Making LLMs Say What They Know: Probe-Targeted Fine-Tuning for Verbal Confidence Calibration

Jon-Paul Cacioli*

May 29, 2026

Abstract

Background: Verbal confidence in instruct-tuned large language models fails because the readout pathway from internal correctness representations to the confidence-token position transmits little of the available signal. Linear probes on hidden states discriminate correct from incorrect responses at $\text{AUROC}_2 = 0.76$–0.88 across seven of eight models spanning four families and three scales (the exception, Qwen 72B, is a probe outlier), yet verbal confidence saturates near ceiling (mean ∼95–99%).

Objectives: We introduce probe-targeted confidence-calibrated supervised fine-tuning (PT-CSFT), which uses a linear probe on a model's own hidden states to generate continuous confidence targets for LoRA fine-tuning. We evaluate whether this method closes the metacognitive gap across model families, scales, and cognitive domains.

Methods: PT-CSFT modifies attention key and output projections and the three MLP projections via LoRA while leaving queries, values, and the language model head unchanged. We evaluate on eight models (7B–72B, four families) across TriviaQA, GSM8K, and ARC-Challenge. A two-stage curriculum addresses the 70B regime. The logit readout reads softmax distributions over confidence tokens as a continuous signal. Pre-registration: OSF Phase 1 https://osf.io/ngkwc/, Phase 0 https://osf.io/mpcr5.

Results: PT-CSFT recovers 91–115% of probe discrimination in verbal confidence at 7–32B. At 70B, a two-stage curriculum closes 66% of the gap on the logit channel ($\text{AUROC}_2 = 0.797$), the first VRS-Valid confidence signal at 70B. On GSM8K, the logit readout achieves $\text{AUROC}_2 = 0.862 \pm 0.013$ across 10 seeds, exceeding the probe. Post-hoc Platt scaling yields ECE = 0.042. The signal transfers out-of-distribution (NQ: 0.757, PopQA: 0.834) and enables confidence-gated retrieval (2.17× accuracy differential).

Conclusions: Controlled activation patching at the confidence-token position supports the interpretation that verbal confidence failure is a routing problem. The intervention is position-specific (mid-question control: chance), bidirectional (91% forward, 89% reverse), selective (83% of answers unchanged), and follows a near-monotonic layer-depth gradient (Spearman $\rho = 0.976$, $p < 10^{-4}$). These observations are consistent with PT-CSFT repairing a position-specific pathway for confidence without disrupting answer computation, though they do not trace the effect to specific model components. Multi-seed replication across three model families confirms stability (Llama 8B: 0.836 ± 0.011; Qwen 7B: 0.801 ± 0.022; Mistral 7B: 0.771 ± 0.017; 6 seeds each). The logit readout universally rescues where text confidence fails.

1 Introduction

This paper investigates why instruct-tuned language models fail to express calibrated verbal confidence despite possessing internal correctness signals, and introduces a method to repair this failure. The argument is organised around a single thesis: verbal confidence fails because

*Independent Researcher, Melbourne, Australia. ORCID: 0009-0000-7054-2014. Correspondence: synthium@hotmail.com.

of a routing bottleneck, not because the model lacks the information or the vocabulary. The following subsections develop this thesis (§ 1.1), characterise the problem (§ 1.2–1.3), motivate the work (§ 1.4), state the research questions (§ 1.5), and summarise the contributions (§ 1.6).

1.1 The Pathway Thesis

This paper makes one claim. Verbal confidence in instruct-tuned language models fails because the readout pathway transmits little of the available correctness signal, not because the signal is absent. The model evaluates its own reliability internally. Linear probes on hidden states discriminate correct from incorrect responses at AUROC_2 in the 0.76 to 0.88 range across seven of eight models spanning four families and three scales (Qwen 72B, the largest, is a probe outlier discussed in § 5.2). The output vocabulary supports graded uncertainty. Digit tokens covering the 0 to 100 confidence range are present in the language model head from pre-training. What is broken is the routing of correctness-relevant features from earlier layers to the position where confidence is verbalised.

Probe-targeted confidence-calibrated supervised fine-tuning (PT-CSFT) repairs this routing. LoRA modifies attention key and output projections and the three MLP projections across transformer layers. The query and value projections and the language model head are unchanged. The intervention does not teach the model new vocabulary for expressing uncertainty. It changes which content arrives at the confidence-token position, where the existing vocabulary geometry is read by the unchanged output mapping. Four converging observations developed in § 8.5 support the routing-not-mapping interpretation: a baseline-hidden-state probe gap, an inference-time steering null, an unchanged output mapping, and training-free baselines at chance. Controlled activation patching at the confidence-token position (§ 8.6) provides further support: the intervention is position-specific, bidirectional, selective for confidence over answer content, and follows a near-monotonic layer-depth gradient. Every experiment in this paper provides evidence about this pathway: where it can be repaired, where it cannot, and what it is made of.

1.2 The Verbal Confidence Problem

When asked to state their confidence as a percentage, instruct-tuned language models overwhelmingly report values near 100% regardless of correctness. Across multiple model families and scales, the verbal confidence channel exhibits ceiling rates above 90%, near-chance Type-2 AUROC (AUROC_2), and response distributions so concentrated that they carry negligible information about correctness (Cacioli 2026c). We term this the *ceiling prior*. It has been documented across 20 frontier models using psychophysical methodology from signal detection theory (Fleming and H. C. Lau 2014; Maniscalco and H. Lau 2012). Every model tested was classified as producing Invalid verbal confidence under the Validity Rating Scale (VRS; Cacioli (2026a)), which classifies confidence signals into three tiers based on ceiling rate (L, the proportion of responses at high confidence), response inconsistency (TRIN), and confidence–correctness correlation (r). Classification follows the VRS reference implementation (Cacioli 2026a): a signal is Invalid if $\text{L} \geq 0.70$, $\text{TRIN} \geq 0.80$, or $|r| < 0.05$ (any one sufficient); Indeterminate if $\text{L} \geq 0.40$ or $\text{TRIN} \geq 0.60$ (without triggering Invalid); and Valid otherwise.

This finding is consistent with broader work on LLM confidence. Tian et al. (2023) showed that verbalized confidence from models trained with reinforcement learning from human feedback (RLHF) is systematically overconfident under naive elicitation. Xiong et al. (2024) found persistent miscalibration across elicitation methods. Groot and Valdenegro-Toro (2024) found that overconfidence persists even with carefully designed prompting strategies. Leng et al. (2024) trace part of this to the RLHF reward itself, which can incentivise overconfident expression. The ceiling prior is not an elicitation failure. It is a systematic property of how instruct-tuned models map internal states to verbal output.

1.3 The Internal Signal Exists

The ceiling prior coexists with informative internal representations. Linear probes trained on hidden states from baseline models achieve AUROC_2 values in the 0.76–0.88 range (seven of eight models; Qwen 72B excepted) on the same items where verbal confidence is near chance (Cacioli 2026c; Kadavath et al. 2022; Burns et al. 2023). Single-pass logit entropy also discriminates correctness better than verbal confidence. The information exists internally. The verbal channel does not transmit it.

Recent mechanistic work characterises why. Kumaran et al. (2026) showed that verbal confidence reflects a second-order evaluation process, computed automatically during answer generation and cached at answer-adjacent positions. The cached representations explain variance in verbal confidence beyond token log-probabilities, confirming that the model performs a genuine correctness evaluation. Yet Miao and Ungar (2026) found that the correctness direction and the confidence verbalization direction are geometrically orthogonal in the residual stream. Zhao et al. (2026) localise the failure further, identifying specific attention heads that causally inflate verbalized confidence on incorrect answers. The model evaluates its own reliability but the generation pathway does not faithfully transmit the result.

This dissociation between internal knowledge and verbal expression defines the *metacognitive gap*: the difference between what a model knows internally about its own reliability and what it reports verbally. Closing this gap is the central problem this paper addresses.

A note on representation levels. Throughout the paper we appeal both to a *signal that exists internally* (probe AUROC_2 0.76 to 0.88 on baseline hidden states) and to a *signal that PT-CSFT creates* (a 2-layer MLP probe on baseline hidden states at the confidence position reaches 0.785, well below the PT-CSFT logit readout at 0.862). These claims are about different representational levels and they are not in tension. We distinguish four levels (illustrated in Figure 1):

1. **Latent correctness information.** The diffuse signal about whether the model's first-order response is correct. Distributed across attention and MLP states in middle and late layers in the baseline.
2. **Probe-decodable representation.** What a linear probe can extract from a single layer at a chosen token position. A constrained projection of level 1. Achieves AUROC_2 0.76 to 0.88 in baseline models at appropriate layers; this is what we mean by "the internal signal exists".
3. **Confidence-position routed content.** What actually arrives at the hidden state above the confidence token at generation time. In the baseline, this is impoverished relative to the probe-decodable representation at non-confidence positions (the MLP probe at the confidence position reaches only 0.785). PT-CSFT modifies attention and MLP weights to bring more of level 1 to this position; this is what we mean by the *trained signal*.
4. **Verbal output.** The text the model emits, read by the unchanged language model head from level 3.

The probe results in § 5.2 establish the existence of level 2. The MLP probe gap in § 8.4 and the logit readout in § 8.1 are claims about level 3. The metacognitive gap is the difference between levels 2 and 4 in baseline models. PT-CSFT addresses level 3.

1.4 Why the Gap Matters

The metacognitive gap has direct practical consequences. Selective prediction, deferral to human experts, risk flagging, and autonomous agent behaviour all rely on model self-assessment. These applications currently require logit-level signals (entropy, softmax margins) or external sampling methods (self-consistency), both of which need model internals access or multiple forward passes. A reliable verbal confidence readout would be cheaper, simpler to deploy, and accessible through any API. For safety-critical deployments, knowing when the model does not know is arguably

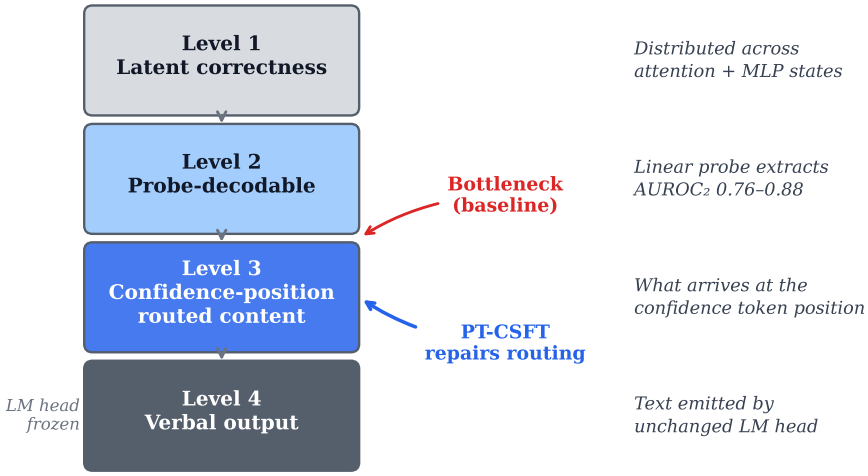

Figure 1: **Four-level representation framework.** Level 1 (latent correctness information) and Level 2 (probe-decodable representation) exist in the baseline. The *bottleneck* between Levels 2 and 3 is where the metacognitive gap arises: correctness-relevant content does not reach the confidence-token position. PT-CSFT repairs the routing at Level 3 while leaving the language model head (Level 4 readout) unchanged.

as important as accuracy itself. Yona et al. (2026) argue that this capability, which they term *faithful uncertainty*, is essential for trustworthy LLMs and becomes the control layer for agentic systems. PT-CSFT provides a concrete method for achieving their objective. The present work focuses on system-side calibration metrics (AUROC_2, ECE, selective prediction coverage, retrieval gating). Whether the resulting confidence signals improve human-AI team accuracy in deferral or trust calibration is an important downstream question that we do not address here.

1.5 Research Questions

This paper addresses five questions:

RQ1 (Method). Can a model's own internal correctness signal, read by a linear probe on hidden states, serve as a training target to improve verbal confidence discrimination? We call this method *probe-targeted confidence-calibrated supervised fine-tuning* (PT-CSFT).

RQ2 (Generality across models). Does PT-CSFT transfer across model families, architecture classes, and scales? We evaluate eight models spanning three architecture classes at scales from 7B to 72B.

RQ3 (Sensitivity). How sensitive is PT-CSFT to LoRA hyperparameters, and does sensitivity vary across architectures? We conduct a systematic ablation across rank (4–128) and learning rate (5×10^{-5} to 2×10^{-4}).

RQ4 (Generality across cognitive domains). Does PT-CSFT transfer beyond factual recall (TriviaQA) to mathematical reasoning (GSM8K) and science reasoning (ARC-Challenge)? We use a probe-gate strategy: if a linear probe on hidden states exceeds $\text{AUROC}_2 = 0.65$, the domain has sufficient internal signal for PT-CSFT to operate on.

RQ5 (Deployment). Does PT-CSFT confidence transfer to out-of-distribution data without retraining, and does it enable downstream metacognitive control? We evaluate the TriviaQA-trained adapter on NaturalQuestions and PopQA, and use it to gate confidence-aware retrieval.

1.6 Contributions

This paper makes four contributions, each drawing on multiple experiments.

1. **Method.** We introduce PT-CSFT, a fine-tuning method that uses a linear probe on a model's own hidden states to generate continuous, model-specific confidence targets. The method requires only a few hundred examples and under 10 minutes of training on consumer hardware. PT-CSFT is evaluated on eight models across four families and three architecture classes at 7B–72B and on three cognitive domains (TriviaQA, GSM8K, ARC-Challenge). Headline results are summarised in § 5.3 and developed across § 5–§ 8.
2. **Mechanism.** The improvement is routing repair, not output remapping. LoRA modifies attention key and output projections and the three MLP projections while leaving the attention queries, attention values, and the language model head unchanged. Under this configuration, the model cannot acquire a new output mapping; the LoRA can only modify how content is integrated across positions and transformed before reaching the confidence-token position. Four converging observations support this interpretation (§ 8.5), and controlled activation patching at the confidence-token position (§ 8.6) provides further evidence: the intervention is position-specific, bidirectional, selective for confidence over answer content, and follows a near-monotonic layer-depth gradient consistent with confidence-relevant content accumulating through the network. The LoRA target analysis and post-training probe analysis in § 9.4 add mechanistic detail.
3. **Boundary map.** We characterise where the pathway breaks and how each break can be addressed (§ 7.3–§ 7.4). The headline boundary cases: a cross-architecture failure on LlamaForCausalLM resolved by a 4× lower learning rate; the 70B regime where standard single-stage LoRA fails and a two-stage curriculum closes most of the gap on the logit channel (§ 7.3.4); and a recipe-specific accuracy-AUROC Pareto on Qwen 7B GSM8K that does not appear on Gemma 12B (§ 7.3.2). The logit readout is the universal rescue wherever the text channel is insufficient (§ 8).
4. **Deployment.** PT-CSFT confidence enables downstream metacognitive control. Confidence-gated retrieval (§ 9.5) stratifies items by accuracy and outperforms naive always-retrieve baselines that degrade under evidence poisoning. The signal transfers out-of-distribution on two factoid QA datasets without retraining (§ 9.6). Post-hoc Platt scaling combines strong discrimination with strong calibration (§ 9.7). The routing repair therefore supports a complete monitoring-to-control loop on consumer hardware.

1.7 Scope and Terminology

We use "metacognitive" operationally to denote Type-2 discrimination, the ability of a readout to distinguish a system's own correct from incorrect first-order responses, following the usage established in signal detection theory (Fleming and H. C. Lau 2014). We do not claim that LLMs possess human-like metacognitive awareness. AUROC_2 is our primary metric throughout. It measures the probability that a confidence signal ranks a correct response higher than an incorrect one, is nonparametric, and is stable across confidence formats (Cacioli 2026e).

We distinguish *model families* (defined by training provider: Gemma, Llama, Mistral, Qwen) from *architecture classes* (defined by HuggingFace model class: GemmaForCausalLM, LlamaForCausalLM, Qwen2ForCausalLM). Mistral shares the LlamaForCausalLM architecture class but is a separate model family. Some patterns we report generalise across families (e.g., the ceiling prior); others generalise across architecture classes (e.g., the LoRA learning-rate sensitivity, which clusters by class); the text distinguishes these explicitly at each result. We use *text channel / text readout* for argmax-parsed verbal output and *logit channel / logit readout* for the softmax distribution at the confidence-token position (full mechanism in § 8).

Reader's guide. The strongest empirical result is the Gemma 12B GSM8K logit readout at $\text{AUROC}_2 = 0.862 \pm 0.013$ across 10 seeds (§ 7.3.2 and § 8). The strongest cross-architecture

finding is the LoRA learning-rate sensitivity on LlamaForCausalLM (§ 6.2), which converts an apparent architecture-intrinsic failure into a recipe issue. The most diagnostically informative section is the 70B scaling story (§ 7.3.3–§ 7.3.4 and § 6.3): standard single-stage LoRA fails across seven configurations, and the two-stage curriculum recovers partial closure on the logit channel. The mechanistic core comprises the four-evidence case for routing (§ 8.5), the controlled activation patching (§ 8.6), and the LoRA target analysis and post-training probe table in § 9.4. Readers short on time can read the abstract, § 1.1, § 5.3, § 7.3.2, § 7.3.4, § 8.5–8.6, and § 10 for the full argument.

2 Related Work

We situate PT-CSFT relative to seven lines of prior work: confidence calibration methods, representation probing, inference-time intervention, signal detection theory applied to metacognition, concurrent fine-tuning approaches, the distinction between discrimination and calibration, and metacognition as an architectural control layer.

2.1 Confidence Calibration in Language Models

The study of LLM confidence has proceeded along several parallel tracks. Token-level approaches use output logit distributions as implicit confidence signals (Kadavath et al. 2022; Kuhn et al. 2023). Verbalized confidence approaches ask models to state their confidence explicitly (Tian et al. 2023; Xiong et al. 2024; Lin et al. 2022). Self-consistency approaches use agreement across multiple samples as a proxy for confidence (X. Wang et al. 2023). Each captures different information: Taubenfeld et al. (2025) demonstrate that verbalized confidence and self-consistency are related but dissociable signals.

Recent work has begun to address confidence quality through targeted training interventions. Y. Li et al. (2025) propose ConfTuner, using a Brier-score-style objective. V. Wang and Stengel-Eskin (2026) use self-generated distractors for calibration. Seo et al. (2025) identify answer-independent confidence generation as a primary driver of overconfidence and introduce ADVICE, a fine-tuning framework that promotes answer-grounded confidence estimation. Hager et al. (2025) introduce uncertainty distillation, using self-consistency estimates with post-hoc calibration as fine-tuning targets for verbalized confidence. This is the closest methodological comparison to PT-CSFT, though their approach inherits the bimodality limitation of self-consistency targets documented in § 4. Jang et al. (2025) apply confidence-supervised fine-tuning (CSFT) with self-consistency targets on LLaMA-3.2-3B, finding emergent self-verification behaviour on GSM8K and ARC-Challenge. Their approach uses the same SC-based targets that fail at larger scales in our Phase 0 analysis and does not employ probe-derived signals. Bani-Harouni et al. (2026) propose an RL approach with a logarithmic scoring rule reward, training models to express calibrated confidence without proxy confidence estimates. Their method requires RL infrastructure, whereas PT-CSFT requires only a linear probe and standard supervised fine-tuning (SFT).

2.2 Representation Probing and Confidence

Linear probing of LLM hidden states has revealed that internal representations carry substantial correctness information that is not expressed verbally. Burns et al. (2023) train probes to identify truthful statements from hidden activations. K. Li et al. (2023) identify "truth directions" in representation space that are causally related to model honesty. Marks and Tegmark (2024) demonstrate geometry-mediated truth representation across models.

