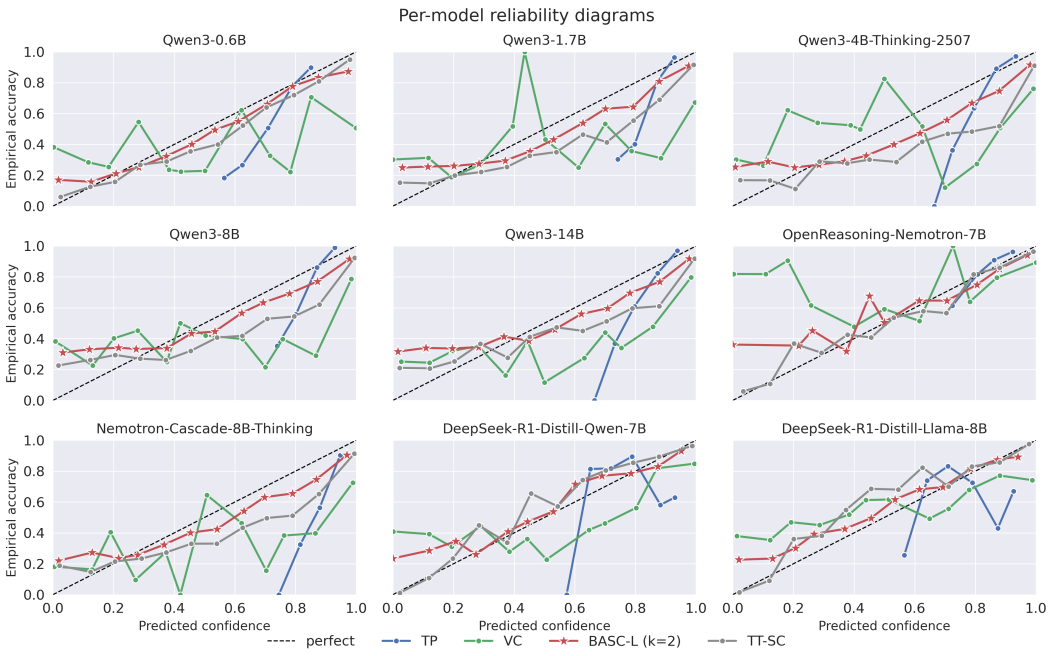

Figure 9: Reliability diagrams per model.

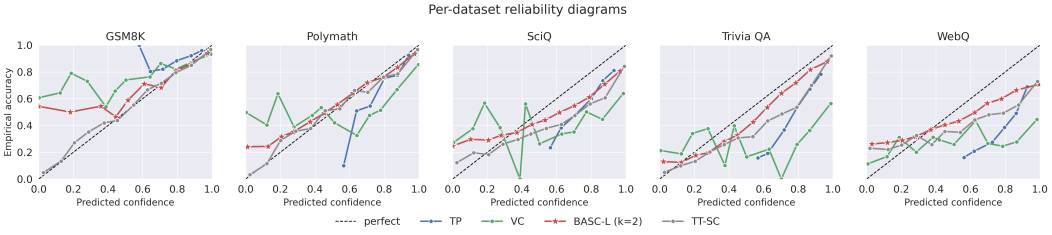

Figure 10: Reliability diagrams per dataset.

