MOST BLACK-BOX UNCERTAINTY SIGNAL COMES FOR FREE: AN ATTRIBUTION STUDY OF SINGLE-PASS LLM CORRECTNESS PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable LLM deployment needs uncertainty estimates that are both faithful and cheap, yet the strongest methods resample the model, while single-pass methods trade accuracy for cost. We ask how much response-correctness signal a single generation's token-wise predictive distributions actually carry once three everyday factors are accounted for: response length, supervision, and model access. To answer, we build PULSE, a supervised sequence probe that reads only the per-step top-k log-probabilities of a single generation, without access to hidden states or weights, and compare it with reference probes that read progressively more of a response, across five LLMs and three mathematical reasoning benchmarks. Most of the black-box signal comes for free: response length alone is a strong predictor, and a zero-shot entropy score comes within about two AUROC points of PULSE. Supervision adds nothing over these free scalars on its own; reading the full top-k channel adds about a point, and modeling sequence order a smaller increment on top. White-box hidden states, by contrast, open a much larger gap. Within the black-box setting, PULSE is the strongest option we test, never significantly worse than the best baseline and matching a five-pass perturbation-based estimator at one-fifth of its cost. The resulting attribution shows where black-box uncertainty signal comes from, and how little of it lies beyond free, zero-shot scores.

1 INTRODUCTION

Large language models (LLMs) are increasingly deployed in settings where an incorrect answer is costly: mathematical and scientific reasoning, code generation, retrieval-augmented question answering, and decision support (Bubeck et al., 2023). In all of these, a fluent response can still be wrong (Huang et al., 2023), so reliable deployment depends not only on raw accuracy but also on *uncertainty quantification* (UQ)—a faithful, response-level estimate of whether a generated answer is likely to be correct (Shorinwa et al., 2024). Such estimates enable abstention, selective verification, and retrieval triggering. A natural question is whether an LLM, like a speaker whose tone can betray doubt behind fluent words, leaves comparable non-verbal traces alongside the tokens it generates.

Two broad families dominate current practice, and they trade faithfulness against efficiency. *Multi-sample* methods estimate uncertainty from disagreement across several generations, for example self-consistency (Wang et al., 2023), semantic entropy (Kuhn et al., 2023), and confidence-filtered reasoning (Fu et al., 2025): often strong, since they operate near the answer or semantic level, but $K\times$ the inference cost. *Single-pass* methods instead read uncertainty from one ordinary generation, using scalar signals such as mean token log-probability, sequence likelihood, predictive entropy (Malinin & Gales, 2021), and self-certainty (Kang et al., 2025), or elicited confidence (Kadavath et al., 2022; Lin et al., 2022). These are cheap, but compress the whole generation into one hand-crafted statistic, discarding the temporal structure of the token-wise predictive distributions.

This paper asks a narrower, more testable question: how much response-correctness signal is actually present in a single generation's top-k log-probability trajectory, and where does it come from—how much is response length, how much is bought by supervision, how much survives dropping sequence order, and how much is left on the table without white-box access? A single decode produces, at every step, a full next-token distribution; scalar summaries capture this trace only in part, and recent trajectory-based methods (Zhu et al., 2026; Li et al., 2026; Xu & Lu, 2025) confirm it is informative but predefine which statistic should matter, so they cannot answer *how much* signal a fixed choice forgoes.

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

We answer by decomposition: comparing four probes that read progressively more of a response—response length, a zero-shot hand-crafted entropy score, an order-agnostic supervised classifier, and a white-box hidden-state probe—against a supervised sequence probe we build for this purpose, PULSE (*Predictive Uncertainty from Logit-Sequence Evidence*), which reads only the per-step top-k token log-probabilities of a single generation, with no access to hidden states, weights, or extra forward passes (architecture in Section 4). The answer is that most of the black-box signal comes for free: the zero-shot entropy score TECP already attains about 94% of PULSE's above-chance AUROC, and supervision, channel width, and sequence order together add under two points. We run this decomposition on exact-match mathematical reasoning, where correctness is unambiguous to grade; a smaller, single-seed check on free-form QA (Appendix S) finds the qualitative picture holds—PULSE remains competitive, and length is no longer a confound there—though it is not a full replication of our 15-cell, 3-seed math results (Section 6).

To compare methods fairly, we adopt a *cost-tier* protocol that separates raw performance from inference cost and scores every method on the same fixed response and correctness label. The decomposition itself is our main contribution; concretely:

- We **attribute the signal rather than assert it**, across four reference probes at 15 model–dataset configurations: response length is a strong baseline, zero-shot TECP nearly matches a trained readout, supervision over the full top-k channel adds about one point, sequence order adds a smaller, mostly non-significant residual, and a white-box probe reopens a much larger gap (Section 5.3, Appendix Q).
- As the instrument for this decomposition, we build **PULSE**, a single-pass ($\times 1$-cost), *black-box* UQ estimator that learns a readout over only the per-step top-k log-probabilities (Section 4); it is never significantly worse than the best black-box baseline and matches the far more expensive perturbation-based TokUR at a fraction of the cost.
- We introduce a **cost-tier, sign-oriented, same-response protocol** that fairly compares single-pass and low-cost methods, including hand-crafted entropy-trajectory baselines (Section 5).
- We show the readout is **partially transferable** across models and datasets, and that the trajectories encode **model-specific structure**—enough to identify the source model—which bounds and explains cross-model transfer (Section 5.4).

2 Related Work

Uncertainty quantification for free-form generation. Unlike classification calibration over a fixed label space (Guo et al., 2017), free-form generation produces long outputs whose correctness may hinge on a few facts or steps, so raw likelihood or average confidence can mislead: fluent continuations may be false, while correct responses may still involve substantial token-level uncertainty.

White-box and model-level UQ. A classical line represents predictive uncertainty through model- or parameter-level mechanisms: Bayesian neural networks (MacKay, 1992; Neal, 1996), variational inference (Blundell et al., 2015), Monte Carlo dropout (Gal & Ghahramani, 2016), Laplace approximations (Ritter et al., 2018), deep ensembles (Lakshminarayanan et al., 2017), and evidential learning (Sensoy et al., 2018; Malinin & Gales, 2018). These give principled estimates but require weights, retraining, ensembling, or repeated stochastic passes—difficult for black-box LLM generation, which our setting avoids by using only the token predictive distributions from a single ordinary generation.

Cross-sample UQ. A major line estimates uncertainty from disagreement across multiple generations: self-consistency aggregates final answers from multiple reasoning paths (Wang et al., 2023), semantic entropy computes uncertainty over semantic equivalence classes (Kuhn et al., 2023), SelfCheckGPT detects hallucination via resample consistency without a knowledge base (Manakul et al., 2023), and DeepConf filters or weights reasoning traces by model-internal confidence (Fu et al., 2025)—often strong, but more expensive and sensitive to sampling/clustering choices.

Single-pass UQ. Single-pass methods extract uncertainty from one response: mean token log-probability, sequence likelihood, probability margins, predictive entropy (Malinin & Gales, 2021), and self-certainty (Kang et al., 2025) (average divergence of the next-token distribution from uniform); others elicit confidence from the model itself, such as P(True) (Kadavath et al., 2022), verbalized confidence (Lin et al., 2022), and targeted calibration prompting (Tian et al., 2023). These are cheap, but compress the generation into one predefined scalar, discarding the temporal structure of the token-wise predictive distributions.

Trained probes and learned readouts. Closest to our approach is a line that *learns* a correctness or uncertainty readout instead of fixing a statistic: P(IK) trains a lightweight head on model features to predict whether the model knows an answer (Kadavath et al., 2022), hidden-state classifiers read truthfulness directly from internal activations (Azaria & Mitchell, 2023), and Semantic Entropy Probes regress semantic entropy from a single generation's hidden states (Kossen et al., 2024). Closest to PULSE's black-box, top-k-only input, Quevedo et al. (2024) classify hallucinations from pooled token-probability statistics—the same family as our order-agnostic pooled control, which we find already recovers most of what PULSE does (Section 5.3)—and APRICOT trains an auxiliary confidence predictor from generated text alone, no token probabilities at all (Ulmer et al., 2024). Apart from these two, almost all trained-readout work needs *white-box* access to hidden states or weights. PULSE keeps the "learn the readout, don't fix it" idea but reads the full per-step top-k log-probability *sequence* of a single generation, not just pooled summaries of it.

Token-level and entropy-trajectory uncertainty. Recent work studies uncertainty as a process distributed across generated tokens. TokUR estimates token-level uncertainty for mathematical reasoning via low-rank random weight perturbations during decoding (Zhang et al., 2025). Others exploit entropy trajectories through hand-crafted diagnostics: EDIS uses entropy-instability patterns (Zhu et al., 2026), entropy-trend methods trigger retrieval from token-level uncertainty changes (Li et al., 2026), and TECP uses token entropy as a conformal nonconformity score (Xu & Lu, 2025). TAD learns a similar per-step readout over richer attention-level inputs, for calibration rather than response-level correctness (Vazhentsev et al., 2024); LM-Polygraph benchmarks a broad suite of such baselines in one framework (Fadeeva et al., 2023; Vashurin et al., 2025), without our top-k-only restriction or length/order/white-box decomposition. These token-level lines show that trajectories carry useful information but predefine which statistic or pattern should matter, leaving open how much a fixed statistic forgoes. We answer this empirically by learning the readout instead, and find that the answer is: not much.

3 LLM Generation and Response-Level Uncertainty

An autoregressive LLM parameterized by θ, with vocabulary $\mathcal{V}$, generates a response $y = (y_1, \ldots, y_T)$ to a prompt x one token at a time: $p_\theta(y \mid x) = \prod_{t=1}^{T} p_\theta(y_t \mid x, y_{<t})$. At each step t the model produces logits $z_t \in \mathbb{R}^{|\mathcal{V}|}$ and a next-token predictive distribution

$$p_t(v) = p_\theta(y_t = v \mid x, y_{<t}) = \frac{\exp(z_{t,v})}{\sum_{u \in \mathcal{V}} \exp(z_{t,u})}, \qquad v \in \mathcal{V}. \tag{1}$$

The full generation therefore induces a distribution trajectory $(p_1, \ldots, p_T)$, one predictive belief per decoding step.

Given a prompt x, response y, and correctness label $c \in \{0, 1\}$ ($c = 1$ correct), our goal is a response-level uncertainty score $u(x, y) \in \mathbb{R}$ whose larger values indicate a higher probability of an incorrect response, ranking incorrect responses above correct ones to support decisions such as abstention or additional computation. We focus on the *single-pass* setting: y is generated once, and u may use only the token predictive distributions from that ordinary generation—no additional sampled responses, perturbations, auxiliary queries, or weight access. We ask how much of the correctness signal in this trajectory a single-pass score can recover, and which properties of the trajectory it comes from.

4 Method

PULSE is the instrument for our decomposition: a supervised sequence readout over a compact top-k summary of the trajectory, so that its gaps to cheaper probes measure what supervision, channel width, and sequence order each contribute. It rests on the view that the sequence of next-token predictive distributions carries useful evidence about response-level correctness, echoing the observation that higher-order structure in a model's own outputs supports metacognitive error estimation (Ma et al., 2025). Learning a UQ functional directly from the full trajectory $(p_1, \ldots, p_T)$ is impractical since each p_t spans the entire vocabulary; we instead summarize each distribution by a compact evidence vector and learn a readout over the resulting trajectory.

4.1 TOP-k LOG-PROBABILITY EVIDENCE

At each decoding step we summarize the predictive distribution p_t by the log-probabilities of its k most likely tokens (its top-k):

$$\phi_t = \left[\log p_t^{(1)},\ \log p_t^{(2)},\ \ldots,\ \log p_t^{(k)}\right]^\top \in \mathbb{R}^{d_\phi}, \qquad d_\phi = k, \tag{2}$$

where $p_t^{(1)} \geq p_t^{(2)} \geq \cdots \geq p_t^{(k)}$ are the k largest values of $p_t(v)$ over $v \in \mathcal{V}$, sorted in descending order. This evidence vector depends only on p_t, not on the realized token y_t: it describes the local shape of the distribution around its mode and contains hand-picked summaries such as the top-1–top-2 gap, as well as a truncated approximation of predictive entropy. The realized token's own log-probability is recovered only when y_t falls within the top-k, which under temperature-0.8 sampling is model-dependent (from $\geq 98\%$ of steps for the stronger models to 45–69% for the weakest, Llama-3.2-1B)—so ϕ_t is a summary of local distributional shape, not a guaranteed proxy for the realized token's own confidence. We use log-probabilities because the top-ranked values span several orders of magnitude across positions and models, and truncate to a small top-k (fixed in Section 5) because these few ranks typically carry most of the probability mass. The full response is then represented by an evidence trajectory

$$\Phi(x, y) = (\phi_1, \ldots, \phi_T). \tag{3}$$

Each ϕ_t is read from the raw temperature-1 softmax equation 1 on the unscaled logits z_t: sampling temperature affects only which tokens are realized, not the softmax ϕ_t is read from. The whole trajectory comes from one ordinary generation—no extra sampling, perturbations, or model access. The sequence-level readout also uses the normalized generation position $\rho_t = (t-1)/\max(T-1, 1) \in [0, 1]$, which enters the model only through the position encoding (Eq. 5), not through ϕ_t.

4.2 PULSE

We propose PULSE (*Predictive Uncertainty from Logit-Sequence Evidence*), a learned UQ functional $u_{\mathrm{PULSE}}(x, y) = g_\psi(\Phi(x, y))$ over the evidence trajectory. Concretely, g_ψ standardizes each channel using the training-set mean μ_j and standard deviation s_j (with a small constant ϵ for numerical stability),

$$\tilde{\phi}_{t,j} = \frac{\phi_{t,j} - \mu_j}{s_j + \epsilon}, \qquad j = 1, \ldots, d_\phi, \tag{4}$$

projects the normalized trajectory $\tilde{\Phi} = (\tilde{\phi}_1, \ldots, \tilde{\phi}_T)$ into a hidden space with layer normalization (Ba et al., 2016), a GELU activation (Hendrycks & Gimpel, 2016), and an additive sinusoidal position encoding $e_{\mathrm{pos}}(\rho_t)$ (the standard Transformer sin/cos schedule, evaluated at the continuous normalized position ρ_t defined above rather than the integer index t),

$$h_t^{(0)} = \mathrm{GELU}\left(\mathrm{LN}(W_{\mathrm{in}}\tilde{\phi}_t + b_{\mathrm{in}})\right) + e_{\mathrm{pos}}(\rho_t), \tag{5}$$

processes the embedded sequence with a causal Transformer (Vaswani et al., 2017) whose mask restricts position t to positions up to t, additionally using rotary position embeddings (Su et al., 2021) inside attention (e_{pos} encodes global progress and RoPE encodes relative distance), and aggregates the contextualized states $(h_1, \ldots, h_T)$ into one response-level representation by attention pooling,

$$\omega_t = \frac{\exp(w_{\mathrm{pool}}^\top h_t)}{\sum_{s=1}^{T} \exp(w_{\mathrm{pool}}^\top h_s)}, \qquad r = \sum_{t=1}^{T} \omega_t h_t, \tag{6}$$

where w_{pool} is a learned pooling vector letting PULSE emphasize the positions whose evidence is most predictive of correctness. The causal mask matches generation order; a bidirectional encoder is a reasonable alternative that we did not ablate. Hyperparameters are given in Section 5.