Our work differs from these probing studies in a crucial respect: we use the probe not as a diagnostic tool but as a *training signal*. The probe's output becomes the target for supervised fine-tuning, creating a closed loop from internal representation to verbal output. This is, to

our knowledge, the first use of a model's own hidden-state probe as a fine-tuning target for confidence calibration.

2.3 Inference-Time Intervention

An alternative to fine-tuning is to intervene on internal representations at inference time. Inference-Time Intervention (ITI; K. Li et al. (2023)) modifies activations along identified truthfulness directions. CORAL (Miao, Cho, et al. 2026) uses logit-level corrections for multiple-choice QA (MCQA) calibration. Representation engineering (Zou et al. 2023) demonstrates that model behaviour can be steered by adding direction vectors to hidden states. Building on their finding that correctness and confidence directions are orthogonal (§ 1.3), Miao and Ungar (2026) propose a two-stage adaptive steering pipeline that reads the model's internal accuracy estimate and steers verbalized output to match. This substantially improves calibration and represents a more sophisticated approach than the single-direction additive steering we test in § 8.5.

2.4 Signal Detection Theory and Metacognition

Type-2 signal detection theory (SDT) provides the psychophysical framework for measuring metacognitive sensitivity (Galvin et al. 2003; Maniscalco and H. Lau 2012). AUROC_2 measures Type-2 discrimination, whether a confidence signal separates correct from incorrect responses. Meta-d' estimates the metacognitive efficiency of a confidence signal relative to optimal use of the available Type-1 information. The M-ratio (meta-d'/d') normalises metacognitive sensitivity by task difficulty (Fleming and H. C. Lau 2014).

2.5 Concurrent Metacognitive Fine-Tuning Approaches

Several concurrent works pursue the same goal of training LLMs to express calibrated uncertainty through different mechanisms. Park et al. (2026) use evolution strategies for metacognitive alignment (ESMA), measuring improvement via d'_type2, equivalent to the SDT framework used here. ESMA operates as a black-box optimiser over the policy, whereas PT-CSFT uses standard LoRA with probe-derived targets, making it simpler to implement and more transparent mechanistically. The logit readout finding, that binary training produces continuous calibrated uncertainty in the softmax distribution, is not explored in their work.

Damani et al. (2025) train with RL from calibration rewards. Their approach and ours converge on the same insight: models need explicit training signal about their own correctness. They differ in how that signal is derived (RL reward vs. probe readout) and delivered (RL policy gradient vs. SFT). RL-based methods are more flexible but require reward model design and RL infrastructure. PT-CSFT achieves comparable results with standard SFT requiring approximately 10 minutes of total computation on consumer hardware. Yaldiz et al. (2026) provide a complementary finding: RLVR (RL with verifiable rewards) systematically produces more overconfident models than supervised fine-tuning, supporting the use of SFT-based methods for calibration.

2.6 Discrimination versus Calibration

A growing body of work distinguishes calibration (knowing the average error rate) from discrimination (knowing which specific items are errors). Savage et al. (2025), Tao et al. (2025), and Taubenfeld et al. (2025) demonstrate that calibration and discrimination are weakly correlated in practice. Our work extends this observation with a constructive contribution: PT-CSFT with logit readout achieves both strong discrimination ($\text{AUROC}_2 = 0.862$) and strong calibration (ECE = 0.042) from binary training labels. The key insight is that discrimination

is the robust property (invariant to seed, recalibration method, and distribution shift), while calibration is tunable post-hoc.

2.7 Metacognition as a Control Layer

Yona et al. (2026) argue that metacognition, the ability to know what one doesn't know, is essential for trustworthy autonomous agents. The metacognitive control layer (routing retrieval, verification, and escalation based on confidence) is the missing component in current agent architectures. Our retrieval-augmented generation (RAG) gating experiment (§ 9.5) provides an empirical demonstration of this architecture using trained verbal confidence. The PT-CSFT confidence signal correctly stratifies items by accuracy (2.17× differential), enabling selective retrieval that outperforms both no-retrieval and always-retrieve baselines. The confidence signal transfers out-of-distribution (§ 9.6), making it deployable without per-dataset retraining.

Prior work in this line (Cacioli 2026c; Cacioli 2026f) has applied these metrics systematically to LLM confidence evaluation, establishing that most frontier models produce Invalid verbal confidence under validity screening, despite carrying informative signals internally. The present study extends this work from diagnosis to intervention.

3 Method

This section describes the PT-CSFT pipeline, from baseline characterisation through probe training, target derivation, and LoRA fine-tuning, followed by the evaluation protocol, model specifications, and pre-registration alignment.

3.1 Overview

The PT-CSFT pipeline has four stages:

1. **Baseline characterisation.** Greedy generation on held-out evaluation items. Record raw responses, parsed answers, parsed verbal confidence, correctness, hidden states at three layer positions, and first-token logit entropy.
2. **Probe training.** Fit a linear probe (L2-regularised logistic regression, 5-fold cross-validation) on hidden states from a separate calibration set. The probe learns to predict correctness from a single hidden-state vector.
3. **Target derivation.** Apply the trained probe to calibration items. The probe's P(correct) output, scaled to 0–100, becomes the confidence target for each item.
4. **Fine-tuning.** LoRA fine-tuning with the probe-derived confidence targets. The model learns to verbalise the probe's estimate of its own correctness.

This pipeline replaces the self-consistency target derivation used in Phase 0 (§ 4) with a probe-based target. The key advantage is that the probe reads the model's internal state directly, providing a continuous, model-specific signal that does not require multiple forward passes and avoids the bimodality problem that limited self-consistency targets at scale.

Terminology. We distinguish four properties of a confidence signal, which are empirically separable: (1) *discrimination* (AUROC_2): whether the signal ranks correct items above incorrect items; (2) *calibration* (ECE; Guo et al. (2017)): whether predicted probabilities match observed frequencies; (3) *variance recovery* (VRS validity): whether the signal uses a range of values rather than saturating at ceiling; (4) *metacognitive efficiency* (M-ratio): the signal's discrimination normalised by task difficulty. A signal can discriminate well while being poorly calibrated (e.g., always ranking correctly but at the wrong probability scale), or be well-calibrated at the aggregate level while failing to discriminate individual items. We report AUROC_2 as the primary metric for discrimination, ECE and Platt-scaled ECE (Platt 1999) for calibration, and VRS tier for variance recovery.

3.2 Data

All items are drawn from TriviaQA rc.nocontext validation (Joshi et al. 2017) and MMLU (Hendrycks et al. 2021). Items are shuffled with a fixed seed (42) and partitioned contiguously:

- **T-eval** (1,000 items): Held-out TriviaQA evaluation. No model sees these items during training.
- **T-cal** (2,000 items): TriviaQA calibration and training. Used for probe training and CSFT.
- **M-eval** (498 items): MMLU cross-benchmark evaluation, stratified across six cognitive domains following the taxonomy in Cacioli (2026d).

Partitions are identical to Phase 0 (Cacioli 2026b). Pairwise disjointness is verified programmatically. T-eval is held out from all training: no probe is fitted on T-eval items, no PT-CSFT targets are derived from T-eval items, and no model sees T-eval during fine-tuning. All reported metrics are computed on T-eval unless otherwise noted.

3.3 Probe Specification

For each model, we extract hidden states at three layer positions (first, middle, last transformer layer) and two token positions (pre-answer token, last-answer token) during greedy generation on T-cal, yielding a 3×2 grid of six probe configurations. Each probe is a logistic regression with L2 regularisation, C selected via 5-fold cross-validation. The probe is trained on T-cal hidden states with binary correctness labels and evaluated on T-eval.

The primary probe configuration is last layer × last answer token, following the convention that later layers carry more task-relevant information. The peak probe is typically at the middle layer, which we use for ablation experiments.

The probe is a supervisory signal, not an upper bound. A linear classifier on one hidden-state slice is a noisy lower-bound estimate of the linearly-decodable correctness information at that point in the network. The probe is constrained in three ways relevant to interpreting PT-CSFT results: (1) it is linear, so it cannot exploit feature interactions a richer readout could; (2) it reads one layer × one position, so it cannot integrate information across layers as the transformer's forward pass does at the confidence-token position; (3) it is trained on a finite sample (T-cal), so its decision boundary is an empirical estimate, not a population maximum. Recovery values above 100% (where verbal AUROC_2 exceeds the probe's training-set AUROC_2) are therefore not paradoxical: they indicate that the trained model accesses correctness-relevant structure that a single-layer linear projection does not capture, consistent with knowledge distillation, where a student model can exceed a teacher's soft labels by integrating richer features (§ 8.4 develops this point empirically with a 2-layer MLP probe baseline). Throughout the paper, probe AUROC_2 is reported as an empirical reference point rather than a theoretical ceiling.

3.4 Target Derivation

The fitted probe is applied to T-cal hidden states. For each item, the probe outputs $P(\text{correct}) \in [0, 1]$, which is scaled to a confidence target in $[0, 100]$ by multiplying by 100 and rounding to the nearest integer. Items where hidden states were not successfully cached (due to answer-token detection failure) are excluded from the training set.

3.5 Fine-Tuning and LoRA Sensitivity Analysis

The primary LoRA configuration is rank 16, scale 2.0 (equivalent to $\alpha = 32$), dropout 0.05, applied to five projection modules per transformer layer: k_proj and o_proj in the attention block, and gate_proj, up_proj, and down_proj in the MLP block. The query projection (q_proj), value projection (v_proj), and the language model head (unembedding matrix) are not modified.

This module selection follows the mlx_lm default and is held constant across all four model families (Gemma, Llama, Mistral, Qwen). Learning rate 2×10^{-4} with cosine schedule, 3 epochs, effective batch size 16 via gradient accumulation. The training input is the chat-formatted prompt with the model's original answer and the probe-derived confidence target in the assistant turn. Labels are masked on prompt tokens.

The choice to leave the language model head unmodified has a mechanistic consequence we develop in § 8: because the unembedding matrix is unchanged, any improvement in the logit readout reflects richer hidden-state routing to the confidence-token position, not a modified output mapping. The model already possessed the vocabulary geometry needed to express graded uncertainty. PT-CSFT teaches it to route the right content to the position where that vocabulary is read.

When the primary configuration failed on all LlamaForCausalLM models, we conducted a systematic LoRA hyperparameter ablation to determine whether the failure was architecture-intrinsic or configuration-specific. Three additional configurations were tested on Llama 8B, Mistral 7B, and Llama 70B:

Table 1: LoRA configurations tested. Scale is held constant at 2.0 across all configurations.

Config	Rank	LR	Scale	Rationale
primary	16	2×10^{-4}	2.0	Pre-registered default
gentle-lr	16	5×10^{-5}	2.0	Isolate learning rate effect
low-rank	4	2×10^{-4}	2.0	Isolate rank effect
gentlest	4	5×10^{-5}	2.0	Maximum restraint

Scale is held constant at 2.0 across all configurations (scale = α/rank), so the effective per-parameter LoRA contribution is constant regardless of rank. For Llama 70B, two additional configurations were tested after initial failure: r16/lr1$\times10^{-4}$ (mid-lr) and r32/lr1$\times10^{-4}$.

A shuffled-target control (seed 43) is trained with identical architecture, hyperparameters, and prompt template, but with confidence targets randomly permuted across items. The real-shuffled target correlation is verified below —r— ¡ 0.05 before training.

3.6 Evaluation

The fine-tuned model and shuffled-target control are evaluated with greedy generation on T-eval and M-eval. Correctness is assessed by flexible substring matching: a response is scored correct if any TriviaQA answer alias appears as a substring of the parsed answer, or vice versa. This accommodates the full-sentence answer format that PT-CSFT models produce (e.g., "The answer is Montreal" matching the alias "Montreal").

We report:

- **AUROC$_2$:** Primary metric. Area under the Type-2 ROC curve.
- **VRS screening:** L (ceiling rate), TRIN (top-response index of non-variability), r(confidence, correct). Models are classified as Valid, Indeterminate, or Invalid.
- **Accuracy and accuracy drop:** Task performance before and after fine-tuning.
- **ECE:** Expected Calibration Error (10 equal-width bins), measuring whether confidence values match empirical accuracy.
- **Probe-relative discrimination ratio:** (PT-CSFT AUROC$_2$ − 0.5) / (probe AUROC$_2$ − 0.5), measuring what fraction of the probe's discrimination is recovered in the verbal channel.
- **Paired-bootstrap CIs:** 10,000 resamples on the AUROC$_2$ delta (baseline vs PT-CSFT).
- **E10 (answer-unchanged items, policy-invariance diagnostic):** AUROC$_2$ computed on the subset of items where the baseline and PT-CSFT models produce the *same answer* (identical greedy answer token sequence). The label "E10" identifies this analysis

Table 2: Models evaluated. All run in bfloat16 on Apple M3 Ultra via MLX.

Model	Architecture class	Parameters	Layers	Hidden dim
Gemma 3 12B-it	GemmaForCausalLM	11.8B	48	3840
Gemma 3 27B-it	GemmaForCausalLM	27B	62	4608
Qwen 2.5 7B-Instruct	Qwen2ForCausalLM	7.6B	28	3584
Qwen 2.5 32B-Instruct	Qwen2ForCausalLM	32.5B	64	5120
Qwen 2.5 72B-Instruct	Qwen2ForCausalLM	72.7B	80	8192
Llama 3.1 8B-Instruct	LlamaForCausalLM	8.0B	32	4096
Llama 3.1 70B-Instruct	LlamaForCausalLM	70B	80	8192
Mistral 7B-Instruct v0.3	LlamaForCausalLM	7.2B	32	4096

in the pre-registration (https://osf.io/ngkwc). The diagnostic separates two reasons AUROC_2 might rise after training: genuine improvement in confidence calibration (the model is better at distinguishing its correct from incorrect answers) versus a policy-shift artifact (the model now answers a different set of items, changing the correct/incorrect mix in a way that artificially inflates AUROC_2). E10 restricts the comparison to items where the answer behaviour is held constant. This is a binding pre-registered control: if AUROC_2 does not improve on answer-unchanged items, the result is interpreted as a policy-shift artifact rather than a calibration improvement.

3.7 Models

We evaluate eight models from four model families spanning three architecture classes:

We distinguish *model families* (defined by training provider: Gemma, Qwen, Llama, Mistral) from *architecture classes* (defined by the HuggingFace model class: GemmaForCausalLM, Qwen2ForCausalLM, LlamaForCausalLM). Mistral is a separate model family from Llama but shares the LlamaForCausalLM architecture class.

All models are run in bfloat16 on Apple M3 Ultra (512GB unified memory) using the MLX framework. The Phase 0 result on Gemma 3 4B-it (Cacioli 2026b) is carried forward as a reference.

3.8 Pre-Registration and Alignment

The Phase 1 protocol was pre-registered on the Open Science Framework prior to baseline data collection (https://osf.io/ngkwc/). The pre-registration specified self-consistency-derived confidence targets on Gemma 3 12B-it and 27B-it. During execution, the method evolved to probe-derived targets (PT-CSFT), and the study expanded to eight models across four families. We document all deviations transparently. The earlier Phase 0 pre-registration (https://osf.io/mpcr5) covers the self-consistency-based work summarised in § 4.

Method deviation. The pre-registered target derivation used 10-sample self-consistency (n_{correct} mapped to confidence percentages). The executed method uses a linear probe on hidden states. Rationale: probe-derived targets directly read the model's internal correctness signal, avoiding the computational cost of multi-sample generation and the distributional biases of self-consistency (smoothing/sharpening fork). The self-consistency approach is reported separately as Phase 0 (Cacioli 2026b).

Scope expansion. The pre-registered design specified two models; the executed study evaluates eight models across four families at 7B–72B, plus domain generalisation to GSM8K and ARC-Challenge. The LoRA sensitivity analysis was conducted post-hoc after the pre-registered configuration failed on LlamaForCausalLM.

The full hypothesis alignment table is in Table 20 (Appendix C).

Table 3: Baseline characterisation. All models exhibit the ceiling prior.

Model	T-eval acc	AUROC$_2$ verbal	VRS	Conf. mean
Gemma 12B	0.735	0.678	Invalid	97.8
Gemma 27B	0.799	0.711	Invalid	97.7
Qwen 7B	0.655	0.577	Invalid	97.6
Qwen 32B	0.753	0.743	Invalid	95.7
Qwen 72B	0.836	0.767	Invalid	96.3
Llama 8B	0.766	0.692	Invalid	95.7
Llama 70B	0.874	0.724	Invalid	94.6
Mistral 7B	0.729	0.583	Invalid	98.9

Unplanned analyses: logit readout (§ 8), recalibration (§ 8.2), out-of-distribution (OOD) generalisation (§ 9.6), retrieval-augmented generation (RAG) gating (§ 9.5), ARC confonly (§ 7.3.1), 70B logit readout (§ 7.3.3), two-stage curriculum at 70B (§ 7.3.4). All clearly labelled as post-hoc.

4 Phase 0: Self-Consistency CSFT (Summary)

Before developing PT-CSFT, we tested whether self-consistency (SC), the agreement rate across multiple samples (X. Wang et al. 2023), could serve as a training signal for verbal confidence. Full details, including method, controls, and analysis, are reported in Cacioli (2026b). We summarise the key findings here.

On Gemma 3 4B-it, SC-CSFT (LoRA fine-tuning with SC-derived confidence targets from 10 samples per item) improved verbal AUROC$_2$ from 0.554 to 0.774. However, the resulting confidence distribution was effectively binary (values near 5% or 95%), and accuracy dropped 7.5 percentage points. A shuffled-target control confirmed the effect was target-dependent (AUROC$_2$ = 0.500).

The method failed at larger scales. Self-consistency becomes increasingly bimodal as models grow. The proportion of items at the extremes ($n_{\text{correct}} \in 0, 10$) rose from 84.6% at 4B to 91.4% at 12B to 94.5% at 27B. With nearly all training targets at 5% or 95%, the fine-tuned models learned the format but not the discrimination: verbal AUROC$_2$ was 0.540 at 12B and 0.499 at 27B (chance). The training signal loses its informativeness precisely when the model most needs it.

This failure established three requirements for a viable training signal: it must be (1) continuous rather than bimodal, (2) model-specific, and (3) available in a single forward pass. The linear probe on hidden states satisfies all three and provides the foundation for Phase 1.

5 Phase 1: Probe-Targeted CSFT Results

This section reports the Phase 1 results in four parts: baseline characterisation confirming the ceiling prior (§ 5.1), probe discrimination establishing the internal signal (§ 5.2), the primary PT-CSFT results across all models (§ 5.3), and controls (§ 5.4).

5.1 Baseline Characterisation

All eight models exhibited the ceiling prior under greedy generation with minimal confidence elicitation.

Every model is classified as Invalid under VRS screening. Confidence means range from 94.6 to 98.9. The ceiling prior is universal across architectures and scales. Qwen 72B has the highest

Table 4: Probe discrimination on T-eval. Gap = peak probe $AUROC_2$ (best across layers and positions, Table 9) minus baseline verbal $AUROC_2$. The metacognitive gap ranges from −0.013 (Qwen 72B) to +0.225 (Mistral 7B); seven of eight models show a positive gap.

Model	Verbal $AUROC_2$	Probe (primary)	Probe (peak)	Peak @	Gap (peak)
Gemma 12B	0.678	0.857	0.880	middle_last	+0.202
Gemma 27B	0.711	0.807	0.865	middle_pre	+0.154
Qwen 7B	0.577	0.762	0.790	middle_pre	+0.213
Qwen 32B	0.743	0.806	0.807	middle_last	+0.064
Qwen 72B	0.767	0.710	0.754	middle_pre	**−0.013**
Llama 8B	0.692	0.843	0.880	middle_pre	+0.188
Llama 70B	0.724	0.803	0.834	middle_last	+0.110
Mistral 7B	0.583	0.762	0.808	middle_pre	+0.225

verbal $AUROC_2$ (0.767), suggesting that larger Qwen models transmit more of their internal signal verbally, a finding that becomes significant in the probe analysis below.