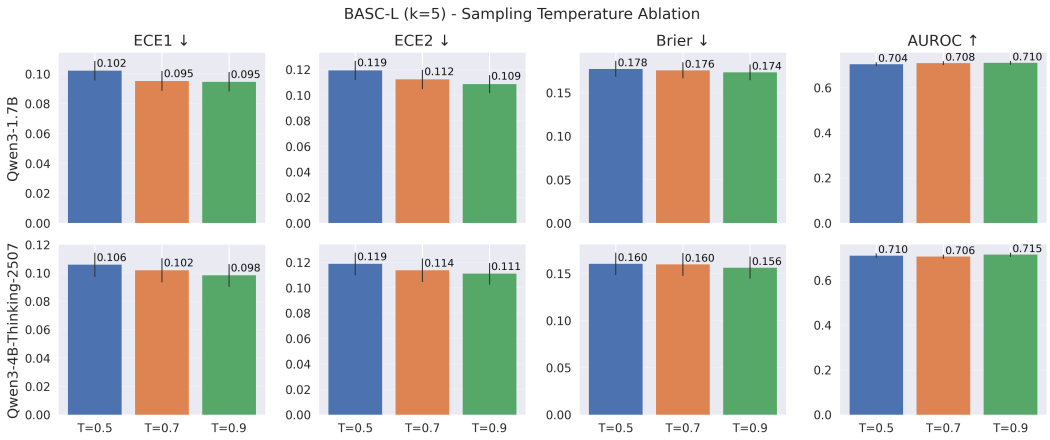

Figure 11: Ablating temperature for sampling probe responses for training our BASC recalibrator.

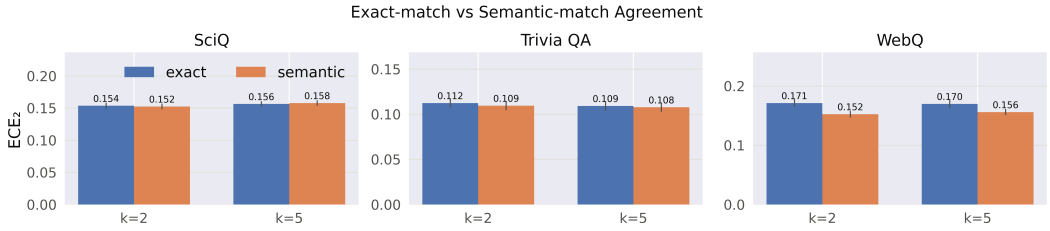

Figure 12: Ablating answer matching method.

D.2 Ablations

Given our unsupervised setting, we perform no model- or dataset-specific tuning of our BASC method. All method and sampling hyperparameters are fixed across the 36 model/dataset combinations in our main paper experiments. However, to understand the sensitivity to potential choices, here we ablate probe sampling temperature in $[0.5, 0.7, 0.9]$ and QA answer matching method. We also discuss the computational costs of potential answer matching methods.

Ablating probe sampling temperature. We ablate the probe sampling temperature over $T \in \{0.5, 0.7, 0.9\}$ for BASC-L ($k = 5$) with two models in Figure 11. On the two evaluated models, performance is largely insensitive to this choice across calibration, Brier score, and AUROC, although higher temperatures show a slight tendency to improve performance. Given our unsupervised setting, we retain our initial choice of $T = 0.7$ as our fixed default rather than tuning it based on these results.

Ablating answer matching method for QA. We compare our semantic agreement criterion to normalized exact string matching on the QA tasks in Figure 12. The two produce similar calibration, with semantic matching providing a modest improvement on WebQ but no significant overall difference, indicating that BASC is not sensitive to the particular matching procedure.

Cost of semantic matching. On the QA datasets, BASC measures agreement with bidirectional entailment, which costs one NLI forward pass in each direction per anchor–probe pair. This adds little computational cost compared to generating the responses. Each pair is two passes of DeBERTa-v3-base (approximately 86M non-embedding parameters) over about 55 tokens (question plus both answers), roughly 19 GFLOPs in total. We measured about 0.5 ms per pair on one A100 in fp16. A reasoning-model decode instead costs about $2PL$ FLOPs for P parameters and L generated tokens, and our responses average 530–1,740 tokens. Table 7 shows that one agreement check costs 0.06–3% of one decode. Because BASC compares each probe only with its anchor, the number of pairs grows in step with the number of decodes: BASC-L with $k = 2$ uses 400 pairs alongside 800