4.3 EVIDENTIAL PREDICTION HEAD

A prediction head maps r to two Dirichlet concentration parameters,

$$\alpha = \mathrm{softplus}(W_2\,\mathrm{GELU}(W_1 r + b_1) + b_2) + 1 = (\alpha_{\mathrm{corr}}, \alpha_{\mathrm{incorr}}), \tag{7}$$

where $\mathrm{softplus}(\cdot) + 1$ ensures $\alpha_{\mathrm{corr}}, \alpha_{\mathrm{incorr}} \geq 1$; following evidential classification (Sensoy et al., 2018; Malinin & Gales, 2018), α parameterizes a two-class Dirichlet distribution over correctness. With $S = \alpha_{\mathrm{corr}} + \alpha_{\mathrm{incorr}}$, the predicted correctness probability is $\hat{p}_{\mathrm{corr}} = \alpha_{\mathrm{corr}}/S$ and the PULSE uncertainty score is

$$u_{\mathrm{PULSE}}(x, y) = 1 - \hat{p}_{\mathrm{corr}} = \frac{\alpha_{\mathrm{incorr}}}{S}, \tag{8}$$

larger values indicating a higher predicted probability of an incorrect response. The same α also yields a Dirichlet vacuity score $v_{\mathrm{PULSE}} = 2/S$ (Sensoy et al., 2018), which we leave to future study, using the evidential parameterization mainly for u_{PULSE}. A plain BCE head performs comparably (Section 5.5), so the evidential form is not the source of PULSE's accuracy.

4.4 Training Objective

PULSE is trained on examples $\mathcal{D} = \{(x_i, y_i, c_i)\}_{i=1}^{N}$. We encode correctness as $\mathbf{e}_i = (1, 0)$ if $c_i = 1$ and $\mathbf{e}_i = (0, 1)$ otherwise. For a single example with label vector $\mathbf{e} = (e_{\mathrm{corr}}, e_{\mathrm{incorr}})$ and Dirichlet parameters α, we minimize an evidential classification objective (Sensoy et al., 2018), averaged over $\mathcal{D}$:

$$\mathcal{L} = - \sum_{m \in \{\mathrm{corr}, \mathrm{incorr}\}} e_m \left[\psi(\alpha_m) - \psi(S)\right] + \lambda_{\mathrm{KL}} \mathrm{KL}\left(\mathrm{Dir}(\tilde{\alpha}) \| \mathrm{Dir}(\mathbf{1})\right), \tag{9}$$

where ψ is the digamma function and

$$\tilde{\alpha} = \mathbf{1} + (\mathbf{1} - \mathbf{e}) \odot (\alpha - \mathbf{1}).$$

The KL term regularizes unsupported evidence toward a uniform Dirichlet prior (Sensoy et al., 2018); training hyperparameters are in Section 5. At test time each prompt is decoded once and $u_{\mathrm{PULSE}}(x, y)$ read off directly, so PULSE stays in the $\times 1$ inference-cost tier while replacing a fixed scalar summary with a learned functional.

5 Experiments

We first compare PULSE with single-pass and higher-cost baselines under a shared protocol (Section 5.2), then decompose its advantage into length, supervision, channel width, sequence order, and white-box access (Section 5.3). The remaining subsections examine transfer across models and datasets, which evidence features matter, and how much labeled data PULSE needs.

5.1 Experimental Setup

Models and datasets. We evaluate PULSE on five instruction-tuned LLMs, grouped by family: Llama-3.2-1B (Grattafiori et al., 2024), Llama-3.1-8B (Grattafiori et al., 2024), Qwen2.5-1.5B (Qwen Team, 2024), Qwen3-1.7B (Qwen Team, 2025), and gemma-4-E2B (Gemma Team, Google DeepMind, 2026). The datasets are GSM8K (Cobbe et al., 2021), MATH5000—the full Hendrycks MATH test split of 5000 competition problems (Hendrycks et al., 2021), a larger, lower-variance superset of MATH500 (Lightman et al., 2023)—and DeepScaler (Luo et al., 2025), spanning grade-school arithmetic, competition mathematics, and harder reasoning traces. All five models run on all three datasets (a fully square grid). Unless noted, responses are sampled at temperature 0.8 and scored against the binary error label $1 - c$. Length caps were chosen per model/dataset to make truncation negligible: none of the $17{,}475$ responses across the 15 test sets are truncated, so labels reflect genuine reasoning outcomes rather than length-cap artifacts.

Architecture, training, and metrics. We instantiate the top-k evidence of Section 4.1 with $k = 10$ and the architecture of Section 4.2 with hidden size $d = 64$, two causal Transformer layers, four attention heads, feed-forward dimension 128, rotary positional attention (Su et al., 2021), and dropout 0.2. We split by question id to avoid leakage across generations of the same prompt, fit the feature normalizer on the training split only, and train with the evidential objective of Section 4.4 ($\lambda_{\mathrm{KL}} = 0.1$) using AdamW (Loshchilov & Hutter, 2019) at learning rate 10^{-3}, batch size 32, gradient clipping 1.0, cosine decay (Loshchilov & Hutter, 2017), and early stopping on validation AUROC. Our primary metric is AUROC for detecting incorrect responses (AUROC$\times$100); we also report AUPRC (incorrect as positive class), sensitive to the base error rate; higher is better for both.

Baselines and cost tiers. We compare methods at their canonical inference cost. Cost 1 (one ordinary generation): PULSE and single-pass baselines self-certainty (SelfCert, next-token divergence from uniform (Kang et al., 2025)), log-probability signals, mean token predictive entropy (PredEnt), verbalized confidence, and the hand-crafted entropy-trajectory baselines EDIS (Zhu et al., 2026) and TECP (Xu & Lu, 2025), computed from the same per-token entropy trajectory (formulas in Appendix G). Note that these entropy-based scores and self-certainty are computed from the full next-token distribution, whereas PULSE sees only its top-10 log-probabilities; the comparison therefore gives the zero-shot baselines strictly more distributional information than PULSE. Cost 2: P(True), one extra query asking whether the answer is true (Kadavath et al., 2022). Cost 5: TokUR (Zhang et al., 2025), five perturbed forward passes over the *same* response, the closest

Model	Dataset	PULSE	SelfCert	LogP	PredEnt	TECP	EDIS	P(True)	Verb.	TokUR
Llama-3.2-1B	GSM8K	**81.8±0.4**	68.9	73.6	75.8	77.4	77.4	52.2	54.9	70.2
Llama-3.2-1B	MATH5000	**82.3±1.0**	72.6	67.3	69.2	82.2	81.8	54.4	57.4	70.1
Llama-3.2-1B	DeepScaler	**81.5±0.4**	67.6	70.3	72.1	81.3	81.0	52.3	61.4	64.8
Llama-3.1-8B	GSM8K	**85.7±0.5**	79.2	81.4	82.4	82.0	82.8	75.6	71.0	75.8
Llama-3.1-8B	MATH5000	**86.1±0.2**	82.5	74.2	75.8	84.4	84.0	80.7	75.6	76.1
Llama-3.1-8B	DeepScaler	83.8±0.7	65.5	74.7	76.0	**84.0**	83.3	80.0	74.4	57.1
Qwen2.5-1.5B	GSM8K	**70.8±0.5**	65.7	65.4	65.7	67.5	67.3	64.3	56.6	66.5
Qwen2.5-1.5B	MATH5000	**82.1±0.9**	73.7	73.4	74.0	80.6	79.4	65.8	58.2	75.9
Qwen2.5-1.5B	DeepScaler	**73.3±1.4**	65.4	65.0	63.5	68.5	67.0	61.1	58.3	66.1
Qwen3-1.7B	GSM8K	**76.1±0.4**	68.2	67.7	70.3	73.4	58.7	58.7	61.5	70.8
Qwen3-1.7B	MATH5000	**83.1±0.1**	77.0	74.1	74.9	82.0	80.6	72.1	68.6	75.4
Qwen3-1.7B	DeepScaler	**83.8±0.3**	80.2	78.0	78.5	81.2	81.9	78.2	62.4	78.6
gemma-4-E2B	GSM8K	**76.3±4.9**	70.5	56.9	73.9	75.3	73.0	71.7	62.8	72.7
gemma-4-E2B	MATH5000	78.3±0.5	76.9	75.4	75.5	**78.7**	78.4	77.7	67.8	75.5
gemma-4-E2B	DeepScaler	**75.9±2.0**	64.3	51.5	64.0	74.3	73.2	75.8	64.8	64.1

Table 1: Canonical-cost AUROC$\times 100$ for detecting incorrect responses. PULSE is mean$\pm$std over three seeds; other columns are deterministic and sign-oriented on a held-out split (Appendix C). The vertical rule separates cost-5 TokUR from the zero-shot baselines. **Bold**/underline = best/second among the nine non-reference estimators; win/tie/loss follows a paired bootstrap test on 2000 resamples per cell (Appendix D), not a fixed margin ($9/15$ wins, 6 ties, 0 losses). SelfCert = self-certainty, LogP = mean log-probability, PredEnt = mean predictive entropy, TECP/EDIS = entropy-trajectory baselines (Appendix G), Verb. = verbalized confidence, TokUR = weight-perturbation uncertainty (`tokur_au`); citations in Section 5.1.

token-level prior method (aleatoric `tokur_au` variant, strongest in preliminary checks). We exclude *cross-sample* methods—self-consistency (Wang et al., 2023), semantic entropy (Kuhn et al., 2023), Monte-Carlo predictive entropy (Malinin & Gales, 2021), and DeepConf (Fu et al., 2025)—since they need $K > 1$ generations and, in standard use, score an aggregated (voted) answer rather than the single fixed response every other estimator scores; TokUR is included at cost-5 since it rescores that same candidate-0 response instead. As a further reference point, we also force semantic entropy and MC predictive entropy to score candidate-0 itself (putting them on equal footing with TokUR/PULSE), and even then both trail PULSE's 79.1 on average (74.5 and 72.5) on the 10 cost-5 cells despite $5\times$ the generations (Appendix O).

5.2 Main Results: Canonical-Cost Comparison

Table 1 compares PULSE against single-, two-, and five-forward-pass baselines under a strict same-candidate protocol (every estimator scores the same candidate-0 response against the same label). Since PULSE is supervised while the scalar baselines are zero-shot, we report PULSE as mean$\pm$std over three seeds and sign-orient each baseline on a held-out split (Appendix C). This table is one piece of the decomposition; the other three reference probes (length, pooled statistics, white-box) are analyzed against PULSE in Section 5.3.

By AUROC, PULSE is best or statistically tied-best on all 15 cells with no losses. Rather than a fixed seed-std margin, we verify this with a paired bootstrap on 2000 resamples per cell (Appendix D), which accounts for baseline sampling variance too, not just PULSE's seed variance: 9 clear wins and 6 statistical ties, no losses. TECP (summed entropy; Appendix G) is the strongest baseline: it edges PULSE by raw point estimate on two cells (gemma-4-E2B/MATH5000, Llama-3.1-8B/DeepScaler), both within the tie band, and trails on the rest. Cost-5 TokUR is matched or exceeded despite PULSE's single pass.

The full AUPRC table and base error rates are in Appendix B; the story is broadly consistent but differs on low-error cells, where TECP/EDIS beat PULSE beyond its seed std on $4/15$ cells by AUPRC. TECP also beats mean-entropy PredEnt on all 15 cells by AUROC despite the identical trajectory (summing vs. averaging partly tracks length); EDIS trails throughout (Appendix G).

Calibration and selective prediction. Since $\hat{p}_{\text{corr}}$ is also meant to support abstention, we check calibration, not just ranking (Appendix M). Raw ECE is uneven (mean 11.5, range 3.3–41.2), as expected since training targets ranking. Selective prediction, though, is where the ranking edge pays off: PULSE is the best *black-box* abstention signal, with the lowest mean AURC (35.2 vs. oracle 25.1) and the highest selective accuracy at 50% coverage (63.5), best on $11/15$ cells and ahead of TECP (36.1), Length (38.7), PredEnt (41.4), and LogP (43.8; Appendix R). Abstention thus stays near-ideal even where raw probabilities are off. For users who do need calibrated probabilities rather than just a ranking, a cheap post-hoc fix helps: 1-D Platt scaling on a held-out validation split reduces

mean ECE from 11.5 to 8.4 and mean Brier from 0.168 to 0.153, improving $11/15$ cells and cutting the worst cell (gemma-4-E2B/GSM8K) from 41.2 to 29.1—a real but partial fix (Appendix M).

5.3 Attribution: Length, Order, and Cheap Baselines

We decompose PULSE's canonical-cost advantage into four questions: how much is response length, how much needs sequence-order modeling versus a supervised probe on pooled statistics, whether a nearly-free 2-feature baseline already gets most of the way there, and how much black-box access leaves on the table. Full per-cell tables are in Appendix K.

Length. Response length alone is a strong single-pass signal (73.9 mean oriented AUROC, above most baselines in Table 1), and PULSE partly reads it: adding length as an explicit PULSE feature does not improve it (79.3 vs. 80.2). But PULSE retains 71.1 AUROC when evaluated *within* response-length quintiles, where length cannot contribute—a genuine, length-independent correctness signal, smaller than the raw numbers suggest but far from zero.

Order vs. pooled statistics. At fixed supervision, an order-agnostic probe on 70 pooled statistics of the top-k channel reaches 79.4, above zero-shot TECP (78.2) and within 0.8 of PULSE's 80.2 (seed-0 checkpoints, as for all supervised controls; Appendix C). Per-cell bootstrap verdicts are in Appendix K. So supervision plus full channel width lifts a probe about $+1.2$ above the free baseline, and adding sequence order (PULSE) buys a further $+0.8$, significant on only $3/15$ cells. Because pooled statistics never see raw sequence length T, one might worry even that residual is hidden length; we test this directly by handing the pooled probe explicit length ($\log T$) and summed entropy (ΣH), which raises it only $+0.3$ ($79.4 \rightarrow 79.7$), still below PULSE (Appendix K)—so explicit length closes only part of the gap, leaving a small residual attributable to sequence order. The missing 2×2 corner (pooled entropy-only stats, 75.0) shows width ($+4.4$) and order ($+3.1$) both contribute in isolation but combine sub-additively (80.2 vs. an 82.5 additive prediction)—overlapping, not complementary, signal.

A cheap alternative. A 2-feature logistic regression on just (length, ΣH)—ΣH being the zero-shot TECP score—averages 78.0, about even with zero-shot ΣH/TECP alone (78.2), so two supervised scalars add nothing over the one free scalar; the full 70-dim pooled probe (79.4) does add about a point on top. *Within* length strata PULSE keeps the highest mean (Appendix K), though it does not lead on every cell, so this is a mean-level, not cell-wise, advantage. PULSE's remaining edge is the top-k evidence itself, not architectural sophistication on top of it.

The black-box ceiling. Since PULSE reads only top-k log-probabilities, we bound this restriction's cost by training a logistic-regression probe (the same family as the pooled control) on the model's last-layer hidden states—a white-box signal outside our single-pass, output-only setting (Appendix Q). That probe averages 86.6 AUROC and scores higher than PULSE on $14/15$ cells ($+6.6$ mean, on comparable but not identical test splits): PULSE is the strongest *black-box* estimator here, not a universal one, and where hidden states are accessible a hidden-state probe is preferable.

5.4 Cross-Model and Cross-Dataset Transfer

We next evaluate whether a PULSE readout trained on one model–dataset pair transfers to others, training a head per source configuration and evaluating it on every target configuration unchanged (reusing the source's fitted feature normalizer, Eq. 4), to test whether it captures reusable patterns rather than an in-distribution calibration curve.

