Self-Calibrating Language Models via Test-Time Discriminative Distillation

Mohamed Rissal Hedna, Jan Strich, Martin Semmann, Chris Biemann
Hub of Computing and Data Science (HCDS)
University of Hamburg, Germany

Correspondence: `{first_name}.{last_name}@uni-hamburg.de`

Abstract

Large language models (LLMs) are systematically overconfident: they routinely express high certainty on questions they often answer incorrectly. Existing calibration methods either require labeled validation data, degrade under distribution shifts, or incur substantial inference costs. Recent work has shown that LLMs already contain a better-calibrated signal than the one they verbalize: the token probability of "True" when the model is asked "Is this answer correct?" ($P(\text{True})$) consistently outperforms their stated confidence, a gap that is theoretically grounded as generative error is lower-bounded by roughly twice the corresponding discriminative error. We introduce **SECL** (**SE**lf-**C**alibrating **L**anguage Models), a test-time training (TTT) pipeline that exploits this gap as self-supervision: adaptation requires no correctness labels or human feedback, with only two hyperparameters set once per model family from a small labeled probe. SECL adapts only when the input distribution shifts, training on just 6–26% of the question stream at lower cost than the baseline it distills from. Across five language models (2B–12B) from four model families and four diverse domains, SECL reduces Expected Calibration Error (ECE) by 56–78%, outperforming its own supervision signal and matching or outperforming recent inference-time methods. SECL is the first method to apply TTT to calibration; seven ablations covering signal quality, gating strategy, weight accumulation, loss design, domain ordering, hyperparameter sensitivity, and layer selection confirm that each component is crucial and robust across configurations. Code: github.com/rissalhedna/SECL

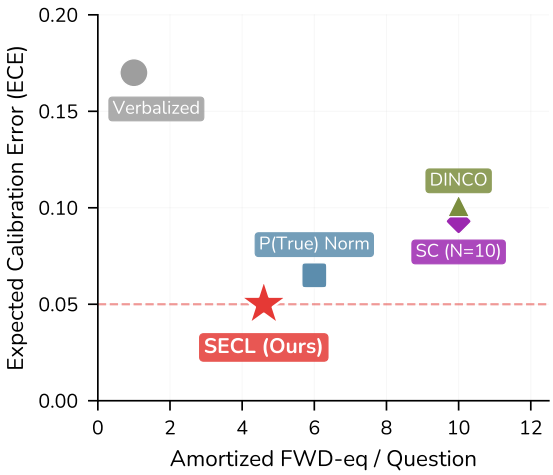

Figure 1: Calibration error vs. inference cost for Llama 3.2-3B (lower-left is better). Each point is a calibration method; the x-axis measures cost in forward-pass equivalents per question, and the y-axis measures Expected Calibration Error (ECE). SECL achieves the lowest calibration error at a fraction of the cost of existing methods.

1 Introduction

LLMs are systematically overconfident (Jiang et al., 2021; Xiong et al., 2024), and alignment procedures such as Reinforcement Learning from Human Feedback (RLHF; Bai et al., 2022) worsen this by rewarding agreement with human preferences over truthfulness (Sharma et al., 2024). The consequences are practical: in healthcare, where LLMs increasingly support triage and diagnostics (Liu et al., 2025), a review of 519 studies found that only 1.2% measured calibration despite 95.4% measuring accuracy (Bedi et al., 2025). A model reporting 90% certainty on questions it answers correctly only 30% of the time erodes clinician trust and risks patient harm. Addressing this problem needs calibration methods that work without labels and can adapt to new domains at test time.

Yet LLMs already contain a better-calibrated signal than the one they verbalize. When asked whether their own answer to a question is correct, LLMs produce probability estimates ($P(\text{True})$) that are substantially better calibrated than the confidence they express during generation (Kadavath

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 2238–2266
October 24-29, 2026 ©2026 Association for Computational Linguistics

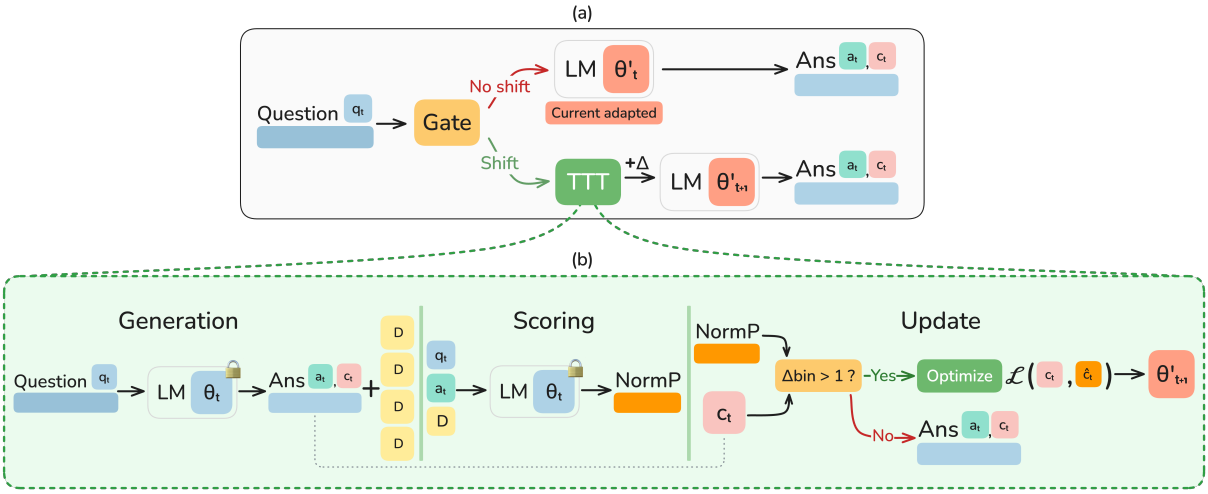

Figure 2: Overview of SECL. **(a)** *Test-Time Inference.* An entropy-based change detector (Section 3.1) monitors the input stream. If no shift is detected, the adapted model θ'_t is used directly; otherwise, a calibration burst updates it to θ'_{t+1}. **(b)** *Calibration Burst.* For each of B=50 questions: the frozen model generates an answer with confidence c_t and distractors, computes $\text{NormP}_{\text{True}}$ (Section 3.2), and applies a LoRA update when the two signals disagree by more than one bin (Section 3.3). Weights accumulate across questions without resetting.

et al., 2022; Tian et al., 2023). This exposes a systematic gap between *discrimination* and *generation*, with theoretical backing: Kalai et al. (2025) shows that a model's generative error is lower-bounded by roughly twice its discriminative error. Intuitively, a model that cannot reliably produce the correct answer can still often recognize when a given answer, or its own, is wrong. This gap provides a continual source of self-supervision that needs no labeled data. Prior work documents this phenomenon in LLMs; our experiments test whether the same mechanism can be exploited effectively in small language models.

Prior work on calibration falls into three broad categories, each with a key limitation that SECL addresses (see Section 2 for more details). Sampling-based methods (Manakul et al., 2023; Kuhn et al., 2023) measure consistency across multiple generations but are expensive and fail on consistent hallucinations. SECL avoids repeated sampling entirely. Static probing methods (Du et al., 2024; Liu et al., 2024) analyze internal representations but degrade under distribution shift. SECL adapts continuously via test-time training. Training-based approaches (Lin et al., 2022a; Stangel et al., 2025; Damani et al., 2025) can improve calibration within a domain, but many either require supervised labels or degrade OOD under standard Reinforcement Learning (RL) training. Test-time training (TTT), which adapts model weights to incoming test data instead of relying on a fixed training distribution, originated in computer vision (Sun et al., 2020; Wang et al., 2021, 2022) and has recently been extended to LLMs for accuracy improvement (Hardt and Sun, 2024; Hu et al., 2025; Sun et al., 2025; Zweiger et al., 2025), but has not yet been applied to calibration due to two obstacles. First, TTT requires a self-supervision signal at test time; existing calibration signals (sampling-based consistency or discriminative probes) are expensive to compute on every question, making continuous adaptation costly. Second, naive weight updates risk catastrophic forgetting or overfitting to noisy targets, especially when the supervision signal itself is imperfect.

We show that both obstacles can be overcome. SECL (SElf-Calibrating Language Models) uses the generation–discrimination gap as a naturally available scalar target: a normalized $P(\text{True})$ signal computed from the frozen base model serves as self-supervision to adjust verbalized confidence via lightweight Low-Rank Adaptation (LoRA; Hu et al., 2022) updates, requiring no correctness labels during adaptation (Figure 1). Entropy-based gating triggers adaptation only on distribution shifts, and a conservative directional loss with bounded updates mitigates catastrophic forgetting. Our contributions are as follows:

- We introduce the first **TTT method for calibration**, using the generation–discrimination gap as self-supervision that requires no correctness labels at adaptation time. Entropy-

based gating limits adaptation to distribution shifts, so SECL trains only on **6–26% of the data** at a lower cost than the signal it distills.

- The adapted model **surpasses** its self-supervision signal and **matches supervised calibration without labels**, showing that SECL generalizes beyond the training signal.
- We provide **seven ablations** that isolate each design choice, showing that signal quality sets the calibration ceiling and that **each component is necessary.** SECL is robust across four architectures, forward and reversed domain orderings, and all tested hyperparameters.

2 Related Work

Sampling-Based Uncertainty Estimation. Self-CheckGPT (Manakul et al., 2023) detects hallucinations by comparing each sentence (S) against N=20 sampled passages at the cost of $\mathcal{O}(S \times N)$. Semantic Entropy (Kuhn et al., 2023) reduces surface sensitivity by clustering semantically equivalent responses before computing entropy, but still requires multiple generations and cannot resolve consistent falsehoods. Ma et al. (2025) improve detection in single-cluster failure cases by working on penultimate-layer logits. Xiong et al. (2024) benchmark these and other black-box elicitation methods comprehensively, finding that systematic overconfidence is inherent to all strategies; Heo et al. (2025) confirm this for instruction-tuned models specifically. All sampling methods share two limitations: high cost at inference time, and failure on consistent hallucinations where the model is confidently wrong across all sampled responses (Lin et al., 2022b).

Static Probing and Lightweight Calibration. White-box methods bypass sampling by analyzing the model's internals directly. HaloScope (Du et al., 2024) shows that hallucinations are geometrically distinct in intermediate-to-late layer embeddings, achieving strong AUROC with reasonable cross-dataset transfer, though the authors note degradation under drastic distribution shift. LitCab (Liu et al., 2024) adds a single linear layer ($<2\%$ of parameters) that predicts a logit bias, reducing ECE by up to 30%. Although these methods are efficient, they are static: because they are trained offline, they cannot adapt when the input distribution shifts at test time.

Training-Based Calibration. Lin et al. (2022a) showed that supervised fine-tuning with calibrated confidence labels produces well-calibrated models, though evaluation did not extend beyond math tasks. Stangel et al. (2025) use Reinforcement Learning (RL) with a logarithmic scoring rule to penalize overconfidence. TruthRL (Wei et al., 2025) uses a ternary reward to incentivize abstention over hallucination when models are uncertain. Damani et al. (2025) use the Brier score (Brier, 1950) as an RL reward, reducing calibration error by up to 90% in-domain, but found that standard RL degrades calibration OOD, directly motivating a test-time approach that can adapt to unseen domains. Prompting methods such as Fact-and-Reflection (Zhao et al., 2024) reduce ECE without training but remain static. A fundamental limitation of training-based approaches is the difficulty of specifying knowledge boundaries in black-box LLMs: RLHF methods rely on human labels that introduce sycophancy (Sharma et al., 2024), and RLHF fine-tuning itself degrades calibration (Tian et al., 2023) despite models retaining well-calibrated internal judgments (Kadavath et al., 2022).

Calibration and Sycophancy. Verbalized miscalibration and sycophancy share a common root: both are amplified by preference optimization, which rewards outputs human raters approve of (Sharma et al., 2024; Tian et al., 2023). They distort different things: sycophancy aligns answer *content* with perceived user preferences, while miscalibration aligns *confidence* with the assertive delivery rewarded during training. The gap SECL exploits follows from this asymmetry, since the discriminative signal is less shaped by preference pressure than generated text, so SECL corrects the rewarded generative confidence using the unrewarded discriminative judgment without altering answer content. Whether mitigating one failure mode benefits the other remains open.

The Generation-Discrimination Gap. Kadavath et al. (2022) established that LLMs' discriminative judgments $P(\text{True})$ are well-calibrated and improve with scale. Tian et al. (2023) extended this to RLHF-tuned models, showing that verbalized confidence is typically better calibrated than conditional token probabilities, even though RLHF degrades the latter. Building on this line of work, Kalai et al. (2025) provided a theoretical basis: generative error is lower-bounded by approximately twice the misclassification rate of the correspond-

ing binary validity problem. Wang and Stengel-Eskin (2025) exploit this gap directly: their DINCO method normalizes verbalized confidence across natural language inference (NLI)-reweighted distractors and integrates self-consistency, outperforming prior baselines. However, DINCO is a static inference-time technique that cannot adapt when the underlying truthfulness signal is brittle under distribution shift (Haller et al., 2025). SECL addresses this by distilling the discriminative signal into the model's weights, so the calibrated state persists and can be re-adapted when the distribution shifts. We provide a direct empirical comparison in Section 4.2 and Appendix V.

Test-Time Adaptation for LLMs. Test-time training (TTT) adapts model weights using unsupervised signals from incoming test data, originating in computer vision (Sun et al., 2020; Wang et al., 2021, 2022) and recently extended to LLMs for accuracy improvement (Hardt and Sun, 2024; Hu et al., 2025; Sun et al., 2025; Zweiger et al., 2025). Snell et al. (2025) showed that adaptive test-time compute allocation can outperform much larger models, and Huang et al. (2025) use calibrated confidence to allocate such compute, but their goal is accuracy, not calibration. SECL is the first method to apply test-time training to improve *calibration* across domains, leveraging a naturally available supervision signal without any task-specific design.