5.2 Probe Discrimination

Linear probes trained on T-cal hidden states substantially outperform verbal confidence on all models.

The metacognitive gap ranges from −0.013 (Qwen 72B, where the peak probe does not exceed verbal) to +0.225 (Mistral 7B), with seven of eight models showing a positive gap. Middle-layer probes consistently outperform last-layer probes, suggesting that correctness information is represented most strongly at intermediate processing stages.

Qwen 72B is a qualitative outlier: the linear probe does not outperform verbal confidence. Within the Qwen family, the gap monotonically closes with scale (Qwen 7B: +0.213, Qwen 32B: +0.064, Qwen 72B: ≈0), suggesting that larger Qwen models increasingly transmit their linearly-decodable correctness signal through the verbal channel. Despite the zero gap, the verbal confidence remains Invalid under VRS screening: the discrimination exists but is compressed into a narrow range near the ceiling (mean 96.3, std 8.1). Under the pathway thesis, this is the interpretation we would expect: the routing bottleneck attenuates with scale within a family, but the residual ceiling-prior is a separate property of the verbal output distribution that persists even when the routing is effectively complete. The Qwen 72B result therefore does not contradict the routing-bottleneck framing; it constrains it. The framing applies most strongly at 7–32B scale; at the upper end of the Qwen family the bottleneck is weaker and the remaining gap is shape (range compression) rather than discrimination (rank ordering). We treat this as a boundary condition of the empirical scope rather than a counterexample to the thesis.

5.3 PT-CSFT: Primary Results (Best Configuration per Model)

Figure 2 summarises the results visually. Table 5 presents the best PT-CSFT result for each model. "Best" here means highest verbal $AUROC_2$ at the M-eval partition; for several models, alternative configurations preserve more accuracy at a modest $AUROC_2$ cost and may be preferable for deployment. The accuracy-AUROC trade-off across configurations is documented in the § 6 hyperparameter ablation. For GemmaForCausalLM and Qwen2ForCausalLM, both the pre-registered default (rank 16, lr 2×10^{-4}) and the gentle configuration (rank 16, lr 5×10^{-5}) succeed; the best is reported. For LlamaForCausalLM models at 7–8B, the best configuration was identified through the hyperparameter ablation described in § 6. Llama 70B did not respond to any standard single-stage configuration; a two-stage curriculum (§ 7.3.4) is the best 70B result.

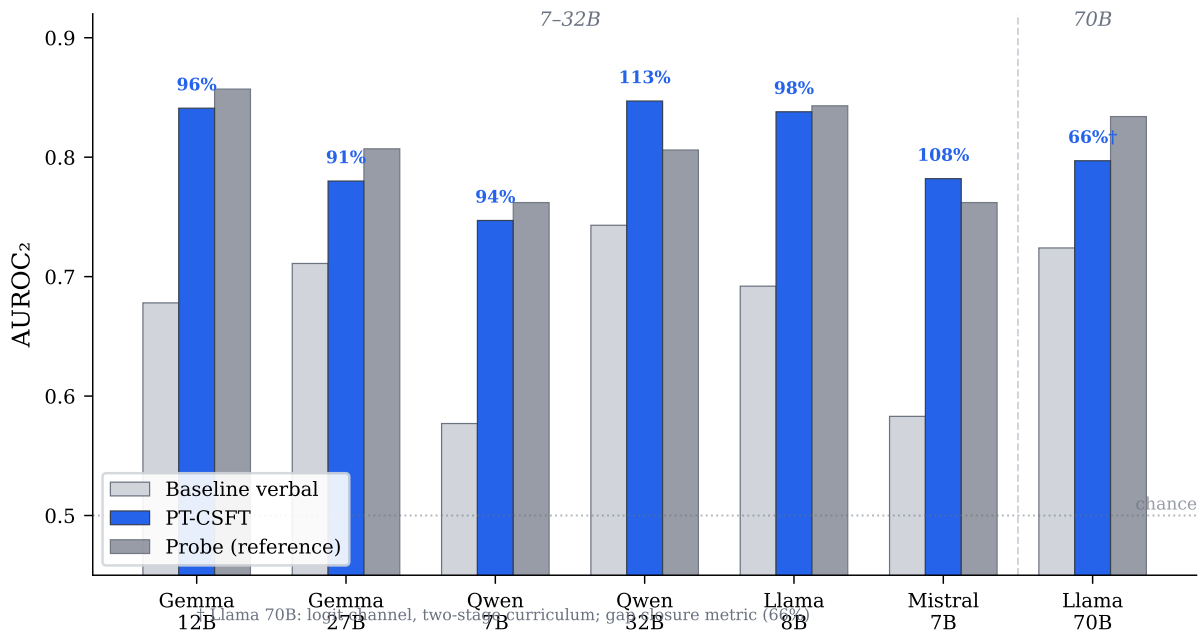

Figure 2: **Probe recovery across models.** Baseline verbal AUROC_2 (light grey), PT-CSFT verbal AUROC_2 (blue), and probe reference AUROC_2 (dark grey) for all models. Recovery percentages (labels above blue bars) range from 91% to 115% at 7–32B. Llama 70B uses the two-stage curriculum logit readout (66% gap closure). The dashed line separates the 7–32B regime (standard PT-CSFT) from the 70B regime (curriculum required).

Table 5: PT-CSFT primary results (best configuration per model). Recovery = (PT-CSFT − 0.5) / (Probe − 0.5). Llama 70B uses the two-stage curriculum (§ 7.3.4).

Model	Best config	AUROC_2	Recovery	Acc drop	ECE	VRS
Gemma 12B	r16/lr2e-4	**0.841**	96%	−9.1pp	0.118	**Valid**
Gemma 27B	r16/lr2e-4	**0.780**	91%	−8.4pp	0.078	Indeterm.
Qwen 7B	r16/lr2e-4	**0.801** ± **0.022**‡	115%	−2.1pp	0.080§	Valid
Qwen 32B	r16/lr5e-5	**0.847**	113%	−0.2pp	—	—
Llama 8B	r16/lr5e-5	**0.838**	98%	−1.5pp	0.128	Invalid
Mistral 7B	r4/lr5e-5	**0.782**	108%	−0.6pp	0.105	Valid
Llama 70B†	Two-stage curriculum	**0.797** (logit) / **0.762** (text)	66% closure (logit)	−0.1pp	—	Invalid (text) / **Valid** (logit)

Numbers above use the last-layer probe as the target. A middle-layer ablation (Appendix E) reaches higher probe recovery (Gemma 12B 102%, Qwen 7B 123%); both the last-layer and middle-layer targets yield VRS-Valid confidence on both models. Layer choice is a tunable target derivation parameter. †*Llama 70B: single-stage standard LoRA failed across seven configurations (§ 6.3). Two-stage curriculum (balanced confidence-only followed by probe-derived continuous targets) is the best 70B result in this study (§ 7.3.4). Recovery is reported as gap closure (probe-baseline gap = 0.110; curriculum logit closes 0.073, or 66%).*

‡*Qwen 7B is the 6-seed mean (± s.d.); seed 42 is a low outlier (0.756). Recovery normalises against the last-layer TriviaQA probe (0.762, Table 4); values >100% arise as in § 3.3 and § 8.4.*

§*ECE computed on seed 42; AUROC_2 and VRS (Valid) are stable across all 6 seeds.*

Seed stability. The TriviaQA rows in this table report single best-configuration seeds for Gemma 12B/27B and the 6-seed mean for the three non-Gemma models. Multi-seed replication on the non-Gemma models (6 seeds each, same configuration) confirms stability: Llama 8B gentle-lr $\text{AUROC}_2 = 0.836 \pm 0.011$ (spread 0.030); Mistral 7B gentlest $\text{AUROC}_2 = 0.771 \pm 0.017$ (spread 0.052); Qwen 7B primary $\text{AUROC}_2 = 0.801 \pm 0.022$, range [0.756, 0.830] (spread 0.074). Qwen 7B is the most seed-sensitive of the three: seed 42 is a low outlier (0.756) relative to the other five (0.804–0.830). No seed in any model produces a catastrophic failure, and accuracy

variance is small (Qwen 7B s.d. ±0.014, the largest of the three). Qwen 7B is also the only one of the three whose VRS tier is stable across all seeds (Valid in all 6); Llama 8B and Mistral 7B cross VRS boundaries as TRIN varies (§ 9.1). The 10-seed validation of Gemma 12B GSM8K (§ 7.3.2) provides a complementary seed-level characterisation in the domain-generalisation setting. The Llama 70B curriculum row is a single training run (Stage 1 then Stage 2).

At 7–32B scale, PT-CSFT produces large improvements in verbal confidence discrimination across all three architecture classes. Recovery rates range from 91% to 115%, meaning the verbal channel captures most or all of the information that was previously accessible only through a linear probe on hidden states. Where measured, the improvement extends beyond discrimination to calibration: ECE drops substantially in the five models with computed values (Qwen 32B ECE was not recorded for the gentle configuration). Three models exceed 100% recovery (Qwen 7B 115%, Qwen 32B 113%, Mistral 7B 108%), the verbal channel matching or surpassing the linear probe (§ 8.4); Qwen 32B pairs high recovery with the smallest accuracy drop (0.2pp) of any model in the study.

These improvements are not policy-shift artifacts. On items where the model gives the same answer before and after training (the E10 diagnostic, a binding pre-registered control), AUROC_2 still improves substantially:

Table 6: E10 diagnostic (AUROC_2 restricted to answer-unchanged items), for the models where E10 data was collected. Positive deltas confirm calibration improvement, not policy shift. Qwen 32B is omitted (E10 not recorded for its gentle configuration).

Model	E10 Δ
Gemma 12B	**+0.184**
Gemma 27B	**+0.074**
Qwen 7B	**+0.209**
Llama 8B	**+0.138**
Mistral 7B	**+0.197**

The E10 deltas are positive and statistically significant under paired bootstrap for every model where E10 was collected (Table 6), ruling out the policy-shift interpretation.

For models where PT-CSFT produces sufficient confidence variance, we computed meta-d' and M-ratio using the Maniscalco and H. Lau (2012) maximum likelihood estimator (MLE) via metadpy (Legrand and Ruby 2022) with four quantile-based confidence bins. Qwen 7B yields d' = 1.049, meta-d' = 0.883, and M-ratio = 0.842, indicating that verbal confidence after PT-CSFT captures 84.2% of the metacognitive information theoretically available given its Type-1 performance. Baseline models preclude this computation entirely due to degenerate confidence distributions.

5.4 Controls

Shuffled-target control. The shuffled-target model shows no improvement over baseline for any model (e.g., Gemma 12B shuffled $\text{AUROC}_2 = 0.501$, VRS = Invalid). The shuffled-minus-baseline delta is consistent with zero across all conditions, confirming that the PT-CSFT improvement is attributable to the probe-derived targets, not the LoRA fine-tuning procedure or format regularisation.

Calibration metrics. Llama 8B low-rank (r4/lr2e-4) achieves ECE = 0.047 and VRS-Valid confidence (L = 0.169, TRIN = 0.419, r = 0.526), the best calibration of any model in this study. Qwen 7B achieves ECE = 0.080 after PT-CSFT. Mistral 7B gentle-lr achieves ECE = 0.086 with VRS-Indeterminate confidence (TRIN = 0.697 ≥ 0.60); note that Table 5 reports Mistral's gentlest configuration (highest AUROC_2), whereas the calibration figure here is for

gentle-lr. For the original Llama 8B configuration (r16/lr2e-4), both metrics worsen, consistent with the AUROC_2 and VRS findings for that configuration.

6 LoRA Sensitivity Analysis

The pre-registered LoRA configuration failed on all LlamaForCausalLM models, raising the question of whether the failure was architecture-intrinsic or configuration-specific. This section reports a systematic hyperparameter sweep that resolves the question. Full per-configuration ablation tables are in Appendix D.

6.1 Motivation

The pre-registered LoRA configuration (rank 16, lr 2×10^{-4}) succeeded on GemmaForCausalLM and Qwen2ForCausalLM but failed on all three LlamaForCausalLM models: Llama 8B showed 3.8% probe recovery, Llama 70B showed a significant negative delta, and Mistral 7B diverged entirely during training (loss spiked to 11.7). The failure appeared to cluster by architecture class, suggesting a fundamental incompatibility. However, the failure could alternatively be configuration-specific. LlamaForCausalLM might simply require different hyperparameters. We tested this by sweeping rank and learning rate on all three LlamaForCausalLM models.

6.2 Results at 7–32B Scale

Llama 8B. All three gentler configurations succeed, with the gentle-lr configuration (r16/lr5e-5) nearly matching the best Gemma result.

The full ablation results are in Table 21 (Appendix D).

The learning rate is the dominant factor. Reducing lr from 2×10^{-4} to 5×10^{-5} at rank 16 transforms recovery from 4% to 98% and accuracy drop from 10.7pp to 1.5pp. Reducing rank to 4 at the original lr also helps substantially (90.7% recovery) but with less accuracy preservation. The best calibration (ECE = 0.047, VRS Valid) is achieved at low-rank with the higher lr. All three gentler configurations pass the E10 diagnostic with positive deltas, confirmed by paired bootstrap (all p ¡ 0.05).

Mistral 7B. The model that fully diverged at r16/lr2e-4 achieves 108% recovery with gentle hyperparameters.

The full Mistral ablation is in Table 22 (Appendix D).

The same pattern holds: lr 5×10^{-5} enables full recovery regardless of rank, while lr 2×10^{-4} either diverges (rank 16) or degrades accuracy (rank 4). Mistral's best result (gentlest, 108% recovery) exceeds the probe, indicating the fine-tuned verbal channel accesses richer structure than the linear probe captures.

6.3 Results at 70B Scale (Standard Single-Stage LoRA)

Llama 70B presents a qualitatively different challenge. Seven configurations spanning rank 4–128 and lr 5×10^{-5} to 2×10^{-4} were tested; none broke the ceiling prior under single-stage standard LoRA. A two-stage curriculum approach that addresses the underlying causes of failure is reported in § 7.3.4.

The full seven-configuration ablation is in Table 23 (Appendix D) and visualised in Figure 3.

The ceiling prior at 70B is deeply entrenched: confidence mean remains above 96 with standard deviation below 12 across all configurations. The best result (gentlest, AUROC_2 = 0.740) represents 79% recovery but retains an Invalid VRS and a near-zero E10 delta (−0.012), indicating that the AUROC_2 improvement comes from marginal use of tiny confidence variance rather than genuine ceiling-breaking.

Increasing LoRA rank to 128 (2.35% trainable parameters, compared to 0.07% at rank 4) degrades accuracy by 20.9pp and pushes AUROC_2 to chance (0.497, ceiling rate 99.5%). The high-capacity adapter collapses confidence to ceiling while damaging task performance, the worst of both worlds. The failure is symmetric with the lr 2×10^{-4} failure on Llama 8B (§ 6.2): too much intervention strength relative to the training signal produces destructive interference rather than targeted modification of the verbalization pathway.

The Llama 70B result establishes that standard single-stage LoRA-based PT-CSFT is insufficient at this scale regardless of rank. The proportional LoRA capacity ranged from 0.07% (rank 4) to 2.35% (rank 128), spanning a 34× range, without any configuration achieving genuine ceiling-breaking. We diagnose the failure as having two compounding causes (data bias at 87% baseline accuracy, and gradient dilution under standard SFT loss) and develop interventions that address both in § 7.3.3 (balanced confidence-only) and § 7.3.4 (two-stage curriculum).

The gentlest configuration achieves the best text-channel AUROC_2 among standard LoRA configurations (0.740, above the balanced confidence-only text result of 0.682 reported in § 7.3.3) but remains below the unmodified baseline (0.724) and produces VRS-Invalid confidence (mean 98.4, std 3.0). We tested whether the logit readout could recover further signal from the gentlest adapter (101 single-token confidence values exist in Llama 70B's vocabulary). The logit readout on the gentlest adapter yields $\text{AUROC}_2 = 0.679$, below baseline. The text-channel improvement in the gentlest configuration reflects marginal use of compressed confidence variance rather than the kind of routed signal that the logit readout recovers in the two-stage curriculum. The two-stage curriculum that resolves this 70B failure is developed in § 7.3.3–§ 7.3.4.

7 Domain Generalisation

The TriviaQA results establish PT-CSFT on factual recall. This section tests whether the method generalises to mathematical reasoning (GSM8K) and science reasoning (ARC-Challenge), characterises what works and what fails across domains and scales, and consolidates the findings into a diagnostic framework for deployment.

7.1 Motivation

We use a probe-first gating strategy for each new domain: if a linear probe on hidden states achieves AUROC_2 ¿ 0.65, the internal signal is sufficient for PT-CSFT to operate on. This gate is necessary because PT-CSFT inherits its training signal from the probe; without a discriminative probe, target derivation cannot succeed regardless of fine-tuning recipe.

The remainder of this section is organised as follows. § 7.2 reports the probe gate results on GSM8K (Cobbe et al. 2021) and ARC-Challenge (Clark et al. 2018). § 7.3 reports what works: confonly on verifiable tasks (§ 7.3.1), end-to-end on reasoning tasks (§ 7.3.2), balanced confonly and the two-stage curriculum at 70B (§ 7.3.3–§ 7.3.4), and target layer choice (§ 7.3.5). § 7.4 reports what fails and why: the verifiability boundary, zero-shot transfer, multi-task SFT, end-to-end at 70B on factoid, and cross-benchmark format mismatch. § 7.5 consolidates results into a diagnostic framework. The logit readout, which is the universal rescue mechanism wherever the text channel is insufficient, is sufficiently central to warrant its own section (§ 8).

7.2 Probe Gate Checks: Does the Internal Signal Exist?

GSM8K (mathematical reasoning). Gemma 12B achieves 88.8% accuracy on GSM8K with chain-of-thought prompting. Verbal AUROC_2 is 0.546 (near chance), confirming that the ceiling prior extends to mathematical reasoning. A linear probe on middle-layer hidden states achieves $\text{AUROC}_2 = 0.769$ (the *gate probe*), yielding a metacognitive gap of +0.223, larger than the TriviaQA gap for the same model (0.179). The probe gate is passed with substantial margin.

Table 7: GSM8K probe gate check (Gemma 12B). The probe gate is passed at the middle layer.

Layer position	Probe $AUROC_2$	Gap vs. verbal
first_last	0.650	+0.104
middle_last	**0.769**	**+0.223**
last_last	0.763	+0.217

ARC-Challenge (science reasoning). We evaluated two models. Gemma 12B achieves 88.8% accuracy on ARC with verbal $AUROC_2 = 0.560$. The probe signal is weak: the best probe achieves $AUROC_2 = 0.653$ at the last layer, marginally above the 0.65 gate threshold. However, the high accuracy leaves only 53 incorrect items in the 472-item evaluation set, limiting the probe's ability to learn a discriminative pattern.

Qwen 7B achieves 78.2% accuracy on ARC, lower than Gemma 12B, providing approximately twice as many incorrect items. Verbal $AUROC_2$ is 0.620. The best probe achieves $AUROC_2 = 0.834$ at the first layer, with a metacognitive gap of +0.214, well above the gate threshold.

Table 8: ARC-Challenge probe gate comparison across models.