This spans a fully square grid: 225 source–target cells, diagonal entries in-distribution, off-diagonal entries measuring transfer (full tables in Appendix E). PULSE often retains useful ranking ability beyond the source (Figure 1 shows the full 15×15 landscape): in-distribution AUROC averages 80.2 versus 73.2 off-diagonal, with 73% of off-diagonal cells above 70, and transfer is strongest within a model family and weakest for Qwen3-1.7B (per-model breakdown in Appendix E). Transfer is uneven, though: 8% of off-diagonal cells fall below 55, concentrated in (but not exclusive to) one source, Qwen3-1.7B/GSM8K (7 of the 17 sub-55 cells, and itself a weak target). A few fall well below 50 (worst: Qwen3-1.7B/GSM8K$\rightarrow$Qwen2.5-1.5B/DeepScaler, 32), a systematic *inversion* that ranks incorrect responses above correct ones. PULSE thus learns reusable UQ signals, but generalization is partial and retains model- and dataset-specific structure—consistent with a companion check finding the same evidence lets a classifier identify *which* model generated a response (93.6% five-way in-dataset accuracy, chance 20%; Appendix H): the signal that makes a head reusable also partly encodes model identity.

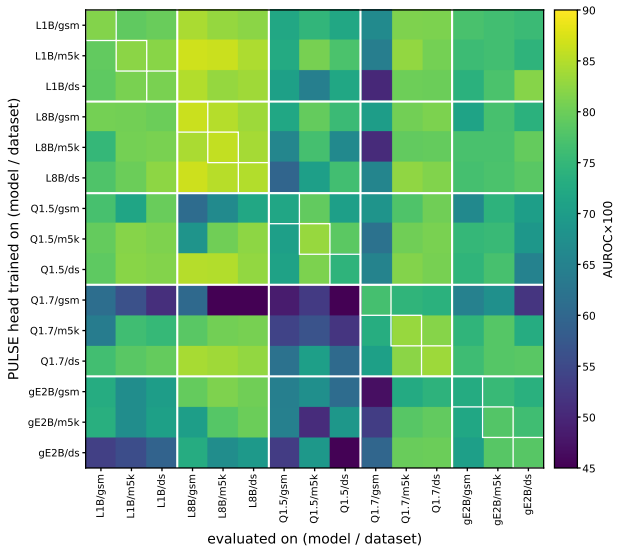


Figure 1: Configuration-level transfer of the PULSE readout across all 15×15 (model, dataset) pairs, coloured by AUROC$\times 100$ (rows = train configuration, columns = eval configuration, diagonal outlined; exact values in Appendix E). Warm cells indicate preserved ranking ability: strong 3×3 model blocks with graded cross-dataset transfer inside each, mostly warm cross-model transfer, except the coldest source Qwen3-1.7B/GSM8K (dipping into inversion on a few targets, worst $\sim$32).

5.5 Feature Ablation

We analyze which components of the top-10 log-probability evidence (Eq. 2) PULSE needs, on Qwen2.5-1.5B across all three datasets: a cumulative top-k truncation sweep ($k \in \{1, 2, 3, 5, 7, 10\}$) and a rank-*position* sweep (single ranks, the full vector minus rank-1, and a tail-only subset of ranks 8–10), both in Appendix F.

Two patterns stand out. First, the cumulative sweep saturates rather than peaks: the full 10-dim vector is the strongest or near-strongest variant on all three datasets, with most ranking signal already in the first few ranks. Second, and more surprising, a *single* low-confidence rank is consistently at least as informative as the highest-confidence rank—rank-10 alone matches or exceeds rank-1 alone on all three datasets (Appendix Table A10)—consistent with rank-10 tracking tail mass (peakedness) whereas rank-1 saturates near 0 once the model is at all confident. So no single rank or truncation is uniquely necessary, and low-confidence ranks being informative rather than noise is new to the evidence representation studied here.

Feature-set robustness, and evidential vs. BCE head. Repeating the truncation/rank-position sweep (10 distinct feature-set variants; two of the original 11 coincide, Appendix Table A8) on all five main-table models across all three datasets (150 cells) confirms the deployed top-10 vector is among the best variants on average (80.2 mean) and never the weakest; the best variant differs per model, so no compact alternative dominates, and a bottom-3 subset strong on one model loses up to 7.7 elsewhere. We retain the full 10-dim vector as the default. Separately, the evidential head adds one hyperparameter (λ_{KL}) over a plain BCE head; retraining every cell with BCE (Appendix L), the Dirichlet head wins only $6/15$ cells, ties 2, loses 7 (mean $+0.7$, median 0.0)—not the source of PULSE's advantage. We keep it mainly because it exposes a vacuity score v_{PULSE} (Section 4.3) for future work, not because it shows a benefit over BCE.

5.6 Sample Efficiency

Unlike the zero-shot scalar baselines, PULSE needs labeled data on the target model/dataset, a cost the cost-tier framing above does not capture. Subsampling each cell's training set to 100/500/2000 labels and retraining over 3 seeds, mean AUROC across all 15 cells is 76.8/79.1/79.4 against 80.1 at full supervision (Figure 2; Appendix N)—96% of full performance at 100 labels, 99% at 2000: a few hundred labels already capture most of PULSE's advantage over the zero-shot baselines.

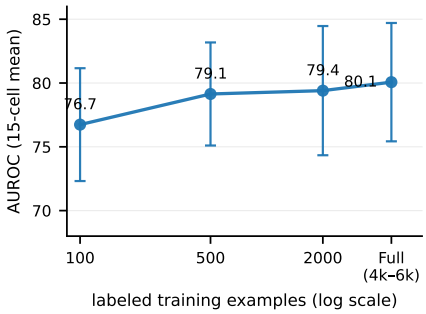


Figure 2: PULSE AUROC vs. labeled training-set size (15-cell mean±std).

6 Discussion

Where the signal comes from. Most of the black-box signal comes for free, and what remains accumulates in small steps, mostly from channel width rather than order. A supervised classifier on length and TECP (78.0 AUROC) does no better than zero-shot TECP alone (78.2), so supervision over free scalars adds nothing by itself. Supervision helps only once it can read the full top-k representation: the order-agnostic pooled probe adds about a point (79.4), and modeling sequence order adds a smaller increment (80.2 with seed-0 checkpoints, Appendix C; significant on only $3/15$ cells). PULSE's margin over TECP is therefore small but systematic: it wins on $13/15$ cells (paired Wilcoxon, $p = 0.0024$), a cross-cell trend distinct from the per-cell bootstrap ties of Appendix D.

Cost and positioning. This gain has a cost that our inference-only cost tiers omit: PULSE requires labeled data and retraining, although 100 labels already reach 96% of full-supervision AUROC. Its clearest practical advantage is matching the five-pass TokUR estimator at single-pass cost without hidden-state access, although that comparison pairs a supervised black-box probe with a zero-shot white-box one. When the full next-token distribution is available, practitioners who can tolerate a gap of about two points should simply use zero-shot TECP; PULSE earns its supervision cost when only top-k log-probabilities are exposed, or when the goal is to approach white-box or sampling-based accuracy cheaply and without hidden-state access. It is not a substitute for sampling-aggregation methods such as self-consistency, semantic entropy, or DeepConf, which score a different, voted-answer label, but a cheap complement where a $K\times$ budget is unavailable.

Limitations. PULSE is black-box by design and pays for it: a white-box hidden-state probe scores higher on $14/15$ cells ($+6.6$ AUROC on average, on comparable but not identical test splits; Appendix Q), and a prompt-only control shows that this gap reflects monitoring of the generation rather than question difficulty. Like any readout of the model's own predictive distribution, PULSE may fail when the model is confidently wrong or when correctness depends on external facts. The learned readout transfers across models and datasets only partially, and trajectories carry enough model-specific structure to identify the source model, so robust deployment likely needs per-target adaptation. Temperature transfer is model-dependent (Appendix I); this analysis, like the rank-position and fingerprint experiments, uses a single model or dataset and should be read as an existence result. Finally, our main results concern exact-match mathematical reasoning and should not be assumed to hold for free-form generation.

7 Conclusion

We decompose the correctness signal in a single generation's output distributions into what response length and zero-shot TECP already capture, what supervision adds with and without sequence order, and the headroom that only white-box hidden states recover. The returns are uneven: most of the usable black-box signal is available without training, PULSE improves on the best zero-shot score only modestly, and a much larger gap separates black-box from white-box access. PULSE is never significantly worse than the best black-box baseline and matches the five-pass TokUR estimator at one-fifth of its cost, given a few hundred labels. These results concern exact-match mathematical reasoning; a smaller check on free-form question answering suggests that they carry over qualitatively, and a full replication is left to future work.

REFERENCES

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pp. 1613–1622, 2015.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*, 2023.

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

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. LM-Polygraph: Uncertainty estimation for language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2023.

Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. Deep think with confidence. *arXiv preprint arXiv:2508.15260*, 2025.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pp. 1050–1059, 2016.

Gemma Team, Google DeepMind. Gemma 4. Model card, 2026. URL `https://ai.google.dev/gemma/docs/core/model_card_4`.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330, 2017.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Zhewei Kang, Xuandong Zhao, and Dawn Song. Scalable Best-of-N selection for large language models via self-certainty. *arXiv preprint arXiv:2502.18581*, 2025. NeurIPS 2025.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. *arXiv preprint arXiv:2406.15927*, 2024.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Bo Li, Tian Tian, Zhenghua Xu, Hao Cheng, Shikun Zhang, and Wei Ye. Modeling uncertainty trends for timely retrieval in dynamic RAG. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31527–31535, 2026.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. *arXiv preprint arXiv:2305.20050*, 2023.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Li Erran Li, Raluca Ada Popa, and Ion Stoica. DeepScaleR: Surpassing O1-Preview with a 1.5B model by scaling RL. Notion blog, Agentica / Berkeley Sky Computing Lab and BAIR, 2025. URL `https://github.com/agentica-project/deepscaler`. Dataset: agentica-org/DeepScaleR-Preview-Dataset.

Hengyuan Ma, Wenlian Lu, and Jianfeng Feng. Spontaneous emergence of metacognition in neuronal computation. *Phys. Rev. Res.*, 7:033188, Aug 2025. doi: 10.1103/f1hv-bf1f. URL `https://link.aps.org/doi/10.1103/f1hv-bf1f`.

David J. C. MacKay. A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992.

Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, 2018.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*, 2021.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Conference on Empirical Methods in Natural Language Processing*, 2023.

Radford M. Neal. *Bayesian Learning for Neural Networks*, volume 118 of *Lecture Notes in Statistics*. Springer, 1996.

Ernesto Quevedo, Jerry Yero, Rachel Koerner, Pablo Rivas, and Tomas Cerny. Detecting hallucinations in large language model generation: A token probability approach. In *Proceedings of the International Conference on Computational Science and Computational Intelligence (CSCE)*, 2024.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable Laplace approximation for neural networks. In *International Conference on Learning Representations*, 2018.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *arXiv preprint arXiv:2412.05563*, 2024.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

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

Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdoo Yun, and Seong Joon Oh. Calibrating large language models using their generations only. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, Maxim Panov, Timothy Baldwin, Preslav Nakov, Alexander Panchenko, and Artem Shelmanov. Benchmarking uncertainty quantification methods for large language models with LM-Polygraph. *Transactions of the Association for Computational Linguistics*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Artem Vazhentsev, Ekaterina Fadeeva, Rui Xing, Gleb Kuzmin, Ivan Lazichny, Alexander Panchenko, Preslav Nakov, Timothy Baldwin, Maxim Panov, and Artem Shelmanov. Unconditional truthfulness: Learning unconditional uncertainty of large language models. *arXiv preprint arXiv:2408.10692*, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023.

Beining Xu and Yongming Lu. TECP: Token-entropy conformal prediction for LLMs. *Mathematics*, 13(20):3351, 2025. Also arXiv:2509.00461.

Tunyu Zhang, Haizhou Shi, Yibin Wang, Hengyi Wang, Xiaoxiao He, Zhuowei Li, Hao Chen, Ligong Han, Kai Xu, Huan Zhang, and Dimitris N. Metaxas. TokUR: Token-level uncertainty estimation for large language model reasoning. *arXiv preprint arXiv:2505.11737*, 2025.

Chenghua Zhu, Siyan Wu, Xiangkang Zeng, Zishan Xu, Zhaolu Kang, Yifu Guo, Yuquan Lu, Junduan Huang, and Guojing Zhou. EDIS: Diagnosing LLM reasoning via entropy dynamics. *arXiv preprint arXiv:2602.01288*, 2026.

A Reproducibility and LLM-Usage Disclosure

Reproducibility Statement. Architecture, training, and evaluation details are in Section 5.1 and this Appendix (hyperparameters, feature and baseline formulas, sign-orientation and bootstrap protocols in Appendices C and D). Code and data will be released upon publication.

Disclosure of LLM Usage. A large language model-based coding and writing assistant (Claude, Anthropic) was used in this project: implementing experiment and analysis scripts, running training and evaluation pipelines, revising manuscript text, and cross-checking numerical results across tables and appendices for consistency. The assistant operated under continuous human direction—every experimental design choice, baseline selection, and reported finding was specified, reviewed, and interpreted by the authors. No claim in this paper rests on unverified AI-generated content, and the authors take full responsibility for all methodology, results, and conclusions reported.

B Additional Main-Results Detail: AUPRC and Base Error Rates

Table A1 reports the AUPRC companion to the main-text AUROC table (Table 1), and Table A2 reports the test-set size and base error rate underlying every cell in both tables, referenced from Section 5.2.