3 Methods

The core idea is simple: when the model encounters a new type of question, it checks whether its stated confidence matches its own self-assessment. If the model claims 90% confidence but its True/False self-check suggests only 30%, a small weight update corrects this mismatch. Over time, these corrections accumulate, producing better-calibrated confidence.

Concretely, SECL operates in three stages (Figure 2): (1) an entropy-based change detector triggers adaptation only on distribution shifts (Section 3.1); (2) a normalized discriminative signal, $\text{NormP}_{\text{True}}$, scores model answers (Section 3.2); (3) when this signal disagrees with verbalized confidence, lightweight LoRA updates reduce the gap (Section 3.3). Each component is validated individually (Section 4.3) and is described below.

3.1 Adaptive Entropy Gating

The calibration procedure (Sections 3.2–3.3) requires computing $\text{NormP}_{\text{True}}$ and running LoRA updates for each question. Once the model has adapted to a domain, these updates are redundant; the current LoRA weights already reflect the calibration characteristics of the current distribution. Therefore, calibration is triggered only when the input distribution shifts.

We track the entropy H_t of the model's output token distribution with an exponential moving average (EMA, smoothing factor α_{ema}) and apply the Page-Hinkley (PH) change detection test (Page, 1954). The PH test maintains a cumulative sum:

$$m_t = \sum_{s=1}^{t} \big(H_s - \bar{H}_t - \epsilon\big), \tag{1}$$

where $\bar{H}_t$ is the running mean entropy, H_s the entropy at step s, and ϵ is a tolerance that suppresses false alarms from minor fluctuations. An alarm fires when $m_t - \min_{s \le t} m_s > \lambda$; a higher λ requires a larger cumulative entropy deviation before triggering, controlling the trade-off between responsiveness and false alarm rate (sensitivity analysis in Section 4.3). Upon detection, cumulative statistics are reset, and a *calibration burst* is initiated: a window of B consecutive questions during which each question is scored by the discriminative signal and used for a potential weight update (Sections 3.2–3.3). Processing multiple questions per burst is crucial since a single LoRA update provides too little signal for stable adaptation, whereas a burst amortizes the cost of entering calibration mode and allows corrections to accumulate across diverse questions from the new distribution.

LoRA weights accumulate across domains without resetting. When a new distribution shift is detected, the subsequent calibration burst adapts the existing accumulated weights rather than starting from scratch, allowing calibration knowledge from earlier domains to compound. Datasets and domain ordering are described in Section 4.1, with ordering sensitivity analyzed in Section 4.3.

3.2 Normalized $P(\text{True})$ as Self-Supervision

Three related quantities appear in this paper and must be kept distinct. *Raw* $P(\textit{True})$ is the token probability of "True" when the model judges a candidate answer; it is better calibrated than verbalized confidence (Kadavath et al., 2022) but suffers from suggestibility bias, and we use it only

diagnostically. *NormP$_{True}$* (Eq. 3) is its distractor-normalized form, which removes this systematic bias and serves as SECL's supervision signal, though it remains noisy on individual questions. The *SECL output* is the model's verbalized confidence after adaptation, which is what a deployed system reports. Throughout, "well-calibrated" is meant operationally: lower ECE on the evaluation stream. The gap between generation and discrimination is structural: Kalai et al. (2025) show that generative error is lower-bounded by roughly twice the misclassification rate of the corresponding binary validity problem, so a model can often recognize wrong answers it cannot avoid producing, and the gap is expected to grow with scale.

Given a question q and the model's generated answer a, we compute $P_{\text{True}}(a \mid q)$: the token probability of "True" when the model is asked "Is the following answer to the question correct? (True/-False)". Following Kadavath et al. (2022), this discriminative signal is better calibrated than the model's own verbalized confidence c and improves with scale; its normalized form (Eq. 3) is what SECL uses as the supervision target.

Following Lin et al. (2022a) and Tian et al. (2023), we elicit verbalized confidence by prompting the model to state a confidence bin (0–9) alongside its answer, where each bin corresponds to a 10-percentage-point interval (e.g., bin 7 $\approx$ 70–80%). We use 10 bins to ensure confidence can be expressed as a single generated token (Naeini et al., 2015; Guo et al., 2017). Rather than taking the argmax bin (hard readout), we compute a *soft* confidence as the expected value over the digit-token probability distribution:

$$c = \sum_{k=0}^{9} P(\text{bin}_k) \cdot \frac{k + 0.5}{10}, \tag{2}$$

where $P(\text{bin}_k)$ is the model's output probability for digit token k. This soft readout preserves information from the full distribution and provides a differentiable signal for the mean squared error (MSE) loss in Section 3.3.

Raw P_{True} suffers from suggestibility bias, the model tends to affirm any answer presented to it, inflating P_{True} regardless of correctness (Wang and Stengel-Eskin, 2025), and degrades under distribution shift (Haller et al., 2025). We address both by normalizing P_{True} across distractor answers. For multiple-choice questions, we use the given answer options as distractors. For open-ended questions, we generate K=4 plausible alternatives. The normalized signal is:

$$\text{NormP}_{\text{True}}(a) = \frac{e_a}{e_a + \sum_{k=1}^{K} e_{d_k}}, \tag{3}$$

where $e_x = \exp\big(P_{\text{True}}(x)/\tau\big)$, with τ as a model-specific temperature (see Section 4.1); the choice of τ is discussed in Section 4.3. This softmax over distractors converts the raw signal into a *relative* confidence that accounts for baseline suggestibility.

Unlike DINCO, we use a simple softmax over distractors without NLI reweighting or self-consistency. The resulting NormP$_{\text{True}}$ provides a continuous training target, which is discretized into the same 10 equal-width bins used for verbalized confidence (Eq. 2), ensuring a common scale between supervision target and model output.

3.3 Test-Time Calibration via LoRA

When the model's verbalized confidence disagrees with NormP$_{\text{True}}$, we update the model to reduce this disagreement. Updates are applied via LoRA (Hu et al., 2022) on intermediate-to-late transformer layers, motivated by the finding that calibration-relevant representations concentrate in these layers (Du et al., 2024). Architecture-specific layer configurations are reported in Appendix K.

Training target. Although distractor normalization suppresses systematic biases (Section 3.2), the discriminative signal can still be noisy on individual questions. We therefore do not use NormP$_{\text{True}}$ directly as the training target. Instead, rather than jumping directly to the discriminative estimate, we nudge confidence toward it in small, bounded steps.

Let c_i be the model's verbalized confidence and c_i^* the NormP$_{\text{True}}$ value for question i. The training target is:

$$\hat{c}_i = c_i + \alpha_{\text{step}} \cdot \text{clip}\big(c_i^* - c_i,\ -\delta,\ \delta\big), \tag{4}$$

where $\text{clip}(x, -\delta, \delta) = \max(-\delta, \min(x, \delta))$ clamps the correction magnitude to the interval $[-\delta, \delta]$, α_{step} controls the correction step size, and δ caps the maximum single-step adjustment. We fix α_{step} and δ across all models (values in Section 4.1).

Why trust the direction but not the magnitude. Eq. 4 resolves an apparent tension: NormP$_{\text{True}}$ is both the supervision signal and too noisy to regress on directly. Its unreliability lives at two scales. Distractor normalization removes the systematic,

population-level bias of raw $P(\text{True})$, so averaged over many questions the signal points toward the model's better-calibrated discriminative judgment; per-question variance remains substantial, since a single discriminative pass over a handful of candidates is a high-variance estimate of correctness. The clipped directional target therefore trusts the direction, where the systematic component dominates, while capping any single update at $\alpha_{\text{step}}\delta$, so no individual noisy estimate moves the accumulating weights far, and zero-mean noise cancels across the burst. Both halves are validated empirically: plain MSE toward $\text{NormP}_{\text{True}}$ degrades ECE from 0.052 to 0.085, and resetting weights per question degrades it to 0.237 (Section 4.3).

Loss and optimization. The training loss is mean squared error between verbalized confidence and the directional target:

$$\mathcal{L}_i = \left(c_i - \hat{c}_i\right)^2. \tag{5}$$

We optimize with AdamW (Loshchilov and Hutter, 2019); learning rate and epoch count are reported in Section 4.1. Crucially, $\text{NormP}_{\text{True}}$ is computed from the *base model without LoRA adapters*, ensuring the supervision signal is not corrupted by ongoing adaptation.

Bin-gate filter. Not every question requires calibration. We skip training when the model is already approximately calibrated: specifically, when $|\text{bin}(c_i) - \text{bin}(c_i^*)| \leq 1$. This avoids gradient updates on questions where verbalized and discriminative confidence already agree, reducing computation and limiting noise from marginal disagreements. We use threshold 1 as default; the effect on training fraction and ECE is shown in Section 4.3.

Deployment model. Although evaluated in a streaming protocol, SECL's intended use is closer to pre-deployment adaptation than continuous live training. Adaptation runs over a stream representative of the target distribution, after which the LoRA weights can be frozen and shipped as a static, better-calibrated checkpoint, analogous to a fine-tuning stage. Alternatively, gating can stay active so a burst re-enters only on a detected shift. Neither mode requires live correctness feedback, and freezing yields a fixed artifact that can be versioned and audited.

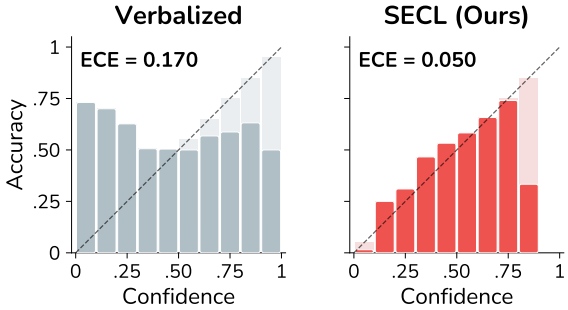

Figure 3: Reliability diagrams for Llama 3.2-3B. **Left:** Verbalized baseline (ECE = 0.170). **Right:** After SECL (ECE = 0.050, 71 % reduction), surpassing its own training signal (ECE = 0.065; Appendix L). The empty top bin shows that SECL eliminates the baseline's false near-100% certainty.

4 Experiments

4.1 Setup

Datasets. We evaluate on four datasets: **GSM8K** (Cobbe et al., 2021), grade-school math word problems (train split, first 500); **MMLU** (Hendrycks et al., 2021), multiple-choice knowledge questions across 57 subjects (test split, 500 sampled round-robin across subjects); **ARC Challenge** (Clark et al., 2018), grade-school science questions (test split, first 500); and **TruthfulQA** (Lin et al., 2022b), adversarially constructed misconception questions (validation split, MC1 variant, first 500). These four domains differ in reasoning type, difficulty, and degree of model overconfidence.

Continual protocol. The four domains are presented sequentially (GSM8K → MMLU → ARC → TruthfulQA), forming a single stream of 2,000 questions with distribution shifts between domains. Per-domain and ordering results are in Appendix I and Appendix M; an additional open-ended variant using TruthfulQA generation answers is reported in Appendix T.

Models. We evaluate four instruction-tuned small language models spanning different architectures: Llama 3.2-3B and Llama 3.1-8B (Grattafiori et al., 2024), Gemma 2-2B (Team et al., 2024), and Phi 3.5-Mini (3.8B) (Abdin et al., 2024). These were selected because all four exhibit a measurable generation–discrimination gap: their $\text{NormP}_{\text{True}}$ signal is better calibrated than their verbalized confidence, which is the prerequisite for SECL.

Baselines. We compare against two baselines representing the cost–quality extremes for label-free

calibration. *Verbalized*: the model's verbalized confidence c (soft readout, Eq. 2), used directly with no adaptation, representing a zero-cost baseline. *P(True) Norm*: the distractor-normalized discriminative signal (Eq. 3) reported directly as the confidence estimate. This serves as an upper-bound reference for signal quality; it requires five discriminative forward passes per question, does not modify model weights, and represents the quality of the supervision signal that SECL distills.

Metrics. We report Expected Calibration Error (ECE; Naeini et al. 2015; Guo et al. 2017) as our primary calibration metric, Brier score (Brier, 1950) as a composite measure of calibration and discrimination, AUROC for discrimination quality, and task accuracy to verify that calibration updates do not degrade correctness. Formal definitions are provided in Appendix B.

Implementation details. All values below are the best configuration determined through the ablation studies in Section 4.3, Appendix H, and Appendix K. LoRA (rank r=8, α=16) is applied to the query and value projection matrices of the last 4–8 transformer layers, targeting intermediate-to-late layers where calibration-relevant representations concentrate (Du et al., 2024). The exact layer range is model-dependent (Appendix K). This modifies 328K–786K parameters ($\sim$0.01–0.02% of total) per model. We optimize with AdamW (learning rate 5×10^{-5}) for 3 epochs per question. The directional target uses $\alpha_{\text{step}} = 0.5$ and $\delta = 0.15$; both are robust across tested ranges (Appendix H). The entropy gate uses $\alpha_{\text{ema}} = 0.05$, $\epsilon = 0.05$, detection threshold $\lambda = 3.0$ (sensitivity analysis in Section 4.3), burst length $B = 50$, and a warmup of 30 questions before first detection. The normalization temperature τ is fixed per model family from preliminary exploratory runs of the $P(\text{True})$ baseline (Section 4.3) and then held constant across datasets: τ=0.7 for Llama 3.2-3B, τ=1.5 for Gemma and Phi, and τ=3.0 for Llama 3.1-8B. All experiments run on NVIDIA A100 (80 GB) and RTX A6000 (48 GB) GPUs. SECL uses no correctness labels during adaptation. Two hyperparameters are set with label access: τ, from a small preliminary grid per model family, and B, tuned once on Llama 3.2-3B and fixed everywhere; a one-off model-level probe, in contrast to the temperature-scaling baseline, which refits via 5-fold cross-validation on the evaluation stream itself, so every question serves in the fitting set for four of five folds (Appendix E). Appendix G gives selection, sensitivity, and defaults for all hyperparameters. Hyperparameters are in Appendix C. The main table reports fixed-seed runs; Llama multi-seed results appear in Appendix D.1.

