Teaching a 3B model to report confidence: what transfers, what fails, and how selective compute cuts inference cost

André Ramos

2026

Abstract

Small language models are most useful where they run cheaply, and least trusted, because they err with a straight face. We ask whether a 3B model (SmolLM3-3B) can be made *operationally reliable* on consumer hardware along two routes: spending inference compute only where it pays, and training the model to state a confidence that discriminates its own errors. The work is an iterative program with pre-registered confirmatory stages, audited throughout by a second AI system. We find four things. **(1)** A frozen *disagreement cascade* (answer twice cheaply; escalate to sampled voting only on disagreement) is confirmatorily non-inferior to uniform 6-sample voting on GSM8K (Δacc +0.013, 95% CI [+0.000, +0.033], margin −0.02), while using **47% of the tokens and 46% of the time**. **(2)** A *reporter-only* adapter, trained with loss restricted to a confidence turn while the frozen base produces every answer, makes the verbal channel discriminative: distilling a cross-fitted joint teacher (internal max-probability plus verbal confidence) yields mean individual-adapter verbal AUROC 0.871, against 0.753 for the best external use of the raw verbal signal (probability ensemble 0.908, secondary), replicated on an independent 400-item holdout *within the training generator*. **(3)** A matched-codebook ablation supports an independent contribution from per-item teacher information beyond target richness (8/10 paired training seeds positive; hierarchical $P(\Delta>0)=0.994$, descriptive), though train-seed heterogeneity is substantial. **(4)** Across domains, trained reporters transfer where a frozen external calibration map degrades, but the *incremental* advantage of the joint teacher is not confirmed (ΔAURC 95% lower bound −0.008), and Brier-calibration superiority over an external Platt competitor was not confirmed across three measurements (two openings of the same holdout and one independent replication). We release the models, the evaluation battery, per-item artifacts with hashes, and the full pre-registration and audit log (immutable URL plus SHA-256).

---

1. Introduction

A 3B model on a laptop answers in seconds and costs nothing per query. Two defects keep it from being *relied upon*. It spends its one scarce resource, compute, evenly across questions whether or not they need it. And when it is wrong, it sounds exactly as confident as when it is right. Neither defect is a capability limit; both are dispositional, which is what suggests they might be fixable at small scale, without retraining for the capability that is the wall small models run into.

We ask three questions, each answered with pre-registered protocols, frozen analysis plans, and adversarial auditing (Section 8):

Q1. Do small models contain signals that predict their own errors? Yes. Internal max-probability (MCQ), verbalized confidence, and self-consistency agreement all discriminate errors — with sharply different profiles per regime (Section 4).

Q2. Can inference compute be allocated by such signals? Cheap *scores* failed as routers, but cheap *disagreement* did not: a frozen cascade matches uniform voting at half the cost, confirmed by a non-inferiority test (Section 5).

Q3. Can the model be trained to say what it knows — without damaging it? Yes, within domain: a reporter-only adapter raises verbal error-discrimination from AUROC ≈0.73 to a mean individual-adapter 0.871 (probability ensemble 0.908, secondary), with capability untouched by construction. What transfers across domains is the reporter itself, not the incremental advantage of its richer teacher (Section 6).

Contributions: (a) a confirmed, frozen cascade policy for small-model inference; (b) a reporter-only training recipe that makes verbal confidence discriminative without touching answers; (c) a matched-codebook mechanism ablation supporting an independent contribution of per-item teacher information beyond target richness; (d) negative and not-confirmed results documented at parity with positive ones; (e) a public evaluation battery ("Padrão") and a complete pre-registration/audit log ("Roteiro").

2. Related work

LLMs partially "know what they know" via internal probabilities [Kadavath et al., 2022]; verbalized uncertainty can be elicited and trained [Lin et al., 2022]. Self-consistency improves reasoning accuracy [Wang et al., 2022] and, as we show, is also a near-ceiling error detector at 3B. Test-time compute allocation can outperform uniform sampling [Snell et al., 2024]; our cascade is a minimal, confirmatory instance whose router is *disagreement between generation modes* rather than a learned score. Distribution-free selective risk control [Angelopoulos et al., 2021; 2022] motivates our risk-coverage reporting; we use proper scoring theory [Gneiting & Raftery, 2007] as a design constraint, not a guarantee.

3. Setup

Model & hardware. SmolLM3-3B (8-bit MLX), single Apple M4, ~3.7 GB peak memory; every experiment in this paper runs on one consumer machine.