Model	Dataset	PULSE	SelfCert	LogP	PredEnt	TECP	EDIS	P(True)	Verb.	TokUR
Llama-3.2-1B	GSM8K	**76.5±0.2**	62.8	47.3	68.2	73.2	74.0	49.1	52.1	66.3
Llama-3.2-1B	MATH5000	92.3±0.5	87.2	85.0	85.9	**93.0**	92.7	77.8	79.5	86.3
Llama-3.2-1B	DeepScaler	96.9±0.3	94.6	94.6	94.4	96.6	**97.3**	92.3	94.0	94.8
Llama-3.1-8B	GSM8K	44.1±2.1	29.9	13.6	22.9	**47.9**	40.3	38.6	23.7	28.1
Llama-3.1-8B	MATH5000	**87.8±0.0**	84.0	77.5	78.8	86.2	85.6	82.4	74.4	79.1
Llama-3.1-8B	DeepScaler	94.4±0.5	85.2	86.1	84.4	**94.8**	93.9	93.4	90.2	85.5
Qwen2.5-1.5B	GSM8K	**67.9±2.1**	59.0	58.3	57.6	60.2	59.2	63.1	51.8	57.9
Qwen2.5-1.5B	MATH5000	**91.7±0.4**	88.4	87.8	88.1	91.2	90.5	85.9	79.4	88.8
Qwen2.5-1.5B	DeepScaler	**95.8±0.3**	94.1	93.5	93.3	94.0	93.9	93.7	92.3	93.5
Qwen3-1.7B	GSM8K	**44.9±0.2**	36.1	34.5	38.2	40.2	27.3	31.6	25.3	38.3
Qwen3-1.7B	MATH5000	**77.0±0.5**	70.4	66.7	67.8	74.9	74.6	63.9	53.0	68.0
Qwen3-1.7B	DeepScaler	**90.8±0.4**	88.8	88.0	88.2	89.8	90.3	86.8	75.1	88.2
gemma-4-E2B	GSM8K	**43.7±7.1**	28.5	18.9	31.9	35.4	33.7	41.5	27.3	30.6
gemma-4-E2B	MATH5000	56.2±0.8	51.1	50.9	51.7	**58.0**	57.7	39.2	28.7	51.2
gemma-4-E2B	DeepScaler	74.1±1.8	62.1	50.3	63.2	71.7	73.3	**74.3**	62.0	63.1

Table A1: Canonical-cost comparison of uncertainty estimators using AUPRC, the companion to Table 1 and computed under the same protocol: PULSE is mean±std over three seeds, every scalar baseline is sign-oriented on a held-out split (Appendix C); the vertical rule separates cost-5 TokUR from the zero-shot baselines. Values are AUPRC$\times 100$ for detecting incorrect responses. Since AUPRC is sensitive to the fraction of incorrect responses (the base error rate), it should be interpreted as complementary to AUROC; cells with low AUPRC despite strong AUROC (e.g. Llama-3.1-8B/GSM8K, gemma-4-E2B/GSM8K, gemma-4-E2B/MATH5000) correspond to low base error rates, where even a good ranker yields low precision on the rare incorrect class (see Table A2). **Bold**/underline mark best/second-best among the nine non-reference estimators (by raw value). Unlike AUROC, PULSE does not win-or-tie everywhere by AUPRC: TECP or EDIS exceed PULSE's mean by more than its seed std on $4/15$ cells (Llama-3.2-1B/MATH5000, Llama-3.2-1B/DeepScaler, Llama-3.1-8B/GSM8K, gemma-4-E2B/MATH5000), consistent with AUPRC's base-rate sensitivity noted above; PULSE leads on $6/15$ with 5 ties elsewhere. Abbreviations follow Table 1: SelfCert = self-certainty, LogP = mean token log-probability, PredEnt = mean token predictive entropy (single-pass), TECP/EDIS = hand-crafted entropy-trajectory baselines (Appendix G), Verb. = verbalized confidence, TokUR = token-level uncertainty via weight perturbation (aleatoric variant, `tokur_au`).

C Baseline Sign Orientation, Seed Variance, and Preprocessing

A scalar uncertainty signal can be anti-correlated with correctness on a given cell (e.g. a log-probability convention that happens to point the wrong way), which would report a sub-50 AUROC and unfairly flatter any method compared against it. A practitioner would fix the sign on validation data, so for the main table we orient every scalar baseline by a *cross-fitted* rule that uses no test labels for the samples it scores: we split the evaluation set into two folds, choose each baseline's sign on one fold (flip iff its AUROC there is below 50), and apply it to the held-out fold, then pool. Cleanly oriented signals are unchanged; only genuinely anti-correlated cells flip. In practice this affects two baselines on the Llama-3.2-1B rows only (Table A3); all other main-table entries are

Model	Dataset	n	Base error rate
Llama-3.2-1B	GSM8K	1495	46.9%
Llama-3.2-1B	MATH5000	1000	75.4%
Llama-3.2-1B	DeepScaler	1000	91.5%
Llama-3.1-8B	GSM8K	1495	8.4%
Llama-3.1-8B	MATH5000	1000	55.5%
Llama-3.1-8B	DeepScaler	1000	80.8%
Qwen2.5-1.5B	GSM8K	1495	46.3%
Qwen2.5-1.5B	MATH5000	1000	75.1%
Qwen2.5-1.5B	DeepScaler	1000	90.6%
Qwen3-1.7B	GSM8K	1495	17.9%
Qwen3-1.7B	MATH5000	1000	37.9%
Qwen3-1.7B	DeepScaler	1000	68.3%
gemma-4-E2B	GSM8K	1495	16.5%
gemma-4-E2B	MATH5000	1000	17.0%
gemma-4-E2B	DeepScaler	1000	50.3%

Table A2: Test-set size and base error rate (fraction of incorrect candidate-0 responses, i.e. the positive-class prevalence for AUPRC) for each of the 15 model–dataset cells in Table 1.

already correctly oriented. PULSE itself needs no orientation—supervision fixes its sign—so this only ever raises baseline scores, i.e. it makes the comparison *harder* for PULSE.

Baseline	Cell	Raw	Oriented
LogP	Llama-3.2-1B/GSM8K	46.0	54.0
LogP	Llama-3.2-1B/DeepScaler	34.7	65.3
P(True)	Llama-3.2-1B/GSM8K	47.8	52.2
P(True)	Llama-3.2-1B/MATH5000	45.6	54.4
P(True)	Llama-3.2-1B/DeepScaler	47.7	52.3

Table A3: The only main-table cells whose baseline sign flips under cross-fitted orientation (raw AUROC < 50). All other baseline entries in Table 1 are already correctly oriented and unchanged.

The PULSE column of Table 1 is a mean over three training seeds (the original run plus two additional model-init/ordering seeds; the train/val split is held fixed). Seed variance is small on MATH5000 (std ≤ 1.0 on all five cells) but larger on several GSM8K and DeepScaler cells (up to ± 4.9 on gemma-4-E2B/GSM8K).

Single-seed checkpoints in secondary experiments. Table 1's PULSE column is a 3-seed mean. For tractability, secondary control experiments that are not themselves seed sweeps (the length-control, cheap-baseline, and supervised-control tables) use each cell's seed-0 PULSE checkpoint, so PULSE values quoted there can differ from the 3-seed mean by ordinary seed variance. The difference is negligible for most cells but pronounced for gemma-4-E2B/GSM8K (72.7 seed-0 vs. 76.3 3-seed mean, std ± 4.9). No qualitative conclusion depends on which checkpoint is used.

Response preprocessing. Unless noted otherwise (Table A8), all per-token feature arrays are truncated at the first end-of-sequence token, so that padding emitted during batched generation does not enter any trajectory-derived quantity.

D Paired Bootstrap Significance Test

A fixed-margin heuristic (PULSE's own seed std) is not a real significance test: it ignores finite-test-set AUROC sampling variance on the *baselines* too (e.g. DeepScaler's high base error rate leaves few positive/correct examples, so even a deterministic baseline's AUROC has a wide confidence interval on n=1000), and a deterministic baseline is not actually noise-free just because it has no seed variance. We replace it with a paired bootstrap: for each cell, we resample the shared test-set index set with replacement $2000\times$ (the same resampled indices are used for PULSE and every baseline, preserving the pairing), recomputing AUROC for PULSE and for every same-cost baseline (sign-oriented per Appendix C, including the EDIS/TECP baselines of Appendix G) on each resample. This gives a 95% percentile confidence interval for PULSE and for the best baseline, and, more directly, a win frequency: the fraction of the 2000 resamples in which PULSE's AUROC exceeds the best same-cost baseline's AUROC on that resample. We classify a cell as a clear win if win frequency $\geq 97.5\%$, a clear loss if $\leq 2.5\%$, and a statistical tie otherwise (symmetric two-sided 95% test). Because the main table reports PULSE as a mean over 3 seeds, we resample each seed's checkpoint too: for every one of the 2000 resamples we compute AUROC separately for each of PULSE's 3 seed

checkpoints on the resampled indices and average the 3 resampled AUROCs, exactly mirroring how the main table's own mean is computed. The point values reported for PULSE and the best baseline in Table A4 are each the mean across the 2000 resamples, not the single full-sample estimate of Table 1, so a baseline's value here can differ from its main-table entry by up to about a point. Table A4 reports the result: 9 clear wins, 6 statistical ties, and 0 losses; using only the seed-0 checkpoint yields the same counts. This per-cell test asks a different question than the headline $+1.9$ mean gap over TECP (best zero-shot baseline): per-cell bootstrap ties test whether PULSE beats TECP *within* one cell's finite test set, while the mean gap is a claim *across* the 15 cells. Testing that directly: PULSE beats TECP on $13/15$ cells (paired Wilcoxon signed-rank across the 15 cells, $p = 0.0024$; paired t-test, $p = 0.0006$), so the small average margin is a systematic cross-cell trend, not an artifact of a few large swings, even though many individual cells are statistical ties.

Model	Dataset	PULSE, 3-seed mean (95% CI)	Best baseline (95% CI)	Win freq.	Verdict
Llama-3.2-1B	GSM8K	81.9 [79.8, 83.9]	EDIS 77.5 [75.2, 79.8]	100%	win
Llama-3.2-1B	MATH5000	82.3 [79.4, 85.2]	TECP 82.2 [79.3, 85.1]	47.7%	tie
Llama-3.2-1B	DeepScaler	81.4 [76.1, 86.5]	TECP 81.3 [75.6, 86.5]	50.6%	tie
Llama-3.1-8B	GSM8K	85.7 [81.8, 89.2]	EDIS 82.8 [78.4, 86.8]	97.6%	win
Llama-3.1-8B	MATH5000	86.2 [83.8, 88.3]	TECP 84.4 [81.9, 86.6]	99.9%	win
Llama-3.1-8B	DeepScaler	83.9 [80.6, 86.9]	TECP 84.0 [80.8, 86.9]	40.4%	tie
Qwen2.5-1.5B	GSM8K	70.8 [68.3, 73.3]	TECP 67.5 [64.7, 70.2]	100%	win
Qwen2.5-1.5B	MATH5000	82.2 [79.1, 85.0]	TECP 80.6 [77.4, 83.6]	99.4%	win
Qwen2.5-1.5B	DeepScaler	73.3 [68.6, 78.1]	TECP 68.6 [62.1, 75.2]	98.5%	win
Qwen3-1.7B	GSM8K	76.2 [73.1, 79.0]	TECP 73.4 [70.2, 76.6]	97.8%	win
Qwen3-1.7B	MATH5000	83.1 [80.5, 85.6]	TECP 82.1 [79.3, 84.8]	97.9%	win
Qwen3-1.7B	DeepScaler	83.8 [81.1, 86.2]	EDIS 82.0 [79.1, 84.5]	99.8%	win
gemma-4-E2B	GSM8K	76.3 [73.9, 78.9]	TECP 75.2 [71.9, 78.5]	76.5%	tie
gemma-4-E2B	MATH5000	78.4 [74.1, 82.5]	TECP 79.2 [75.0, 83.5]	6.3%	tie
gemma-4-E2B	DeepScaler	76.0 [73.1, 78.7]	P(True) 75.8 [72.9, 78.6]	51.9%	tie

Table A4: Paired bootstrap (2000 resamples/cell) comparison of PULSE against the strongest same-cost baseline on the identical resample, PULSE resampled under the same 3-seed-mean protocol as the main table. Reported point values are resample means, so they can differ slightly from Table 1's full-sample estimates. Win frequency $\geq 97.5\%$ = clear win, $\leq 2.5\%$ = loss, otherwise a statistical tie. 9 wins, 6 ties, 0 losses.

E Full Cross-Model and Cross-Dataset Transfer Matrices

Tables A5 and A6 report the full 15×15 cross-model/cross-dataset transfer matrices summarized in Section 5.4: rows are training configurations, columns are evaluation configurations, and diagonal entries (bold) are in-distribution results.

Train/Test	L1B/gsm	L1B/m5k	L1B/ds	L8B/gsm	L8B/m5k	L8B/ds	Q1.5/gsm	Q1.5/m5k	Q1.5/ds	Q1.7/gsm	Q1.7/m5k	Q1.7/ds	gE2B/gsm	gE2B/m5k	gE2B/ds
L1B/gsm	**81.7**	79.1	79.7	84.3	82.8	82.1	72.0	74.6	71.9	66.6	81.3	81.2	77.0	76.3	75.7
L1B/m5k	79.2	**82.2**	82.0	86.8	86.3	84.4	72.5	80.8	77.2	64.2	82.6	80.5	75.9	76.3	74.2
L1B/ds	78.8	80.9	**81.0**	84.9	83.0	83.6	70.6	64.4	71.8	50.1	80.0	79.8	73.7	77.2	81.7
L8B/gsm	80.5	80.5	79.9	**86.4**	85.0	83.5	71.9	79.4	75.7	70.0	80.4	81.1	71.0	76.8	73.8
L8B/m5k	74.9	80.5	81.0	84.3	**85.9**	83.9	65.6	76.7	66.3	50.5	79.2	79.4	77.1	77.1	79.4
L8B/ds	77.7	79.9	82.3	86.6	85.2	**84.8**	59.6	70.4	76.4	65.5	82.5	81.5	76.8	77.3	78.5
Q1.5/gsm	77.0	71.4	79.7	60.9	66.5	72.0	**71.6**	79.3	70.4	68.3	77.6	80.2	66.3	74.1	70.3
Q1.5/m5k	79.3	81.9	81.3	68.2	80.3	82.3	70.5	**83.0**	78.7	62.0	80.3	80.9	75.0	75.5	68.8
Q1.5/ds	78.8	82.0	81.7	85.0	84.9	82.7	70.9	81.1	**74.2**	65.2	81.6	80.7	74.5	76.9	64.9
Q1.7/gsm	61.3	56.1	51.0	60.4	43.7	41.0	48.4	52.7	32.0	**76.8**	74.5	73.7	64.7	67.6	51.9
Q1.7/m5k	64.0	76.3	75.1	78.5	80.4	80.8	54.1	56.5	52.0	73.1	**83.1**	81.8	74.4	78.2	72.8
Q1.7/ds	76.4	78.5	79.3	84.2	83.6	82.7	62.0	70.3	60.5	70.5	82.2	**83.4**	76.1	78.2	78.5
gE2B/gsm	72.8	67.0	70.1	79.4	81.3	80.2	64.5	67.4	60.9	46.9	72.5	74.1	**72.7**	75.3	73.5
gE2B/m5k	73.4	67.1	70.5	70.3	78.2	79.9	64.7	50.8	68.8	53.2	78.3	79.2	71.8	**78.0**	76.1
gE2B/ds	53.5	55.6	59.4	72.9	67.1	69.4	52.9	68.9	45.3	59.9	79.6	79.8	70.3	78.3	**78.3**

Table A5: Cross-model and cross-dataset transfer of PULSE measured by AUROC. Each row denotes the model–dataset configuration used to train the PULSE readout head, and each column denotes the target configuration used for evaluation. The grid is fully square: all five models are run on all three datasets, so the 15 source and 15 target configurations yield 225 populated source–target cells with no blank entries. Values are AUROC$\times$100. Diagonal entries are in-distribution evaluations and are shown in bold. Abbreviations: L1B = Llama-3.2-1B, L8B = Llama-3.1-8B, Q1.5 = Qwen2.5-1.5B, Q1.7 = Qwen3-1.7B, gE2B = gemma-4-E2B, gsm = GSM8K, m5k = MATH5000, and ds = DeepScaler.