4.2 Main Results

SECL improves calibration on every model tested. Table 1 reports overall metrics across the full 2,000-question stream. Compared with the verbalized baseline, ECE reductions range from 56% on Phi to 78% on Gemma. SECL also surpasses its own supervision signal, P(True) Norm, on all four models despite training on only 6–26% of the stream (Table 3), showing that the model internalizes the discriminative signal well enough to generalize beyond the questions it trains on.

Scaling and the calibration–discrimination trade-off. On Mistral-Nemo 12B, the full 2,000-question protocol reduces ECE from 0.238 to 0.074 (69%), with Brier improving from 0.266 to 0.223 and accuracy preserved (66.3% vs. 66.0%), so SECL does not depend on small model scale. The AUROC change there ($0.658 \rightarrow 0.590$) reflects a general property: SECL's objective is calibration, not discrimination, and updates targeting the confidence token can improve bin-level calibration while reducing ranking quality. AUROC improves on Llama 3.2-3B (+0.077) and is preserved or slightly degraded elsewhere (Table 1); Brier, which combines both, improves on all models (see Limitations).

Largest gains appear where miscalibration is most severe. MMLU and TruthfulQA show the largest drop; GSM8K and ARC show smaller gains. Per-domain breakdowns appear in Appendix I.

Comparison with supervised post-hoc calibration. Supervised baselines (temperature and Platt scaling (Platt, 1999), fitted with ground-truth labels via 5-fold cross-validation) achieve lower ECE but at the cost of severe confidence compression, collapsing predictions into narrow ranges (e.g., $[0.44, 0.56]$ for Llama). SECL preserves a wide confidence range without any labels (Appendix E).

Accuracy is preserved. Accuracy differs by $<$ 1 percentage point overall (Table 1); per-domain shifts are at most 3 percentage points. Unlike RL-based methods that can harm task performance (Damani et al., 2025; Stangel et al., 2025), SECL

		Llama 3.2-3B			Llama 3.1-8B			Gemma 2-2B			Phi 3.5-Mini		
Method	**Cost**↓	ECE↓	Brier↓	AUROC↑	ECE↓	Brier↓	AUROC↑	ECE↓	Brier↓	AUROC↑	ECE↓	Brier↓	AUROC↑
Verbalized	**1**	.170	.292	.510	.225	.258	.684	.256	.314	.558	.251	.275	.600
P(True) Norm	6	.065	.223	.694	.120	.211	.718	.141	.259	**.650**	.154	.227	.675
DINCO†	∼10	.101	**.207**	**.762**	.117	**.210**	**.756**	.408	.410	.566	**.110**	**.212**	**.749**
SECL (Ours)	1.8–4.6	**.050**	.241	.587	**.083**	.222	.643	**.056**	**.254**	.548	**.110**	.251	.521

Table 1: Overall results across the full 2,000-question stream for the four fully evaluated models; Mistral-Nemo 12B (Verbalized vs. SECL) is reported in Section 4.2. Best values per column in **bold** (lowest for ECE/Brier/Cost, highest for AUROC). Cost is amortized forward-pass equivalents per question; SECL's range reflects model-dependent gating rates (Appendix N). Accuracy is preserved within 1 percentage point for all models (Appendix D). †Wang and Stengel-Eskin (2025).

updates only the confidence representation, not the model's task behavior.

Computational cost. SECL is cheaper than P(True) Norm and 2–5× cheaper than DINCO. Entropy gating is key: models that trigger infrequently (Gemma: 5.9%, Phi: 8.0%) reduce cost to less than a third of the P(True) Norm baseline (Appendix N).

Comparison with DINCO. SECL also outperforms DINCO (Wang and Stengel-Eskin, 2025), the closest inference-time calibration method, which combines NLI-weighted distractor normalization with self-consistency sampling at ∼10 forward passes per question. SECL achieves lower or equal ECE on all four models at 2–5× lower amortized cost. On Gemma, DINCO fails entirely: its ECE is worse than the unadapted baseline, and accuracy drops substantially, confirming that NLI-based reweighting is not robust across architectures. DINCO achieves stronger discrimination on Llama and Phi through multiple sampling passes, though its default beam-search answer selection yields lower task accuracy than greedy decoding. Figure 1 visualizes the cost–calibration trade-off; details in Appendix V.

Selective calibration. We additionally report Selective ECE, computed on predictions above a threshold α, at the highest α with sufficient coverage across methods ($\alpha \leq 0.7$ for Llama 3.2-3B, $\alpha \leq 0.6$ otherwise); SECL achieves the lowest selective calibration error on all four models (Appendix U). Marginal accuracies are 47–69% on GSM8K, MMLU, and TruthfulQA and 73–83% on ARC. Above $\alpha = 0.8$, SECL retains few predictions on three models: the bounded updates deliberately limit pushes into the very-high range, so predictions concentrate where accuracy lies, while output still spans a wide range (e.g., $[.05, .95]$) rather than compressing as supervised scaling does (Appendix E). Applications needing aggressive selective prediction at $\alpha \geq 0.8$ should prefer the supervised upper bound.

4.3 Ablation Studies

We validate every design decision via seven targeted ablations, each isolating one component while keeping all others fixed. Results are on Llama 3.2-3B unless otherwise noted; domain ordering ablations cover all four models. Together, these ablations establish three key insights: (1) the adaptation mechanism is cheap, requiring only 25.6% of the stream; (2) signal quality determines the ceiling of calibration improvement; and (3) each component is individually necessary.

Entropy gating matches always-on adaptation. Table 3 compares four gating strategies on the full 2,000-question stream.

Entropy gating determines *when* to enter a calibration burst; the bin-gate filters *which* questions within the burst receive updates. The combined entropy gate + bin-gate processes only 25.6% of the stream while matching always-on ECE (0.050 vs. 0.047): a 4× reduction in calibration compute with no measurable loss in calibration quality. This makes SECL cheaper per question than the raw P(True) Norm baseline it learns from.

Directional loss improves over plain MSE. Replacing the directional target with plain MSE toward $\text{NormP}_{\text{True}}$ degrades ECE from 0.052 to 0.085 (Appendix H). Conservative clipping prevents overshooting noisy targets, confirming that the update direction matters more than its magnitude.

Model	Verbalized	SECL $\mathcal{O}$	SECL $\mathcal{R}$
Llama 3.2-3B	.170	**.050** (−71%)	.068 (−60%)
Llama 3.1-8B	.225	**.083** (−63%)	.189 (−16%)
Gemma 2-2B	.256	**.056** (−78%)	.149 (−42%)
Phi 3.5-Mini	.251	**.110** (−56%)	.173 (−31%)

Table 2: Domain order robustness (ECE). $\mathcal{O}$: GSM8K→MMLU→ARC→TruthfulQA (default). $\mathcal{R}$: reversed. SECL improves over the verbalized baseline under both orderings for all models; per-domain breakdowns in Appendix M.

Calibration knowledge must accumulate across questions. Resetting LoRA weights after each question yields ECE 0.237 and degrades AUROC to 0.484, *below chance*, indicating that isolated single-question updates destroy the model's confidence signal entirely (Table 15). Weight accumulation achieves ECE 0.050: a 4.7× improvement, demonstrating that calibration knowledge compounds across questions and that the continual learning design is essential, not optional.

Distractor normalization is essential for signal quality. Raw $P(\text{True})$ already improves over the verbalized baseline (ECE: 0.161 vs. 0.170), confirming the generation–discrimination gap. Distractor normalization reduces ECE further to 0.065, a 60% improvement, by converting absolute affirmation into relative preference among candidates and suppressing the suggestibility bias identified by Wang and Stengel-Eskin (2025) (Appendix F).

Signal quality determines the calibration ceiling. This is the most revealing ablation. Replacing $\text{NormP}_{\text{True}}$ with Self-Consistency (Wang et al., 2023) (N=10) as SECL's training target, while keeping LoRA adaptation, entropy gating, and all hyperparameters identical, degrades ECE from 0.050 to 0.432, which is 2.5× *worse* than the untrained baseline (0.170). The adaptation mechanism faithfully distills whatever signal it receives: Self-Consistency's systematic overconfidence propagates directly into verbalized confidence, while $\text{NormP}_{\text{True}}$'s tighter correspondence to correctness probability yields well-calibrated outputs (Appendix R).

Calibration is robust to domain ordering. Table 2 confirms that SECL improves over the baseline under both forward and reversed orderings on all four models, with ECE reductions of 16–71%.

Strategy	ECE↓	Brier↓	AUROC↑	Trained
Always-on MSE	0.047	0.240	0.593	100%
Bin-gate (≤ 1)	0.052	0.242	0.585	55.8%
Bin-gate (≤ 2)	0.044	0.242	0.578	32.6%
Entropy-gated ($B=50$)	0.050	0.241	0.587	25.6%

Table 3: Gating strategy ablation (Llama 3.2-3B). Entropy-gated bursts achieve comparable ECE to always-on training while requiring only 25.6% of questions.

Hyperparameter sensitivity. Appendix H reports sensitivity to each hyperparameter on Llama 3.2-3B. SECL is robust to most hyperparameters: α_{step} and δ have minimal impact on ECE (0.064–0.067 across tested values). The single consequential choice is burst length B: $B = 20$ yields ECE 0.114 versus 0.050 for $B = 50$, confirming that sufficient distillation per trigger is critical. This robustness simplifies deployment: only one hyperparameter requires careful tuning.

5 Conclusion

We introduced SECL, a test-time training pipeline that exploits the generation–discrimination gap as self-supervision to improve calibration without correctness labels or human feedback during adaptation. Across five language models (2B–12B) from four model families and four diverse domains, SECL reduces ECE by 56–78% while preserving task accuracy, training on only 6–26% of the question stream via entropy-gated adaptation. Seven ablations show the method is robust to domain ordering and hyperparameter settings, and that its success rests on signal quality, weight accumulation, and entropy-gated adaptation.

Our results suggest a broader principle: when a model's ability to evaluate exceeds its ability to generate, the gap can be distilled back into its outputs. Calibration is a natural first target because the discriminative signal is scalar and cheap to compute, but the same approach applies wherever an analogous evaluation–generation gap exists (Kalai et al., 2025), and that gap is theoretically expected to grow with scale. Scaling SECL beyond 12B parameters, extending it toward conditionally calibrated estimates, and integrating richer discriminative signals are the direct next steps.

Limitations

Signal quality bounds improvement. SECL's calibration gains are bounded by the quality of the $\text{NormP}_{\text{True}}$ supervision signal. When the discriminative signal is only marginally better than verbalized confidence, distillation yields smaller improvements. On Gemma TruthfulQA, for example, SECL reduces ECE from 0.267 (Verbalized) to 0.210, but P(True) Norm achieves 0.104, indicating that the distillation captures only part of the available signal on adversarially constructed questions. Integrating richer discriminative signals beyond binary P(True), such as multi-step verification or ensemble-based judgments, could close this gap. SECL requires a measurable generation–discrimination gap, which we verified for all four evaluated model families. For models where this gap is absent, such as Qwen 2.5-3B (Qwen Team et al., 2025) where $\text{NormP}_{\text{True}}$ underperforms the verbalized baseline at all tested temperatures (Appendix O), SECL correctly produces no improvement, since no useful signal exists to distill. Characterizing when and why this gap closes is an important direction for future work.

Per-domain calibration is not uniformly improved. ARC ECE increases slightly in the forward ordering across all three smaller models (e.g., Llama: 0.095 $\rightarrow$ 0.112), but this effect reverses under alternative orderings (Appendix M), indicating a domain-sequencing artifact from cumulative LoRA weight transfer rather than a systematic limitation. Aggregate ECE remains improved under every ordering tested.

Calibration–discrimination trade-off. SECL improves AUROC on Llama 3.2-3B (+0.077) but degrades it on Phi (-0.079) and on Llama 3.1-8B (-0.041; Appendix Q). Since the LoRA updates target the confidence token, they can redistribute probability mass in ways that improve bin-level calibration at the cost of per-question ranking quality. Brier score, which captures both components, improves across all models, but applications requiring fine-grained discrimination should weigh this trade-off.

Hyperparameter sensitivity. The burst size B is the most important hyperparameter: B=20 yields ECE 0.114 versus 0.050 for B=50 on Llama (Section 4.3), indicating that sufficient distillation per trigger is necessary. In deployment settings with very rapid distribution shifts (fewer than 50 questions per domain), the method may not accumulate enough training signal.

Scale. We evaluate models from 2B to 12B parameters; the Mistral-Nemo 12B result (Section 4.2) shows relative ECE reductions consistent with the smaller models. Beyond this range, effectiveness is unverified: the generation–discrimination gap is theoretically expected to widen with scale (Kalai et al., 2025), which would favor SECL, but heavily RLHF-tuned frontier models have degraded token-level probability estimates (Tian et al., 2023), and we have not established that a usable discriminative signal survives at that scale. The direction of the calibration–discrimination trade-off also does not move monotonically with size in our results, so it cannot be extrapolated.