Model	Accuracy	Verbal $AUROC_2$	Best probe	Layer	Gap
Gemma 12B	0.888	0.560	0.653	last	+0.093
Qwen 7B	0.782	0.620	**0.834**	first	**+0.214**

Methodological implications. Two findings from the gate checks bear on PT-CSFT deployment. First, probe discriminability is not solely a function of the benchmark or the model size. It depends on the interaction between model accuracy and error-set size. Gemma 12B's marginal probe signal on ARC is a consequence of its high accuracy leaving too few errors, not an ARC-specific limitation. The gating check should be model-specific, not benchmark-specific. Second, the optimal probe layer varies across domains: middle for TriviaQA and GSM8K, first for ARC on Qwen 7B (first-layer probe 0.834 vs last-layer 0.797). Target derivation should include a layer sweep rather than assuming a fixed layer position.

Three cognitive domains are now confirmed viable:

Table 9: Probe gate results across three cognitive domains. ‡The TriviaQA row covers seven of eight models; Qwen 72B is a probe outlier (peak probe 0.75, verbal 0.77, gap −0.01) and is discussed in § 5.2.

Domain	Benchmark	Model	Best probe	Verbal $AUROC_2$	Gap
Factual recall	TriviaQA	7 of 8 models‡	0.76–0.88	0.58–0.74	+0.06–0.19
Mathematical reasoning	GSM8K	Gemma 12B	0.769	0.546	+0.223
Science reasoning	ARC	Qwen 7B	0.834	0.620	+0.214

7.3 What Works

This subsection catalogues the conditions under which PT-CSFT succeeds: confidence-only training on verifiable tasks (§ 7.3.1), end-to-end training on reasoning tasks (§ 7.3.2), the balanced confidence-only and two-stage curriculum interventions at 70B (§ 7.3.3–§ 7.3.4), and the choice of target layer (§ 7.3.5).

7.3.1 Confidence-Only on Verifiable Tasks

ARC-Challenge (Qwen 7B). Given the strong ARC gate result, we trained a confidence-only PT-CSFT adapter. Following the two-pass architecture described in § 7.4.1, the model generates an answer without the adapter, then rates its confidence with the adapter loaded. Training used 450 items from a 500-item calibration split with probe-derived confidence targets, LoRA rank 16, lr 5×10^{-5}, and 1,350 iterations. Evaluation on a held-out 500-item eval split:

Table 10: ARC-Challenge confidence-only PT-CSFT (Qwen 7B).

Metric	Baseline	PT-CSFT confonly
Accuracy	0.824	**0.824**
$AUROC_2$	0.609	**0.813**
Conf mean	—	86.8
Conf std	—	33.7
Probe $AUROC_2$ (eval)	—	0.725
L (ceiling)	—	0.868
TRIN	—	0.808
r(conf, correct)	—	+0.689
VRS	Invalid	Invalid

PT-CSFT improves $AUROC_2$ by +0.204 with zero accuracy loss. The verbal output (0.813) exceeds the probe that generated its training targets (0.725), a pattern discussed in § 8.4. (This experiment uses a held-out 500-item ARC split distinct from the § 7.2 gate-check set, so baseline accuracy (0.824) and verbal $AUROC_2$ (0.609) differ from the gate-check values in Table 8; the 0.725 training-target probe is the last-layer probe on this split, not the first-layer gate probe of 0.834.) The signal discriminates strongly — $AUROC_2 = 0.813$ and confidence–correctness correlation $r = 0.689$, the strongest correlation in the study — yet remains VRS-Invalid because the confidence distribution is peaked near ceiling (L = 0.868, TRIN = 0.808). This is the clearest single instance of the discrimination/shape dissociation we discuss in § 9.1: VRS keys on distributional shape, $AUROC_2$ on discrimination, and the two come apart here as starkly as anywhere in the paper. The remaining variation nonetheless carries substantial discriminative information. This extends PT-CSFT to science reasoning as a third cognitive domain.

TriviaQA confidence-only. The two-pass confidence-only architecture also works on TriviaQA for models where it has been tested (Llama 8B: text $AUROC_2 = 0.833$, logit $AUROC_2 = 0.846$, where the logit readout reads the softmax distribution over confidence-token digits rather than parsing the argmax text output; mechanism developed in § 8), preserving accuracy by construction since the answer generation pass runs without the adapter.

7.3.2 End-to-End on Reasoning Tasks

The two-pass confidence-only approach fails on computation tasks (§ 7.4.1). Following Jang et al. (2025), who showed that end-to-end CSFT produces calibrated confidence at 3B scale, we tested single-pass training on GSM8K, where the model generates CoT, answer, and confidence together in one pass with the adapter loaded. The training format appends "Confidence: X" (X $\in$ 0, 9, binary target) to the model's chain-of-thought. The coarse 0–9 scale is required because Gemma 3's tokenizer splits multi-digit numbers into per-character tokens.

Training on 282 balanced items (50% correct, 50% incorrect) with LoRA rank 16 on Gemma 12B:

At lr = $\mathbf{2}\times10^{-4}$ (initial). Single-seed argmax $AUROC_2 = 0.694$ [0.629, 0.756], a +0.148 improvement over the 0.546 baseline. A 10-seed validation yields mean argmax $AUROC_2 = 0.637 \pm 0.062$ (range: 0.500–0.711). End-to-end training reliably breaks the ceiling but with

substantial seed-dependent variance. One seed collapsed entirely ($\text{AUROC}_2 = 0.500$), two exhibited partial ceiling retention. Accuracy was stable across all seeds (0.877 ± 0.003).

At lr = 5×10^{-5} (gentle). A 10-seed validation yields mean argmax $\text{AUROC}_2 = 0.743 \pm 0.062$ (range: 0.580–0.803), with four seeds exceeding the probe (0.769) from argmax alone. No seeds collapse. Every seed improves compared to its lr = 2×10^{-4} counterpart (gains of +0.012 to +0.235).

The argmax output is binary (TRIN = 1.0). The model outputs "0" or "9" exclusively. However, the softmax distribution over confidence tokens tells a different story: the logit readout achieves $\text{AUROC}_2 = 0.862 \pm 0.013$ (95% CI [0.855, 0.871]; 10,000-resample bootstrap), exceeding the probe (0.769) by +0.093 per seed, with all 10 seeds above 0.85. The logit readout is the study's strongest result. Its mechanism is detailed in § 8.

Cross-family replication on Qwen 7B (Pareto trade-off across recipe space). To test whether the E2E logit readout pattern generalises beyond Gemma, we applied the same recipe to Qwen 7B GSM8K at two learning rates.

Table 11: Cross-family replication of E2E training on Qwen 7B GSM8K.

LR	Accuracy	Acc drop	Text AUROC_2	Logit AUROC_2	Δ vs. baseline (logit)
5×10^{-5}	0.404	−39.4pp	0.500	0.740	+0.240
1×10^{-5}	0.666	−13.2pp	0.500	0.641	+0.141

At lr = 5×10^{-5}, the logit readout achieves $\text{AUROC}_2 = 0.740$ versus text $\text{AUROC}_2 = 0.500$ (chance), a Δ of +0.240. The readout-channel advantage replicates across architectures: the logit readout exposes correctness signal that the text channel does not transmit, in both GemmaForCausalLM and Qwen2ForCausalLM. However, accuracy dropped substantially under the E2E recipe (0.798 baseline → 0.404 post-PT-CSFT), in contrast to Gemma 12B where accuracy was stable across all 10 seeds. The dropped predictions are well-formed numerical answers but systematically truncated relative to gold (for example, pred=27 vs gold=89; pred=4 vs gold=34), suggesting that under Qwen's tokenizer or chat template, E2E training perturbs the chain-of-thought completion rather than the confidence assignment.

At lr = 1×10^{-5}, accuracy is better preserved (0.666, −13.2pp) but the logit advantage shrinks (0.641 vs 0.500 text, $\Delta = +0.141$). The two learning rates define a Pareto trade-off: gentler training preserves more accuracy at the cost of weaker logit discrimination, while more aggressive training extracts the logit signal at the cost of substantial accuracy loss. Both LRs produce text AUROC_2 at chance (0.500), confirming the logit-over-text channel advantage at both points on the Pareto front.

Two cross-architecture conclusions follow. First, the logit-readout *mechanism* generalises across architectures. The text channel collapses to argmax on both Gemma and Qwen, and the logit distribution exposes a correctness signal that text parsing cannot recover. Second, the E2E training *recipe* is architecture-dependent. Where Gemma 12B preserves accuracy at lr = 5×10^{-5}, Qwen 7B does not, and the Pareto front does not contain a configuration matching Gemma's combination of strong accuracy preservation and strong logit discrimination. Per-model recipe validation is necessary. Whether a different target schema, learning rate schedule, or training duration moves Qwen 7B's Pareto front toward Gemma's regime is left to follow-up work.

7.3.3 Balanced Confidence-Only at Scale (Llama 70B, Stage 1)

Standard single-stage PT-CSFT failed on Llama 70B across seven LoRA configurations (§ 6.3). The failure reflects two compounding causes: at 87% baseline accuracy, 87% of probe targets are near 100% (data bias), and the confidence tokens receive negligible gradient in full-response formats (gradient dilution). Balanced confidence-only training addresses both. Items are binned

by probe-derived target and resampled to equal representation, and the confidence-only format concentrates 100% of training gradient on the confidence tokens.

Training on 353 balanced items (5 bins × 70 items, 58 gradient updates, rank 16, lr 5×10^{-5}):

Table 12: Balanced confidence-only training on Llama 70B (Stage 1).

Metric	Baseline	Std. PT-CSFT (best of 7)	Bal. confonly
Accuracy	0.874	0.867 (gentlest)	**0.873**
$AUROC_2$ (text)	0.724	0.740 (gentlest)	0.682
Conf mean	94.6	98.4	**63.8**
VRS	Invalid	Invalid	**Indeterm.**
L (ceiling)	—	0.985	**0.638**

The ceiling prior was broken. VRS transitioned from Invalid to Indeterminate at 70B for the first time in this study — the first 70B condition in which the verbal channel transmits non-trivial discrimination signal. Concretely, three things changed together. L dropped from 0.985 (Invalid, ceiling-locked) to 0.638 — above 0.40 and so in Indeterminate territory, but below the Invalid threshold of 0.70; TRIN fell to 0.349, below the 0.60 Indeterminate gate and so not itself disqualifying; and r reached 0.274, above the Invalid threshold of 0.05. The classification is Indeterminate, driven by the residual ceiling rate. The response distribution stopped being a near-delta at 100% and became a graded distribution covering the 0–100 range. In substantive terms: under baseline, almost every answer the model gave reported the same near-ceiling confidence regardless of whether the answer was correct, so confidence carried no information about correctness; after Stage 1 balanced confonly, the model began varying its reported confidence in a way that correlated with whether it was right. Text-based $AUROC_2$ decreased from 0.724 to 0.682, but the logit readout (reading the softmax distribution over Llama's 101 single-token confidence values) achieves $AUROC_2 = 0.746$, exceeding both the text confidence (0.682) and the baseline (0.724). This is a modest +0.022 improvement over baseline. The probe-verbal gap at 70B is 0.110 in absolute terms (probe 0.834 vs verbal 0.724); the balanced confonly logit readout closes 0.022 of this gap, a 20% closure.

The balanced confonly result is *Stage 1* of a two-stage process. The ceiling is broken but the routing remains under-trained. The model uses the confidence range but does not yet route correctness-relevant content to the confidence position with high fidelity, which a refinement stage using continuous probe-derived targets can address (§ 7.3.4).

Sensitivity to training volume. A v2 experiment with 3× more data (1,000 items, 200 per bin, 168 gradient updates) regressed: $AUROC_2$ dropped to 0.538, confidence mean returned to 98.0, and VRS reverted to Invalid. The lowest-confidence bin contained only 13 unique items, oversampled 15.4× to reach the target of 200. The model memorised those items rather than learning the distributional shift. The v1 result succeeded because it underfitted. Brief training learns the general pattern without overfitting to specific items. The critical factor is the ratio of unique items to oversampling rate in low-confidence bins.

7.3.4 Two-Stage Curriculum at 70B (Stage 2: Routing Refinement)

The single-stage failure (§ 6.3) and the Stage 1 ceiling break (§ 7.3.3) together motivate a two-stage curriculum. Stage 1 (balanced confonly) unblocks the channel by addressing data bias and gradient dilution (Figure 3). Stage 2 (probe-derived continuous targets) then refines what content is routed to the now-open confidence position. Under the pathway thesis, this decomposition is principled. Routing repair at 70B is decomposable into *opening the channel* (Stage 1) and *training what arrives* (Stage 2).

Method. Stage 1 adapter loaded as resume point. Stage 2 training uses 1,778 T-cal items with continuous targets from the middle layer × last-answer-token probe (T-eval probe

Table 13: Two-stage curriculum results on Llama 70B (full T-eval, $n = 1000$).

Metric	Baseline	Bal. confonly (S1)	Curriculum (S1+S2)
Accuracy	0.874	0.873	**0.873**
Text $AUROC_2$	0.724	0.682	**0.762**
Logit $AUROC_2$	—	0.746	**0.797**
Text conf mean	94.6	63.8	95.4
Text conf std	—	—	16.2
Logit conf mean	—	—	84.6
Logit conf std	—	—	10.4
L (text, $\geq$95%)	—	0.638	0.968
VRS (text)	Invalid	Indeterm.	Invalid
VRS (logit)	—	—	**Valid**

$AUROC_2 = 0.834$), formatted as confidence-only (gradient concentrated on confidence tokens). Probe scores distribution: mean 89.1, std 19.3, quantiles 10%/50%/90% = 65/97/99. Training: 80 iters, LoRA rank 16, lr 5×10^{-5}, batch size 2 (effective 16 with gradient accumulation). Approximately 7 minutes wall-clock on M3 Ultra. Validation loss trajectory clean: 2.503 $\rightarrow$ 0.446 across 80 iters, no memorisation spike.

Results (full T-eval, n = 1000):

Gap closure. The probe-verbal baseline gap is 0.110 (probe 0.834 vs baseline 0.724). The curriculum closes 0.073 of this gap on the logit channel (66% closure) and 0.038 on the text channel (35% closure). Compared to Stage 1 alone, Stage 2 adds +0.051 on logit (0.746 $\rightarrow$ 0.797) and +0.080 on text (0.682 $\rightarrow$ 0.762). For the first time in this study, a 70B intervention produces text-channel $AUROC_2$ above baseline (+0.038). Accuracy is fully preserved (0.873 vs 0.874 baseline).

Pathway-thesis interpretation.

The curriculum result demonstrates the routing-not-mapping decomposition. Stage 1 opens the channel: balanced data plus concentrated gradient on the confidence position break the ceiling prior. Stage 2 refines what arrives: continuous probe-derived targets train the routing to bring graded correctness content to the now-open position. Single-stage standard LoRA fails because it attempts both jobs at once with insufficient gradient signal on the confidence token. The two-stage decomposition predicts and produces a substantially larger improvement than either stage alone.

Honest framing of the 70B result. The curriculum is the best 70B result in this study. It substantially exceeds the previous best (balanced confonly logit, 0.746) and is the only 70B result with text $AUROC_2$ above baseline. The text channel remains ceiling-locked: 96.8% of items report $\geq$95% confidence (L = 0.968, TRIN = 0.968), classifying as Invalid despite text $AUROC_2 = 0.762$ above baseline. On the same model, same items, same training run, the logit readout achieves the first VRS-Valid confidence signal at 70B in this study (L = 0.021, TRIN = 0.430, r = 0.434, $AUROC_2 = 0.797$). The dissociation is mechanistically informative: the graded correctness content reached the confidence position (the logit distribution is Valid-quality), but the argmax that produces the verbal surface collapses that distribution onto a high digit, because its mode sits at a high confidence token even when the expected value is graded (§ 8.1). The 70B bottleneck that survives PT-CSFT is therefore the discrete decoding step, not the routing — the routed signal is present and Valid; the text surface cannot express it. However, the gap is not fully closed. 0.037 $AUROC_2$ remains between the curriculum logit and the probe (0.797 vs 0.834). The intervention works at 70B, with reduced effectiveness compared to 7–32B where recovery reaches 91–115%. Whether the residual gap reflects deeper RLHF entrenchment, insufficient LoRA capacity for a fully routed pathway at this scale, or a fundamental scale ceiling on LoRA-based routing intervention is unresolved. We report this as partial closure via two-stage training, where standard single-stage LoRA failed across seven

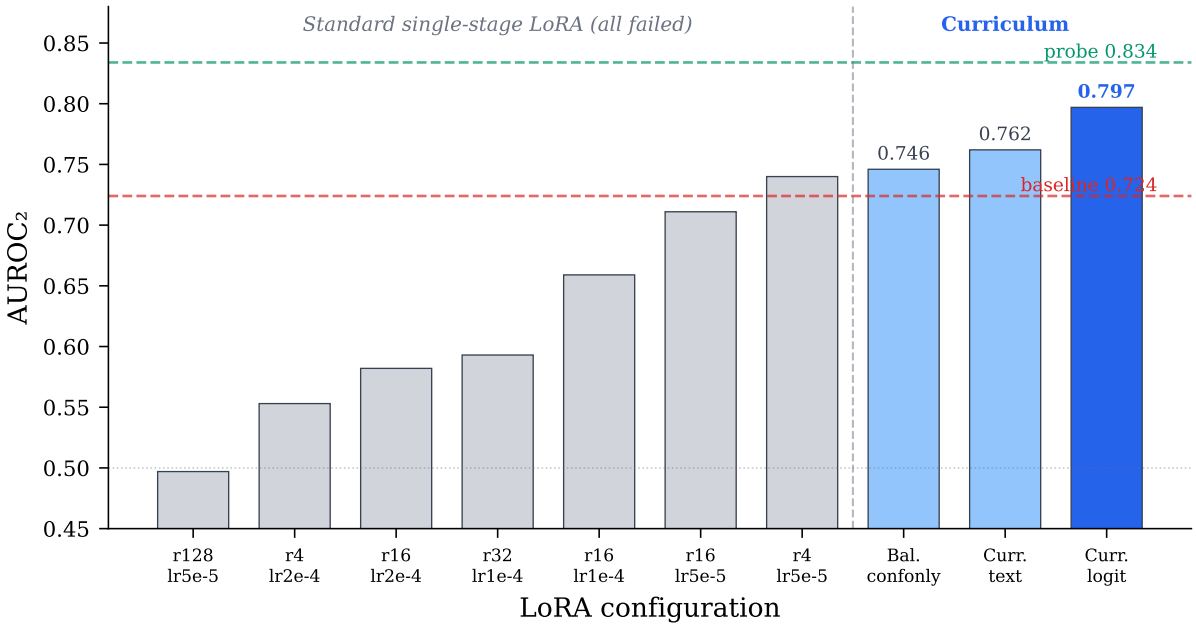

Figure 3: **Llama 70B scaling story.** Seven standard single-stage LoRA configurations (grey) all fail to close the probe–verbal gap; the best (r4/lr5e-5) reaches $\text{AUROC}_2 = 0.740$ but remains below baseline (dashed red). The two-stage curriculum (blue) succeeds: balanced confidence-only (Stage 1) opens the channel, and Stage 2 refines routing to logit $\text{AUROC}_2 = 0.797$ (66% gap closure toward probe at 0.834, dashed green).

configurations, not as evidence that PT-CSFT scales unchanged to 70B.

7.3.5 Target Layer Choice

The pre-registered probe configuration uses the last layer of the model as the source of training targets; the peak probe across all models on TriviaQA is at the middle layer, raising the question of whether middle-layer targets would improve PT-CSFT outcomes. We tested this on Gemma 12B and Qwen 7B. Middle-layer targets yield higher probe recovery than last-layer targets on both models (Gemma 12B: 102% vs 98%; Qwen 7B: 123% vs 113%), and both layers reach VRS-Valid confidence on both models. The optimal layer is domain-specific however — the first layer is the best probe on Qwen 7B ARC — so target derivation should include a layer sweep rather than assume a fixed position. Full per-model results and the target-layer comparison table are in Appendix E.