Train/Test	L1B/gsm	L1B/m5k	L1B/ds	L8B/gsm	L8B/m5k	L8B/ds	Q1.5/gsm	Q1.5/m5k	Q1.5/ds	Q1.7/gsm	Q1.7/m5k	Q1.7/ds	gE2B/gsm	gE2B/m5k	gE2B/ds
L1B/gsm	**78.3**	90.6	96.7	48.5	83.8	93.3	71.0	88.4	96.0	32.3	74.7	89.4	44.2	49.7	75.4
L1B/m5k	75.7	**92.4**	97.1	53.9	87.0	94.6	71.7	90.7	96.8	29.0	75.5	89.1	41.4	43.8	72.8
L1B/ds	75.0	92.1	**97.5**	43.5	83.4	94.3	67.6	80.0	94.7	21.5	72.7	89.1	42.1	49.0	82.0
L8B/gsm	77.3	91.8	97.2	**50.7**	86.1	94.5	71.7	90.6	96.4	35.2	73.6	89.3	31.5	50.8	73.7
L8B/m5k	70.6	91.9	97.5	47.9	**87.8**	94.9	60.2	88.9	94.3	18.9	72.7	88.7	43.9	48.1	79.5
L8B/ds	74.3	91.6	97.7	49.9	86.9	**95.1**	62.5	87.9	96.7	28.8	76.4	89.5	42.0	48.4	75.8
Q1.5/gsm	73.0	87.5	97.3	14.1	69.9	90.5	**69.0**	90.4	94.6	36.4	69.5	88.5	28.0	41.0	69.9
Q1.5/m5k	75.4	92.2	97.4	25.3	83.8	94.7	67.5	**91.9**	97.0	30.2	72.7	89.3	36.5	48.1	66.5
Q1.5/ds	75.3	92.5	97.3	50.2	85.4	93.8	69.2	91.0	**95.9**	31.7	73.8	89.0	33.3	45.9	59.3
Q1.7/gsm	57.1	79.8	91.8	13.0	54.9	79.1	46.8	77.2	85.1	**45.1**	59.8	82.0	21.7	27.8	47.3
Q1.7/m5k	62.7	89.0	96.6	37.7	81.7	93.6	47.9	78.4	91.1	39.5	**76.5**	89.7	33.9	49.1	72.1
Q1.7/ds	73.7	90.5	97.1	45.5	84.4	94.3	60.4	86.2	93.4	36.7	76.1	**90.2**	41.3	51.8	77.1
gE2B/gsm	69.4	83.2	95.2	40.5	82.8	92.5	61.7	84.5	93.6	16.3	64.6	85.2	**40.8**	43.1	73.5
gE2B/m5k	70.5	85.6	95.7	32.5	81.1	92.9	60.4	75.8	94.7	21.3	71.3	88.5	35.2	**55.8**	73.8
gE2B/ds	50.2	78.6	93.7	33.1	73.4	89.6	46.2	84.5	87.0	29.0	72.2	88.4	30.1	58.0	**75.5**

Table A6: Cross-model and cross-dataset transfer of PULSE measured by AUPRC. Rows and columns follow the same 15-source, 15-target square layout and abbreviations as Table A5. Values are AUPRC$\times$100 for detecting incorrect responses. Because AUPRC depends on the target error rate, cross-target comparisons should be interpreted with care; the table is mainly intended to show whether a trained readout retains useful ranking ability under model and dataset shift.

	GSM8K		MATH5000		DeepScaler	
Model	Full	B3	Full	B3	Full	B3
L1B	81.7	81.7	82.2	82.3	81.0	81.8
L8B	86.4	86.0	85.9	85.5	84.8	83.9
Q1.5	71.6	71.7	83.0	81.3	74.2	76.1
Q1.7	76.8	76.5	83.1	82.5	83.4	83.3
gE2B	72.7	65.8	78.0	78.3	78.4	70.7

Table A7: Feature-set robustness across all five models (AUROC$\times$100, same-candidate protocol on the candidate-0 responses of Table 1; referenced from Section 5.5). Full = 10-dim top-10 default; B3 = bottom-3 subset (ranks 8–10); Full averages 80.2 vs. 79.2 for B3. L1B = Llama-3.2-1B, L8B = Llama-3.1-8B, Q1.5 = Qwen2.5-1.5B, Q1.7 = Qwen3-1.7B, gE2B = gemma-4-E2B.

F Additional Feature-Ablation Results

Table A8 extends the full 11-variant sweep above (Qwen2.5-1.5B only) to the remaining four main-table models, giving per-model, per-dataset-mean AUROC for every truncation and rank-position variant, referenced from Section 5.5.

Feature set	L1B	L8B	Q1.5	Q1.7	gE2B	Mean
Top-1 (rank-1 only)	78.4	81.3	73.3	79.3	76.5	77.8
Top-2	78.6	82.7	76.0	80.5	74.5	78.4
Top-3	82.0	83.7	75.0	79.3	74.9	79.0
Top-5	80.3	84.6	75.9	80.4	78.8	80.0
Top-7	79.8	85.4	73.6	80.9	74.9	78.9
Top-10 (full, default)	81.6	85.7	76.3	81.1	76.4	80.2
Rank-5 only	80.1	83.9	73.7	80.6	76.0	78.9
Rank-10 only	80.9	83.8	75.9	80.9	76.2	79.5
Drop rank-1 (keep 2–10)	81.9	84.8	75.4	80.5	75.0	79.5
Bottom-3 only (ranks 8–10)	81.9	85.1	76.4	81.0	73.0	79.5

Table A8: Full feature-ablation sweep (six cumulative truncations, five rank-position variants, 10 distinct after deduplication) on all five main-table models, values are AUROC$\times$100 averaged over the three datasets (per-model, per-dataset scores follow the same protocol as Tables A9–A10, which report the Qwen2.5-1.5B row of this table broken out by dataset). Only 10 rows are shown because "Top-1 (rank-1 only)" and the rank-position sweep's "Rank-1 only" variant are the identical feature set, so the 11th variant is not a distinct row. "Top-10 (full, default)" is the deployed PULSE feature set; the "Mean" column averages across the five models. L1B = Llama-3.2-1B, L8B = Llama-3.1-8B, Q1.5 = Qwen2.5-1.5B, Q1.7 = Qwen3-1.7B, gE2B = gemma-4-E2B. For the Llama-3.2-1B and Llama-3.1-8B GSM8K/DeepScaler entries, variants other than Top-10 and Bottom-3 are computed without end-of-sequence truncation and run about one point low there; the relative ranking of variants is unaffected.

Table A9 sweeps the cumulative truncation length k, i.e., keeping only the k highest-probability ranks $[\log p^{(1)}, \ldots, \log p^{(k)}]$ for $k \in \{1, 2, 3, 5, 7, 10\}$, on Qwen2.5-1.5B across all three datasets. Table A10 instead probes individual rank *positions*: a single rank kept alone, the full vector with

rank-1 removed, and a compact tail-only subset (ranks 8–10). Both are referenced from Section 5.5. All numbers use the same generations and candidate-0 test set as the main table.

Ranks kept	AUROC	AUPRC
GSM8K		
Top-1 (rank-1 only)	67.4	64.6
Top-2	69.5	65.3
Top-3	69.1	63.6
Top-5	69.8	66.0
Top-7	66.7	62.0
Top-10 (full, default)	71.6	69.0
MATH5000		
Top-1 (rank-1 only)	80.4	91.4
Top-2	82.6	92.3
Top-3	82.0	92.0
Top-5	82.0	92.0
Top-7	82.4	92.2
Top-10 (full, default)	83.0	91.9
DeepScaler		
Top-1 (rank-1 only)	72.1	94.6
Top-2	75.7	96.3
Top-3	73.9	95.9
Top-5	76.1	96.1
Top-7	71.7	95.3
Top-10 (full, default)	74.2	95.9

Table A9: Cumulative top-k truncation ablation on Qwen2.5-1.5B, grouped by dataset. Values are AUROC$\times$100 and AUPRC$\times$100. "Top-10 (full, default)" is the deployed PULSE feature set.

Rank subset	AUROC	AUPRC
GSM8K		
Rank-1 only	67.4	64.6
Rank-5 only	67.6	65.1
Rank-10 only	71.4	69.2
Drop rank-1 (keep 2–10)	71.0	66.3
Bottom-3 only (ranks 8–10)	71.7	70.0
All 10 (full, default)	71.6	69.0
MATH5000		
Rank-1 only	80.4	91.4
Rank-5 only	82.1	91.9
Rank-10 only	82.3	92.0
Drop rank-1 (keep 2–10)	82.3	92.1
Bottom-3 only (ranks 8–10)	81.3	92.0
All 10 (full, default)	83.0	91.9
DeepScaler		
Rank-1 only	72.1	94.6
Rank-5 only	71.3	95.0
Rank-10 only	74.0	95.6
Drop rank-1 (keep 2–10)	72.7	95.2
Bottom-3 only (ranks 8–10)	76.1	96.4
All 10 (full, default)	74.2	95.9

Table A10: Rank-position ablation on Qwen2.5-1.5B, grouped by dataset. Values are AUROC$\times$100 and AUPRC$\times$100. "All 10 (full, default)" is the deployed PULSE feature set.

Dropping rank-1 while keeping the rest costs little on any of the three datasets (at most 1.5 points, on DeepScaler; 0.7 on MATH5000 and 0.6 on GSM8K), so given the remaining ranks the single greedy log-probability is largely redundant. The compact bottom-3 subset (ranks 8–10) is the strongest or tied-strongest single-feature-family variant on GSM8K (AUPRC) and DeepScaler (both metrics) and stays within about a point of the full vector on MATH5000, again showing that a handful of tail ranks alone recovers most of full-vector performance. We do not read the exact optimal k per dataset in Table A9 as meaningful—with a single training run per cell the differences are of similar size to the noise documented elsewhere in the paper—but the near-flat, only-weakly-increasing trend points to redundancy among the ten channels.

Robustness to smaller k. The feature ablation above also shows that PULSE does not depend on the full top-10 vector: Top-5 (80.0 mean, Table A8) is within noise of the deployed Top-10 (80.2 mean), so the evidence remains usable when fewer ranks are available.

G Hand-Crafted Trajectory Baselines: EDIS and TECP

Table 1 adds EDIS (Zhu et al., 2026) and TECP (Xu & Lu, 2025) as hand-crafted alternatives that also read the per-token predictive-distribution trajectory; neither is implemented or compared in prior work in our setting. We implement both directly from the same per-token entropy trajectory $H = (H_1, \ldots, H_T)$, $H_t = -\sum_v p_t(v) \log p_t(v)$, used to compute the PredEnt baseline, so both are cost-1: no re-generation or extra forward pass is needed. TECP sums entropy over the response, $U_{\mathrm{TECP}} = \sum_t H_t$ (Xu & Lu, 2025). EDIS combines a spike-instability score with trajectory variance, $\mathrm{EDIS} = S(H) \cdot (1 + \mathrm{Var}(H))$, $S(H) = \frac{1}{2}(S_{\mathrm{burst}} + S_{\mathrm{rebound}})$, $S_{\mathrm{burst}} = \sum_{t=1}^{T-w} \mathbb{1}[H_{t+w} - H_t > \tau_b]$, $S_{\mathrm{rebound}} = \sum_{t=2}^{T} \mathbb{1}[H_t - \min_{s<t} H_s > \tau_r]$, using the original paper's fixed thresholds $\tau_b = 1.36, \tau_r = 1.33$ (Zhu et al., 2026); the window size w is not specified in that paper, so Table 1 reports $w = 5$ and we separately checked $w = 10$, which differs by at most 0.9 AUROC on any cell—results are insensitive to this undocumented choice. Both scores use the paper's cross-fit sign orientation (Appendix C).

H Model-Fingerprint Details

Section 5.4 summarizes a model-identification classifier trained on per-rank summary statistics (mean, spread, percentiles, local variation) computed over each of the 10 log-probability channels (Eq. 2), fit with a random forest on the concatenated feature vector. The task is five-way classification over Llama-3.2-1B, Llama-3.1-8B, Qwen2.5-1.5B, Qwen3-1.7B, and gemma-4-E2B, where chance accuracy is 20%. Table A11 reports the full results.

Setting	Accuracy
Train GSM8K / test held-out GSM8K	93.6
Train GSM8K → test MATH5000	80.1
Train GSM8K → test DeepScaler	74.5
In-dataset, without response-length feature	93.1

Table A11: Model-fingerprint classification from top-10 log-probability trajectory features. Values are classification accuracy in percent. Chance accuracy is 20%.

I Full Temperature-Generalization Matrices

Section 6 summarizes two 7×7 train/test temperature sweeps, both on main-table models: we train one PULSE head at each of the seven temperatures $t \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2\}$ on GSM8K and evaluate every head on a fixed held-out test set ($n = 500$) generated at each temperature. Tables A12 and A13 report the full matrices for Llama-3.2-1B and Qwen2.5-1.5B respectively. This is not in tension with Section 4.1's claim that ϕ_t is read from the temperature-1 softmax regardless of sampling temperature: that claim is only about how a *single* evidence vector is computed from a given response, not about whether responses (and their correctness rates) sampled at different temperatures induce the same distribution of evidence trajectories. They do not—higher temperature changes which tokens are realized and how often responses are correct, so a head trained on one temperature's trajectory distribution can fail to generalize to another's, exactly as observed below.

Overall AUROC spans a narrow 76.0–84.4 band (mean 80.3), with only weak dependence on the train/test temperature mismatch, consistent with the near-invariance summarized in Section 6.

For Qwen2.5-1.5B the picture is different: AUROC spans 53.6–88.1 (mean 72.7), and heads trained at high temperature ($t \geq 0.8$) transfer poorly to low-temperature test sets and vice versa—the single worst cell (train $t = 0.6$, test $t = 1.2$, AUROC 53.6) is a large-mismatch pair, and the correlation between mismatch $|t_{\mathrm{train}} - t_{\mathrm{test}}|$ and AUROC is $r = -0.46$, confirming the degradation tracks the mismatch itself rather than being noise. As elsewhere, these come from single runs without confidence intervals, so we do not read sub-point differences as meaningful, but the qualitative gap between the two models' matrices is far larger than seed noise.

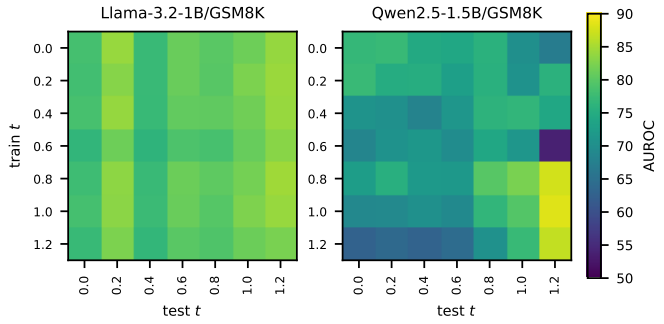


Figure A1: Temperature-generalization matrices: near-uniform for Llama-3.2-1B, structured for Qwen2.5-1.5B.

Train \ Test	$t = 0.0$	$t = 0.2$	$t = 0.4$	$t = 0.6$	$t = 0.8$	$t = 1.0$	$t = 1.2$
$t = 0.0$	**78.2 / 67.3**	83.9 / 80.0	77.1 / 69.7	80.8 / 74.7	79.8 / 76.7	81.1 / 82.0	83.6 / 87.3
$t = 0.2$	77.9 / 67.0	**82.8 / 78.9**	77.1 / 69.8	80.9 / 75.7	79.6 / 76.5	82.1 / 82.7	83.9 / 87.9
$t = 0.4$	78.4 / 67.9	83.7 / 79.3	**77.0 / 69.9**	80.5 / 74.3	80.3 / 76.6	81.3 / 81.8	83.8 / 87.6
$t = 0.6$	76.4 / 64.1	81.2 / 76.5	76.0 / 68.6	**78.8 / 72.8**	78.3 / 75.1	80.7 / 81.6	82.9 / 87.7
$t = 0.8$	77.6 / 67.0	83.2 / 79.8	77.1 / 70.0	81.0 / 76.1	**80.1 / 77.1**	81.6 / 82.3	84.4 / 88.2
$t = 1.0$	78.2 / 68.1	83.0 / 80.4	77.3 / 70.5	80.9 / 75.8	79.4 / 75.8	**82.3 / 83.0**	83.8 / 88.1
$t = 1.2$	76.9 / 66.7	82.0 / 77.8	76.5 / 69.1	79.7 / 73.8	79.0 / 76.5	81.0 / 81.9	**82.0 / 86.8**

Table A12: Full temperature-generalization matrix on Llama-3.2-1B/GSM8K. Rows are the training temperature of the PULSE head; columns are the test-set decoding temperature. Each cell reports AUROC / AUPRC, multiplied by 100, on a fixed held-out test set ($n = 500$). Diagonal (in-temperature) cells are shown in bold.