Adversarial inputs. We evaluate under benign distribution shift only. Three attack surfaces follow from the design: the entropy gate is observable, so an adversary shaping the input stream could force frequent bursts or suppress adaptation on a genuinely shifted distribution; burst content is unfiltered, so an attacker controlling the B questions after a detected shift controls the supervision signal, and because weights accumulate, such an update persists into later domains; and $\text{NormP}_{\text{True}}$ depends on generated distractors for open-ended questions, which crafted inputs could degrade. TruthfulQA targets misconceptions but does not model an adversary with stream access. Auditing burst contents and bounding the total accumulated update are natural mitigations, which we leave to future work.

Marginal calibration only. SECL is a marginal calibration method: it targets aggregate reliability over the stream, not calibration conditional on subpopulations or inputs, which is the more demanding standard the calibration literature is moving toward. Relatedly, SECL provides no finite-sample coverage guarantees of the kind studied in the online calibration literature; connecting entropy-gated adaptation to sample-size-aware bounds is a direct next step.

Ethics and Societal Impact

SECL reduces ECE by 56–78% without human labels or held-out data. This directly lowers the barrier to calibrated deployment: only 1.2% of 519 healthcare LLM studies measured calibration (Bedi et al., 2025), partly because existing methods

require labeled validation sets that are expensive to curate. By operating entirely at test time with self-generated supervision, SECL makes calibration feasible in settings where annotation budgets are limited or domain expertise is scarce.

Our target ablation (Section 4.3) carries an important ethical lesson: substituting Self-Consistency as the supervision signal makes calibration 2.5× worse than the unadapted baseline. This demonstrates that self-supervised calibration methods can *amplify* overconfidence when the supervision signal is poorly chosen. Practitioners deploying any self-supervised calibration approach, including SECL, should verify that the underlying signal genuinely correlates with correctness before trusting the resulting confidence scores.

Two further risks merit attention. First, calibration does not imply correctness: a well-calibrated model reporting 60% confidence is still wrong 40% of the time. Our AUROC results (Section 4.2) show that SECL can degrade per-question ranking quality on some models while improving bin-level calibration, meaning users should not assume that higher confidence always identifies more reliable answers. Second, when gating is left active in deployment, SECL's confidence characteristics evolve with the input stream. This complicates auditing and reproducibility compared to static calibration methods; logging the gating decisions and LoRA checkpoints would be necessary for post-hoc review in regulated domains, and freezing the adapter after an initial adaptation pass (Section 3.3) avoids the issue by producing a fixed, auditable checkpoint.

All experiments use public benchmarks (GSM8K, MMLU, ARC-Challenge, TruthfulQA) with standard splits. No human data was collected.

Acknowledgement

This work is supported by the Genial4KMU project, Universität Hamburg, funded by BMBF (grant no. 16IS24044B).

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *Preprint*, arXiv:2404.14219.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova Dassarma, Dawn Drain, Stanislav Fort, Deep Ganguli, T. J. Henighan, Nicholas Joseph, Saurav Kadavath, John Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *Preprint*, arXiv:2204.05862.

Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A. Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, Hyo Jung Hong, Mehr Kashyap, Akash R. Chaurasia, Nirav R. Shah, Karandeep Singh, Troy Tazbaz, Arnold Milstein, Michael A. Pfeffer, and Nigam H. Shah. 2025. Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. *JAMA*, 333(4):319–328.

Glenn W. Brier. 1950. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*, 78(1):1–3.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think You Have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *Preprint*, arXiv:1803.05457.

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. 2021. Training Verifiers to Solve Math Word Problems. *Preprint*, arXiv:2110.14168.

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. 2025. Beyond Binary Rewards: Training LMs To Reason About Their Uncertainty. *Preprint*, arXiv:2507.16806.

Xuefeng Du, Chaowei Xiao, and Yixuan Li. 2024. HaloScope: Harnessing Unlabeled Llm Generations For Hallucination Detection. In *Advances In Neural Information Processing Systems*, volume 37, pages 102948–102972, Vancouver, Canada.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 1321–1330, Sydney, Australia.

Patrick Haller, Mark Ibrahim, Polina Kirichenko, Levent Sagun, and Samuel J. Bell. 2025. LLM Knowledge Is Brittle: Truthfulness Representations Rely On Superficial Resemblance. *Preprint*, arXiv:2510.11905.

Moritz Hardt and Yu Sun. 2024. Test-Time Training on Nearest Neighbors for Large Language Models. In *The Twelfth International Conference on Learning Representations, ICLR 2024*, Vienna, Austria.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations, ICLR 2021*, Vienna, Austria.

Juyeon Heo, Miao Xiong, Christina Heinze-Deml, and Jaya Narain. 2025. Do LLMs Estimate Uncertainty Well In Instruction-Following? In *The Thirteenth International Conference On Learning Representations, ICLR 2025*, Singapore, Singapore.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation Of Large Language Models. In *Proceedings Of The International Conference On Learning Representations, ICLR 2022*, Virtual Event.

Jinwu Hu, Zitian Zhang, Guohao Chen, Xutao Wen, Chao Shuai, Wei Luo, Bin Xiao, Yuanqing Li, and Mingkui Tan. 2025. Test-Time Learning for Large Language Models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 24823–24849, Vancouver, Canada. PMLR.

Chengsong Huang, Langlin Huang, Jixuan Leng, Jiacheng Liu, and Jiaxin Huang. 2025. Efficient Test-Time Scaling Via Self-Calibration. In *Proceedings of the NeurIPS 2025 Workshop on Efficient Reasoning (ER)*, San Diego, USA.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How Can We Know When Language Models Know? On The Calibration Of Language Models For Question Answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

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, and 17 others. 2022. Language Models (Mostly) Know What They Know. *Preprint*, arXiv:2207.05221.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. Why Language Models Hallucinate. *Preprint*, arXiv:2509.04664.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *Proceedings of the International Conference on Learning Representations, ICLR 2023*, Kigali, Rwanda.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022a. Teaching Models to Express Their Uncertainty in Words. *Transactions on Machine Learning Research*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022b. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Fenglin Liu, Hongjian Zhou, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S. Chen, Yining Hua, Peilin Zhou, Junling Liu, Chengfeng Mao, Chenyu You, Xian Wu, Yefeng Zheng, Lei Clifton, Zheng Li, Jiebo Luo, and David A. Clifton. 2025. Application Of Large Language Models In Medicine. *Nature Reviews Bioengineering*, 3(6):445–464.

Xin Liu, Muhammad Khalifa, and Lu Wang. 2024. LitCab: Lightweight Language Model Calibration Over Short- And Long-Form Responses. In *The Twelfth International Conference On Learning Representations, ICLR 2024*, pages 13033–13047, Vienna, Austria.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *Proceedings of the International Conference on Learning Representations, ICLR 2019*, New Orleans, USA.

Huan Ma, Jiadong Pan, Jing Liu, Yan Chen, Joey Tianyi Zhou, Guangyu Wang, Qinghua Hu, Hua Wu, Changqing Zhang, and Haifeng Wang. 2025. Semantic Energy: Detecting LLM Hallucination Beyond Entropy. *Preprint*, arXiv:2508.14496.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore, Singapore. Association for Computational Linguistics.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. 2015. Obtaining Well Calibrated Probabilities Using Bayesian Binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 2901–2907, Austin, Texas. AAAI Press.

E. S. Page. 1954. Continuous Inspection Schemes. *Biometrika*, 41(1-2):100–115.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

John C. Platt. 1999. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advances in Large Margin Classifiers*.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 Technical Report. *Preprint*, arXiv:2412.15115.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards Understanding Sycophancy in Language Models. In *The Twelfth International Conference on Learning Representations, ICLR 2024*, Vienna, Austria.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling LLM Test-Time Compute Optimally Can Be More Effective Than Scaling Model Parameters. In *Proceedings of the International Conference on Learning Representations, ICLR 2025*, Singapore, Singapore.

Paul Stangel, David Bani-Harouni, Chantal Pellegrini, Ege Özsoy, Kamilia Zaripova, Matthias Keicher, and Nassir Navab. 2025. Rewarding Doubt: A Reinforcement Learning Approach to Calibrated Confidence Expression of Large Language Models. *Preprint*, arXiv:2503.02623.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. 2025. Learning To (Learn At Test Time): RNNs With Expressive Hidden States. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 57503–57522, Vancouver, Canada. PMLR.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. 2020. Test-Time Training with Self-Supervision for Generalization Under Distribution Shifts. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9229–9248, Online. PMLR.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving Open Language Models At A Practical Size. *Preprint*, arXiv:2408.00118.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just Ask For Calibration: Strategies For Eliciting Calibrated Confidence Scores From Language Models Fine-Tuned With Human Feedback. In *Proceedings Of The 2023 Conference On Empirical Methods In Natural Language Processing*, pages 5433–5442, Singapore, Singapore. Association for Computational Linguistics.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. 2021. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *International Conference on Learning Representations, ICLR 2021*, Vienna, Austria.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. 2022. Continual Test-Time Domain Adaptation. In *2022 IEEE/CVF Conference On Computer Vision and Pattern Recognition (CVPR)*, pages 7191–7201, New Orleans, USA.

Victor Wang and Elias Stengel-Eskin. 2025. Calibrating Verbalized Confidence With Self-Generated Distractors. *Preprint*, arXiv:2509.25532.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *Proceedings of the International Conference on Learning Representations, ICLR 2023*, Kigali, Rwanda.

Zhepei Wei, Xiao Yang, Kai Sun, Jiaqi Wang, Rulin Shao, Sean Chen, Mohammad Kachuee, Teja Gollapudi, Tony Liao, Nicolas Scheffer, Rakesh Wanga, Anuj Kumar, Yu Meng, Wen-tau Yih, and Xin Luna Dong. 2025. TruthRL: Incentivizing Truthful Llms Via Reinforcement Learning. *Preprint*, arXiv:2509.25760.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can Llms

Express Their Uncertainty? An Empirical Evaluation Of Confidence Elicitation In LLMs. In *The Twelfth International Conference On Learning Representations, ICLR 2024*, Vienna, Austria.

Xinran Zhao, Hongming Zhang, Xiaoman Pan, Wenlin Yao, Dong Yu, Tongshuang Wu, and Jianshu Chen. 2024. Fact-And-Reflection (FaR) Improves Confidence Calibration Of Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8702–8718, Bangkok, Thailand. Association for Computational Linguistics.

Adam Zweiger, Jyo Pari, Han Guo, Yoon Kim, and Pulkit Agrawal. 2025. Self-Adapting Language Models. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, San Diego, USA.

A Experimental Setup Details

Datasets and splits. We evaluate on four domains presented sequentially: GSM8K (Cobbe et al., 2021) (`train` split), MMLU (Hendrycks et al., 2021) (`test` split), ARC-Challenge (Clark et al., 2018) (`test` split), and TruthfulQA (Lin et al., 2022b) (`validation` split, MC1 variant). We sample 500 questions per domain for a total of 2,000 questions per run. The default domain order is GSM8K → MMLU → ARC → TruthfulQA.

Models. We evaluate four instruction-tuned small language models spanning 2–8B parameters: Llama 3.2-3B-Instruct and Llama 3.1-8B-Instruct (Grattafiori et al., 2024), Gemma 2-2B-IT (Team et al., 2024), and Phi 3.5-Mini-Instruct (3.8B) (Abdin et al., 2024).

Trainable parameters. Table 4 reports the exact LoRA parameter counts per model. The variation reflects architectural differences (Phi uses a fused `qkv_proj` module rather than separate `q_proj`/`v_proj`).

Model	Total params	LoRA layers	LoRA params	% of total
Llama 3.2-3B	3.21B	4	327,680	0.010%
Llama 3.1-8B	8.03B	8	1,048,576	0.013%
Gemma 2-2B	2.61B	8	491,520	0.019%
Phi 3.5-Mini	3.82B	8	786,432	0.021%

Table 4: LoRA trainable parameters per model (rank r=8). Llama 3.2-3B: last 4 layers (24–27 of 28); Llama 3.1-8B: last 8 layers; Gemma and Phi: last 8 layers (Gemma: 18–25 of 26; Phi: 24–31 of 32).

Compute and software. All experiments were conducted on a shared workstation equipped with NVIDIA A100 (80 GB) and RTX A6000 (48 GB) GPUs. Each 2,000-question SECL run takes approximately 2–4 GPU-hours on a single GPU, depending on model size. We use PyTorch 2.9 (Paszke et al., 2019), HuggingFace Transformers 4.57 (Wolf et al., 2020), and the PEFT library (v0.17) with CUDA 12.8.

B Metric Definitions

Expected Calibration Error (ECE; Naeini et al. 2015; Guo et al. 2017; lower is better). Predictions are divided into M=10 equal-width bins $B_1, \ldots, B_M$ by confidence value:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|, \quad (6)$$

where N is the total number of predictions, $\mathrm{acc}(B_m)$ is the fraction of correct predictions in bin B_m, and $\mathrm{conf}(B_m)$ is the mean predicted confidence. A perfectly calibrated model achieves $\mathrm{ECE} = 0$.

Brier score (Brier, 1950) (lower is better). The mean squared error between predicted confidence and binary correctness:

$$\mathrm{BS} = \frac{1}{N} \sum_{i=1}^{N} (c_i - y_i)^2, \quad (7)$$

where $c_i \in [0, 1]$ is the predicted confidence and $y_i \in \{0, 1\}$ indicates correctness. The Brier score decomposes into calibration and refinement components.

AUROC (area under the ROC curve; higher is better). The area under the receiver operating characteristic curve, treating predicted confidence as a score for binary classification of correctness. AUROC measures discrimination, whether higher confidence corresponds to higher correctness, independent of calibration.