Domains. (D-syn) a parameterized synthetic ordering/transitivity MCQ generator (held-out seeds per split; bodies disjoint across splits, hashed); (ARC) ARC-Challenge 4-option items; (HS) HellaSwag formatted as 4-option MCQ (disjoint slices by eligible position, dataset fingerprint recorded); (GSM8K) free-form math with a tolerant semantic scorer validated against blind annotation (98.1% binary agreement, 0 false-corrects; in-sample — tuned and validated on the same 106-output challenge set).

Signals. S1: max softmax probability over single-token option logits (normalized for option count); S6: verbalized confidence ("Confiança: N%") elicited in a second turn conditioned on the given answer; S4: fraction of M=6 temperature-0.7 samples whose final answer matches the frozen greedy answer.

Methodology. Every confirmatory experiment was pre-registered (sample, policy, estimand, gate) before measurement; holdouts opened once (single exception: the seed-4 holdout was re-elicited

once for audit reproducibility, which exposed the process-state caveat of Section 7.3); teachers and calibration maps frozen before evaluation; artifacts dumped per item with SHA-256 manifests; final reporter evaluations conducted after detection use fresh-process elicitation (process-state nondeterminism was detected and neutralized this way — Section 7.3; earlier exploratory measurements predate this practice). A second AI system (OpenAI Codex), acting as an adversarial auditor, reviewed every protocol and result and forced three retractions of our own over-claims; the complete log is released.

4. Which signals predict errors? (the "bake-off")

In-domain, internal signals are strong: on D-syn, S1 reaches AUROC 0.863 [0.799, 0.918]; on a paired ARC confirmatory set, S1 = 0.709 and S6 = 0.677 (both permutation $p \leq 0.0001$, Holm-corrected). In generation (GSM8K, n=150, 16 operational errors), **S4 is near-ceiling: AUROC 0.998 [0.994, 1.000]**, and even failure-mode signals (sample truncation, parse failures, modal agreement) reach 0.82–0.87, while S6 is weak (0.612).

Two transfer facts shape everything downstream: **ranking transfers, zero-shot calibration does not** (a Platt map fitted on D-syn degrades on ARC: fitted slope 1.93 vs ~0.2 descriptive ideal; transferred Brier worse than a constant), and ~40 destination labels usually restore calibration ($P(\Delta>0)\approx 0.93$–0.98 across K=200 random label subsets — subset stability, not a guarantee). A joint logistic combination S1+S6 was always the best MCQ candidate descriptively but its incremental gain over S6-alone did not replicate at the pre-registered gate (ΔAUROC one-sided LB95 $= -0.000$).

Cheap signals as *escalation routers* failed in generation: verbal confidence over a no-thinking answer captured only 4–6 of 19 rescuable errors at relevant budgets (within the random-routing interval), and free logprob features plus a linear router were at chance out-of-fold.

5. The disagreement cascade (confirmatory)

Policy, frozen before the confirmation block: generate A0 (thinking-disabled) and A1 (thinking); if their normalized answers agree, answer; on disagreement (including any unparseable output), draw up to six thinking samples with sequential early stopping (stop when no rival can catch the current leader) and answer by vote.

On a fresh GSM8K block (items 190–339, full counterfactual: A0, A1 and all six samples generated for every item), the pre-registered primary — non-inferiority vs uniform 6-sample voting, margin -0.02 accuracy — **passed**:

policy	accuracy	tokens (total)	time
A0 (thinking-disabled)	0.807	—	median 10.0 s/item
A1 (thinking)	0.927	—	—
uniform 6-vote	0.933	719,181	423 min
cascade (frozen)	**0.947**	**334,588 (47%)**	**196 min (46%)**

Δ(cascade − uniform) = +0.0133, 95% CI [+0.0000, +0.0333]; superiority is *not* claimed (2 favorable discordant pairs; McNemar p=0.25). 32/150 items escalated; sequential stopping used 4.7/6 samples on average. Auxiliary finding: thinking alone is a risky step (A0→A1: 18 rescues, 10 regressions); the value concentrates in the vote (A0→A2: 19 rescues, 2 regressions). See Fig. 2.

Frozen claim: on SmolLM3-3B/GSM8K, the disagreement cascade preserved uniform-voting accuracy within 2 points using ≈47% of the tokens and ≈46% of the time. No generalization to other models, domains or prompts is claimed.