J Per-Panel Discussion of PULSE Score Separability

Figure A2 shows the PULSE score distribution for every model–dataset cell; this appendix gives the full per-model walkthrough. Separation is sharpest for Llama-3.1-8B and Llama-3.2-1B, where the correct-response mass concentrates near $u_{\mathrm{PULSE}} = 0$ while incorrect responses pile up near 1. On MATH5000 this split is clean on all five models—for gemma-4-E2B, whose 83% base accuracy leaves most responses correct, roughly 90% of the correct mass falls below $u_{\mathrm{PULSE}} = 0.2$ while the incorrect responses spread out toward higher scores, and PULSE narrowly leads this cell over the $\times 2$ P(True) baseline. Qwen3-1.7B fits the same per-dataset pattern as the other models. Its MATH5000 and DeepScaler panels show clean separation—on MATH5000 the correct mass concentrates low (84% below $u_{\mathrm{PULSE}} = 0.2$) while incorrect responses shift high (median $u \approx 0.55$, 83.1 AUROC), and on DeepScaler, its lowest-accuracy cell, the split is sharpest of the three (85% of incorrect responses above 0.5, 83.4 AUROC). On its highest-accuracy GSM8K cell both classes cluster at low scores, as they do for the other easy, high-accuracy GSM8K cells, but the tails still separate (incorrect median $u \approx 0.057$ against correct $u \approx 0.028$), consistent with the solid 76.8 AUROC there. The weakest separation is instead on Qwen2.5-1.5B/GSM8K, where both classes lie closer together and are harder to tell apart—consistent with the more modest AUROC (71.6) there.

K Response Length, Cheap Single-Pass Baselines, and Supervised Controls

Response length is a strong baseline, and PULSE partly reads it. On math reasoning, correct and incorrect solutions differ systematically in length, so response length alone is a strong single-pass correctness signal, and any trajectory readout that can observe length—through positional encoding, attention pooling, or sequence length—partly inherits it. Table A14 reports, per cell, the oriented AUROC of length alone, of the strongest cheap window statistic (window-min confidence: the most-uncertain fixed-width entropy window), of PULSE, and of PULSE evaluated *within* response-length quintiles so length cannot contribute. Length alone averages 73.9—above most single-pass baselines in Table 1—and on the strongest cells (e.g. Llama-3.1-8B) nearly matches PULSE; adding length to PULSE does not improve it (79.3 vs. 80.2), confirming PULSE already encodes length. Crucially, PULSE still retains 71.1 AUROC *within* length strata ($+21.1$ over chance), so it carries a genuine length-independent correctness signal rather than being a length proxy; the $80.2 \rightarrow 71.1$ gap

Train \ Test	$t = 0.0$	$t = 0.2$	$t = 0.4$	$t = 0.6$	$t = 0.8$	$t = 1.0$	$t = 1.2$
$t = 0.0$	**76.6 / 67.2**	77.0 / 65.6	74.3 / 64.0	73.8 / 63.1	75.5 / 70.7	69.9 / 70.8	66.9 / 88.9
$t = 0.2$	77.1 / 67.7	**74.7 / 62.7**	74.9 / 63.3	72.5 / 64.4	75.4 / 70.0	70.5 / 74.7	75.6 / 92.6
$t = 0.4$	70.5 / 59.9	70.2 / 57.9	**67.9 / 59.3**	70.9 / 65.3	75.8 / 72.6	76.4 / 80.2	73.8 / 91.8
$t = 0.6$	68.2 / 57.4	70.5 / 58.1	71.2 / 59.5	**69.4 / 61.0**	73.8 / 68.6	71.0 / 71.1	53.6 / 83.3
$t = 0.8$	72.2 / 61.6	75.2 / 64.4	71.6 / 61.8	71.4 / 66.7	**79.6 / 79.3**	81.9 / 86.9	87.3 / 96.8
$t = 1.0$	68.6 / 57.7	68.8 / 56.1	70.1 / 60.5	69.0 / 64.2	76.3 / 76.3	**79.3 / 83.1**	88.1 / 97.4
$t = 1.2$	62.5 / 47.6	63.6 / 48.1	62.7 / 48.3	64.1 / 51.9	70.3 / 69.3	76.9 / 83.9	**86.4 / 96.5**

Table A13: Full temperature-generalization matrix on Qwen2.5-1.5B/GSM8K. Same protocol as Table A12.

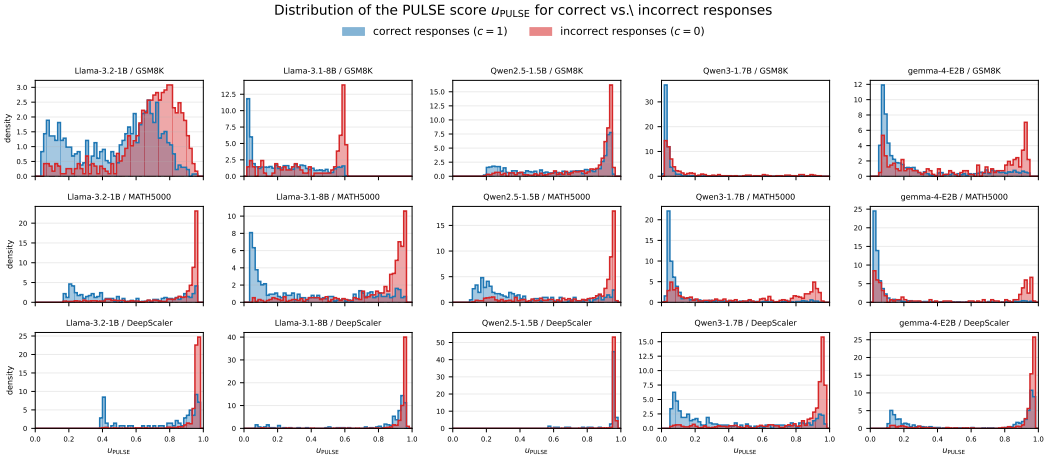


Figure A2: Distribution of the PULSE score u_{PULSE} for correct (blue) versus incorrect (red) responses, one panel per model–dataset cell on the same grid as Table 1. Each histogram is density-normalized so the two classes are comparable despite differing sizes. Because $u_{\text{PULSE}} = 1 - \hat{p}_{\text{corr}}$, larger scores indicate responses predicted more likely to be incorrect; better separability appears as correct mass concentrated at low scores and incorrect mass at high scores.

quantifies how much of its raw ranking is length-mediated. The trajectory signal is real but smaller than the raw numbers suggest. Other cheap single-pass baselines the main table omits (15-cell mean oriented AUROC): window-min confidence 74.9, mean top1–top2 margin 66.9, tail (bottom-10%) confidence 64.2, min token probability 61.7.

Model	Dataset	Length	Win-min	PULSE	PULSE|len-strata
Llama-3.2-1B	GSM8K	65.5	75.3	81.7	79.2
Llama-3.2-1B	MATH5000	76.5	79.5	82.2	69.4
Llama-3.2-1B	DeepScaler	77.9	80.1	81.0	67.1
Llama-3.1-8B	GSM8K	73.9	83.1	86.4	81.5
Llama-3.1-8B	MATH5000	79.8	81.5	85.9	76.4
Llama-3.1-8B	DeepScaler	80.9	81.0	84.8	68.1
Qwen2.5-1.5B	GSM8K	65.6	63.7	71.6	67.1
Qwen2.5-1.5B	MATH5000	74.4	77.2	83.0	72.6
Qwen2.5-1.5B	DeepScaler	71.7	58.5	74.2	70.3
Qwen3-1.7B	GSM8K	65.8	69.4	76.8	72.1
Qwen3-1.7B	MATH5000	78.4	82.5	83.1	70.3
Qwen3-1.7B	DeepScaler	73.7	81.0	83.4	76.8
gemma-4-E2B	GSM8K	71.9	73.4	72.7	66.6
gemma-4-E2B	MATH5000	77.0	76.9	78.0	62.1
gemma-4-E2B	DeepScaler	76.0	60.7	78.3	66.1
Mean		73.9	74.9	80.2	71.1

Table A14: Length as a baseline and length-controlled PULSE (oriented AUROC$\times$100). Win-min = window-min confidence. PULSE|len-strata = PULSE evaluated within response-length quintiles (length cannot contribute). Length alone averages 73.9; PULSE retains 71.1 after controlling length.

A cheap 2-feature alternative. Response length and the TECP score ($\Sigma_t H_t$, Appendix G) are both zero-shot and nearly free to compute; we ask whether a classifier on just these two numbers

already captures most of what the 70-dim pooled probe or PULSE itself achieves. Table A15 fits a 2-feature logistic regression on (length, ΣH) per cell with the same grouped train/val/test protocol as the pooled probe below. It averages 78.0—about even with zero-shot ΣH alone (78.2), so two supervised scalars add essentially nothing over the single free one. The full 70-dim pooled probe does better (79.4), so the extra top-k channels are worth about $+1.4$ over the two scalars, and PULSE's sequence model adds about another $+0.8$ (80.2). The gains are real but small and diminishing at each step.

Model	Dataset	Length	ΣH	Len+ΣH (LR)	Pooled-LR	PULSE
Llama-3.2-1B	GSM8K	65.5	77.4	78.1	80.0	81.7
Llama-3.2-1B	MATH5000	76.5	82.3	82.1	80.8	82.2
Llama-3.2-1B	DeepScaler	77.9	81.3	81.4	81.7	81.0
Llama-3.1-8B	GSM8K	73.9	82.0	79.8	86.5	86.4
Llama-3.1-8B	MATH5000	79.8	84.4	84.3	85.0	85.9
Llama-3.1-8B	DeepScaler	80.9	84.0	84.0	83.9	84.8
Qwen2.5-1.5B	GSM8K	65.6	67.6	67.4	70.0	71.6
Qwen2.5-1.5B	MATH5000	74.4	80.5	80.1	79.9	83.0
Qwen2.5-1.5B	DeepScaler	71.7	68.5	69.2	74.5	74.2
Qwen3-1.7B	GSM8K	65.8	73.4	73.3	75.4	76.8
Qwen3-1.7B	MATH5000	78.4	82.0	81.8	82.2	83.1
Qwen3-1.7B	DeepScaler	73.7	81.2	81.3	82.3	83.4
gemma-4-E2B	GSM8K	71.9	75.2	74.8	75.5	72.7
gemma-4-E2B	MATH5000	77.0	79.1	79.2	79.4	78.0
gemma-4-E2B	DeepScaler	76.0	74.2	73.7	73.4	78.4
Mean		73.9	78.2	78.0	79.4	80.2

Table A15: Cheap length/entropy baselines vs. the supervised controls (oriented AUROC$\times$100). ΣH = TECP's summed-entropy score, zero-shot. Len+ΣH (LR) = a 2-feature logistic regression on (length, ΣH), same grouped train/val/test protocol as Pooled-LR. Pooled-LR/PULSE columns match Table A18. The 2-feature probe (78.0) is about even with zero-shot ΣH (78.2); the 70-dim pooled probe (79.4) and PULSE (80.2) each add roughly $+1$ on top.

Does the edge over these baselines survive length control? Table A16 extends Table A14's within-length-quintile protocol from PULSE alone to TECP and EDIS. Un-stratified, PULSE leads TECP by $+1.9$ (Table 1); *within* length strata the gap *widens* to $+3.9$ (mean within-strata AUROC: PULSE 71.1, TECP 67.2, EDIS 65.6), consistent with TECP's summed-entropy score being itself a partial length proxy that loses more from length control than PULSE does. PULSE's within-strata edge is real on average but not uniform across cells.

Model	Dataset	TECP|strata	EDIS|strata	PULSE|strata
Llama-3.2-1B	GSM8K	74.2	73.5	79.2
Llama-3.2-1B	MATH5000	72.5	72.8	69.4
Llama-3.2-1B	DeepScaler	66.3	64.9	67.1
Llama-3.1-8B	GSM8K	73.2	73.8	81.5
Llama-3.1-8B	MATH5000	71.9	70.7	76.4
Llama-3.1-8B	DeepScaler	67.1	65.6	68.1
Qwen2.5-1.5B	GSM8K	57.5	57.4	67.1
Qwen2.5-1.5B	MATH5000	70.0	69.6	72.6
Qwen2.5-1.5B	DeepScaler	54.6	54.3	70.3
Qwen3-1.7B	GSM8K	69.3	55.1	72.1
Qwen3-1.7B	MATH5000	69.3	66.2	70.3
Qwen3-1.7B	DeepScaler	73.9	74.6	76.8
gemma-4-E2B	GSM8K	66.4	59.8	66.6
gemma-4-E2B	MATH5000	63.2	65.2	62.1
gemma-4-E2B	DeepScaler	58.0	60.4	66.1
Mean		67.2	65.6	71.1

Table A16: Length-quintile-stratified AUROC$\times$100 (length cannot contribute), extending Table A14's PULSE|len-strata protocol to the other methods. PULSE retains the largest within-strata AUROC on $13/15$ cells and the highest mean (71.1 vs. 67.2/65.6 for TECP/EDIS).

Fixing the supervision budget: what the top-k channel is worth. Because PULSE is trained while the scalar baselines are zero-shot, we compare it against two supervised controls sharing its protocol and, for one, its exact architecture: (i) an order-agnostic *pooled-statistics* probe

(seven pooled statistics—mean, standard deviation, min, max, and the 25th/50th/75th percentiles—computed over the response's per-step values for each of the k ranks, 70-dim, logistic regression); and (ii) an *entropy-only Transformer*—the identical causal Transformer and Dirichlet head fed only the 1-dim per-step entropy sequence. With supervision held fixed (Table A18), the top-k logprobability *channel* is the main lever: PULSE beats the entropy-only Transformer by $+2.1$ on average (same architecture, wider input) and the pooled probe by only $+0.8$. A paired bootstrap on the same shared test indices (identical protocol to Appendix D, $B = 2000$) shows PULSE beats pooled-LR outright on just $3/15$ cells and is statistically tied on the remaining 12, with zero losses (Sig. column, Table A18): once supervision and full channel width are granted, adding sequence order buys little on top. The pooled probe (79.4) itself sits above zero-shot TECP (78.2), so supervision plus width—not order—is what lifts a probe above the free baseline. Because pooled statistics (means, percentiles) never see the raw sequence length T, we test whether PULSE's residual $+0.8$ over the pooled probe is really hidden length by augmenting the pooled probe's inputs with $\log T$ and $\Sigma_t H_t$. The augmented probe averages 79.7 (per cell: Qwen2.5-1.5B 71.5/81.4/74.5, gemma-4-E2B 75.6/79.8/69.7, Llama-3.2-1B 80.5/82.8/82.0, Qwen3-1.7B 75.9/82.3/82.9, Llama-3.1-8B 86.4/85.6/84.5; GSM8K/MATH5000/DeepScaler), a mere $+0.3$ over the 79.4 pooled baseline and still below PULSE's 80.2. Explicit length thus closes only part of the gap, leaving a small residual attributable to sequence order.