Adaptive ECE (AdaECE). Same as ECE but with equal-mass bins (lower is better); we report it to verify robustness to binning choice.

Accuracy. Fraction of correct answers; we report it to verify that calibration updates do not degrade task performance.

C Hyperparameter Settings

Component	Parameter	Value
LoRA	Rank r	8
	Scaling factor α	16
	Target layers (Llama 3.2-3B)	Last 4
	Target layers (Llama 3.1-8B, Gemma, Phi)	Last 8
	Target modules (Llama, Gemma)	`q_proj`, `v_proj`
	Target modules (Phi)	`qkv_proj`
Optimization	Optimizer	AdamW
	Learning rate	5×10^{-5}
	Epochs per question	3
Directional loss	Step size α_{step}	0.5
	Clip bound δ	0.15
Bin-gate	Threshold	1 bin
Page-Hinkley	Tolerance ϵ	0.05
	Detection threshold λ	3.0
	EMA smoothing α_{ema}	0.05
	Warmup period	30 questions
Burst	Burst size B	50
Normalization τ	Llama 3.2-3B	**0.7**
	Gemma 2-2B	**1.5**
	Phi 3.5-Mini (3.8B)	**1.5**
	Llama 3.1-8B	**3.0**
Other	Weight accumulation	On
	Distractors k	4

Table 5: Full hyperparameter settings used across all experiments. Model-specific values are noted where applicable.

D Main Results

Table 6 reports overall metrics for all methods and all four models. We report both standard ECE (10 equal-width bins) and Adaptive ECE (AdaECE; 10 equal-mass bins; see Section B) to verify that improvements are robust to the choice of binning strategy. Table 1 in the main paper summarizes the same four models with ECE/Brier/AUROC and cost; this table adds AdaECE and supervised baselines.

D.1 Llama Multi-Seed Robustness

To quantify seed sensitivity for the main SECL setting, we report Llama 3.2-3B results for seeds 42, 43, and 44 on the same 2,000-question sequential stream for both the verbalized baseline and SECL.

E Post-Hoc Calibration Baselines

We compare SECL against standard post-hoc calibration methods: **temperature scaling** (Guo et al., 2017) (single parameter T fitted to minimize NLL) and **Platt scaling** (Platt, 1999) (logistic regression $\sigma(a \cdot c + b)$). Both are applied to the soft verbalized confidence and fitted via 5-fold cross-validation (each fold uses 1,600 labeled examples for fitting).

Key difference from SECL. Temperature and Platt scaling are *supervised* post-hoc methods: they require ground-truth correctness labels to fit their parameters. SECL is *unsupervised* at test time; it uses only the model's own P(True) signal, with no access to ground truth. Post-hoc methods are also monotone transforms that *cannot* improve discrimination (AUROC); SECL's AUROC improves on Llama 3.2-3B (+0.077) and is preserved or slightly degraded on the other three models (Section 4.2).

Discussion. For Llama, temperature scaling on the soft verbalized confidence requires T=17.6, compressing all predictions to $[0.44, 0.56]$, essentially predicting the base rate. Platt scaling is more extreme: its output range $[0.56, 0.61]$ is near-constant, yielding ECE=0.021 but *degrading* AUROC to 0.481 (below chance). These methods achieve low ECE by destroying discriminative information. By contrast, SECL achieves ECE=0.050 while *improving* AUROC from 0.510 to 0.587 (+0.077), i.e., genuine calibration rather than confidence compression.

For Phi, where the verbalized confidence carries more signal (baseline AUROC=0.600), temperature scaling uses a more moderate T=3.4 and achieves good calibration (ECE=0.047). The strongest supervised baseline across all models is P(True) Norm + Temp, which combines the discriminative P(True) signal with post-hoc recalibration.

Complementarity: SECL + Temperature Scaling. SECL and post-hoc calibration are complementary: applying temperature scaling to SECL's output further reduces ECE, since SECL improves the underlying confidence signal while temperature scaling recalibrates it.

Calibration–discrimination trade-off. SECL's primary objective is calibration (ECE), not discrimination (AUROC). Table 6 reveals a model-dependent trade-off: for Llama, SECL *improves* AUROC from .510 to .587 (+.077) while reducing ECE from .170 to .050, i.e., the adapted confidence carries better signal. For Gemma, AUROC is roughly preserved (.558 → .548, −.010) despite a large ECE reduction (.256 → .056). For Phi, the trade-off is more pronounced: AUROC drops from .600 to .521 (−.079) as ECE improves from .251 to .110. Brier score, which jointly captures calibration and discrimination, improves for all four models (Llama 3B: .292 → .241; Llama 8B: .258 → .222; Gemma: .314 → .254; Phi: .275 → .251), so the net effect of SECL is positive. By comparison, supervised temperature scaling also reduces AUROC (Llama: −.006; Gemma: −.008; Phi: −.017) and Platt scaling degrades it further for Llama (−.029, below chance at .481), demonstrating that AUROC loss is not unique to SECL but is an inherent cost of recalibrating overconfident models.

F P(True) Normalization and the Generation–Discrimination Gap

A key premise of SECL is that LLMs possess a *discriminative* confidence signal, P(True), that is better calibrated than their *generative* verbalized confidence. Table 10 demonstrates this gap and the effect of normalization.

Raw P(True) vs. the verbalized baseline. For Llama, raw P(True) (the model's probability that its own answer is correct, without normalization) already improves calibration over the verbalized baseline (ECE: .161 vs. .170) and dramatically improves discrimination (AUROC: .673 vs. .510). This confirms that the model "knows more than it says"; its internal judgment of correctness is more reliable than its expressed confidence.

Model	Method	ECE↓	AdaECE↓	Brier↓	AUROC↑	Acc
Llama 3.2-3B	Verbalized	.170	.167	.292	.510	.576
	Self-Consistency†	.093	.096	**.211**	**.728**	.624
	+ Temp Scaling*	.047	.075	.250	.504	.576
	P(True) Norm	.065	.067	.223	.694	.576
	+ Temp Scaling*	**.029**	**.030**	.218	.692	.576
	SECL (Ours)	<u>.050</u>	<u>.060</u>	.241	.587	.577
Gemma 2-2B	Verbalized	.256	.252	.314	.558	.516
	+ Temp Scaling*	.047	.037	.249	.550	.516
	P(True) Norm	.141	.142	.259	**.650**	.516
	+ Temp Scaling*	**.037**	**.033**	**.233**	.647	.516
	SECL (Ours)	<u>.056</u>	<u>.060</u>	.254	.548	.515
Phi 3.5-Mini (3.8B)	Verbalized	.251	.240	.275	.600	.667
	+ Temp Scaling*	**.047**	**.049**	.215	.583	.667
	P(True) Norm	.154	.151	.227	**.675**	.667
	+ Temp Scaling*	.059	.060	**.205**	.664	.667
	SECL (Ours)	<u>.110</u>	<u>.119</u>	.251	.521	.665
Llama 3.1-8B	Verbalized	.225	.225	.258	.684	.644
	+ Temp Scaling*	**.080**	.089	.213	.681	.644
	P(True) Norm	.120	.117	.211	**.718**	.646
	+ Temp Scaling*	.108	.103	**.208**	.716	.646
	SECL (Ours)	<u>.083</u>	**.080**	.222	.643	.646

*Requires ground-truth labels; fitted via 5-fold cross-validation. See Section E.
†Self-Consistency requires N=10 sampling passes per question. We report it for Llama 3.2-3B only to contextualize SECL's single-pass efficiency; it is omitted for other models as it measures generation consistency rather than calibration directly.

Table 6: Overall results across all models and methods (2,000 questions each). **Bold**: best overall per model for ECE, AdaECE, Brier, and AUROC. <u>Underline</u>: best among label-free methods.

Metric	Seed 42	Seed 43	Seed 44	Mean±Std
Verbalized ECE	.1695	.1757	.1757	.1736±.0029
SECL ECE	.0496	.0502	.0392	**.0464±.0051**
Verbalized AUROC	.5095	.5119	.5119	.5111±.0011
SECL AUROC	.5870	.6005	.5920	**.5932±.0056**

Table 7: Llama 3.2-3B multi-seed robustness (seeds 42, 43, 44). SECL remains strongly better calibrated than the verbalized baseline across seeds and also improves AUROC on this model.

Normalization with distractors. P(True) Norm applies a temperature-scaled softmax over the correct answer and k=4 distractors, converting a raw probability into a calibrated confidence score. This yields a large further improvement (Llama ECE: .161 → .065 at τ=0.7), demonstrating that relative comparison against plausible alternatives sharpens the discriminative signal.

Temperature selection. The normalization temperature τ controls sharpening (τ<1) or smoothing (τ>1) of the softmax over candidates. Table 11 shows the P(True) Norm ECE at different τ values for each model. Llama 3.2-3B benefits from sharpening (τ=0.7), Gemma and Phi require smoothing (τ=1.5), and Llama 3.1-8B requires stronger smoothing (τ=3.0), likely reflecting differences in the models' internal confidence distributions. The selected τ is used in all subsequent experiments. Robustness to τ varies by family: on Gemma and Phi, every tested value yields P(True) Norm ECE below the verbalized baseline, whereas on the Llama models a poorly chosen τ can leave the signal no better than verbalized confidence (Table 11).

G Hyperparameter Guidance

Table 12 summarizes how each hyperparameter was selected, whether labeled data was used, and its sensitivity.

H Hyperparameter Sensitivity

Table 13 reports ECE and Adaptive ECE on Llama 3.2-3B when varying individual hyperparameters. The α_{step} and δ ablations use the full pipeline *without* entropy gating (all questions trained) to isolate the loss function's effect; the burst size B ablation uses the entropy-gated pipeline.

Model	Method	ECE↓	AdaECE↓	Brier↓	AUROC↑	Conf. range
Llama 3.2-3B	Verbalized + Temp (T=17.6)	.047	.075	.250	.504	[.44, .56]
	Verbalized + Platt	.021	.057	.244	.481	[.56, .61]
	TS on log P(True) (T=1.8)	.029	.030	.218	.692	[.14, .86]
	SECL (no labels)	.050	.060	.241	.587	[.05, .85]
Gemma 2-2B	Verbalized + Temp (T=11.2)	.047	.037	.249	.550	[.41, .59]
	Verbalized + Platt	.035	.050	.250	.550	[.42, .55]
	TS on log P(True) (T=3.0)	.037	.033	.233	.647	[.26, .74]
	SECL (no labels)	.056	.060	.254	.548	[.05, .95]
Phi 3.5-Mini (3.8B)	Verbalized + Temp (T=3.4)	.047	.049	.215	.583	[.29, .71]
	Verbalized + Platt	.052	.050	.216	.585	[.16, .69]
	TS on log P(True) (T=2.4)	.059	.060	.205	.664	[.21, .79]
	SECL (no labels)	.110	.119	.251	.521	[.05, .95]
Llama 3.1-8B	Verbalized + Temp (T=2.8)	.080	.089	.213	.681	[.35, .74]
	Verbalized + Platt	.063	.052	.206	.677	[.02, .75]
	TS on log P(True) (T=0.7)	.108	.103	.208	.716	[.02, .99]
	SECL (no labels)	.083	.080	.222	.643	[.25, .95]

Table 8: Post-hoc calibration baselines (5-fold CV). Temperature scaling compresses the confidence range substantially, achieving low ECE by predicting near-constant values for models with weak verbalized discrimination (fitted T shown).

Model	Method	ECE↓	AdaECE↓	Brier↓	AUROC↑
Llama 3.2-3B	P(True) Norm + Temp*	.029	.030	.218	.692
	SECL	.050	.060	.241	.587
	SECL + Temp*	.049	.053	.241	.584
Gemma 2-2B	P(True) Norm + Temp*	.037	.033	.233	.647
	SECL	.056	.060	.254	.548
	SECL + Temp*	.011	.037	.249	.542
Phi 3.5-Mini (3.8B)	P(True) Norm + Temp*	.059	.060	.205	.664
	SECL	.110	.119	.251	.521
	SECL + Temp*	.097	.082	.232	.516
Llama 3.1-8B	P(True) Norm + Temp*	.108	.103	.208	.716
	SECL	.083	.080	.222	.643
	SECL + Temp*	.062	.045	.213	.638

*Requires ground-truth labels for temperature fitting.

Table 9: SECL combined with temperature scaling (5-fold CV). SECL+Temp achieves the best or near-best ECE overall while preserving SECL's discrimination improvements.

Method	ECE↓	Brier↓	AUROC↑	Acc
Verbalized	.170	.292	.510	.576
Raw P(True)	.161	.248	.673	.576
P(True) Norm τ=0.3	.118	.238	.685	.576
P(True) Norm τ=0.7	**.065**	**.223**	**.694**	.576
P(True) Norm τ=1.5	.169	.254	.658	.579

Table 10: Effect of P(True) normalization (Llama 3.2-3B, N=2,000). Raw P(True) improves over the verbalized baseline. Normalization with temperature τ yields further gains. The optimal τ varies by model (see Table 11).

τ	Llama 3.2-3B	Gemma 2-2B	Phi 3.5-Mini	Llama 3.1-8B
0.3	.118	.255	.172	.270
0.7	**.065**	.215	.168	.234
1.0	.083	.180	.161	.195
1.5	.169	**.141**	**.154**	.153
3.0	.247	.143	.245	**.120**

Table 11: P(True) Norm ECE as a function of normalization temperature τ. Bold: selected value per model.