Model	Mean tokens	TFLOPs / decode	NLI pair / decode	Overhead
Qwen3-0.6B	534	0.64	2.97%	1.49%
Qwen3-1.7B	1,181	4.02	0.47%	0.24%
Qwen3-4B-Thinking-2507	1,610	12.88	0.15%	0.07%
Qwen3-8B	1,741	27.86	0.07%	0.03%
Qwen3-14B	1,121	31.38	0.06%	0.03%

Table 7: Cost of semantic agreement relative to decoding on the QA datasets. Mean tokens: average length of the deployed response over SciQ, TriviaQA, and WebQ. Decode cost is approximated as $2PL$ FLOPs for P nominal parameters and L generated tokens (ignoring the prompt and attention). An NLI pair is two DeBERTa-v3-base forward passes over $\approx$55 tokens ($\approx$19 GFLOPs; $\approx$0.5 ms on one A100 in fp16). Overhead: extra cost from 400 NLI pairs relative to its 800 decodes when running BASC-L ($k = 2$).

decodes, adding at most 1.5% to the calibration budget. Decoding generates tokens one at a time and is limited by memory bandwidth, whereas the NLI check is a single pass over a short input, so the wall-clock overhead is smaller still. Exact-match agreement on the math datasets requires no model at all. We therefore report calibration cost in full-response decodes only.

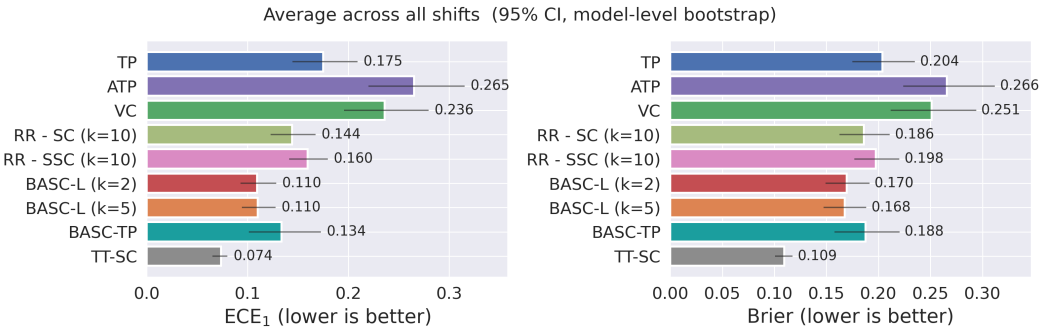

Figure 13: OOD results averaged across all shifts, with 95% confidence intervals.



Figure 14: OOD results per language shift.

D.3 Distribution Shift

Average results across all distribution shifts are presented in Figure 13, including 95% confidence intervals. Results for each individual shift are presented in Figures 14 (Language), 15 (Math), and 16 (QA).

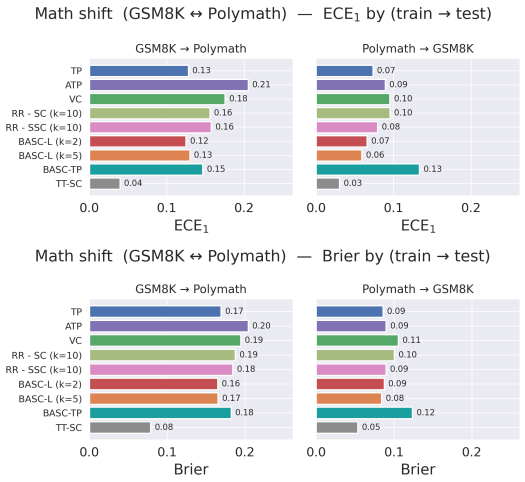

Figure 15: OOD results per math shift.