6. Teaching the verbal channel (reporter-only training)

Architecture. The main answer is always produced by the frozen base; a LoRA adapter is activated only for the turn "what is your confidence?", with loss restricted to the confidence completion (~10 supervised tokens/example). This removes, by construction, the catastrophic failure modes we measured in naive confidence SFT (capability collapse, thinking suppression, style drift — all documented in the released log). An automatic architectural gate verifies byte-identical main answers across arms.

Teachers (cross-fitted, 5-fold; targets never use the item's own label). Arm *s6* (control): Platt-calibrated verbal confidence — a monotone transform of what the model already says. Arm *joint*: logistic S1+S6. 200 labeled items (acc 0.53; 180 LoRA-train + 20 validation, all 200 used for cross-fitting), 3 training seeds per arm, identical items/prompts/budget (90 iterations = exactly one epoch), checkpoint selection on a dev split only.

In-domain result (holdout seed 4, opened once, n=150). Joint reporters: Brier 0.136–0.143, AUROC 0.88–0.90 across all three seeds (s6 control: 0.186–0.205 / 0.72–0.75; raw S6 0.358; constant 0.249; frozen external S6+Platt 0.181; frozen teacher reference 0.115). The pre-registered paired contrast **joint-reporter > s6-reporter** was later replicated on an independent 400-item holdout (seed 5, 3 fresh processes per model, per-seed mean-loss estimand): ΔBrier +0.0338, 95% CI [+0.0101, +0.0565]; per-run mean AUROC 0.871 vs 0.728 (external 0.753). The pre-registered primary against the *external* competitor (Brier of joint-reporter < frozen S6+Platt) was **marginal in all three measurements** — two openings of the same seed-4 holdout and one independent 400-item replication (Δ +0.041/+0.038/+0.020; one-sided LB95 +0.0015/−0.0010/−0.0004): we report **no confirmed calibration superiority; the data are compatible with a null or small benefit**. The reporter's confirmed value is *discrimination*, not absolute calibration.

Mechanism (matched-codebook ablation, descriptive causal language). To separate per-item information from target richness, both arms were retrained on the same 200 labeled items (180 LoRA-train + 20 validation; all 200 in cross-fitted ranking construction) with the *identical* 9-value target codebook and frequencies (equal token lengths per item; identical length-sorted batching, asserted programmatically; deterministic SHA-256 tie-breaking), differing only in which item gets which value (ranked by each teacher). The pre-registered item-conditional gate passed (ΔAUROC +0.051, one-sided 5th percentile +0.020). With only the three original training seeds, a hierarchical (items × seeds) re-analysis did NOT support robustness to training randomness (5th percentile −0.061, P(Δ>0)=0.767); extending to 10 paired training seeds, the same hierarchical analysis gives mean +0.104, 5th percentile +0.035, P(Δ>0)=0.994 — *descriptive*, as the synthetic holdout had been opened previously. With these training runs fixed, equalizing target richness did not eliminate the joint teacher's average advantage — supporting an independent contribution from per-item teacher information — with substantial train-seed heterogeneity (exactly 2/10 pairs negative in-generator; 3/10 in the cross-domain confirmation; Fig. 3).

Cross-domain, part A — zero-shot transfer (HellaSwag, eligible items 501–650, n=150; original tt4 arms). Both trained reporters beat the frozen external map zero-shot: it degrades to Brier 0.282 (worse than constant 0.255) while the reporters — mean individual-adapter, the estimand tested — reach 0.1898 (s6) and 0.1954 (joint) (vs external: s6 Δ+0.092, LB95 +0.052;

joint Δ+0.087, LB95 +0.034); probability ensembles, secondary: 0.1749/0.1828. Destination recalibration with 20/40 labels does *not* improve the trained reporters — their calibration travels in the weights — while raw S6 recalibrated with 40 labels (0.210) still trails the zero-shot reporters.

Cross-domain, part B — incremental confirmation (HellaSwag, eligible items 651–1050, n=400, frozen sample, 10 matched-codebook pairs, fresh processes). The pre-registered AURC gate for the *incremental* advantage of joint-9 over s6-9 was **not confirmed**: Δ +0.026, 7/10 pairs positive, hierarchical (items × seeds) one-sided 5th percentile −0.008. AURC means (no formal test): joint-9 0.4263, s6-9 0.4526, frozen external 0.4624; oracle 0.2389, random 0.6055. Exploratorily, joint ranks better (AUROC 0.751 vs 0.723) and calibrates worse (Brier 0.218 vs 0.205). See Fig. 1.