Param	Default	Labels?	Sensitivity	Guidance
B (burst)	50	once	high	keep ≥ 50
τ (temp.)	per family	once	moderate	small grid on P(True)
λ (PH)	3.0	no	low	default
α_{step}	0.5	no	negligible	default
δ (clip)	0.15	no	negligible	default
LoRA r, lr	8, 5×10^{-5}	no	low	default
layers	last 4–8	no	moderate	App. K

Table 12: Hyperparameter guidance. "Labels?" indicates whether selection used labeled data (τ: one small grid per model family; B: tuned once on Llama 3.2-3B, fixed everywhere). Sensitivity reflects ECE variation across tested ranges (Appendix H). On a new model, only τ and B warrant attention; all other values transfer.

Parameter	Value	ECE↓	AdaECE↓
$\alpha_{\text{step}}{}^{\dagger}$	0.2	.067	.083
	0.3	.066	.069
	0.5	**.052**	**.073**
$\delta^{\dagger}$	**0.15**	**.052**	**.073**
	0.20	.064	.067
Loss†	Plain MSE (no directional)	.085	.088
Burst $B^{\ddagger}$	20	.114	–
	50	**.050**	**.060**
τ (per model)	Llama: **0.7**	**.065**	**.064**
	Gemma: **1.5**	.141	.144
	Phi: **1.5**	.154	.151

Table 13: Hyperparameter sensitivity (Llama 3.2-3B). Bold indicates the selected configuration. †Without entropy gating. ‡With entropy gating.

In additional non-default sweeps, we observed further gains for some models (e.g., Llama in Table 18 and Phi under a tuned λ/learning-rate pair, best observed ECE 0.097). We keep the canonical defaults in all headline comparisons to preserve a single cross-model protocol and avoid per-model post-hoc tuning.

I Per-Domain Results

Model	Method	Domain	ECE↓	AdaECE↓	Brier↓	AUROC↑	Acc
Llama 3.2-3B	Verbalized	GSM8K	.218	.215	.294	.565	.472
		MMLU	.106	.109	.253	.601	.562
		ARC	**.095**	.112	.207	.568	.726
		TruthfulQA	.372	.371	.414	.301	.544
	P(True) Norm	GSM8K	.133	.138	.264	.594	.476
		MMLU	.091	.090	.239	.652	.566
		ARC	.117	.132	.210	.624	.728
		TruthfulQA	.089	**.092**	.180	.818	.532
	SECL	GSM8K	**.070**	**.079**	.249	.573	.488
		MMLU	**.067**	**.060**	.249	.549	.558
		ARC	.112	**.106**	.207	.616	.714
		TruthfulQA	**.068**	.114	.260	.454	.546
Gemma 2-2B	Verbalized	GSM8K	.549	.549	.446	.658	.186
		MMLU	.194	.178	.271	.608	.576
		ARC	**.082**	**.085**	.201	.534	.738
		TruthfulQA	.267	.263	.338	.417	.566
	P(True) Norm	GSM8K	.395	.394	.326	.685	.186
		MMLU	.163	.175	.263	.616	.576
		ARC	.185	.182	.229	.597	.738
		TruthfulQA	**.104**	**.104**	.213	.729	.566
	SECL	GSM8K	**.356**	**.356**	.271	.609	.180
		MMLU	**.054**	**.064**	.238	.594	.578
		ARC	.144	.136	.213	.591	.728
		TruthfulQA	.210	.198	.293	.343	.574
Phi 3.5-Mini	Verbalized	GSM8K	.290	.290	.304	.563	.650
		MMLU	.264	.257	.286	.602	.648
		ARC	**.109**	**.105**	.150	.588	.826
		TruthfulQA	.343	.336	.362	.592	.546
	P(True) Norm	GSM8K	**.261**	**.257**	.303	.559	.650
		MMLU	.171	.158	.243	.648	.648
		ARC	.129	.119	.157	.680	.826
		TruthfulQA	**.146**	**.141**	.205	.771	.546
	SECL	GSM8K	.283	.280	.297	.575	.658
		MMLU	**.054**	**.058**	.223	.604	.644
		ARC	.129	.129	.167	.499	.826
		TruthfulQA	.229	.224	.319	.407	.532

Table 14: Per-domain results for Llama 3.2-3B, Gemma 2-2B, and Phi 3.5-Mini. Best ECE and AdaECE per domain are in **bold**.

J Ablation Studies

Weight accumulation. By default, SECL accumulates LoRA weights across the stream (each question's adaptation builds on the previous state). Table 15 shows that disabling accumulation (resetting LoRA to the base model after each question) is catastrophic: ECE degrades from 0.050 to 0.237, and AUROC drops below chance (0.484), i.e., isolated per-question updates destroy the model's confidence signal.

Setting	ECE↓	Brier↓	AUROC↑	Acc
SECL (accumulate)	**.050**	**.241**	**.587**	.577
Reset per question	.237	.327	.484	.576

Table 15: Weight accumulation ablation (Llama 3.2-3B).

Entropy gating vs. always-train. Training on every question (no entropy gating) achieves comparable ECE to SECL (0.047 vs. 0.050), but at 4× the computational cost since all 2,000 questions receive TTT updates. Entropy gating triggers calibration bursts when distribution shift is detected; under our protocol this results in 25.6% of questions receiving TTT updates, achieving the same calibration with 75% fewer gradient steps.

Setting	ECE↓	Brier↓	AUROC↑	Trained %
Always-train	**.047**	**.240**	**.593**	100%
Entropy-gated (SECL)	.050	.241	.587	25.6%

Table 16: Entropy gating vs. always-train (Llama 3.2-3B). Entropy gating matches always-train ECE at a fraction of the cost.

KL divergence regularization. We experiment with adding a KL divergence term $\beta \cdot D_{\mathrm{KL}}(p_{\text{base}} \| p_{\text{adapted}})$ to the training loss to preserve the base model's output distribution. Table 17 shows that a small β=0.01 slightly improves Llama ECE (0.044 vs. 0.050) but degrades Gemma and Phi, while β=0.1 is too aggressive for all models. We use β=0 (no KL term) in all reported experiments.

K Ablation: LoRA Layer Position

To justify placing LoRA adapters on late transformer layers, we compare five layer configurations on Llama 3.2-3B while holding all other hyperparameters fixed (top block of Table 18). Mid and late layers achieve the same ECE (0.039), while early layers remain competitive (0.046). Adapting all 28 layers yields the best AUROC and Brier but the worst ECE (0.058), suggesting that adapting too many layers introduces noise that hurts bin-level calibration despite improving per-question discrimination. The robustness across layer positions indicates that SECL's calibration mechanism is not tightly coupled to a specific layer group; we use late layers as the default following evidence that calibration-relevant representations concentrate in intermediate-to-late layers (Du et al., 2024).

Model	β	ECE↓	Brier↓	AUROC↑	Acc
Llama 3.2-3B	**0** (default)	.050	**.241**	.587	**.577**
	0.01	**.044**	.242	**.591**	.573
	0.1	.149	.279	.527	.569
Gemma 2-2B	**0** (default)	**.056**	**.254**	.548	.515
	0.01	.127	.271	.486	.514
	0.1	.238	.305	**.551**	**.520**
Phi 3.5-Mini (3.8B)	**0** (default)	**.110**	**.251**	.521	.665
	0.01	.144	.252	.506	.669
	0.1	.248	.272	**.598**	**.673**

Table 17: KL regularization ablation across models (SECL + entropy gating).

We further verify the effect of layer *count* on Gemma 2-2B and Phi 3.5-Mini (3.8B) (bottom block of Table 18). Halving the adapter count from 8 to 4 layers roughly doubles ECE on both models (Gemma: 0.056→0.116; Phi: 0.110→0.222) with negligible accuracy change, so 8 late layers provide sufficient capacity for calibration across architectures.

Model	Layers	Range	ECE↓	Brier↓	AUROC↑	Acc
Llama 3.2-3B	Early	0–3	.046	.235	.619	.591
	Mid	12–15	**.039**	.233	.628	.573
	Late (default)	24–27	**.039**	.241	.592	.570
	Last 8	20–27	.044	.240	.609	.558
	All	0–27	.058	**.232**	**.640**	**.588**
Gemma 2-2B	Last 4	22–25	.116	.277	.464	.513
	Last 8 (default)	18–25	**.056**	**.254**	**.548**	**.515**
Phi 3.5-Mini (3.8B)	Last 4	28–31	.222	.277	**.524**	**.669**
	Last 8 (default)	24–31	**.110**	**.251**	.521	.665

Table 18: LoRA layer ablation. **Top:** layer position on Llama 3.2-3B (4 layers each except "All"). **Bottom:** layer count on Gemma 2-2B and Phi 3.5-Mini (3.8B).

L Reliability Diagrams and Normalization Effect

Figure 4 compares reliability diagrams for raw $P(\text{True})$ and distractor-normalized $\text{NormP}_{\text{True}}$ on Llama 3.2-3B. Raw $P(\text{True})$ exhibits upward bias in mid-to-high confidence bins, consistent with the suggestibility effect documented by Wang and

Stengel-Eskin (2025): the model tends to affirm plausible-sounding answers regardless of correctness. Distractor normalization removes this bias by converting absolute affirmation into relative preference among candidate answers, reducing ECE from 0.161 to 0.065.

Figure 5 extends this comparison to all three models. Gemma 2-2B benefits the most (ECE: 0.369 → 0.141), while Phi 3.5-Mini (3.8B) shows a similarly large reduction (ECE: 0.248 → 0.154).

Reliability diagrams for the verbalized baseline vs. SECL on all three models are shown in Figure 6.

M Domain Order Sensitivity

To assess sensitivity to the presentation order of domains, we run SECL with the reversed sequence (TruthfulQA → ARC → MMLU → GSM8K). Table 19 compares overall metrics; Table 20 gives per-domain detail. SECL improves over the verbalized baseline under both orderings for all models, though the original order yields lower ECE. The gap is smallest for Llama 3.2-3B (0.050 vs. 0.068); Llama 3.1-8B shows a larger gap (0.083 vs. 0.189), partly because the reversed order triggers gating on 62.6% of questions (vs. 12.6% forward), amplifying cumulative LoRA drift on the longer model.

Model	Order	ECE↓	AdaECE↓	Brier↓	AUROC↑	Acc
Llama 3.2-3B	Verbalized	.170	.167	.292	.510	.576
	SECL $\mathcal{O}$	**.050**	**.060**	.241	.587	.577
	SECL $\mathcal{R}$	.068	.082	.248	.578	.575
Gemma 2-2B	Verbalized	.256	.252	.314	.558	.516
	SECL $\mathcal{O}$	**.056**	**.060**	.254	.548	.515
	SECL $\mathcal{R}$	.149	.133	.258	.625	.520
Phi 3.5-Mini (3.8B)	Verbalized	.251	.240	.275	.600	.667
	SECL $\mathcal{O}$	**.110**	**.119**	.251	.521	.665
	SECL $\mathcal{R}$	.173	.170	.250	.575	.674
Llama 3.1-8B	Verbalized	.225	.225	.258	.684	.644
	SECL $\mathcal{O}$	**.083**	**.080**	.222	.643	.646
	SECL $\mathcal{R}$	.189	.191	.272	.580	.629

Table 19: Domain order sensitivity (overall metrics). $\mathcal{O}$: GSM8K→MMLU→ARC→TruthfulQA. $\mathcal{R}$: TruthfulQA→ARC→MMLU→GSM8K.

Model	Order	GSM8K	MMLU	ARC	TruthfulQA
Llama 3.2-3B	$\mathcal{O}$	.070	.067	.112	.068
	$\mathcal{R}$	.117	.032	.081	.134
Gemma 2-2B	$\mathcal{O}$	.356	.054	.144	.210
	$\mathcal{R}$	.396	.096	.046	.189
Phi 3.5-Mini (3.8B)	$\mathcal{O}$	.283	.054	.129	.229
	$\mathcal{R}$	.068	.213	.099	.343
Llama 3.1-8B	$\mathcal{O}$	.066	.120	.198	.038
	$\mathcal{R}$	.206	.217	.234	.223

Table 20: Per-domain ECE for original ($\mathcal{O}$) vs. reversed ($\mathcal{R}$) domain order.

ARC domain ordering effect. ARC-Challenge is the only domain where SECL consistently increases ECE relative to the verbalized baseline in the forward order (Llama 3B: .095 → .112; Gemma: .082 → .144; Phi: .109 → .129; Llama 8B: .108 → .198). In the reversed order, ARC ECE *improves* for the three smaller models (Llama 3B: .095 → .081; Gemma: .082 → .046; Phi: .109 → .099), though Llama 8B degrades further (.108 → .234). For the smaller models, the ARC degradation is therefore an artifact of domain sequencing, specifically, the LoRA weights adapted on preceding domains transfer poorly to ARC's scientific reasoning format, not a limitation of SECL itself. Overall ECE remains improved under both orderings for all models, so the per-domain redistribution does not undermine the method's aggregate benefit.

N Computational Cost Analysis

Table 21 compares the computational cost of SECL against running P(True) Norm on every question. We report costs in forward-pass equivalents (FWD-eq), where one backward pass ≈ 2 FWD. For questions where entropy gating skips TTT, SECL requires only 1 FWD (the generation pass); the P(True) Norm baseline requires $1 + (k{+}1) = 6$ FWD per question ($k{=}4$ distractors plus the correct answer).

Model	Trained	Skipped	SECL	P(True)
Llama 3.2-3B	512 (25.6%)	1,488	9,168	12,000
Llama 3.1-8B	251 (12.6%)	1,749	5,514	12,000
Gemma 2-2B	119 (5.9%)	1,881	3,666	12,000
Phi 3.5-Mini (3.8B)	160 (8.0%)	1,840	4,240	12,000

Table 21: Computational cost comparison (in FWD-equivalents). *Trained* = questions receiving TTT; *Skipped* = entropy-gated questions (generation only). Per trained question: ≈15 FWD-eq (1 gen + 5 P(True) + 3 epochs × 3 FWD-eq/step). Per skipped question: 1 FWD-eq. P(True) Norm baseline: 6 FWD-eq × 2,000 = 12,000.