These two controls each change *one* axis relative to PULSE (order or channel width) but leave the other fixed at PULSE's setting, so neither alone tells us whether order and channel width are independent, complementary, or redundant sources of signal. To test this we add the missing fourth corner of the 2×2 (order $\times$ channel-width) design: pooled statistics of the *entropy channel alone* (order-agnostic, 1-channel, same seven pooled statistics as the pooled probe but over 1 channel instead of 10, 7-dim logistic regression). Table A17 reports all four corners. Both single-axis effects are sizeable—adding order at fixed 1-channel width gains $+3.1$ ($75.0 \to 78.1$), and adding the other 9 channels at fixed order-agnostic pooling gains $+4.4$ ($75.0 \to 79.4$)—so channel width contributes somewhat more than order but both are informative. What is genuinely small is their combination: an additive model summing both single-axis effects onto the 1-channel, order-agnostic floor predicts $75.0 + 3.1 + 4.4 = 82.5$ for PULSE, but PULSE's actual mean is 80.2, a -2.3 interaction (mean of the 15 per-cell interaction terms -2.3, median -1.2). Order and channel width are thus mildly *redundant* rather than synergistic: each recovers an overlapping share of the same underlying confidence signal, so combining them returns a little less than their sum, not more. PULSE's advantage over zero-shot scalars is carried by the supervised top-k logprob representation as a whole—order and width together—rather than by either axis being the "real" source and the other incidental; the pooled probe, and even its 2-feature reduction above, remains a genuinely competitive, cheap-to-train control.

	1-channel (entropy only)	10-channel (top-k)
Order-agnostic (pooled stats)	75.0	79.4
Order-aware (sequence model)	78.1	80.2 (PULSE)

Table A17: The 2×2 order $\times$ channel-width design (mean AUROC$\times$100 over 15 cells). Single-axis effects from the 1-channel order-agnostic floor: width $+4.4$, order $+3.1$; additive prediction $75.0 + 4.4 + 3.1 = 82.5$; actual PULSE mean 80.2 (interaction -2.3, per-cell median -1.2)—width and order recover overlapping, not complementary, signal.

Is logistic regression an artificially weak pooled classifier? The pooled-statistics control above uses logistic regression on the 70-dim pooled vector; a weak classifier there would inflate the apparent value of PULSE's sequence modeling. We refit the identical 70-dim input with a small MLP (one hidden layer, same protocol as the main pooled-LR control) and with a histogram gradient-boosted tree (`HistGradientBoostingClassifier`, the same family as LightGBM/XGBoost), across all 15 cells. Neither non-linear classifier beats logistic regression on average: mean AUROC is 78.4 (MLP) and 79.2 (GBDT), both *below* logistic regression's 79.4. If anything, the simplest classifier is the strongest of the three on this 70-dim representation, so the pooled-LR control is not an artificially weak stand-in for a stronger non-sequential baseline.

L Evidential Head vs. a Plain BCE Head

Section 5.5 asks whether the Dirichlet (evidential) parameterization of Section 4.2 earns the extra hyperparameter λ_{KL} it introduces over a plain calibrated-probability head.

Model	Dataset	Pooled-LR	Entropy-only	PULSE (top-k)	Sig. vs. PULSE
Llama-3.2-1B	GSM8K	80.0	78.6	81.7	win
Llama-3.2-1B	MATH5000	80.8	82.9	82.2	tie
Llama-3.2-1B	DeepScaler	81.7	79.0	81.0	tie
Llama-3.1-8B	GSM8K	86.5	82.6	86.4	tie
Llama-3.1-8B	MATH5000	85.0	84.0	85.9	tie
Llama-3.1-8B	DeepScaler	83.9	83.4	84.8	tie
Qwen2.5-1.5B	GSM8K	70.0	67.0	71.6	tie
Qwen2.5-1.5B	MATH5000	79.9	80.7	83.0	win
Qwen2.5-1.5B	DeepScaler	74.5	68.5	74.2	tie
Qwen3-1.7B	GSM8K	75.4	74.0	76.8	tie
Qwen3-1.7B	MATH5000	82.2	82.4	83.1	tie
Qwen3-1.7B	DeepScaler	82.3	82.7	83.4	tie
gemma-4-E2B	GSM8K	75.5	75.8	72.7	tie
gemma-4-E2B	MATH5000	79.4	79.5	78.0	tie
gemma-4-E2B	DeepScaler	73.4	70.8	78.4	win
Mean		79.4	78.1	80.2	3 win / 12 tie

Table A18: Supervised controls at fixed supervision (AUROC$\times$100). Pooled-LR: order-agnostic 70-dim pooled statistics with logistic regression. Entropy-only: PULSE's exact architecture fed only the 1-dim per-step entropy sequence. PULSE: 10-dim top-k log-probabilities (seed-0 checkpoints). Sig.: paired bootstrap verdict (Appendix D protocol, $B = 2000$), win if PULSE's resampled AUROC exceeds pooled-LR's on $\geq 97.5\%$ of resamples. PULSE beats pooled-LR by $+0.8$ on average and is a bootstrap tie on all but 3 cells (never a loss), and beats the entropy-only sequence model by $+2.1$. Adding explicit $\log T$ and $\Sigma_t H_t$ to the pooled probe raises it only to 79.7 (per-cell values in the text).

`TransformerDirichlet` already supports a 1-dimensional sigmoid output head as a drop-in alternative (identical backbone: input projection, causal Transformer, attention pooling, two-layer MLP head; only the output width and loss—BCE vs. the evidential KL-regularized loss—differ), so we retrain every main-table cell with the BCE head on the identical train/val split and evaluate on the identical test set. Table A19 reports the full comparison.

Model	Dataset	BCE head	Dirichlet head	Δ
Llama-3.2-1B	GSM8K	82.2	81.7	-0.5
Llama-3.2-1B	MATH5000	83.4	82.3	-1.1
Llama-3.2-1B	DeepScaler	81.8	81.0	-0.8
Llama-3.1-8B	GSM8K	86.5	86.4	-0.1
Llama-3.1-8B	MATH5000	86.3	86.1	-0.1
Llama-3.1-8B	DeepScaler	84.4	84.8	$+0.4$
Qwen2.5-1.5B	GSM8K	69.1	70.8	$+1.7$
Qwen2.5-1.5B	MATH5000	82.6	82.1	-0.4
Qwen2.5-1.5B	DeepScaler	75.1	73.3	-1.8
Qwen3-1.7B	GSM8K	76.1	76.1	0.0
Qwen3-1.7B	MATH5000	82.3	83.1	$+0.8$
Qwen3-1.7B	DeepScaler	83.3	83.8	$+0.5$
gemma-4-E2B	GSM8K	67.0	76.3	$+9.3$
gemma-4-E2B	MATH5000	78.3	78.3	0.0
gemma-4-E2B	DeepScaler	73.4	75.9	$+2.5$
Mean		79.5	80.1	$+0.7$
Median				0.0

Table A19: Evidential (Dirichlet) head vs. a plain BCE head, same backbone and protocol (AUROC$\times$100). The Dirichlet head wins $6/15$ cells, ties on 2, and loses 7; the mean gap is inflated by one outlier (gemma-4-E2B/GSM8K, $+9.3$), and the median gap (0.0) shows the typical cell has no difference at all between heads. This is a small, inconsistent edge, not the source of PULSE's advantage over the zero-shot baselines in Table 1.

M Calibration and Selective-Prediction Metrics

Table A20 reports, per main-table cell, the expected calibration error (ECE, 15-bin), Brier score, and area under the risk–coverage curve (AURC, retaining the most-confident coverage levels first) of the evidential head's $\hat{p}_{\text{corr}}$ output, computed offline by a forward pass of the already-trained checkpoint on the same test set as Table 1; no retraining or recalibration is applied. Oracle AURC

is the best AURC achievable given the cell's base error rate (perfect ranking), included as a lower-bound reference.

Model	Dataset	ECE	Brier	AURC	Oracle AURC
Llama-3.2-1B	GSM8K	3.6	0.184	25.1	13.3
Llama-3.2-1B	MATH5000	5.9	0.140	54.2	40.9
Llama-3.2-1B	DeepScaler	3.3	0.065	80.1	70.6
Llama-3.1-8B	GSM8K	4.9	0.069	2.0	0.4
Llama-3.1-8B	MATH5000	7.5	0.159	27.7	19.5
Llama-3.1-8B	DeepScaler	4.3	0.111	60.0	49.2
Qwen2.5-1.5B	GSM8K	21.2	0.277	30.2	12.9
Qwen2.5-1.5B	MATH5000	7.7	0.145	53.0	40.5
Qwen2.5-1.5B	DeepScaler	6.1	0.088	81.9	68.4
Qwen3-1.7B	GSM8K	12.0	0.144	7.8	1.7
Qwen3-1.7B	MATH5000	11.7	0.178	16.7	8.3
Qwen3-1.7B	DeepScaler	9.0	0.157	44.7	31.9
gemma-4-E2B	GSM8K	41.2	0.396	7.8	1.5
gemma-4-E2B	MATH5000	7.0	0.114	7.7	1.5
gemma-4-E2B	DeepScaler	26.7	0.294	29.3	15.6
Mean		11.5	0.168	35.2	25.1

Table A20: Calibration (ECE, Brier) and selective-prediction (AURC) of $\hat{p}_{\text{corr}}$, raw and unrecalibrated. Calibration is uneven across cells (ECE 3.3–41.2); the risk–coverage ranking that drives abstention is closer to the oracle bound than raw calibration would suggest (mean AURC gap 10.1).

Does a cheap post-hoc fix actually repair calibration? We apply 1-dimensional Platt scaling: a logistic regression on the logit of $\hat{p}_{\text{corr}}$, fit on a held-out validation split of the *training* data (never test labels) and applied to the test set. Table A21 reports the per-cell result. Mean ECE drops from 11.5 to 8.4 and mean Brier from 0.168 to 0.153; 11 of 15 cells improve, including the worst raw-ECE cell (gemma-4-E2B/GSM8K, $41.2 \rightarrow 29.1$). The fix is real but partial—four already well-calibrated cells get marginally worse (largest: Llama-3.1-8B/GSM8K, $4.9 \rightarrow 5.9$), and the worst cell remains far from perfectly calibrated even after correction. We take this as evidence for, not proof against, the paper's position: $\hat{p}_{\text{corr}}$ is not perfectly calibrated out of the box, a cheap fix meaningfully narrows the gap on average and on the worst cases, but does not fully close it, so calibration-sensitive deployments should apply and re-validate such a fix rather than trust the raw probability.

Model	Dataset	Raw ECE	Platt ECE	Raw Brier	Platt Brier
Llama-3.2-1B	GSM8K	3.6	5.3	0.184	0.186
Llama-3.2-1B	MATH5000	5.9	4.1	0.140	0.138
Llama-3.2-1B	DeepScaler	3.3	2.8	0.065	0.064
Llama-3.1-8B	GSM8K	4.9	5.9	0.069	0.068
Llama-3.1-8B	MATH5000	7.5	5.5	0.159	0.155
Llama-3.1-8B	DeepScaler	4.3	5.0	0.111	0.112
Qwen2.5-1.5B	GSM8K	21.1	16.6	0.277	0.248
Qwen2.5-1.5B	MATH5000	7.7	3.2	0.145	0.136
Qwen2.5-1.5B	DeepScaler	6.1	3.9	0.087	0.082
Qwen3-1.7B	GSM8K	12.0	4.7	0.144	0.128
Qwen3-1.7B	MATH5000	11.7	5.9	0.178	0.163
Qwen3-1.7B	DeepScaler	9.0	3.3	0.157	0.149
gemma-4-E2B	GSM8K	41.2	29.1	0.396	0.250
gemma-4-E2B	MATH5000	7.0	2.3	0.113	0.105
gemma-4-E2B	DeepScaler	26.7	28.3	0.294	0.304
Mean		11.5	8.4	0.168	0.153

Table A21: Post-hoc Platt scaling, fit on a held-out training-side validation split and applied unchanged to the test set (no test-label leakage). Improves $11/15$ cells; the fix is real but partial. Raw ECE/Brier here are recomputed independently of Table A20 and agree to within floating-point rounding (e.g. Qwen2.5-1.5B/GSM8K: 21.1 here vs. 21.2 there).

N Sample-Efficiency Curve

Table A22 reports, per main-table cell, PULSE's AUROC when trained on a random subsample of 100, 500, or 2000 labeled examples from that cell's original training set (4000–5978 examples), evaluated on the same held-out test set as Table 1. Each entry is a mean±std over 3 seeds (varying

both the subsample draw and model init/ordering, matching the main table's own seed protocol); the "Full" column reproduces the main table's 3-seed PULSE AUROC for reference.

Model	Dataset	n=100	n=500	n=2000	Full
Llama-3.2-1B	GSM8K	77.7±3.1	79.8±0.5	80.7±0.8	81.8
Llama-3.2-1B	MATH5000	77.7±2.9	80.2±2.2	81.9±0.2	82.3
Llama-3.2-1B	DeepScaler	75.4±1.5	81.4±0.8	82.1±0.3	81.5
Llama-3.1-8B	GSM8K	82.8±2.4	84.1±1.3	85.7±0.7	85.7
Llama-3.1-8B	MATH5000	81.0±2.9	85.5±0.2	86.2±0.1	86.1
Llama-3.1-8B	DeepScaler	82.5±0.4	82.6±0.4	82.7±1.4	83.8
Qwen2.5-1.5B	GSM8K	66.4±5.2	72.5±0.4	71.4±1.1	70.8
Qwen2.5-1.5B	MATH5000	79.1±1.5	81.3±0.3	81.7±0.9	82.1
Qwen2.5-1.5B	DeepScaler	72.9±2.3	72.6±7.2	74.7±1.7	73.3
Qwen3-1.7B	GSM8K	71.4±1.7	72.8±2.9	72.9±4.3	76.1
Qwen3-1.7B	MATH5000	78.8±3.0	80.2±0.9	82.6±0.6	83.1
Qwen3-1.7B	DeepScaler	76.8±1.8	80.6±0.8	82.8±0.6	83.8
gemma-4-E2B	GSM8K	73.7±4.0	74.9±2.5	69.8±6.5	76.3
gemma-4-E2B	MATH5000	77.2±1.2	77.1±1.0	78.8±0.1	78.3
gemma-4-E2B	DeepScaler	78.5±3.0	80.2±1.0	76.2±1.7	75.9
Mean		76.8	79.1	79.4	80.1

Table A22: Sample-efficiency curve: PULSE AUROC vs. labeled training-set size, mean±std over 3 seeds. The 15-cell mean rises monotonically toward the full-data value ($76.8 \rightarrow 79.1 \rightarrow 79.4 \rightarrow 80.1$), and per-cell curves are similarly well-behaved; n=100 already recovers 96% of full-data AUROC.

O Cross-Sample Reference Baselines: Semantic Entropy and MC Predictive Entropy

Section 5.1 excludes cross-sample methods from the bold comparison because they score a different variable (a voted/aggregated answer over K samples) than the single fixed candidate-0 response every other method scores. To situate PULSE's "cheap but competitive" positioning against these stronger, more expensive methods rather than only omitting them, we compute two of them the standard way (Kuhn et al., 2023; Malinin & Gales, 2021) on the 10 main-table cells (GSM8K, DeepScaler) where cost-5 candidate data exists (MATH5000 was generated at cost-1 only): for each question, candidate-0 is the response being scored, and the *other* 4 sampled candidates—never candidate-0 itself—are used to build the uncertainty estimate. Semantic entropy clusters the 4 reference candidates' extracted answers by exact match (the standard simplification when labels are already exact-match, as here) and takes the Shannon entropy of the resulting cluster distribution; MC predictive entropy averages the 4 reference candidates' own length-normalized negative log-likelihoods. Table A23 reports both against PULSE on the same 10 cells.