7. Claims, limitations, and a measurement caveat

7.1 Claims table (frozen with the auditor)

claim	status	domain
Joint reporter improves verbal discrimination over the S6-trained control	**Confirmed & replicated**	synthetic generator
Per-item teacher information contributes even with target richness equalized; training variance is real	**Descriptive support** (strong; 10 pairs)	synthetic generator
Disagreement cascade non-inferior to uniform voting at 47% of tokens	**Confirmed** (operational)	SmolLM3 / GSM8K
Brier superiority over external S6+Platt	**Not confirmed** (3 attempts, all marginal)	synthetic generator
Incremental joint advantage cross-domain (AURC)	**Not confirmed**	HellaSwag
Joint ranks better / calibrates worse cross-domain	Exploratory	HellaSwag
Trained reporters transfer zero-shot where external map degrades	Descriptive	HellaSwag

7.2 Limitations

One 3B model; capability preservation is established **only for the reporter-only pipeline** (the adapter was never active while answering, and we did not test it so); confirmatory training results live in a synthetic MCQ generator; train-seed heterogeneity is substantial and all training results are reported with per-seed uncertainty; GSM8K prompts are Portuguese (scorer validated against blind annotation); the cascade is confirmed for one model/dataset/prompting configuration.

7.3 Measurement caveat: process-state non-determinism

Identical greedy elicitations differ across *process histories* (per-seed Brier shifts ~±0.01 flipped a marginal gate from pass to fail), while fresh processes are byte-reproducible (0 differing outputs

across 8,400 repeated elicitations). All final reporter evaluations, conducted after detecting this issue, use fresh-process elicitation; the root cause (suspected Metal/MLX state) is unproven. We recommend fresh-process evaluation as standard practice on this stack.

8. The audited workflow ("Roteiro")

Every confirmatory step was pre-registered and the full lab log is released: pre-registrations, gates, three forced retractions of our own announced milestones, two date-keeping errors, one statistical error (bootstrap-null → permutation tests, with robustness re-checks), and one memory-safety incident with its fix. The adversarial auditor (a second AI system) reproduced our numbers from artifacts and rejected over-claims; we believe the resulting claims table is unusually load-bearing for its size, and we offer the log itself as a case study in AI-audited empirical work.

9. Reproducibility & artifacts

Code, JOURNAL (full log), pre-registrations, per-item dumps, SHA-256 manifests, frozen teachers/calibrators, evaluation battery ("Padrão"), and figures. Selected adapter checkpoints (`bojador_adapters_v1.tar.gz`, sha256 `22aad81c…`) are deposited at an immutable Hugging Face revision: `huggingface.co/ajdramos/bojador-reporter-smollm3-3b`, revision `fc65bd67cb17d3e6383435e7cda7226270cffdca` (made public at launch). Seeds and splits are exact; every table number maps to a versioned artifact.

Acknowledgments

This project was executed by a human researcher with two AI systems in adversarial roles (assistant/executor and auditor). All claims were verified against artifacts; errors that survived are the human's and the assistant's.

Figures

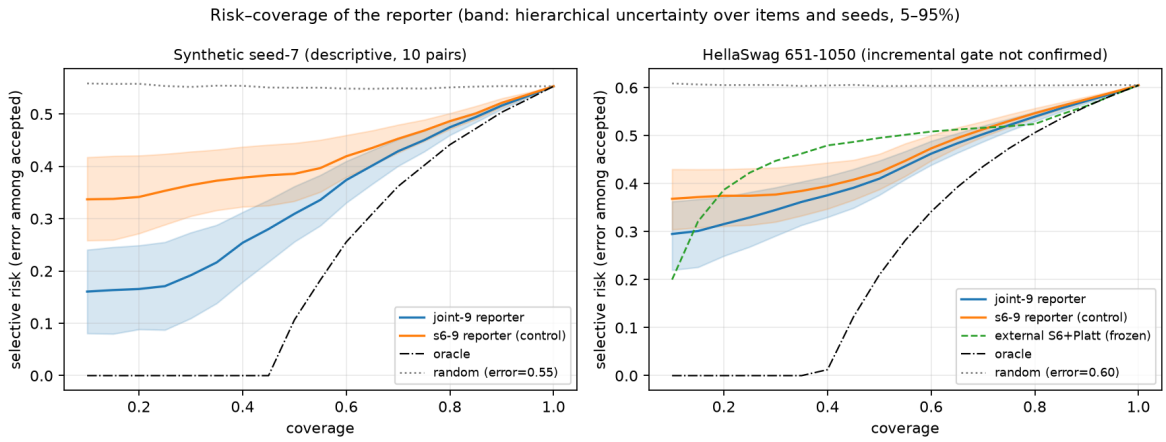

Figure 1: Risk–coverage of the reporter on synthetic seed-7 (descriptive, 10 pairs) and HellaSwag (incremental gate not confirmed); bands = hierarchical uncertainty over items and seeds (5–95%), SHA tie-breaking; reference curves: frozen external S6+Platt, oracle, random.