SECL is cheaper than P(True) Norm baseline for all four models. Gemma and Phi trigger infrequently (6–8%), reducing cost to roughly a third of the baseline while achieving substantially better calibration. Even Llama, which triggers most often (25.6%), remains below the baseline cost while improving ECE from .065 to .050. The temperature scaling and Platt scaling add negligible compute (a single parameter fit) but require ground-truth labels, which are unavailable in the test-time setting.

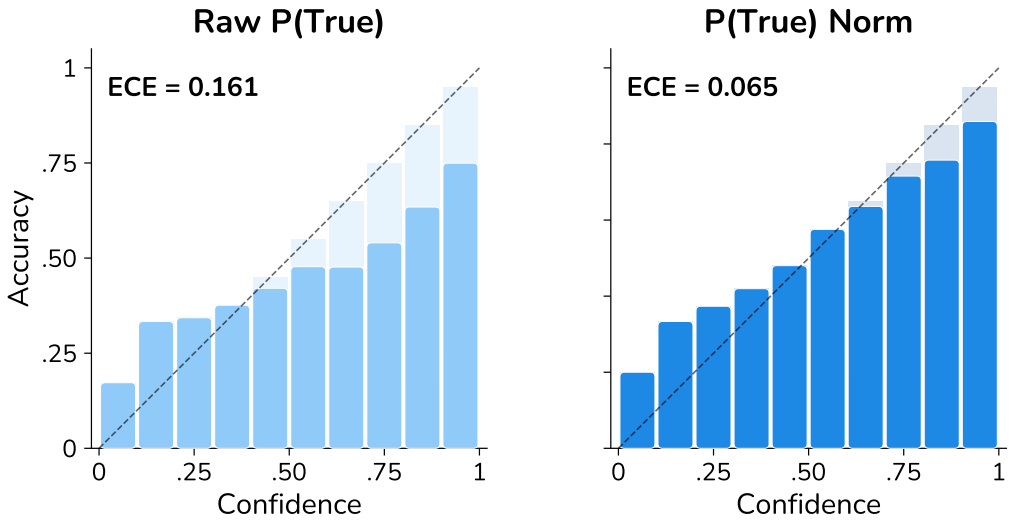

Figure 4: Raw $P(\text{True})$ (**left**) vs. distractor-normalized $\text{NormP}_{\text{True}}$ (**right**) on Llama 3.2-3B. Normalization suppresses suggestibility bias, producing a better-calibrated supervision signal (ECE: 0.161 $\rightarrow$ 0.065).

O Negative Control: Qwen 2.5-3B

We evaluate Qwen 2.5-3B-Instruct (Qwen Team et al., 2025) as a negative control. Unlike the other four models, Qwen's P(True) Norm baseline (*best* τ=1.0, ECE=0.257) is *worse* than its verbalized baseline (ECE=0.247), indicating that the discriminative signal provides no calibration advantage over generation. Since SECL distills the discriminative signal into generative confidence, the method cannot improve calibration when the discriminative signal itself is uninformative. This confirms a key prerequisite of our approach: SECL succeeds precisely when a generation–discrimination gap exists.

Method	**ECE**$\downarrow$
Verbalized	**.247**
P(True) Norm τ=0.3	.290
P(True) Norm τ=0.7	.263
P(True) Norm τ=1.0	.257
P(True) Norm τ=1.5	.265
P(True) Norm τ=2.0	.267
P(True) Norm τ=3.0	.291

Table 22: Qwen 2.5-3B: no generation–discrimination gap. All P(True) Norm temperatures yield worse ECE than the verbalized baseline.

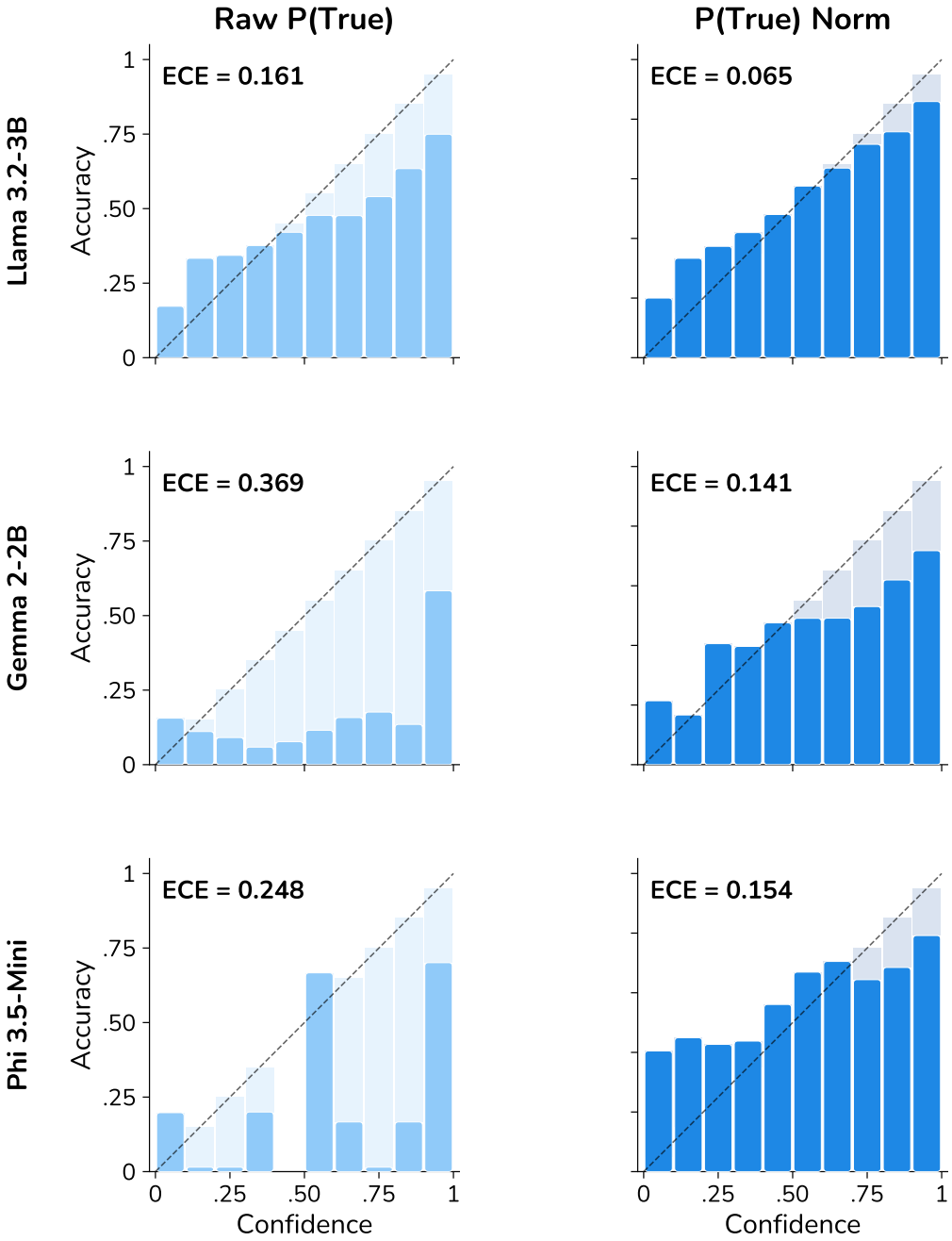

Figure 5: Raw $P(\text{True})$ (**left**) vs. $\text{NormP}_{\text{True}}$ (**right**) for the three non-8B models. Distractor normalization consistently reduces calibration error across model families.

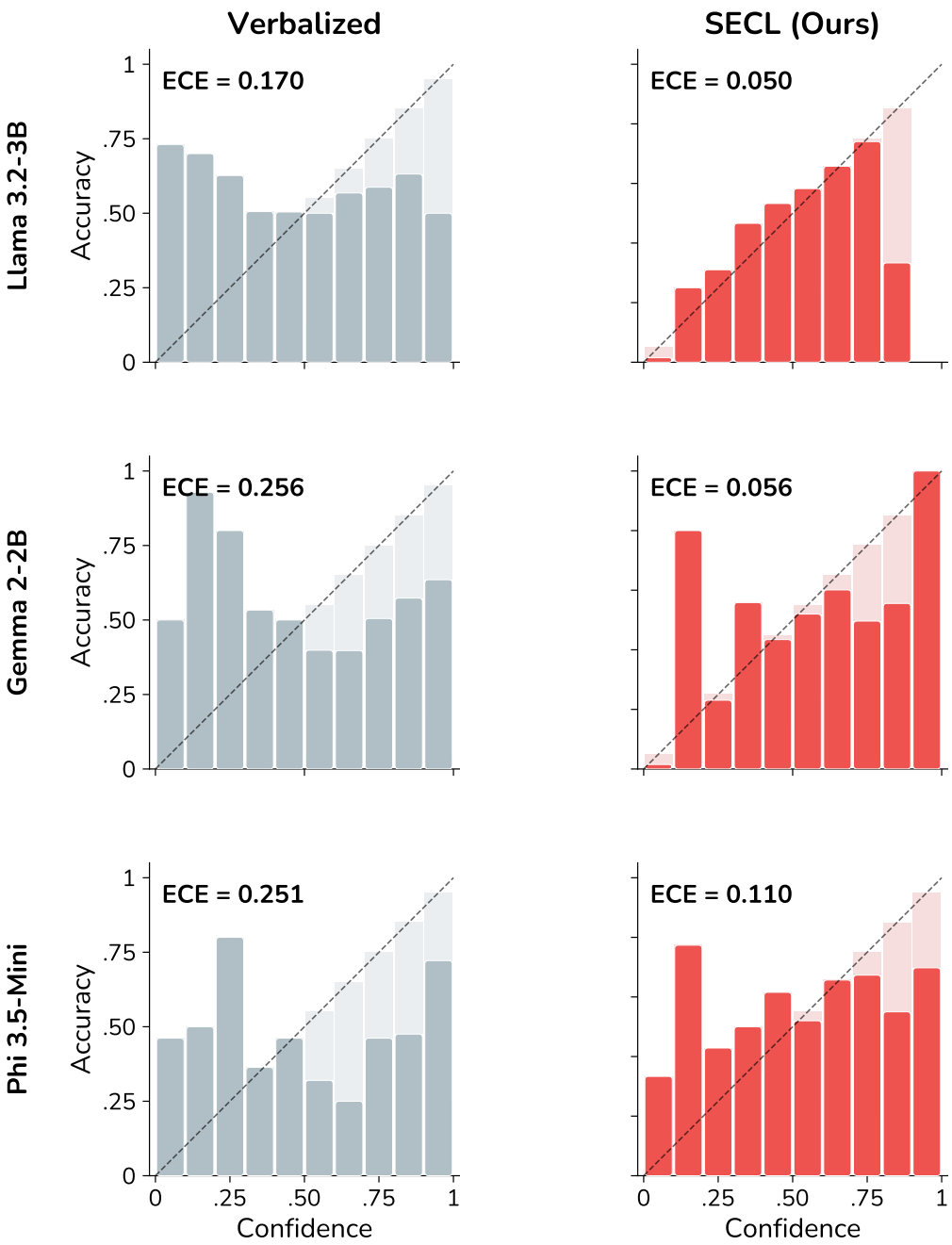

Figure 6: Reliability diagrams: Verbalized baseline (**left**) vs. SECL (**right**) for Llama 3.2-3B (ECE: 0.170 → 0.050), Gemma 2-2B (ECE: 0.256 → 0.056), and Phi 3.5-Mini (3.8B) (ECE: 0.251 → 0.110). SECL consistently shifts predictions toward the diagonal, eliminating the systematic overconfidence visible in the verbalized baselines.

P Entropy Gating Trigger Statistics

Model	Total	Trained (%)	Skipped (%)
Llama 3.2-3B	2,000	512 (25.6%)	1,488 (74.4%)
Gemma 2-2B	2,000	119 (5.9%)	1,881 (94.1%)
Phi 3.5-Mini (3.8B)	2,000	160 (8.0%)	1,840 (92.0%)
Llama 3.1-8B	2,000	251 (12.6%)	1,749 (87.4%)

Table 23: Entropy gating trigger statistics per model. *Trained%* is the fraction of questions receiving TTT adaptation.

Q Scaling to Llama 3.1-8B

To test whether SECL scales to larger models, we evaluate Llama 3.1-8B-Instruct (8.03B parameters) on the same 2,000-question stream. We use LoRA on the last 8 layers (rank 8), τ=3.0, and the same entropy gating parameters as the 3B models.

Method	ECE↓	AdaECE↓	Brier↓	AUROC↑	Acc
Verbalized	.225	.225	.258	.684	.644
+ Temp Scaling*	**.080**	.089	.213	.681	.644
P(True) Norm	.120	.117	.211	**.718**	.646
+ Temp Scaling*	.108	.103	**.208**	.716	.646
SECL (Ours)	<u>.083</u>	**.080**	.222	.643	.646

Table 24: Llama 3.1-8B results. SECL reduces ECE by 63% relative to the verbalized baseline while adapting only 12.6% of questions. *Requires ground-truth labels (5-fold CV).