7.4 What Fails and Why

This subsection catalogues where the pathway breaks and why: the verifiability boundary (§ 7.4.1), zero-shot domain transfer (§ 7.4.2), multi-task training (§ 7.4.3), end-to-end training at 70B on factoid QA (§ 7.4.4), and cross-benchmark format mismatch (§ 7.4.5). In each case some part of the pipeline cannot support the routing intervention; § 9.2 synthesises the common structure.

7.4.1 The Verifiability Boundary

The two-pass confidence-only architecture fails on tasks where the model cannot assess answer correctness from the text alone.

Unbalanced confonly on GSM8K (Gemma 12B). With unbalanced probe-derived targets in confidence-only format, accuracy was preserved (88.8%) but AUROC_2 improved only from 0.546 to 0.617, and confidence mean remained at 99.7. Solving gradient dilution alone was insufficient because data bias remained (87% of targets near 100%).

Balanced confonly on GSM8K (Gemma 12B). With balanced resampling (184 items, two populated bins, 432 gradient updates), confidence remained at ceiling: mean 100.0, AUROC_2 = 0.500. In controlled testing with trivially verifiable examples (e.g., "What is 2+2? Your answer: 7"), the adapter correctly output 0% confidence. On real GSM8K items with 200–400-token chain-of-thought derivations, it defaulted to 100%.

Stripped-CoT control. To test whether the failure was caused by chain-of-thought length or inherent unverifiability, we removed the chain of thought, presenting only the final numeric answer ("Your final answer: 72"). AUROC_2 remained at chance (0.500). The model cannot verify numeric answers by inspection. "Is 72 correct for this word problem?" requires re-computing the answer, unlike factoid answers where the model can assess plausibility from world knowledge.

Brier score loss. We tested whether replacing cross-entropy with a tokenised Brier score (Y. Li et al. 2025) on the confidence token could recover graded calibration. The model collapsed to a uniform low-confidence output (mean 42.2, AUROC_2 = 0.594). The Brier gradient on a single confidence token is dominated by the cross-entropy (CE) gradient on 300 answer tokens; increasing the weight produced unstable oscillation.

The boundary is verifiability, not format: factoid recall is verifiable (the model knows whether "Paris" is a plausible capital); numeric computation is not. End-to-end single-pass training (§ 7.3.2) breaks this boundary by routing generation-time uncertainty directly to the confidence token.

7.4.2 Zero-Shot Domain Transfer

We tested whether TriviaQA-trained adapters transfer zero-shot to GSM8K and ARC.

Table 14: Zero-shot domain transfer. TriviaQA-trained adapters fail on GSM8K and ARC.

Test	Model	Accuracy	AUROC_2	Δ vs. baseline	Conf mean
GSM8K	Gemma 12B	0.023	0.548	+0.002	35.2
ARC	Qwen 7B	0.004	0.549	−0.071	83.5

Both tests show catastrophic accuracy loss and no confidence transfer. The adapter overrides the answer format, producing TriviaQA-style short answers instead of chain-of-thought (GSM8K) or letter-selection (ARC). The confidence channel does change (confidence drops from ceiling) but is uncorrelated with correctness because task performance has collapsed. The adapter teaches "how to express uncertainty" but not "what to be uncertain about" on a new domain. The adapter is format-specific.

7.4.3 Multi-Task Training

To address the format mismatch, we combined TriviaQA and GSM8K probe-derived targets into a single multi-task training set (1,870 TriviaQA + 800 GSM8K items, LoRA rank 16, lr 2×10^{-4}, 3 epochs on Gemma 12B). Accuracy was preserved on GSM8K (88.2%) and TriviaQA confidence remained calibrated, but GSM8K confidence stayed at ceiling: mean 100.0, AUROC_2 = 0.508. The model compartmentalised: TriviaQA format → varied confidence; GSM8K chain-of-thought format → ceiling prior.

The mechanism is loss dilution. GSM8K chain-of-thought responses are 200–400 tokens; the confidence tag is 5 tokens at the end. Under standard SFT loss, the confidence tokens

contribute less than 2% of the total loss and receive negligible gradient. The model learns to generate correct chain-of-thought but does not learn to vary confidence. PT-CSFT via standard SFT requires the confidence tokens to constitute a non-trivial fraction of the training loss. Short-answer formats satisfy this condition; chain-of-thought formats do not.

7.4.4 End-to-End at 70B on Factoid

End-to-end single-pass training on factoid QA at 70B (Llama 70B, TriviaQA) destroys factual recall. The model's answer accuracy collapses. This is the symmetric failure to the multi-task loss dilution case (§ 7.4.3) and the two are reconciled by the same gradient-allocation principle. When confidence tokens are a small fraction of the response (chain-of-thought, § 7.4.3), they receive negligible gradient and confidence learning fails while accuracy is preserved. When confidence tokens are a substantial fraction of a short factoid response, the gradient on confidence learning is large enough that, combined with LoRA's limited capacity at 70B, it interferes destructively with factual recall and accuracy collapses while confidence learning proceeds. Gemma 12B GSM8K succeeds end-to-end (§ 7.3.2) because the chain-of-thought reasoning process tolerates LoRA perturbation in a way that factoid recall does not: reasoning is recomputed at inference time; factoid knowledge is recalled from weights that LoRA partially overwrites. The architecture choice is therefore task-type-dependent: end-to-end for reasoning tasks where the chain-of-thought process transfers, confidence-only (single-stage or two-stage curriculum) for factoid QA where answer generation must not be perturbed.

7.4.5 Cross-Benchmark Format Mismatch (TriviaQA → MMLU)

We pre-registered an MMLU cross-benchmark transfer test to assess whether PT-CSFT confidence calibration generalises beyond the training format. The pre-registered configuration (r16/lr2e-4) was applied to TriviaQA-trained adapters and evaluated zero-shot on MMLU. The confidence pathway shows signs of transfer: paired AUROC_2 Δ on matched items ranges from +0.027 (Llama 8B, not significant) to +0.104 (Llama 70B, significant), with Gemma 27B and Qwen 7B also reaching statistical significance. However, accuracy collapses 31–69pp across all models because the LoRA teaches free-text answers that destroy the MCQA letter format required by MMLU. The two pathways — confidence calibration and answer format — transfer somewhat independently, but multi-task training over both formats would be required for practical cross-benchmark deployment. Full per-model results are in Appendix F.

7.5 Diagnostic Framework

The comprehensive results table consolidating all models, tasks, and methods is in Table 19 (Appendix B).

Operating envelope. The triage framework that emerges from these results: standard PT-CSFT for verifiable short-answer tasks at 7–32B (§ 7.3.1); two-stage curriculum at 70B+ (§ 7.3.4); end-to-end single-pass training for reasoning tasks, with per-model recipe validation (§ 7.3.2, § 7.4.4); the logit readout (§ 8) as universal rescue where text confidence is insufficient. The known failure modes are zero-shot format transfer (§ 7.4.2), multi-task SFT on chain-of-thought tasks (§ 7.4.3), end-to-end at 70B on factoid (§ 7.4.4), and end-to-end on architectures with weak correctness encoding at the confidence position (probe AUROC_2 ¡ 0.7).

8 The Logit Readout

The logit readout reads the softmax probability distribution over confidence tokens rather than parsing the generated text. It is the universal rescue mechanism, recovering discrimination at every scale where the text channel fails. This section details its mechanism, calibration

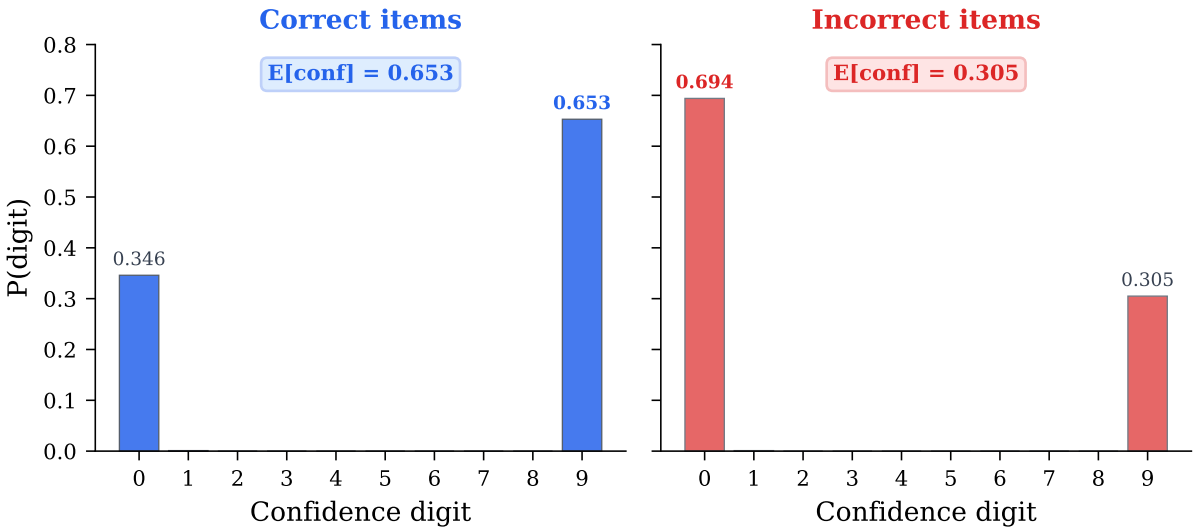

Figure 4: **Location-based logit mechanism.** Softmax probability over confidence digits (0–9) for correct items (left, blue) and incorrect items (right, red) on Gemma 12B GSM8K. Both distributions are peaked (>99% mass on digits 0 and 9), but mass concentrates at opposite poles: digit 9 for correct, digit 0 for incorrect. The expected value captures this location shift as a continuous confidence score.

properties, and boundary conditions, and articulates the four-evidence case for the routing-not-mapping thesis on which this paper turns. The readout has one structural prerequisite: confidence values must be single tokens in the model's vocabulary. This is tokeniser-dependent and varies across model families (Llama: 0–100 all single-token; Gemma coarse 0–9 single-token; Mistral: none single-token). We return to this constraint in § 8.4 and discuss its deployment implications in § 9.8.

8.1 The Mechanism: Location, Not Spread

The logit readout computes expected confidence as $\sum_{d=0}^{9}(d/9)\cdot P(d)$, where P(d) is the softmax probability of digit d at the confidence token position. This continuous signal (expected-value standard deviation 9.6 on a 0–100 scale for this run) achieves $\text{AUROC}_2 = 0.862 \pm 0.013$ on Gemma 12B GSM8K, compared to the binary argmax output (std 45.9).

The training creates the signal. Three training-free baselines computed from the same digit distributions (token entropy ($-\Sigma\ p_i \log_2 p_i$, negated), maximum softmax probability, and softmax margin (top-1 minus top-2)) all achieve AUROC_2 near chance: entropy 0.471 ± 0.250, max probability 0.472, margin 0.472. The logit expected value exceeds the best training-free baseline by +0.39 AUROC_2. None of these distributional features carries the trained signal. The signal the readout produces is not present in the same digit distributions before training.

The mechanism is location-based, not spread-based (Figure 4).

Correct and incorrect items both produce peaked distributions with ¿99% of probability mass concentrated on digits 0 and 9. For correct items, P(digit=9) = 0.653 and P(digit=0) = 0.346; for incorrect items, the ratio reverses (P(digit=9) = 0.305, P(digit=0) = 0.694). Both cases have low entropy because the distribution is peaked, but the mass is at different poles. Entropy, max probability, and margin measure distributional shape (how peaked), not location (where the mass sits). The expected value captures the 0-to-9 mass ratio. This explains why binary training targets produce graded confidence: the model learns to shift probability mass between two poles proportionally to its uncertainty, and the expected value reads this ratio as a continuous score.

Robustness to seed instability. The logit readout is stable even when the argmax

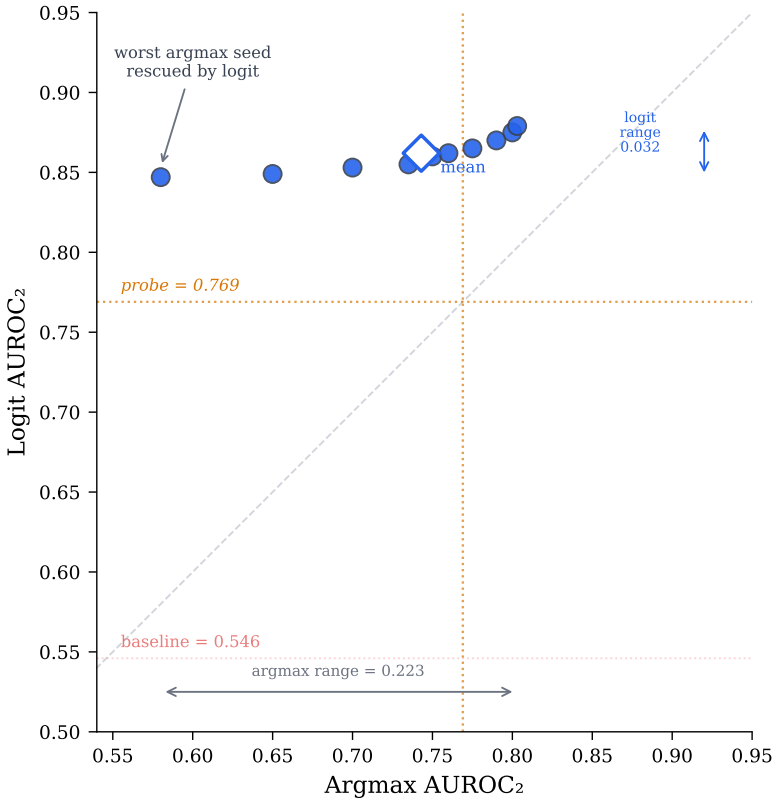

Figure 5: **Logit readout rescues seed instability.** Gemma 12B GSM8K end-to-end training at lr $= 5\times10^{-5}$, 10 seeds. Each point is one seed; the diamond marks the mean. Argmax AUROC_2 ranges from 0.580 to 0.803 (range $= 0.223$); logit AUROC_2 ranges from 0.847 to 0.879 (range $= 0.032$). The worst argmax seed is rescued from 0.580 to 0.853 by the logit readout. Orange dotted line: probe reference (0.769); all logit values exceed the probe.

degenerates (Figure 5). At lr $= 5\times10^{-5}$ across 10 seeds, the worst argmax seed ($\text{AUROC}_2 = 0.580$, floor-biased confidence) is rescued to logit $\text{AUROC}_2 = 0.853$. Argmax AUROC_2 ranges from 0.580 to 0.803 (range $= 0.223$); logit AUROC_2 ranges from 0.847 to 0.879 (range $= 0.032$). The logit readout is completely robust to the discrete token bottleneck.

Where the signal is created. LoRA targets five projection matrices per transformer layer (k_proj, o_proj, gate_proj, up_proj, down_proj). The query and value attention projections (q_proj, v_proj) are unchanged, as is the language model head (unembedding matrix). Two consequences follow. First, the model does not change what it attends to: the attention pattern itself, computed from unchanged queries and values, is largely preserved. Instead, PT-CSFT modifies how the attended-to information is transformed (key and output projections) and how it is integrated across positions (MLP block). Second, the language model head is the same matrix at evaluation time as at pre-training time. The vocabulary geometry of digit tokens, including the relative positions of "0" through "9" in the output embedding space, was already in place before PT-CSFT. The training did not need to teach the model how to express confidence values. It needed to teach the model to route correctness-relevant content to the confidence-token position, where the existing unembedding could read it.

This explains why an MLP probe on baseline hidden states ($\text{AUROC}_2 = 0.785$; § 8.4) cannot match the PT-CSFT logit readout ($\text{AUROC}_2 = 0.862$). The MLP probe extracts what is present in baseline hidden states at the confidence position. PT-CSFT changes which content arrives at that position. The probe and the readout are not reading the same underlying signal; the readout reads a hidden state that the probe cannot anticipate from the baseline distribution.

PT-CSFT does not surface existing information at the confidence position; the routing is created by training. The controlled activation-patching results in § 8.6 provide converging support for this interpretation: patching PT-CSFT hidden states at the confidence position into a baseline forward pass shifts confidence in the predicted direction, while patching at a control position does not.

8.2 Calibration

Raw logit expected value calibration varies substantially across seeds: per-seed ECE ranges from 0.054 to 0.540 (mean 0.261 ± 0.161), strongly correlated with logit mean. Seeds whose logit mean matches the base-rate accuracy (87%) calibrate naturally (ECE ¡ 0.10 for 3 of 10 seeds). Seeds with lower logit mean are systematically underconfident.

Post-hoc Platt scaling (logistic regression fit on 260 held-out items) reduces ECE to 0.042 ± 0.009 across all 10 seeds (84% reduction), with every seed below 0.07. Isotonic regression achieves comparable results (0.043 ± 0.016). Even a simple additive bias shift reduces ECE to 0.065 ± 0.019. Discrimination (AUROC_2) is preserved by all methods since they apply monotone transformations.

The logit readout separates two properties: discrimination is the robust property (AUROC_2 = 0.862 ± 0.013, invariant to seed, recalibration method, and distribution shift), while calibration is tunable post-hoc with a small held-out set. After Platt scaling, the readout achieves both excellent discrimination and calibration from binary training labels on 282 items.

8.3 Selective Prediction

Using the logit expected value as a selective prediction threshold, the model achieves 95% accuracy on retained items at 67.8% coverage, retaining roughly two-thirds of all answers. The argmax text confidence achieves similar coverage (66.2%), but the logit readout is more stable across seeds. Token entropy cannot reach the 95% accuracy target at any coverage level, confirming that the trained signal is necessary for practical selective prediction.

8.4 Why the Student Exceeds the Teacher

Recovery values above 100%, where the fine-tuned confidence discriminates better than the probe that generated its training targets, occur because the linear probe is a constrained readout: it reads one layer at one token position using a linear decision boundary. The fine-tuned model integrates information across all layers through the transformer's forward pass at the confidence token position and can exploit nonlinear feature combinations. The probe provides a noisy but directionally correct training signal; the model generalises beyond the specific probe predictions. This is analogous to knowledge distillation, where a student model can exceed a teacher's soft labels.

Nonlinear probe does not close the gap. A 2-layer MLP probe (128/64 hidden units) on Gemma 12B GSM8K hidden states achieves $\text{AUROC}_2 = 0.785$ (last layer, 5-fold CV), marginally above the linear probe (0.768, the *confidence-position probe*, distinct from the 0.769 gate probe of § 7.2) but still substantially below the logit readout (0.862). The +0.077 gap between MLP probe and logit readout cannot be attributed to linear-probe underextraction. In the level vocabulary introduced in § 1.3, the MLP probe extracts level 2 from baseline hidden states at the confidence position; the logit readout reads level 3 from PT-CSFT hidden states at the same position. The fact that nonlinear probing of level 2 in the baseline cannot reach the PT-CSFT readout is the most direct evidence we have that PT-CSFT operates on level 3, not by surfacing more of level 2. This is the routing interpretation: PT-CSFT does not surface information that was already present in baseline hidden states at the confidence position. It changes which content arrives at that position. The MLP probe extracts what is present in

baseline hidden states; the logit readout reads a hidden state that PT-CSFT produces but the baseline does not. The two readouts are not reading the same underlying signal. The 0.077 gap is a lower bound on the signal that PT-CSFT *introduces* through routing modification, beyond what richer probing of baseline hidden states can recover.