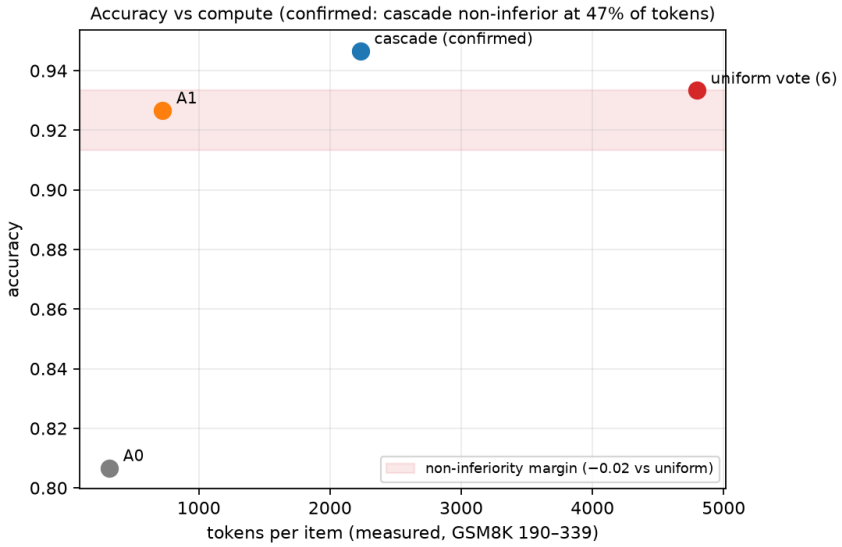

Figure 2: Accuracy vs compute on GSM8K (190–339): A0, A1, uniform 6-vote, and the frozen cascade; tokens per item on the x-axis; the non-inferiority margin (-0.02 vs uniform) shaded.

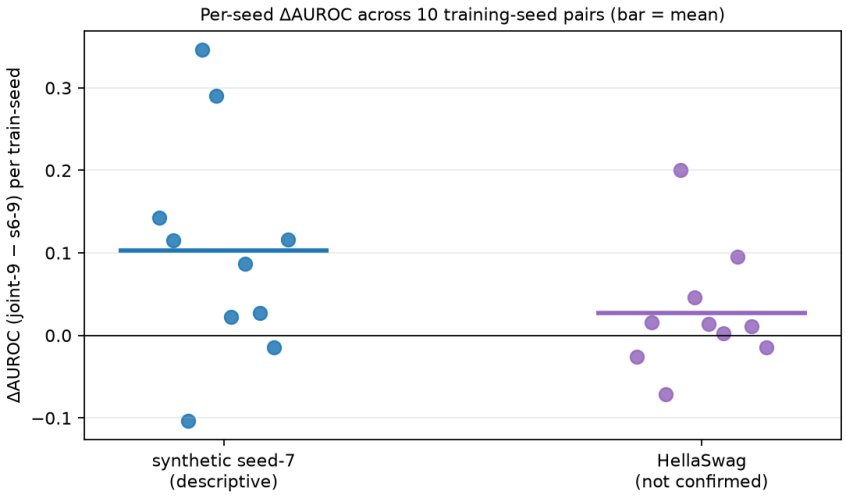

Figure 3: Per-train-seed AUROC difference (joint minus control) across 10 pairs, in the synthetic generator and cross-domain; bar = mean.

References

- Kadavath et al. (2022). *Language Models (Mostly) Know What They Know.* arXiv:2207.05221.
- Lin et al. (2022). *Teaching Models to Express Their Uncertainty in Words.* arXiv:2205.14334.
- Wang et al. (2022). *Self-Consistency Improves Chain-of-Thought Reasoning in Language Models.* arXiv:2203.11171.
- Snell et al. (2024). *Scaling LLM Test-Time Compute Optimally…* arXiv:2408.03314.
- Angelopoulos et al. (2021). *Learn then Test.* arXiv:2110.01052.
- Angelopoulos et al. (2022). *Conformal Risk Control.* arXiv:2208.02814.
- Gneiting & Raftery (2007). *Strictly Proper Scoring Rules, Prediction, and Estimation.* JASA 102(477).