Method	Domain	ECE↓	Brier↓	AUROC↑	Acc
Verbalized	GSM8K	.298	.305	.714	.602
	MMLU	.208	.251	.644	.670
	ARC	.108	.164	.633	.798
	TruthfulQA	.287	.310	.669	.506
SECL	GSM8K	**.066**	.238	.602	.572
	MMLU	**.120**	.223	.591	.686
	ARC	.198	.178	.632	.824
	TruthfulQA	**.038**	.249	.552	.504

Table 25: Per-domain results for Llama 3.1-8B.

SECL achieves ECE=0.083 on the 8B model, a 63% reduction from the verbalized baseline (0.225), confirming that the method scales effectively. The pattern mirrors the 3B results: large improvements on GSM8K, MMLU, and TruthfulQA, with ARC remaining resistant to improvement (see domain ordering analysis in Section M). The entropy gate triggers on 12.6% of questions, intermediate between the 3B models' 5.9–25.6% range.

R Ablation: Training Target Signal

A natural question is whether P(True) Norm is the right pseudo-label for SECL training. Self-Consistency (SC; Wang et al., 2023) is an alternative unsupervised signal that measures agreement among N=10 temperature-sampled answers. We run the full SECL pipeline (LoRA adaptation, entropy gating, burst training, MSE loss) with SC as the training target, keeping all other hyperparameters identical to the canonical configuration.

Training Target	ECE↓	Brier↓	AUROC↑	Acc
None (Verbalized baseline)	.170	.292	.510	.576
Self-Consistency (N=10)	.432	.470	.443	.564
P(True) Norm (ours)	**.050**	**.241**	**.587**	**.577**

Table 26: Effect of training target on Llama 3.2-3B. All rows use the same SECL pipeline; only the pseudo-label differs. Self-Consistency as a target degrades calibration below the untrained baseline.

Why the target signal determines calibration quality. SECL minimizes $\mathcal{L} = \mathbb{E}[(c(q) - t(q))^2]$, where $c(q)$ is the model's verbalized confidence and $t(q)$ is the training target. With sufficient model capacity and training convergence, $c^*(q) \to t(q)$, so the trained model *inherits the calibration of its target*: $\mathrm{ECE}(c^*) \to \mathrm{ECE}(t)$. The choice of target therefore bounds the achievable calibration.

SC is a biased proxy for correctness. As $N \to \infty$, $\mathrm{SC}(q) \to p_{\mathrm{gen}}(\mathrm{mode} \mid q)$, the probability mass the generation distribution places on its modal answer. For modern LLMs, the generation distribution is low-entropy: $p_{\mathrm{gen}}(\mathrm{mode} \mid q)$ is typically high (>0.7) regardless of whether the answer is correct. Among questions where SC ≈ 1, the actual fraction correct can be significantly below 1, violating the calibration condition $P(Y{=}1 \mid c{=}p) = p$. Training on SC therefore teaches the model to report generation concentration as confidence, producing systematic overconfidence.

P(True) Norm is a less biased proxy. P(True) Norm computes $\mathrm{softmax}(\log P(\mathrm{True} \mid q, a_i)/\tau)_0$ over the model's answer and k distractors. When the model cannot distinguish its answer from distractors (equal $P(\mathrm{True})$ scores), P(True) Norm $\to 1/(k{+}1)$, correctly indicating low confidence. When the model strongly prefers its answer, P(True) Norm $\to$ 1. This normalization maps the discriminative signal to $[0, 1]$, aligning it more

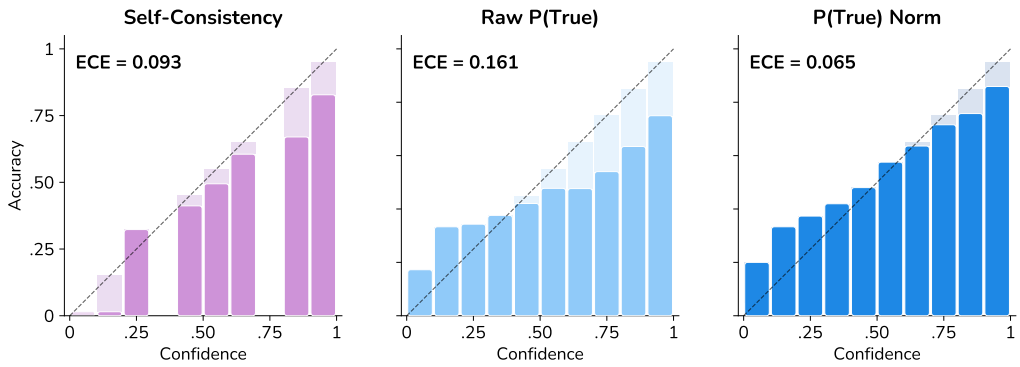

Figure 7: Reliability diagrams comparing candidate training targets on Llama 3.2-3B. Self-Consistency (**left**, ECE=0.093) produces a reasonable calibration curve but with systematic bias. Raw P(True) (**center**, ECE=0.161) exhibits upward bias from suggestibility. P(True) Norm (**right**, ECE=0.065) tracks the diagonal most closely, confirming it as the best-calibrated supervision signal for SECL training.

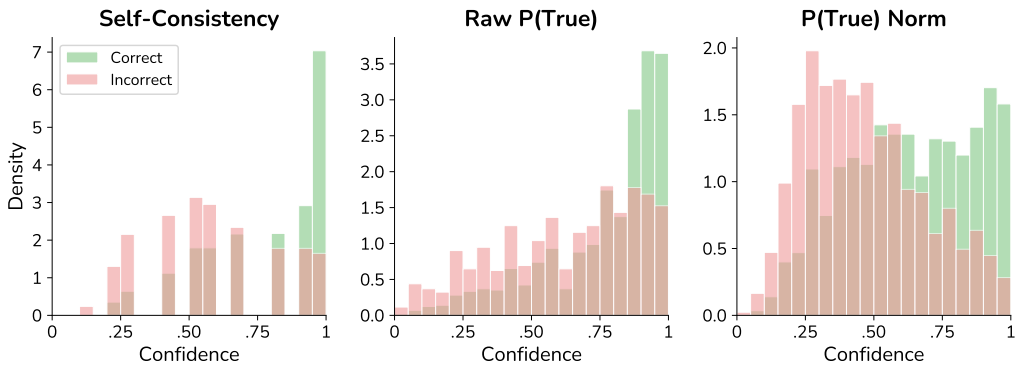

Figure 8: Confidence score distributions for correct (green) and incorrect (red) predictions on Llama 3.2-3B. Self-Consistency and Raw P(True) both cluster near 1.0 with poor class separation; P(True) Norm spreads scores across $[0, 1]$ with better separation between correct and incorrect predictions.

closely with the correctness probability and reducing both raw and trained ECE.

The raw P(True) Norm signal (ECE=0.065) is better calibrated than the raw Self-Consistency signal (ECE=0.093) *before any training*. After SECL training with the P(True) Norm target, the model's verbalized confidence inherits this calibration (ECE=0.050), while training With the SC target, SC's biases are amplified (ECE=0.432). Figure 7 provides direct visual evidence via reliability diagrams: SC produces a monotonically increasing but biased calibration curve, Raw P(True) suffers from suggestibility bias in mid-to-high bins, while P(True) Norm tracks the diagonal most closely. Figure 8 complements this with score distributions, showing that P(True) Norm achieves the best separation between correct and incorrect predictions.

Calibration quality is thus determined by the training target. P(True) Norm is a better proxy for $P(\text{correct})$ than Self-Consistency; the adaptation mechanism distills whatever signal it receives, so signal quality determines the ceiling of SECL's effectiveness.

S Extended Stream Length (4,000 Questions)

To test whether SECL's calibration improvements are stable over longer streams, we double the stream length from 2,000 to 4,000 questions (1,000 per domain) on Llama 3.2-3B, Gemma 2-2B, and Phi 3.5-Mini (3.8B). Table 27 reports the results.

At 4,000 questions, SECL improves over verbalized confidence on Llama (ECE: 0.190→0.051; Brier: 0.313→0.239; AUROC: 0.508→0.607) and Gemma (ECE: 0.276→0.086; Brier: 0.328→0.263), though Gemma shows a modest AUROC trade-off (0.548→0.500). For Phi, SECL improves AUROC/accuracy (0.593→0.712; 0.687→0.696) but remains worse on calibration (ECE/Brier: 0.259/0.260 vs. 0.192/0.245), indicating under-distillation.

Model	Method	ECE↓	Brier↓	AUROC↑	Acc
Llama 3.2-3B	Verbalized (4,000q)	.190	.313	.508	.577
	SECL (4,000q)	**.051**	**.239**	**.607**	**.581**
Gemma 2-2B	Verbalized (4,000q)	.276	.328	**.548**	.508
	SECL (4,000q)	**.086**	**.263**	.500	**.512**
Phi 3.5-Mini (3.8B)	Verbalized (4,000q)	**.192**	**.245**	.593	.687
	SECL (4,000q)	.259	.260	**.712**	**.696**

Table 27: 4,000-question comparison against same-stream verbalized baselines. SECL substantially improves calibration on Llama and Gemma. For Phi, stronger long-stream calibration requires sufficient distillation; with the 4,000q run, calibration remains worse than verbalized while discrimination and accuracy improve.

Model	Method	FWD-eq	ECE↓	Brier↓	AUROC↑	Acc
Llama 3.2-3B	Verbalized	1	.170	.292	.510	.576
	DINCO	∼10	.101	**.207**	**.762**	.466
	SECL (ours)	4.6	**.050**	.241	.587	**.577**
Phi 3.5-Mini (3.8B)	Verbalized	1	.251	.275	.600	**.667**
	DINCO	∼10	**.110**	**.212**	**.749**	.560
	SECL (ours)	2.1	**.110**	.251	.521	.665
Gemma 2-2B	Verbalized	1	.256	.314	.558	**.516**
	DINCO	∼10	.408	.410	.566	.327
	SECL (ours)	1.8	**.056**	**.254**	.548	.515
Llama 3.1-8B	Verbalized	1	.225	.258	.684	.644
	DINCO	∼10	.117	**.210**	**.756**	.522
	SECL (ours)	2.8	**.083**	.222	.643	**.646**

FWD-eq: amortized forward-pass equivalents per question over the full stream (see Section N).

Table 28: SECL vs. DINCO across all models. SECL achieves better calibration (ECE) at lower amortized cost, while DINCO provides superior discrimination (AUROC) at higher cost.

T Additional Open-Ended Evaluation (TruthfulQA-Gen)

To reduce the gap between multiple-choice evaluation and open-ended generation, we replace the final domain with the TruthfulQA generation split and keep the rest of the stream unchanged. We report the forward and reversed streams (500 questions per domain, 2,000 total) and an extended 4,000-question forward stream (1,000 per domain) to test long-horizon stability on open-ended answers.

U Selective ECE

Selective ECE is computed on predictions above the highest confidence threshold with common coverage across methods ($\alpha \leq 0.7$ for Llama 3.2-3B, $\alpha \leq 0.6$ otherwise); see Table 30. SECL achieves the lowest selective calibration error on all four models.

Method	Order	Stream	ECE↓	Brier↓	AUROC↑	Acc
Verbalized	Forward	2k	.195	.287	.586	**.506**
P(True) Norm	Forward	2k	.173	.273	**.625**	**.506**
SECL (Ours)	Forward	2k	**.047**	**.239**	.618	.500
SECL (Ours)	Reversed	2k	**.047**	.241	**.625**	.497
Verbalized	Forward	4k	.179	.278	.598	**.525**
SECL (Ours)	Forward	4k	**.061**	**.237**	**.625**	.522

Table 29: TruthfulQA generation-domain evaluation (Llama 3.2-3B). SECL reduces ECE by 76% on the 2k forward stream (0.195 → 0.047), remains stable under reversed ordering, and at 4k still improves over the same-stream verbalized baseline by 66% (0.179 → 0.061).

Model	Verb.	P(T) Norm	DINCO	SECL
Llama 3.2-3B	.130	.050	.119	**.023**
Llama 3.1-8B	.224	.149	.122	**.094**
Gemma 2-2B	.255	.197	.463	**.059**
Phi 3.5-Mini	.251	.147	.132	**.100**

Table 30: Selective ECE at the highest threshold with common coverage. SECL is lowest on all four models.

V Comparison with DINCO

DINCO (Wang and Stengel-Eskin, 2025) is a recent inference-time calibration method that normalizes verbalized confidence against self-generated distractors and combines it with Self-Consistency. Unlike SECL, DINCO does not adapt the model; it computes a calibrated confidence score at each inference step using ∼10 forward passes (beam search + P(True) per candidate + NLI + SC sampling).

SECL and DINCO address different trade-offs. Across all four models, SECL achieves lower or equal ECE (Llama 3B: 0.050 vs. 0.101; Llama 8B: 0.083 vs. 0.117; Phi: 0.110 vs. 0.110; Gemma: 0.056 vs. 0.408) with lower total computational cost and no external models. DINCO achieves stronger AUROC on all four models using multiple inference passes and an NLI model, at the cost of ∼10× inference compute. The calibration–cost trade-off is visualized in Figure 1 (main paper).

Under DINCO's default configuration, accuracy is consistently lower than SECL's (Llama 3B: 46.6% vs. 57.7%; Llama 8B: 52.2% vs. 64.6%; Phi: 56.0% vs. 66.5%; Gemma: 32.7% vs. 51.5%), as beam-search answer selection yields lower task accuracy than greedy decoding; using greedy answers with beam-search distractors could mitigate this. On Gemma, DINCO essentially fails: ECE of 0.408 is worse than the verbalized baseline (0.256), and accuracy drops to 32.7%. Gemma's internal representations appear ill-suited to DINCO's NLI-

based reweighting, whereas SECL's distillation of the discriminative signal still works (ECE 0.056).