Cross-family validation. The logit readout generalises across architectures. Llama 8B TriviaQA confonly achieves logit $\text{AUROC}_2 = 0.846$ versus text $\text{AUROC}_2 = 0.833$ (+0.013) using fine-mode confidence (0–100) where all 101 values are single tokens in the Llama tokeniser. Qwen 7B GSM8K E2E achieves logit $\text{AUROC}_2 = 0.740$ versus text $\text{AUROC}_2 = 0.500$ (+0.240) at lr=5e-5, the largest text-to-logit gap in the study (§ 7.3.2). The logit readout is tokeniser-dependent: it requires confidence values to be single tokens (Llama: 101/101; Gemma 0–9 coarse: 10/10; Mistral: 0/101). For models where multi-digit values are multi-token, text-based parsing remains the readout method. These models already achieve strong text-based discrimination after PT-CSFT (Gemma 12B: 0.841; Mistral 7B: 0.782; Qwen 32B: 0.847).

Boundary condition. The logit readout does not generalise unconditionally. End-to-end GSM8K training on Llama 8B produces ceiling confidence ($\text{AUROC}_2 = 0.544$), and the logit distribution shows no signal (0.546). A diagnostic probe at the confidence position achieves $\text{AUROC}_2 = 0.666$ for Llama 8B (weak encoding) versus 0.768 for Gemma 12B (5-fold CV, n = 222). The routing efficiency, how much of the probe signal reaches the output, is the critical differentiator. For Gemma 12B, routing efficiency exceeds 100% (the model integrates richer cross-layer structure than a single-layer probe reads). For Llama TriviaQA, routing efficiency is 98% with more modest encoding (probe 0.808). The same architecture that fails on GSM8K (probe 0.666) succeeds on TriviaQA (probe 0.808). The bottleneck is the interaction of training method and task, not the model family. Empirically, the logit readout works when the training procedure concentrates sufficient correctness information at the confidence position (probe AUROC_2 ¿ 0.7).

8.5 The Four-Evidence Case for a Routing-Based Interpretation

The logit readout result, combined with the structural choices in PT-CSFT, makes a converging four-evidence case for the pathway thesis. The case draws on the structural constraints of the LoRA configuration, the behaviour of training-free baselines, and the gap between probes on baseline states and the PT-CSFT readout. Controlled activation patching (§ 8.6) provides additional, partially independent support.

1. **MLP probe gap (0.785 vs 0.862).** A 2-layer MLP probe on baseline hidden states cannot match the PT-CSFT logit readout. If the readout were merely surfacing information already present in baseline activations, a sufficiently expressive probe on those activations should match it. It does not. The 0.077 gap is the lower bound on signal that PT-CSFT creates through routing modification.
2. **Steering null.** If the internal correctness signal and the confidence verbalization pathway shared a common representational direction, then adding the probe's weight vector to hidden states at inference time should shift verbal confidence; this would imply PT-CSFT works by amplifying an existing signal. To test this, we extracted the probe's coefficient vector (trained on middle-layer pre-answer-token hidden states), normalised it to unit length, and added $\alpha \times$ direction to the target layer's output during forward pass, sweeping $\alpha \in -5, -2, -1, -0.5, 0, 0.5, 1, 2, 5, 10$ on the full T-eval set ($\alpha = 0$ reproduces the unsteered baseline). On Gemma 12B, accuracy, confidence mean, confidence standard deviation, and AUROC_2 are invariant across α (AUROC_2 varies by at most 0.011, within noise). On Qwen 7B, the confidence mean remains frozen at 97.3–97.8 across all α; an apparent AUROC_2 shift at $\alpha = -5$ reflects accuracy-induced changes in the eval set composition rather than confidence verbalization changes. The probe direction is invisible to the verbalization pathway. Adding a vector to hidden states modifies what is present after computation. PT-CSFT modifies the computation that produces the hidden

state. Inference-time additive interventions cannot reach the failure mode that PT-CSFT addresses. This is consistent with the Miao and Ungar (2026) finding of orthogonality between correctness and confidence directions.

3. **Unchanged LM head and Q/V projections.** LoRA targets k_proj, o_proj, gate_proj, up_proj, down_proj. q_proj, v_proj, and the language model head are unchanged across every model in the study. The output vocabulary geometry that expresses graded uncertainty was in place before PT-CSFT. The training does not modify what the model *can say*. It modifies what reaches the position where it says it. One caveat: modifying upstream representations can functionally alter the effective input to a frozen output mapping, so an unchanged LM head does not by itself rule out output-side reinterpretation. This evidence item is therefore supportive rather than decisive; it is the conjunction with items #1, #2, and #4 that motivates the routing interpretation (developed further in § 9.4).
4. **Training-free baselines at chance.** Token entropy, maximum softmax probability, and softmax margin all sit at $\text{AUROC}_2 \approx 0.47$ on the same digit distributions that the logit readout reads at 0.862. If the trained signal amplified an existing distributional feature of the baseline output, at least one of these training-free measures would track it. None do. The signal at the readout is introduced by training, not amplified from a baseline distributional feature.

The four pieces of evidence are independent and converge on the routing-based interpretation: the existing output mapping is intact while routing of correctness-relevant content to the confidence position is altered. The following subsection reports a controlled activation-patching experiment that provides additional support. We return to the full evidence base in § 9.4 alongside the LoRA target analysis and the post-training probe analysis, where one caveat to the unchanged-LM-head argument is discussed.

8.6 Controlled Activation Patching at the Confidence Position

To test whether PT-CSFT changes the content available in residual states at the confidence-token position, we conducted a controlled activation-patching experiment on Llama 8B (gentle-lr configuration). At each of eight evenly spaced layers (0, 4, 8, 12, 16, 20, 24, 28 of 32), we replaced the baseline model's hidden state at the confidence-token position with the corresponding PT-CSFT hidden state at that position and layer, then continued the baseline forward pass and measured the shift in confidence-digit logit distribution. We define the normalised shift as $s = (p_{\text{patched}} - p_{\text{bl}})/(p_{\text{pt}} - p_{\text{bl}})$, where p is the softmax probability assigned to the confidence digit favoured by PT-CSFT; $s = 1$ indicates full recovery of the PT-CSFT confidence value, $s = 0$ indicates no effect. Items where the PT-CSFT and baseline confidence values are indistinguishable ($|p_{\text{pt}} - p_{\text{bl}}| < 0.01$) are excluded ($n = 79$ per layer after exclusion).

Table 15 shows a near-monotonic layer-depth gradient: patching at layer 0 produces no shift (41.8% positive, below chance), while patching at layer 28 shifts 91.1% of items toward the PT-CSFT confidence value with a median normalised shift of 0.96 (Spearman ρ between layer depth and mean shift = 0.976, $p < 10^{-4}$). The early layers (0–8) produce a mean shift of 0.001 with 52.7% positive; the late layers (24–28) produce a mean shift of 0.740 with 89.2% positive (Mann-Whitney $p < 10^{-6}$). This gradient is consistent with confidence-relevant content accumulating through the network, with the strongest contributions from the upper quarter.

Three controls establish the specificity of the effect:

1. **Bidirectional.** Reverse patching (injecting baseline states into the PT-CSFT forward pass at layer 28) shifts 88.6% of items back toward baseline confidence (mean shift toward baseline: 0.876). The effect operates symmetrically in both directions.
2. **Position-specific.** Patching PT-CSFT states at a mid-question token position at layer 28 produces a mean shift of 0.007 with 54.4% positive — indistinguishable from chance. The same layer at the confidence-token position produces a mean shift of 0.947 with

Table 15: Activation patching at the confidence-token position (Llama 8B gentle-lr, $n = 79$ per layer). The normalised shift measures the fraction of the baseline-to-PT-CSFT gap closed by patching a single layer. p values are one-sided binomial tests against chance (50%).

Layer	Mean shift	Median	Frac > 0	Frac > 0.5	p (binomial)
0	-0.012	-0.009	41.8%	1.3%	0.94
4	-0.002	$+0.022$	55.7%	3.8%	0.18
8	$+0.018$	$+0.082$	60.8%	15.2%	0.036
12	$+0.127$	$+0.116$	57.0%	29.1%	0.13
16	$+0.242$	$+0.204$	63.3%	34.2%	0.012
20	$+0.235$	$+0.277$	64.6%	32.9%	0.006
24	$+0.534$	$+0.617$	87.3%	65.8%	<0.001
28	$+0.947$	$+0.960$	91.1%	84.8%	<0.001

91.1% positive. The effect is specific to the confidence position, not a general property of PT-CSFT hidden states.

3. **Answer-preserving.** Of the 100 items patched at layer 28 at the confidence position, 83% produce unchanged first-answer tokens. Among those, 72% also show a confidence shift > 0.02. The patching modifies confidence output while largely preserving answer content.

Interpretation. These results demonstrate that PT-CSFT changes the content available in late residual states specifically at the position where confidence is generated, that the change is bidirectional, and that it is largely selective for confidence over answer computation. This is consistent with the routing-repair interpretation: PT-CSFT modifies what arrives at the confidence-token position without disrupting the answer pathway. However, the experiment operates at the level of the full hidden state at a single position and does not identify which specific features, attention heads, or MLP neurons are responsible for the effect. The layer-depth gradient is consistent with progressive accumulation of routing modifications through the network, but does not distinguish between this interpretation and one in which upper layers alone create the signal. Component-level dissection (per-head and per-MLP patching) would be needed to trace the full pathway from specific LoRA modifications to the confidence output.

9 Discussion

We organise the discussion around the pathway thesis introduced in § 1.1 and developed throughout the empirical sections. The thesis claims that verbal confidence failure in instruct-tuned LLMs is a routing problem (correctness-relevant content does not reach the confidence-token position), not an output-mapping problem (the model lacks the vocabulary to express graded uncertainty). PT-CSFT works by repairing the routing while leaving the output mapping intact. The remainder of this section organises the results into four parts under this frame: where the pathway breaks (§ 9.2), how it can be repaired (§ 9.3), what it is made of (§ 9.4), and how its repair translates to deployment (§ 9.5–9.7), followed by limitations (§ 9.8).

9.1 The Routing-Not-Mapping Thesis

The thesis (§ 1.1) and the representational levels (§ 1.3) frame everything that follows. Briefly: level 1 (latent correctness information) and level 2 (probe-decodable representations) exist in the baseline; level 3 (confidence-position routed content) does not, and PT-CSFT's job is to create it; level 4 (verbal output) is read from level 3 by an unchanged language model head.

VRS tier and discrimination are dissociable. A methodological observation runs through these results: VRS tier and AUROC_2 come apart, and they do so in both directions.

The most striking case is the ARC confidence-only adapter on Qwen 7B (§ 7.3.1): its verbal confidence correlates with correctness at r = 0.689 and discriminates at $AUROC_2$ = 0.813 — the strongest correlation and among the strongest discrimination in the study — yet it classifies as VRS-Invalid, because the confidence distribution is peaked near ceiling (L = 0.868, TRIN = 0.808). The same pattern, less extreme, appears in Llama 8B gentle-lr ($AUROC_2$ between 0.82 and 0.85 across its five replication seeds, VRS moving between Invalid, Indeterminate, and Valid as TRIN crosses 0.80) and Mistral 7B gentlest ($AUROC_2$ 0.74–0.80, bouncing across the 0.60 boundary). Conversely, the MMLU transfer condition produces VRS-Valid confidence on three models (Gemma 12B, Gemma 27B, Qwen 7B) whose absolute $AUROC_2$ is near chance (0.56–0.59): the format mismatch collapses task accuracy, the model becomes diffusely uncertain, the confidence distribution spreads, and that spread satisfies the VRS shape criteria even though the signal barely tracks correctness. VRS keys on distributional shape (ceiling rate L and concentration TRIN); $AUROC_2$ keys on discrimination (whether correct responses are ranked above incorrect). A signal can discriminate well while peaked (Invalid), or be well-spread while uninformative (Valid). This is why $AUROC_2$ is the pre-registered primary metric: it is invariant to the distributional shape that VRS — and, for the same reason, M-ratio — is sensitive to (Cacioli 2026e). We report VRS tiers throughout as a screening summary, not as the quality measure; where tier and $AUROC_2$ disagree, $AUROC_2$ is the metric that reflects whether the confidence signal separates correct from incorrect responses.

The remainder of the discussion is organised as evidence about levels 3 and 4: § 9.2 catalogues where the routing intervention breaks, § 9.3 catalogues how each break can be addressed, and § 9.4 catalogues what the modified pathway is made of, drawing the converging evidence from § 8.5–8.6 together with the LoRA target analysis.

9.2 Where the Pathway Breaks

The empirical break points are catalogued in § 7.4 (verifiability boundary, zero-shot domain transfer, multi-task training, end-to-end at 70B on factoid, cross-benchmark format mismatch) and § 6 (LlamaForCausalLM hyperparameter sensitivity, 70B single-stage failure).

Each break has a common shape: the routing intervention is asked to deliver under conditions where some other part of the pipeline cannot support it. The verifiability boundary case is cleanest. The two-pass architecture requires the model to evaluate its own answer from text alone; on GSM8K arithmetic, that condition is not met. At 70B, the failure is not of *form* but of *capacity*: standard single-stage LoRA cannot simultaneously unblock the channel and train what arrives under the gradient available, which is why the two-stage curriculum succeeds where seven single-stage configurations fail. The LlamaForCausalLM hyperparameter sensitivity is a recipe issue, not an architecture issue: a 4× lower learning rate restores the intervention. The E2E Pareto on Qwen 7B GSM8K is the same pattern in a different costume: the mechanism transfers; the recipe does not. Where the internal signal exists and the LoRA has the capacity to modify routing without destroying generation, the intervention works. The break points map onto violations of one of these two conditions.

9.3 How the Pathway Repairs

The empirical repair strategies are catalogued in § 7.3 (confidence-only PT-CSFT, end-to-end PT-CSFT, balanced confidence-only at 70B Stage 1, two-stage curriculum at 70B); the logit readout (§ 8) is a rescue across all of them.

Successful repairs cluster into two distinct *jobs* that the LoRA can perform: **unblocking the channel** (addressing data bias and gradient dilution so any signal at the confidence position can be modified) and **shaping the content** (altering which features arrive at that position). At 7–32B these two jobs can be done together in a single training pass. At 70B they cannot. The two-stage curriculum demonstrates the decomposition: Stage 1 (balanced confonly) unblocks

Table 16: Post-training probe analysis. Original probes lose power on PT-CSFT states; retrained probes reveal distinct patterns across models and configurations. Deltas are relative to baseline probe $AUROC_2$. Verbal $AUROC_2$ values are for the single adapters on which the post-training probes were computed (Qwen 7B: seed 42, 0.747; cf. the 6-seed mean 0.801 in Table 5).

	Gemma 12B	Qwen 7B	Llama 8B (failed)	Llama 8B (gentle-lr)
Verbal $AUROC_2$ change	0.678 → **0.841**	0.577 → **0.747**	0.692 → 0.513	0.692 → **0.838**
Retrained probe, last (post-PT, CV)	**0.844** (+0.037)	0.658 (−0.070)	0.749 (−0.057)	**0.864** (+0.058)
Retrained probe, middle (post-PT)	**0.843** (+0.021)	0.679 (−0.035)	**0.828** (−0.004)	**0.855** (+0.024)

without much content training; Stage 2 (probe-derived continuous targets) shapes content into the now-open channel. Adding Stage 2 to Stage 1 yields +0.051 logit and +0.080 text — the gains one would predict if the two jobs are separable and the second becomes effective only after the first. End-to-end training for reasoning (§ 7.3.2) is a different solution to a different problem: when the two-pass architecture cannot succeed because the answer cannot be evaluated from a post-hoc input, end-to-end training routes the signal during chain-of-thought generation instead. The logit readout is not a separate repair strategy; it is the more sensitive way to read the result of any of the above when the discrete text channel adds noise.

9.4 What the Pathway Is Made Of

Two analyses unique to this section, when combined with the observations in § 8.5 and the controlled activation patching in § 8.6, characterise what the modified pathway is made of.

LoRA target modules. PT-CSFT modifies five projection matrices per transformer layer: k_proj and o_proj in the attention block, and gate_proj, up_proj, and down_proj in the MLP block. This selection is the mlx_lm default and is held constant across all four model families. The unchanged matrices include q_proj and v_proj (attention queries and values) and the language model head (the unembedding matrix). The model's attention queries and values are preserved, so the attention pattern itself is largely unchanged. The output vocabulary is preserved, so the mapping from final hidden states to tokens is unchanged. What changes is how attended information is transformed (key and output projections) and how it is integrated across positions (the MLP block). These are precisely the components that carry out routing between positions. One caveat to the unchanged-LM-head argument: modifying upstream representations can functionally alter the effective input to a frozen output mapping, so an unchanged head does not by itself rule out output-side reinterpretation. The case for routing rests on the combination of the § 8.5 observations plus this target analysis, not on any single one.

Post-training probe analysis. To distinguish whether PT-CSFT surfaces existing information or reorganises representations, we re-applied the original baseline-trained probes to PT-CSFT-adapted hidden states (extracted with the LoRA loaded), then trained fresh probes on the post-PT states via 5-fold CV. For Gemma 12B and Qwen 7B, the pre-registered (r16/lr2e-4) configuration was used. For Llama 8B we report both the failed configuration (r16/lr2e-4) and the successful gentle-lr configuration (r16/lr5e-5) to characterise the mechanistic basis of the learning-rate sensitivity reported in § 6.

The original probe directions lose discriminative power on post-PT states across all models and configurations. This rules out a "transparent readout" interpretation in which PT-CSFT merely opens a channel; the internal representations are reorganised, not simply surfaced. When fresh probes are retrained, four patterns emerge across the columns.

Gemma 12B shows information *amplified*: the retrained last-layer probe reaches 0.844, exceeding the baseline (0.807). PT-CSFT creates additional linearly decodable structure. Qwen 7B shows *partial preservation with geometric shift*: the middle-layer probe is nearly preserved (δ

$= -0.035$) while the last-layer shifts more, consistent with redirected representational structure. Llama 8B under the failed configuration shows *preserved information with blocked verbalization*: the middle-layer probe recovers virtually all baseline discrimination ($\delta = -0.004$) yet the verbal channel collapsed to 0.513. The failed configuration disrupts the verbalization mapping without creating a replacement route.

The Llama 8B gentle-lr column resolves the learning-rate question mechanistically. Where the failed configuration *destroys* last-layer linear separability (0.749, $\delta = -0.057$), the gentle configuration *amplifies* it (0.864, $\delta = +0.058$). The 0.115 AUROC_2 gap between the two configurations at the last layer is the mechanistic basis of the learning-rate sensitivity: aggressive fine-tuning collapses precisely the upper-layer representational structure that the activation-patching results (§ 8.6) identify as carrying the strongest confidence-relevant signal. The gentle configuration preserves and enhances this structure, enabling the verbal channel to reach 0.838.

9.5 Deployment: From Monitoring to Control

The results above demonstrate metacognitive *monitoring*: the model can express calibrated confidence. Metacognition also requires *control*: acting on that assessment to regulate behaviour (Nelson and Narens 1990). We use confidence-gated retrieval as a proof-of-concept for the complete monitoring-to-control loop.

The gating problem. The RAG experiment evaluates on the 966-item subset of T-eval with parseable confidence in both baseline and PT-CSFT conditions. On this subset, baseline accuracy is 0.751 and PT-CSFT accuracy is 0.655. Without PT-CSFT, Gemma 12B reports $\geq$95% confidence on 960 of 966 items (99.4%): a retrieval system cannot distinguish items that need external evidence from items that don't, and must either retrieve for all items (expensive) or none (unreliable). With PT-CSFT, only 112 items (11.6%) report $\geq$95% confidence and 222 (23.0%) fall below the 50% threshold. Calibrated confidence makes selective retrieval architecturally possible. (An oracle analysis of theoretical accuracy improvement under varying retrieval quality is reported in Appendix A.)