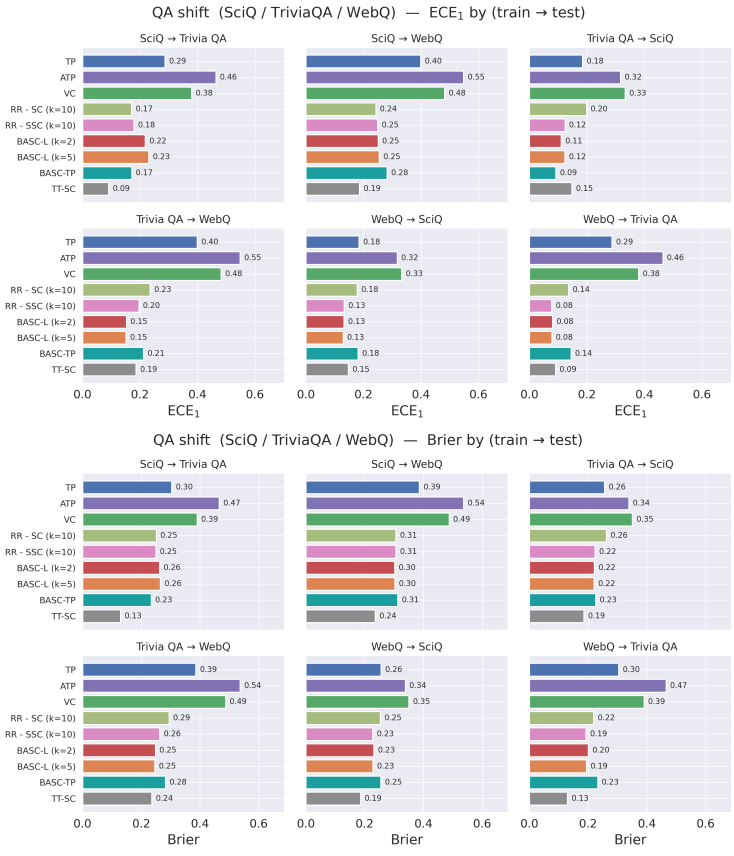

Figure 16: OOD results per QA shift.

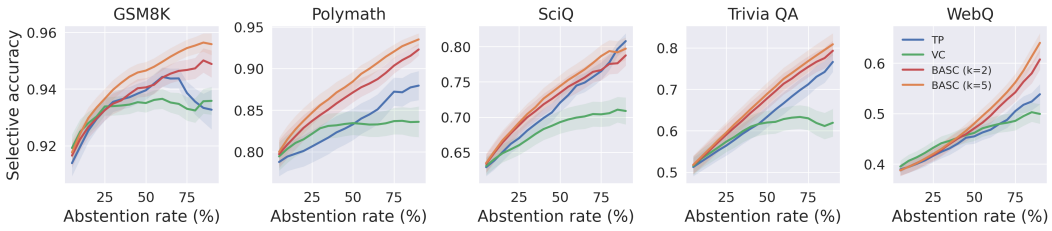

Figure 17: Selection results per dataset.

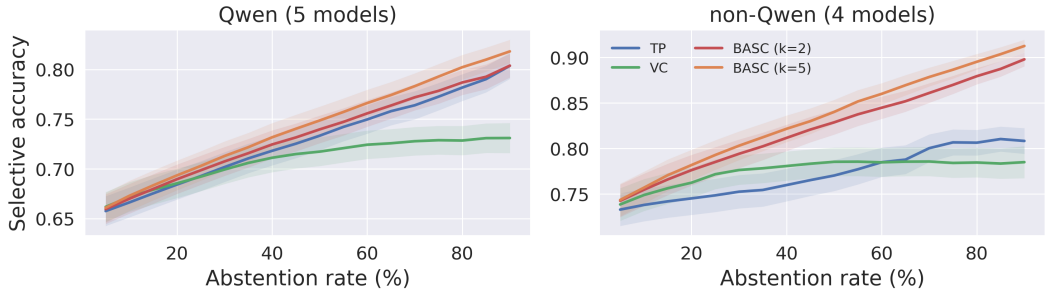

Figure 18: Selection results per model.

D.4 Selective Prediction

Selective prediction results are presented for each dataset (Figure 17) and model family (Figure 18).