Model	Dataset	PULSE	Semantic Ent. (K=5)	MC PredEnt (K=5)
Llama-3.2-1B	GSM8K	81.7	74.4	69.4
Llama-3.2-1B	DeepScaler	81.0	72.2	74.9
Llama-3.1-8B	GSM8K	86.4	77.7	74.4
Llama-3.1-8B	DeepScaler	84.8	82.0	72.6
Qwen2.5-1.5B	GSM8K	71.5	64.3	64.7
Qwen2.5-1.5B	DeepScaler	74.2	67.6	75.4
Qwen3-1.7B	GSM8K	76.8	76.4	64.8
Qwen3-1.7B	DeepScaler	83.4	78.9	77.2
gemma-4-E2B	GSM8K	72.8	72.1	69.6
gemma-4-E2B	DeepScaler	78.4	79.8	82.1
Mean		79.1	74.5	72.5

Table A23: PULSE (cost-1) vs. two cross-sample reference baselines (cost-5, 4 extra generations per question) on the 10 cells with cost-5 data. Despite using $5\times$ the generations and having access to 4 additional full samples, semantic entropy trails PULSE on $9/10$ cells and MC predictive entropy on $8/10$ (MC predictive entropy edges PULSE on the two DeepScaler cells for Qwen2.5-1.5B and gemma-4-E2B); both are strong—above most cost-1 scalar baselines in Table 1—so this is a genuine reference upper bound, not a straw man.

P WITHIN-QUESTION AUROC: RESPONSE-LEVEL VS. QUESTION-LEVEL SIGNAL

The prompt-only white-box control (Appendix Q) scores 77.2, close to PULSE's 80.1 and ahead of it on DeepScaler, raising a direct question: how much of PULSE's global AUROC is *this response is wrong* versus *this question is hard*, a signal already present before any token is generated? Length-quintile stratification (Appendix K) controls one such confound; the direct test is *within-question* AUROC, ranking only the correct-vs.-incorrect responses to the *same* question against each other and pooling across questions. Any signal that is constant across a question's responses (pure difficulty) cannot contribute here—only genuine per-response signal can.

We reuse the same 5-candidates-per-question files as the cross-sample reference baselines above (Appendix O, available for GSM8K/DeepScaler only), score every one of the 5 candidates with the identical canonical checkpoint used in the main table, and pool all (correct, incorrect) response pairs sharing a question id (skipping the 14–76% of questions per cell whose 5 candidates are unanimous and contribute no pair). Table A24 reports the result against response length, TECP, and EDIS, computed from the same pooled candidates.

Model	Dataset	PULSE|w-q	PULSE glob.	Length|w-q	TECP|w-q	EDIS|w-q
Llama-3.2-1B	GSM8K	74.2	81.4	51.1	64.5	66.5
Llama-3.2-1B	DeepScaler	59.7	81.2	55.4	54.8	56.2
Llama-3.1-8B	GSM8K	74.2	86.3	57.8	69.1	71.1
Llama-3.1-8B	DeepScaler	66.6	83.9	61.8	66.9	66.1
Qwen2.5-1.5B	GSM8K	70.0	76.0	53.2	63.3	64.5
Qwen2.5-1.5B	DeepScaler	64.8	80.7	57.1	64.6	63.9
Qwen3-1.7B	GSM8K	62.4	76.8	46.6	55.5	54.0
Qwen3-1.7B	DeepScaler	61.4	82.7	53.2	57.7	57.9
gemma-4-E2B	GSM8K	69.8	73.4	56.5	57.3	57.2
gemma-4-E2B	DeepScaler	65.5	84.9	60.0	64.8	64.6
Mean		66.8	80.7	55.3	61.9	62.2

Table A24: Within-question AUROC (pooling only correct-vs.-incorrect pairs that share a question id) with, for PULSE, the global AUROC on the identical pooled candidate set, on the 10 cells with 5-candidates-per-question data. PULSE retains a mean 66.8 within-question AUROC, above every baseline's within-question mean and above chance on all 10 cells; length is the weakest (55.3), TECP and EDIS in between.

Every method loses a share of its global AUROC once question identity is controlled (PULSE $80.7 \to 66.8$, TECP $78.9 \to 61.9$, length $70.9 \to 55.3$): a real portion of raw ranking ability is question-level, not response-level, consistent with the prompt-only control above. But PULSE does not collapse to chance—among responses to the *same* question it still ranks the incorrect one higher $\sim 67\%$ of the time, above every baseline on average and above chance on all 10 cells. In particular response length, the dominant global confound, is the *weakest* within-question signal (55.3, barely above chance), so PULSE's within-question edge is genuinely per-response and not a length artifact. This is the direct rebuttal to the prompt-only control's implication: PULSE is not *only* reading question difficulty.

Q WHITE-BOX HIDDEN-STATE PROBE (REFERENCE UPPER BOUND)

PULSE is deliberately black-box: it reads only the per-step top-k log-probabilities of a single generation. To quantify what this restriction costs, we compare against a white-box probe with access to the model's hidden states. For each response we take the model's last-layer hidden states over the generated tokens (mean-pooled and last-token) and train a logistic-regression probe on them to predict correctness—the same probe family as our pooled control, but on activations rather than top-k log-probs (the Azaria & Mitchell / Semantic Entropy Probe setting). We use a grouped train/test split (by question, no leakage) on up to 1500 responses per cell; capping training only *under*-states the white-box probe, so the gap below is conservative.

Table A25 reports the stronger of the mean-pooled and last-token probes against PULSE. *One caveat on the comparison*: the white-box probe is evaluated on its own grouped held-out split of the training-domain generations, whereas the PULSE column is its canonical-test-set AUROC (Table 1), so the two columns are scored on comparable but not identical response sets. The per-cell Δ and the "$14/15$" count below should therefore be read as an aggregate reference-bound comparison, not a strict paired test—a same-test-set re-extraction is future work. With that caveat, the white-box

probe averages 86.6 AUROC versus PULSE's 80.1—a +6.6 gap—and is higher on 14 of 15 cells. The gap conflates two changes at once—the input (activations vs. top-k log-probs) and the model (logistic regression vs. PULSE's sequence Transformer)—so it should be read as a black-box *vs.* white-box headroom number, not a clean ablation of either factor alone; the closest same-model comparison is white-box logistic regression (86.6) against the order-agnostic pooled-statistics probe on top-k log-probs (79.4, Section 5.3), which isolates the input change and still leaves a large +7.2 gap. The top-k log-probability channel thus leaves substantial correctness signal on the table relative to hidden states: PULSE is the strongest *black-box* estimator we find (never significantly worse than any black-box baseline in Table 1), but when activations are available a hidden-state probe is clearly preferable. This gap is the headroom of the black-box setting, not a property of the task, and motivates future black-box readouts that narrow it.

Model	Dataset	PULSE (black-box)	Hidden-state (white-box)	Δ
Llama-3.2-1B	GSM8K	81.8	83.7	+1.9
Llama-3.2-1B	MATH5000	82.3	83.3	+1.0
Llama-3.2-1B	DeepScaler	81.5	90.3	+8.8
Llama-3.1-8B	GSM8K	85.7	82.7	−3.0
Llama-3.1-8B	MATH5000	86.1	89.8	+3.7
Llama-3.1-8B	DeepScaler	83.8	87.9	+4.1
Qwen2.5-1.5B	GSM8K	70.8	80.6	+9.8
Qwen2.5-1.5B	MATH5000	82.1	82.9	+0.8
Qwen2.5-1.5B	DeepScaler	73.3	87.6	+14.3
Qwen3-1.7B	GSM8K	76.1	92.2	+16.1
Qwen3-1.7B	MATH5000	83.1	86.0	+2.9
Qwen3-1.7B	DeepScaler	83.8	92.1	+8.3
gemma-4-E2B	GSM8K	76.3	87.1	+10.8
gemma-4-E2B	MATH5000	78.3	86.9	+8.6
gemma-4-E2B	DeepScaler	75.9	86.5	+10.6
Mean		80.1	86.6	+6.6

Table A25: Black-box PULSE vs. a white-box last-layer hidden-state probe (logistic regression on activations, stronger of mean-pooled / last-token), oriented AUROC×100. The white-box probe wins on 14/15 cells by +6.6 on average; PULSE is the best black-box estimator but trails when activations are accessible. Grouped, leakage-free split; training capped at 1500 responses per cell, which under-states the white-box probe.

Is the white-box gap generation-monitoring, or prompt leakage? The probe above pools hidden states over the *generated* response, but a model's representation of the *prompt alone*—before it produces a single token—may already encode how hard or familiar a problem is, the same kind of confound response length is for the black-box side (Appendix K). We test this directly: the identical probe trained on only the hidden state at the *last prompt token* (one forward pass, no generation, same 1500-response cap and split). If this matched the full white-box probe, the +6.6 gap over PULSE would be mostly prompt-level leakage rather than a benefit of monitoring generation.
It does not. Table A26 shows the prompt-only probe averages 77.2, *below* PULSE's 80.1, and 9.4 below the generation-pooled white-box probe's 86.6—a gap significant on every single cell (paired Wilcoxon signed-rank, $p = 6\times10^{-5}$, 15/15). Reading the prompt alone is a real but modest signal: above the length baseline (77.2 vs. 73.9, +3.3, paired Wilcoxon $p = 0.03$), but not significantly different from PULSE overall ($p = 0.18$, prompt-only wins only 6/15 cells, concentrated on DeepScaler). The white-box probe's advantage is therefore substantially a property of monitoring the generation, not a static read of problem difficulty available before generation starts, supporting the "ceiling from activations" framing in the main text rather than undercutting it.

R Selective Prediction (Downstream Abstention)

We turn each uncertainty signal into an abstention rule: rank the responses of a cell by uncertainty, retain the most-confident fraction (coverage), and measure error on the retained set. Table A27 reports the resulting AURC (area under the risk–coverage curve, lower is better) and the selective accuracy at 50% coverage, averaged over the 15 cells, for PULSE and the cheap black-box baselines it can be computed alongside (all use only the single response, no extra cost). PULSE gives the best (lowest) mean AURC and the best selective accuracy, and is the single best signal on 11 of 15 cells—so the AUROC edge in Table 1 converts into a concrete deployment benefit (which responses to trust or defer). The margin over the summed-entropy TECP baseline is small, consistent with the main table.

Model	Dataset	Length	Prompt-only	PULSE	White-box (gen.)
Llama-3.2-1B	GSM8K	65.5	70.8	81.8	83.7
Llama-3.2-1B	MATH5000	76.5	77.8	82.3	83.3
Llama-3.2-1B	DeepScaler	77.9	84.3	81.5	90.3
Llama-3.1-8B	GSM8K	73.9	65.9	85.7	82.7
Llama-3.1-8B	MATH5000	79.8	81.6	86.1	89.8
Llama-3.1-8B	DeepScaler	80.9	81.8	83.8	87.9
Qwen2.5-1.5B	GSM8K	65.6	65.3	70.8	80.6
Qwen2.5-1.5B	MATH5000	74.4	74.1	82.1	82.9
Qwen2.5-1.5B	DeepScaler	71.7	83.3	73.3	87.6
Qwen3-1.7B	GSM8K	65.8	76.7	76.1	92.2
Qwen3-1.7B	MATH5000	78.4	81.5	83.1	86.0
Qwen3-1.7B	DeepScaler	73.7	84.5	83.8	92.1
gemma-4-E2B	GSM8K	71.9	68.8	76.3	87.1
gemma-4-E2B	MATH5000	77.0	80.2	78.3	86.9
gemma-4-E2B	DeepScaler	76.0	80.8	75.9	86.5
Mean		73.9	77.2	80.1	86.6

Table A26: The prompt-only hidden-state control (last-prompt-token activation, zero generated tokens, same probe family and split as Table A25), against length, PULSE, and the generation-pooled white-box probe. Ordering is length $<$ prompt-only $<$ PULSE $<$ white-box on average; the white-box probe beats prompt-only on $15/15$ cells ($p = 6 \times 10^{-5}$), so its advantage is mainly from monitoring the generation, not from problem-difficulty already visible in the prompt.

Signal	Mean AURC ↓	Selective acc. @50% ↑
PULSE	**35.2**	**63.5**
TECP	36.1	62.3
Length	38.7	60.1
PredEnt	41.4	58.5
LogP	43.8	55.4

Table A27: Selective prediction over the 15 cells: mean AURC (area under risk–coverage, lower is better) and mean selective accuracy at 50% coverage. PULSE is the best black-box abstention signal (best on $11/15$ cells).

S Free-Form QA Generalization Check: TriviaQA

All main-table results are on exact-match mathematical reasoning; we run a smaller, single-seed check on free-form factual QA to see whether the attribution qualitatively holds elsewhere. We use TriviaQA's closed-book `rc.nocontext` configuration (no supporting document or search results, just the question), the setting closest to a model answering from parametric knowledge alone—and the setting where confidently-wrong answers (hallucination) are most at stake. Responses are graded by the standard SQuAD/TriviaQA protocol: normalize (lowercase, strip punctuation and articles) and check whether any of the question's reference answer aliases appears as a substring of the normalized response, since models frequently answer in a full phrase rather than an exact string match. We run three representative models spanning an $8\times$ parameter range and two families (Llama-3.2-1B, Llama-3.1-8B, gemma-4-E2B), 2000 training and 500 test questions per model, one seed, at the same temperature (0.8) and with the same top-k evidence and architecture as the main table.

Table A28 reports PULSE against the same cheap black-box baselines used throughout the paper. PULSE remains the best or near-best signal (mean 86.3, best on $2/3$ cells), so the top-k channel's usable signal is not an artifact of exact-match math grading. But the domain differs qualitatively from math reasoning in exactly the way the length attribution (Section 5.3) predicts: response length, the dominant confound on math (73.9 mean AUROC there), collapses to near-chance here (53.2 mean)—short factual answers do not confound with correctness the way long reasoning traces do. The one exception is the weakest model (Llama-3.2-1B, base error 83.8%), where plain mean log-probability edges PULSE narrowly (85.8 vs. 85.1). Given the small scale (one seed, three models, 500-example test sets, no bootstrap), we read this as an existence result—the attribution's qualitative shape (PULSE competitive, length domain-dependent) survives leaving math reasoning—not as a claim that our 15-cell, 3-seed math results themselves generalize quantitatively.

Model	Base error	PULSE	Length	TECP	PredEnt	LogP
Llama-3.2-1B	83.8%	85.1	50.1	81.3	82.9	<u>85.8</u>
Llama-3.1-8B	32.8%	**88.9**	57.2	86.7	85.9	82.4
gemma-4-E2B	62.6%	**84.8**	52.2	79.8	76.3	73.9
Mean		**86.3**	53.2	82.6	81.7	80.7

Table A28: TriviaQA (closed-book, `rc.nocontext`) generalization check, oriented AUROC$\times$100, single seed. **Bold** = best in row, <u>underline</u> = best baseline where it edges PULSE. Length's near-chance mean (53.2) contrasts sharply with its role as the dominant confound on math reasoning (73.9; Table A14).