Empirical validation. For items where PT-CSFT confidence fell below 50% (n = 222, 23% of items), we re-generated answers using the base model with Wikipedia evidence passages from the TriviaQA RC corpus. Low-confidence items had baseline accuracy 0.414, confirming the confidence signal correctly identifies weak items (vs 0.900 for high-confidence items, a 2.17× differential). RAG corrected 49 of 130 baseline errors ($p_{\text{fix}} = 0.378$) while introducing 42 new errors, yielding net RAG accuracy of 0.446 on the flagged subset. The confidence-gated system (keep baseline answers above threshold, use RAG below) achieved overall accuracy 0.758, exceeding the no-retrieval baseline (0.751) by +0.007 while retrieving for only 23% of items. Naive always-retrieve RAG achieves only 0.509: evidence poisoning degrades performance when applied indiscriminately. Confidence gating prevents this, protecting high-confidence items from unnecessary evidence injection. The modest net improvement (+0.007) reflects the noisy quality of TriviaQA's built-in evidence, not the confidence signal quality; the monitoring signal is proven (2.17× differential) and the control benefit scales with retrieval quality.

Cross-model replication on Qwen 7B. Running the same evaluation on the Qwen 7B PT-CSFT adapter reproduces the pattern: baseline reports $\geq$95% confidence on 93.9% of items, PT-CSFT on 0.0%; high-confidence items have accuracy 0.766 versus low-confidence 0.289, a 2.65× differential (vs 2.17× for Gemma 12B); oracle gating at matched $p_{\text{fix}} = 0.5$ yields gated accuracy 0.735, an improvement of +0.104 over PT-CSFT alone. Ceiling collapse plus stratification plus oracle gain replicates across both model families tested.

9.6 Out-of-Distribution Generalisation

A critical question for deployment is whether PT-CSFT confidence transfers to questions outside the training distribution. We evaluated the TriviaQA-trained PT-CSFT adapter (Gemma 12B)

Table 17: Out-of-distribution transfer. TriviaQA-trained adapter evaluated on NQ and PopQA without retraining.

Dataset	Baseline $AUROC_2$	PT-CSFT $AUROC_2$	Δ	Transfer ratio	Base. acc	PT-CSFT acc
NaturalQuestions	0.551	0.757	+0.206	137.7%	0.366	0.294
PopQA	0.604	0.834	+0.231	154.7%	0.294	0.248

on two OOD factoid QA datasets, NaturalQuestions Open and PopQA, using the identical prompt format. No retraining, no adaptation.

On both OOD datasets, PT-CSFT produces large $AUROC_2$ improvements that exceed the in-distribution improvement on the matched T-eval partition ($\Delta = +0.149$; baseline 0.684, PT-CSFT 0.833, a 500-item partition distinct from the full T-eval) in absolute terms. The transfer ratios above 100% indicate that the adapter learned generalizable factoid QA uncertainty rather than TriviaQA-specific patterns. (The transfer ratio normalises by the in-distribution improvement; its magnitude is therefore sensitive to the in-distribution denominator. The absolute $AUROC_2$ values, 0.757 on NQ and 0.834 on PopQA, are the primary headline; the ratio describes the *direction* of transfer, not its magnitude on a calibrated scale.) The ceiling-breaking effect transfers fully across both datasets: baseline ceiling rates of 94.5% (NQ) and 97.0% (PopQA) drop to 5.2% and 8.8% respectively post-PT-CSFT, enabling confidence-based gating on new question distributions without retraining.

The PopQA result is particularly informative because both baseline and PT-CSFT accuracy are low (0.294 and 0.248 respectively). The confidence signal remains discriminative ($AUROC_2$ = 0.834) in the regime where the model is mostly wrong. A model that knows which 25% of its answers it got right is operationally useful for deferral and routing, even when overall accuracy is poor. The confidence pathway operates independently of the task-accuracy regime.

Under the pathway thesis, this OOD generality follows from the mechanism. The features the routing recovers (hidden-state activations associated with retrieval success) are largely task-invariant across factoid QA, so the same routing pattern that exposes signal on TriviaQA exposes the same kind of signal on NQ and PopQA. The accuracy cost on OOD (0.366 $\rightarrow$ 0.294 for NQ, 0.294 $\rightarrow$ 0.248 for PopQA) tracks the single-pass architecture trade-off observed on TriviaQA (0.751 $\rightarrow$ 0.655); a two-pass confonly architecture trained specifically for factoid QA would preserve accuracy while providing calibrated confidence.

9.7 Calibration Transfer

Post-hoc Platt scaling requires a small held-out calibration set per adapter deployment. We tested whether Platt parameters transfer across seeds and domains. Cross-seed transfer (fitting on seed 42, applying to seeds with similar logit means) achieves ECE comparable to per-seed fitting for seeds with matching bias points (ECE ¡ 0.05) but fails for seeds with substantially different logit means (ECE ¿ 0.30). Per-adapter calibration remains necessary. Cross-domain transfer (GSM8K Platt parameters applied to ARC-Challenge text confidence) modestly improves calibration (raw ECE 0.083 $\rightarrow$ 0.052), despite differences in model family, domain, and confidence format. While per-deployment calibration is recommended, even crude cross-domain scaling provides some benefit.

9.8 Limitations

The limitations cluster into three groups: measurement and methodology choices, scope and boundary conditions of the method, and open questions we leave for future work.

Methodological and measurement.

Probe bias inheritance. The probe-target approach inherits any biases in the probe's correctness assessment. If the probe systematically misjudges certain question types, the fine-tuned

model will learn to express miscalibrated confidence on those items.

Seed validation. Multi-seed replication on three non-Gemma TriviaQA models (6 seeds each; § 5.3) confirms stability: Llama 8B 0.836 ± 0.011, Mistral 7B 0.771 ± 0.017, Qwen 7B 0.801 ± 0.022 (most seed-sensitive; seed 42 low outlier at 0.756). The Gemma 12B GSM8K E2E result is validated across 10 seeds. The Llama 70B curriculum result and Gemma 12B/27B TriviaQA results remain single-seed. The RAG gating and OOD transfer experiments are reported on Gemma 12B (single training seed) plus a Qwen 7B replication. Cross-model replication for ARC confonly is limited to Qwen 7B alone.

Post-training probe coverage. The post-training probe analysis (§ 9.4) now covers both the failed and gentle-lr configurations on Llama 8B, but has not been replicated on Mistral 7B or on the 70B model. Whether the information-amplification pattern generalises beyond Llama 8B gentle-lr and Gemma 12B remains an open question.

Qwen 72B probe interpretation. The Qwen 72B probe result ($\text{AUROC}_2 = 0.754$, below verbal) may be partially affected by a 75% hidden-state cache rate, which biases the probe sample toward easier items. On matched items, the gap narrows to −0.006 (within noise). A nonlinear probe might extract additional signal that the linear probe cannot.

Argmax granularity on E2E. All end-to-end argmax outputs remain binary (TRIN = 1.0), though the logit expected-value distribution is graded (per-seed std spanning 12–28 across the 10 seeds; the single representative run in § 8.1 has std 9.6). The gentle learning rate (5×10^{-5}) produces mean logit $\text{AUROC}_2 = 0.862 \pm 0.013$ across 10 seeds, all above 0.85 and all exceeding the probe, making the binary argmax a cosmetic rather than substantive limitation when logit access is available.

Scope and boundary conditions.

End-to-end at small scale (Llama 8B). The end-to-end GSM8K result is demonstrated on Gemma 12B with 10-seed validation. Llama 8B fails on the same task with the same procedure. A diagnostic probe at the confidence token position is consistent with a representational bottleneck (Llama probe $\text{AUROC}_2 = 0.666$ vs Gemma 0.768, 5-fold CV). The logit readout exceeds the probe for Gemma (0.862 vs 0.768) but cannot rescue the Llama encoding failure. However, the logit readout generalises cross-family on TriviaQA (Llama 8B logit $\text{AUROC}_2 = 0.846$), suggesting the bottleneck is training-method-dependent rather than architecture-dependent.

End-to-end on factoid at 70B. End-to-end training at 70B on factoid QA (TriviaQA) destroys factual recall regardless of rank or iteration count. The adapter memorises training answers. The two-stage curriculum architecture (§ 7.3.4) is the correct choice for factoid QA at scale, with end-to-end reserved for reasoning tasks where the chain-of-thought process transfers.

Tokenizer dependency. The logit readout requires each confidence value to be a single token in the model's vocabulary. This is tokenizer-dependent. Llama supports 101 single-token values (0–100), enabling fine-mode logit readout. Gemma (10/101 in fine mode) and Mistral (0/101) are limited to coarse-mode or text-based readout.

E2E recipe is not plug-and-play. The E2E training recipe is not universally accuracy-preserving. Cross-family replication on Qwen 7B GSM8K reproduces the logit-channel advantage but exposes an accuracy-AUROC Pareto trade-off across learning rates that is not observed on Gemma 12B. Neither tested LR matches Gemma's combination of strong accuracy preservation and strong logit discrimination. Whether a different target schema, learning rate schedule, or training duration moves Qwen 7B's Pareto front toward Gemma's regime is unresolved. The implication for practitioners is that the E2E recipe should be validated for accuracy preservation per model rather than treated as plug-and-play.

70B gap not fully closed. The two-stage curriculum at Llama 70B closes 66% of the probe-verbal gap on the logit channel and 35% on the text channel. 0.037 AUROC_2 remains between the curriculum logit and the probe. Whether the residual gap reflects entrenched RLHF priors, insufficient LoRA capacity at this scale, or a fundamental scale ceiling on LoRA-based routing intervention is unresolved. A scaled-up balanced confonly v2 (3× more data) regressed to

$\text{AUROC}_2 = 0.538$ due to oversampling-induced memorisation, confirming that the sweet spot of brief training at Stage 1 is narrow.

Probe gate is accuracy-dependent. Probe discriminability depends on the interaction between model accuracy and error-set size (§ 7.2). On high-accuracy benchmarks with few errors, the probe may not find enough signal to generate useful PT-CSFT targets. The gating strategy in § 7.2 addresses this, but it means PT-CSFT may be less applicable to tasks where the target model is already near ceiling accuracy.

Open questions and future work.

Component-level mechanistic dissection. The activation patching in § 8.6 operates at the level of the full hidden state at a single position. It demonstrates that PT-CSFT changes the content at the confidence-token position in a way that is position-specific, bidirectional, and largely selective for confidence over answer content. However, it does not identify which specific features, attention heads, or MLP neurons are responsible for the effect. The layer-depth gradient is consistent with progressive accumulation of routing modifications, but does not distinguish between this interpretation and one in which upper layers alone create the signal. Per-head and per-MLP patching, along with probing for specific feature directions, would be needed to trace the full pathway from LoRA weight modifications to the confidence output. This component-level dissection is the natural next step for establishing a complete mechanistic account.

Extension to other architectures. The activation patching was conducted on Llama 8B only. Extending the patching and controls to Gemma, Qwen, and the 70B regime would test whether the layer-depth gradient and position specificity are architecture-general properties of PT-CSFT or specific to the LlamaForCausalLM transformer.

Human-AI team evaluation. The paper evaluates model-centric metrics (AUROC_2, ECE, selective prediction coverage) but does not include a human-subject evaluation of deferral or trust calibration. Whether PT-CSFT confidence improves human-AI team accuracy when humans use the confidence signal to accept, defer, or reject model answers remains an important open question for deployment (cf. LACIE; the deferral literature).

Generality of the pathway thesis. The pathway thesis describes the verbal confidence case specifically. Whether the same routing-not-mapping decomposition applies to other LLM behaviours where the failure could plausibly be in routing rather than output mapping (e.g., refusal calibration, instruction following on rare formats, certain hallucination patterns) is an open empirical question. Failure modes that operate on the model's answer-token distribution rather than the verbal-confidence pathway (e.g., regime-induced compression in reasoning models) are mechanistically distinct from the one addressed here and PT-CSFT is not expected to apply to them without modification.

Scope of the activation patching. The controlled activation patching (§ 8.6) demonstrates that PT-CSFT changes confidence-relevant content at the confidence-token position in a position-specific, bidirectional, and answer-preserving manner. However, the experiment operates on the full hidden state at a single position and does not identify which specific features, attention heads, or MLP neurons carry the signal. Component-level patching (per-head, per-MLP) and probing for specific feature directions would be needed to establish a complete mechanistic account of the routing pathway.

10 Conclusion

Verbal confidence in instruct-tuned language models fails because the readout pathway from internal correctness representations to the confidence-token position transmits little of the available signal, not because the signal is absent. The model evaluates its own reliability internally; the output vocabulary already supports graded uncertainty; what is broken is the routing. PT-CSFT repairs this routing through targeted LoRA modification of attention key and output

projections and the three MLP projections, leaving attention queries, attention values, and the language model head unchanged. The intervention requires a few hundred examples and under ten minutes of training on consumer hardware. Across eight models at 7–72B, three architecture classes, and three cognitive domains (factual recall, mathematical reasoning, science reasoning), the method substantially closes the probe-verbal gap; full headline results are in the abstract and § 5.3, with the 70B curriculum in § 7.3.4 and the GSM8K end-to-end result in § 7.3.2.

The mechanism is consistent with routing repair rather than output remapping. Four converging observations (§ 8.5) plus controlled activation patching (§ 8.6) support this interpretation: the patching intervention is position-specific, bidirectional, selective for confidence over answer content, and follows a near-monotonic layer-depth gradient (Spearman $\rho = 0.976$). The LoRA target analysis and post-training probe analysis (§ 9.4) add mechanistic detail, showing that successful configurations amplify last-layer linear separability while failed configurations destroy it. The logit readout, reading the softmax distribution over confidence tokens rather than parsing the generated text, is the universal rescue mechanism: it recovers discrimination at every scale where the text channel is insufficient, and exposes a location-based, not spread-based, account of how the model expresses graded uncertainty. The confidence signal supports downstream metacognitive control: confidence-gated retrieval (§ 9.5), out-of-distribution transfer (§ 9.6), and post-hoc Platt scaling (§ 9.7) together demonstrate a complete monitoring-to-control loop. The pathway thesis is a substantive empirical claim: the difference between what a model knows and what it says is a routing problem more than an output-mapping problem, and a routing-style intervention closes the gap.

The pathway has limits, and the evidence base, while substantially strengthened by the activation patching, does not constitute a complete mechanistic dissection. The 70B gap is not fully closed (§ 7.3.4); cross-family GSM8K training is recipe-sensitive (§ 7.3.2); the unchanged-LM-head argument has a caveat about upstream representations functionally altering a frozen output mapping (§ 9.4); and the activation patching operates on the full hidden state rather than identifying specific components. Component-level dissection and extension to additional architectures are the natural next steps. The remaining limitations are catalogued in § 9.8.

References

Bani-Harouni, D., C. Pellegrini, P. Stangel, E. Özsoy, K. Zaripova, N. Navab, and M. Keicher (2026). "Rewarding Doubt: A Reinforcement Learning Approach to Calibrated Confidence Expression of Large Language Models". In: *Proceedings of the International Conference on Learning Representations (ICLR)*. arXiv: 2503.02623.

Burns, C., H. Ye, D. Klein, and J. Steinhardt (2023). "Discovering Latent Knowledge in Language Models Without Supervision". In: *Proceedings of the International Conference on Learning Representations (ICLR)*. arXiv: 2212.03827.

Cacioli, J.-P. (2026a). "Before You Interpret the Profile: Validity Scaling for LLM Metacognitive Self-Report". In: *arXiv preprint*. arXiv: 2604.17707.

— (2026b). "Distilling Self-Consistency into Verbal Confidence: A Pre-Registered Negative Result and Post-Hoc Rescue on Gemma 3 4B". In: *arXiv preprint*. arXiv: 2604.24070.

— (2026c). "Do LLMs Know What They Know? Measuring Metacognitive Efficiency with Signal Detection Theory". In: *arXiv preprint*. arXiv: 2603.25112.

— (2026d). "Domain-Level Metacognitive Monitoring in Frontier LLMs: A 33-Model Atlas". In: *arXiv preprint*. arXiv: 2605.06673.

— (2026e). "Quantisation Reshapes the Metacognitive Geometry of Language Models". In: *arXiv preprint*. arXiv: 2604.08976.

— (2026f). "The Metacognitive Monitoring Battery: A Cross-Domain Benchmark for LLM Self-Monitoring". In: *arXiv preprint*. arXiv: 2604.15702.

Clark, P., I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord (2018). "Think You Have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge". In: *arXiv preprint*. arXiv: 1803.05457.

Cobbe, K., V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, D. Prafulla, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman (2021). "Training Verifiers to Solve Math Word Problems". In: *arXiv preprint*. arXiv: 2110.14168.

Damani, M., I. Puri, S. Slocum, I. Shenfeld, L. Choshen, Y. Kim, and J. Andreas (2025). "Beyond Binary Rewards: Training LMs to Reason About Their Uncertainty". In: *arXiv preprint*. arXiv: 2507.16806.

Fleming, S. M. and H. C. Lau (2014). "How to Measure Metacognition". In: *Frontiers in Human Neuroscience* 8, p. 443.

Galvin, S. J., J. V. Podd, V. Drga, and J. Whitmore (2003). "Type 2 Tasks in the Theory of Signal Detectability". In: *Psychonomic Bulletin & Review* 10.4, pp. 843–876.

Groot, T. and M. Valdenegro-Toro (2024). "Overconfidence is Key: Verbalized Uncertainty Evaluation in Large Language and Vision-Language Models". In: *TrustNLP Workshop at ACL 2024*. arXiv: 2405.02917.

Guo, C., G. Pleiss, Y. Sun, and K. Q. Weinberger (2017). "On Calibration of Modern Neural Networks". In: *Proceedings of the International Conference on Machine Learning (ICML)*. arXiv: 1706.04599.

Hager, S., D. Mueller, K. Duh, and N. Andrews (2025). "Uncertainty Distillation: Teaching Language Models to Express Semantic Confidence". In: *arXiv preprint*. arXiv: 2503.14749.

Hendrycks, D., C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt (2021). "Measuring Massive Multitask Language Understanding". In: *Proceedings of the International Conference on Learning Representations (ICLR)*.

Jang, C., M. Choi, Y. Kim, H. Lee, and J.-y. Lee (2025). "Verbalized Confidence Triggers Self-Verification: Emergent Behavior Without Explicit Reasoning Supervision". In: *arXiv preprint*. arXiv: 2506.03723.

Joshi, M., E. Choi, D. S. Weld, and L. Zettlemoyer (2017). "TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension". In: *Proceedings of the Association for Computational Linguistics (ACL)*.

Kadavath, S., T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma, E. Tran-Johnson, S. Johnston, S. El-Showk, A. Jones, N. Elhage, T. Hume, A. Chen, Y. Bai, S. Bowman, S. Fort, D. Ganguli, D. Hernandez, J. Jacobson, J. Kernion, S. Kravec, L. Lovitt, K. Ndousse, C. Olsson, S. Ringer, D. Amodei, T. Brown, J. Clark, N. Joseph, B. Mann, S. McCandlish, C. Olah, and J. Kaplan (2022). "Language Models (Mostly) Know What They Know". In: *arXiv preprint*. arXiv: 2207.05221.

Kuhn, L., Y. Gal, and S. Farquhar (2023). "Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation". In: *Proceedings of the International Conference on Learning Representations (ICLR)*. arXiv: 2302.09664.

Kumaran, D., A. Conmy, F. Barbero, S. Osindero, V. Patraucean, and P. Veličković (2026). "How Do LLMs Compute Verbal Confidence?" In: *arXiv preprint*. arXiv: 2603.17839.

Legrand, N. and P. Ruby (2022). "metadpy: A Python Package for Metacognitive Signal Detection Theory". In: *Journal of Open Source Software* 7.79, p. 4655.

Leng, J., C. Huang, B. Zhu, and J. Huang (2024). "Taming Overconfidence in LLMs: Reward Calibration in RLHF". In: *arXiv preprint*. arXiv: 2410.09724.

Li, K., O. Patel, F. Viégas, H. Pfister, and M. Wattenberg (2023). "Inference-Time Intervention: Eliciting Truthful Answers from a Language Model". In: *Advances in Neural Information Processing Systems (NeurIPS)*. arXiv: 2306.03341.

Li, Y., M. Xiong, J. Wu, and B. Hooi (2025). "ConfTuner: Training Large Language Models to Express Their Confidence Verbally". In: *Advances in Neural Information Processing Systems (NeurIPS)*. arXiv: 2508.18847.

Lin, S., J. Hilton, and O. Evans (2022). "Teaching Models to Express Their Uncertainty in Words". In: *Transactions on Machine Learning Research*. arXiv: 2205.14334.

Maniscalco, B. and H. Lau (2012). "A Signal Detection Theoretic Approach for Estimating Metacognitive Sensitivity from Confidence Ratings". In: *Consciousness and Cognition* 21.1, pp. 422–430.

Marks, S. and M. Tegmark (2024). "The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets". In: *arXiv preprint*. arXiv: 2310.06824.

Miao, M. M., Y.-M. Cho, and L. Ungar (2026). "Correctness-Optimized Residual Activation Lens (CORAL): Transferrable and Calibration-Aware Inference-Time Steering". In: *Proceedings of the International Conference on Machine Learning (ICML)*. arXiv: 2602.06022.

Miao, M. M. and L. Ungar (2026). "Closing the Confidence-Faithfulness Gap in Large Language Models". In: *arXiv preprint*. arXiv: 2603.25052.

Nelson, T. O. and L. Narens (1990). "Metamemory: A Theoretical Framework and New Findings". In: *The Psychology of Learning and Motivation*. Vol. 26, pp. 125–173.

Park, S., E. Meyerson, X. Qiu, and R. Miikkulainen (2026). "Fine-Tuning Language Models to Know What They Know". In: *arXiv preprint*. arXiv: 2602.02605.

Platt, J. C. (1999). "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods". In: *Advances in Large Margin Classifiers*. Vol. 10. 3, pp. 61–74.

Savage, T., J. Wang, R. Gallo, A. Boukil, V. Patel, S. A. A. Safavi-Naini, A. Soroush, and J. H. Chen (2025). "Large Language Model Uncertainty Proxies: Discrimination and Calibration for Medical Diagnosis and Treatment". In: *Journal of the American Medical Informatics Association* 32.1, pp. 139–149. DOI: 10.1093/jamia/ocae254.

Seo, K.-J., S. Lim, and T. Kim (2025). "ADVICE: Answer-Dependent Verbalized Confidence Estimation". In: *arXiv preprint*. arXiv: 2510.10913.

Tao, L., Y.-F. Yeh, M. Dong, T. Huang, P. Torr, and C. Xu (2025). "Revisiting Uncertainty Estimation and Calibration of Large Language Models". In: *arXiv preprint*. arXiv: 2505.23854.

Taubenfeld, A., T. Sheffer, E. Ofek, A. Feder, A. Goldstein, Z. Gekhman, and G. Yona (2025). "Confidence Improves Self-Consistency in LLMs". In: *Findings of the Association for Computational Linguistics (ACL)*. arXiv: 2502.06233.

Tian, K., E. Mitchell, A. Zhou, A. Sharma, R. Rafailov, H. Yao, C. Finn, and C. D. Manning (2023). "Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Wang, V. and E. Stengel-Eskin (2026). "Calibrating Verbalized Confidence with Self-Generated Distractors (DiNCo)". In: *Proceedings of the International Conference on Learning Representations (ICLR)*. arXiv: 2509.25532.

Wang, X., J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou (2023). "Self-Consistency Improves Chain of Thought Reasoning in Language Models". In: *Proceedings of the International Conference on Learning Representations (ICLR)*. arXiv: 2203.11171.

Xiong, M., Z. Hu, X. Lu, Y. Li, J. Fu, J. He, and B. Hooi (2024). "Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs". In: *Proceedings of the International Conference on Learning Representations (ICLR)*. arXiv: 2306.13063.

Yaldiz, D. N., E. Spiliopoulou, Z. Qi, S. Varia, S. Doss, and N. Pappas (2026). "Balancing Classification and Calibration Performance in Decision-Making LLMs via Calibration-Aware Reinforcement Learning". In: *arXiv preprint*. arXiv: 2601.13284.

Yona, G., M. Geva, and Y. Matias (2026). "Hallucinations Undermine Trust; Metacognition is a Way Forward". In: *arXiv preprint*. arXiv: 2605.01428.

Zhao, T., Y. He, W. Zheng, Y. Zhang, and C. Chen (2026). "Wired for Overconfidence: A Mechanistic Perspective on Inflated Verbalized Confidence in LLMs". In: *arXiv preprint*. arXiv: 2604.01457.

Zou, A., L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, X. Yin, M. Mazeika, A.-K. Dombrowski, S. Goel, N. Li, M. J. Byun, Z. Wang, A. Mallen, S. Basart, S. Koyejo, D. Song, M. Fredrikson, J. Z. Kolter, and D. Hendrycks (2023). "Representation Engineering: A Top-Down Approach to AI Transparency". In: *arXiv preprint*. arXiv: 2310.01405.

A Oracle Analysis of Confidence-Gated Retrieval

The empirical RAG result in § 9.5 reflects one specific retrieval-quality regime (TriviaQA's built-in Wikipedia evidence, $p_{\text{fix}} = 0.378$ on Gemma 12B). This appendix characterises how the confidence-gated system's overall accuracy varies with retrieval quality and confidence threshold, holding the PT-CSFT confidence signal fixed.

Let τ be the confidence threshold (items with PT-CSFT confidence below τ are sent to retrieval). Let p_{fix} be the probability that retrieval converts an incorrect baseline answer to a correct answer (assuming retrieval does not break correct answers, an idealised limit). Let acc_{high} and acc_{low} be PT-CSFT accuracy on the high-confidence ($\geq \tau$) and low-confidence (¡ τ) subsets, and let f_{high} and $f_{\text{low}} = 1 - f_{\text{high}}$ be the corresponding population fractions. The retrieval policy retrieves only for the low-confidence subset (retrieval rates $r_{\text{high}} = 0$, $r_{\text{low}} = 1$); the population fractions, not the retrieval rates, weight each subset's contribution to overall accuracy. The overall gated accuracy is:

$$\text{gated\_acc} = f_{\text{high}} \cdot \text{acc}_{\text{high}} + f_{\text{low}} \cdot [\text{acc}_{\text{low}} + p_{\text{fix}} \cdot (1 - \text{acc}_{\text{low}})]$$

The oracle assumes the PT-CSFT model's own answers are retained for high-confidence items; using baseline answers instead would shift the break-even threshold slightly.

For Gemma 12B on the 966-item RAG eval subset:

Table 18: Oracle analysis of confidence-gated retrieval (Gemma 12B, 966-item subset).

Threshold τ	Retrieval rate	acc_{low}	p_{fix} needed to match baseline (0.751)	Oracle gated acc at $p_{\text{fix}} = 0.7$
50%	23.0%	0.414	0.41	0.786
75%	38.4%	0.470	0.32	0.821
90%	50.6%	0.530	0.27	0.863

At a 50% threshold, retrieval needs to correct only 41% of errors among flagged items to match the no-retrieval baseline; the empirical Wikipedia-evidence p_{fix} of 0.378 falls just below this break-even (which is why the empirical gain is small, +0.007). At higher retrieval thresholds (more aggressive flagging), the break-even threshold falls and oracle gated accuracy rises substantially. For deployments where retrieval quality reaches $p_{\text{fix}} = 0.7$ (achievable with curated knowledge bases and stronger retrievers), the combined system (PT-CSFT + confidence-gated retrieval) exceeds the no-retrieval baseline by 5 to 11 percentage points while retrieving for fewer than half of items.

The corresponding analysis for Qwen 7B with empirical $p_{\text{fix}} = 0.5$ yields gated accuracy 0.735 at the 50% threshold, an improvement of +0.104 over PT-CSFT alone (PT-CSFT baseline accuracy 0.631).

The control benefit therefore scales with retrieval quality; the monitoring signal (the 2.17× and 2.65× accuracy differentials between confidence strata) is the necessary precondition that makes selective gating possible.

B Comprehensive Results

Table 19: Comprehensive results across all models, tasks, and methods. Recovery = (PT-CSFT − 0.5) / (Probe − 0.5) at 7–32B; gap closure = (PT-CSFT − Baseline) / (Probe − Baseline) at 70B+. †GSM8K probe is the CV estimate at the confidence position (0.768), distinct from the training target (0.769). ‡Llama 8B confonly probe is the peak probe (middle layer, T-eval), distinct from the 0.808 confidence-position diagnostic in § 8.4.

Model	Scale	Task	Method	Baseline	PT-CSFT	Probe	Recovery / Closure	Acc Δ
Gemma 12B	12B	TriviaQA	PT-CSFT	0.678	0.841	0.857	96%	−9.1pp
Gemma 27B	27B	TriviaQA	PT-CSFT	0.711	0.780	0.807	91%	−8.4pp
Qwen 7B	7B	TriviaQA	PT-CSFT	0.577	0.801 ± 0.022	0.762	115%	−2.1pp
Qwen 32B	32B	TriviaQA	PT-CSFT (gentle)	0.743	0.847	0.806	113%	−0.2pp
Llama 8B	8B	TriviaQA	PT-CSFT (gentle)	0.692	0.838	0.843	98%	−1.5pp
Mistral 7B	7B	TriviaQA	PT-CSFT (gentlest)	0.583	0.782	0.762	108%	−0.6pp
Llama 70B	70B	TriviaQA	Std. LoRA (×7)	0.724	FAILED	0.834	—	—
Llama 70B	70B	TriviaQA	Bal. confonly logit	0.724	0.746	0.834	20% closure	0pp
Llama 70B	70B	TriviaQA	Curriculum text	0.724	**0.762**	0.834	35% closure	−0.1pp
Llama 70B	70B	TriviaQA	Curriculum logit	0.724	**0.797**	0.834	**66% closure**	−0.1pp
Qwen 72B	72B	TriviaQA	Post-hoc isotonic	0.767	0.767	0.754	Gap closed	0pp
Gemma 12B	12B	GSM8K	E2E logit	0.546	**0.862 ± 0.013**	0.768†	>100%	−1.1pp
Qwen 7B	7B	ARC	Confonly	0.609	0.813	0.725	>100%	0pp
Llama 8B	8B	TriviaQA	Confonly logit	0.692	0.846	0.880‡	—	—
Gemma 12B	12B	NQ (OOD)	No retrain	0.551	0.757	—	137.7% transfer	−7.2pp
Gemma 12B	12B	PopQA (OOD)	No retrain	0.604	0.834	—	154.7% transfer	−4.6pp
Qwen 7B	7B	GSM8K E2E	E2E logit (lr=5e-5)	0.500	0.740	—	—	−39.4pp
Qwen 7B	7B	GSM8K E2E	E2E logit (lr=1e-5)	0.500	0.641	—	—	−13.2pp

Operating envelope. Standard PT-CSFT for verifiable short-answer tasks at 7–32B; two-stage curriculum at 70B+; end-to-end single-pass for reasoning tasks with per-model recipe validation; the logit readout (§ 8) as universal rescue where text confidence is insufficient. Known failure modes: zero-shot format transfer, multi-task SFT on chain-of-thought tasks, end-to-end at 70B on factoid, and end-to-end on architectures with weak correctness encoding (probe $\text{AUROC}_2 < 0.7$).

C Pre-Registration Alignment

Table 20: Pre-registered hypothesis alignment. Phase 1 protocol registered at https://osf.io/ngkwc/.

Pre-registered prediction	Status	Evidence
AUROC_2 exceeds baseline by ≥0.03	**Confirmed**	6/6 models at 7–32B (Δ = +0.06 to +0.22)
Shuffled-target control shows no improvement	**Confirmed**	Gemma 12B shuffled = baseline ± noise
Accuracy drop < 2pp	**Confirmed (gentle)**	−0.2 to −1.5pp
MMLU confidence transfer	Tested (§ 7.4.5)	Positive Δ (+0.027 to +0.104), accuracy collapse
MMLU accuracy preservation	Not met (§ 7.4.5)	Drops 31–69pp (format mismatch)
Policy-invariance (E10)	**Confirmed**	AUROC_2 improves on answer-unchanged items
Logit entropy as baseline comparator	**Confirmed**	Token entropy $\text{AUROC}_2 = 0.47$
Diagnostic probe comparison	Conducted	Documented in § 8.4

D LoRA Sensitivity Ablation Tables

This appendix contains the full per-configuration LoRA ablation tables referenced in § 6. Key findings are stated in the main text; these tables provide the complete numerical evidence.

D.1 Llama 8B

Table 21: LoRA sensitivity on Llama 8B. Learning rate is the dominant factor: reducing lr from 2×10^{-4} to 5×10^{-5} at rank 16 transforms recovery from 4% to 98%.

Config	Rank	LR	Acc	Drop	AUROC$_2$	Recovery	ECE	VRS	E10 Δ
primary	16	2e-4	.659	−10.7pp	.513	4%	.287	Invalid	−0.202
gentle-lr	16	5e-5	.751	−1.5pp	**.838**	**98%**	.128	Invalid	**+0.138**
low-rank	4	2e-4	.747	−1.9pp	.811	91%	**.047**	**Valid**	+0.132
gentlest	4	5e-5	.758	−0.8pp	.796	86%	.088	Invalid	+0.116

D.2 Mistral 7B

Table 22: LoRA sensitivity on Mistral 7B. The model that diverged at r16/lr2e-4 achieves 108% recovery with gentle hyperparameters.

Config	Rank	LR	Acc	Drop	AUROC$_2$	Recovery
primary	16	2e-4	—	—	DIVERGED	—
gentle-lr	16	5e-5	.723	−0.6pp	.758	98%
low-rank	4	2e-4	.636	−9.3pp	.712	81%
gentlest	4	5e-5	.723	−0.6pp	**.782**	**108%**

D.3 Llama 70B (Standard Single-Stage)

Table 23: Llama 70B: seven standard single-stage LoRA configurations spanning rank 4–128. None breaks the ceiling prior. See also Figure 3 for a visual summary.

Config	Rank	LR	AUROC$_2$	Recovery	Conf mean	Conf std	Acc drop
primary	16	2e-4	.582	neg.	—	—	—
gentlest	4	5e-5	.740	79%	98.4	3.0	−0.7pp
gentle-lr	16	5e-5	.711	70%	98.3	5.6	−0.7pp
mid-lr	16	1e-4	.659	52%	98.6	0.5	−3.3pp
r32/mid-lr	32	1e-4	.593	31%	98.1	2.9	−1.6pp
low-rank	4	2e-4	.553	18%	98.9	0.4	−1.1pp
r128/gentle	128	5e-5	.497	chance	97.2	6.9	−20.9pp

E Target Layer Choice

The peak probe is consistently found at the middle layer across all models on TriviaQA. We tested whether using the middle-layer probe as the PT-CSFT training signal produces better results than the standard last-layer probe. On Gemma 12B, middle-layer PT-CSFT achieved AUROC$_2$ = 0.817 with 102.0% probe recovery, VRS-Valid confidence (L = 0.287, TRIN = 0.504), and an accuracy drop of 7.5 percentage points (vs 9.1pp for last-layer). On Qwen 7B, the pattern replicated and strengthened: AUROC$_2$ = 0.793 with 122.7% probe recovery, VRS-Valid confidence (L = 0.006, TRIN = 0.388, r = 0.521), and a well-distributed confidence profile (mean 72.6, std 17.2).

Table 24: Target layer comparison for PT-CSFT training signal.

	Gemma 12B last	Gemma 12B middle	Qwen 7B last	Qwen 7B middle
$AUROC_2$	0.842	0.817	0.754	0.793
Recovery	98.2%	102.0%	112.6%	122.7%
VRS	**Valid**	**Valid**	**Valid**	**Valid**
Accuracy drop	−9.1pp	−7.5pp	−4.2pp	−8.6pp

Note. Recovery in this table normalises against the per-condition target-layer probe, estimated on this appendix's calibration split (for the Qwen 7B last-layer condition, 0.725); this differs from the last-layer TriviaQA probe used for the Table 5 recovery figures (0.762, Table 4). The Qwen 7B last-layer column is a single-seed (seed 42) target-layer experiment: $AUROC_2 = 0.754$ on the original adapter and 0.756 on the retrained adapter.

Both models show higher probe recovery with middle-layer targets, and all four conditions (both layers, both models) reach VRS-Valid confidence (Table 24); the middle-layer advantage is in recovery and robustness, not in VRS tier. This suggests the middle-layer probe provides a cleaner calibration signal, possibly because the middle layer's correctness representation is more robust to the representational reorganisation that PT-CSFT induces (as supported by the post-training probe analysis in § 9.4, where the middle-layer probe drops least after training). The ARC domain generalisation results (§ 7.2) add a caveat: on Qwen 7B ARC, the first-layer probe outperforms the middle layer, suggesting that the optimal layer for target derivation should be determined empirically per domain rather than assumed from TriviaQA results.

F MMLU Cross-Benchmark Format Mismatch

We tested whether PT-CSFT confidence calibration transfers from TriviaQA (training domain) to MMLU (unseen benchmark) under severe answer-format mismatch. PT-CSFT trains on free-text QA but MMLU requires letter-selection (A/B/C/D), creating a stringent test of whether the confidence calibration pathway generalises independently of the task format. Evaluation uses the 498-item M-eval partition (§ 3.2), stratified across six MMLU domains.

Table 25: MMLU cross-benchmark transfer under format mismatch. Paired-bootstrap $AUROC_2$ deltas (10,000 resamples) on matched items.

Model	Paired $AUROC_2$ Δ	Sig?	Abs. $AUROC_2$	VRS after	Acc drop
Gemma 12B	+0.031	NS	0.565	Valid	−31.7pp
Gemma 27B	+0.080	Yes	0.559	Valid	−36.5pp
Qwen 7B	+0.070	Yes	0.592	Valid	−33.3pp
Llama 8B	+0.027	NS	0.501	Invalid	−56.8pp
Llama 70B	+0.104	Yes	0.594	Invalid	−68.9pp

Note: these results use the pre-registered configuration (r16/lr2e-4); cross-benchmark evaluation of the successful LlamaForCausalLM configurations is left to future work.

Under this format mismatch, confidence discrimination shows signs of transfer for the successful architecture classes, despite no MMLU-specific training. Gemma 27B and Qwen 7B show statistically significant positive deltas on matched items. All three successful models achieve VRS-Valid confidence profiles on MMLU. These VRS-Valid classifications reflect distributional shape, not discrimination: absolute $AUROC_2$ on the three Valid models is near chance (0.565, 0.559, 0.592). The format mismatch collapses accuracy and the confidence distribution spreads accordingly; the spread satisfies the VRS shape criteria (low L, low TRIN) without the signal meaningfully tracking correctness. The positive paired deltas indicate a small improvement over

baseline on matched items, but these are not deployment-ready calibration on MMLU. The dissociation between VRS-Valid shape and near-chance discrimination here is itself an instance of the general point developed in § 9.1.

Accuracy drops substantially across all models (31–69 percentage points) because the LoRA teaches TriviaQA-style free-text answers, destroying the MCQA letter format. These results should be interpreted as evidence that PT-CSFT modifies the confidence pathway somewhat independently of the answer-format pathway, not as evidence of deployment-ready cross-domain calibration. Multi-task training on both formats would be needed for practical deployment.