Graded Metacognition in LLMs

Anonymous ACL submission

Abstract

Metacognition, the ability to monitor one's own knowing, is measured in humans through graded confidence ratings that track accuracy. LLM metacognition typically uses a binary Yes/No probe evaluated with Type-2 d'. We show this is structurally fragile: across eleven LLMs from five families (Qwen2.5, Qwen3, Llama-3.2, Mistral, Gemma-2) on MMLU, the binary probe collapses to near-constant "no" on most models, a failure that calibration scores hide. We propose a graded measurement framework pairing bias-resistant metrics from human metacognition research (Goodman–Kruskal γ, Type-2 AUROC, grade separation) with a probe-health diagnostic that flags response collapse. We extend ESMA, a gradient-free evolution-strategies training method, with a continuous graded reward across a four-level (A–D) confidence scale that rewards confident-correct responses more than tentative ones. Graded metacognition emerges as a scale-dependent capability appearing only at mid-sized models and replicating across all five families. Graded ESMA substantially improves rank-calibration with concurrent accuracy gains, rescues smaller models from anti-calibration, and aligns verbalised confidence with internal token probabilities, addressing the implicit–explicit confidence gap from prior work. Binary ESMA collapses the graded readout out-of-domain and damages in-domain calibration at smaller scales.

1 Introduction

Metacognition, the ability to monitor one's own cognitive processes, is foundational to human intelligence (Flavell, 1979; Fleming, 2024). Humans do not merely distinguish knowing from not knowing; they experience graded confidence states, "I'm certain," "I think so," "no idea," that reliably predict accuracy (Hart, 1965; Koriat, 1993; Nelson, 1984; Fleming and Lau, 2014).

Recent LLM metacognition work has produced mixed results: some find poor discrimination (Yin et al., 2023; Chhikara, 2025), others demonstrate access to internal states (Betley et al., 2025; Lindsey, 2026). Steyvers and Peters (2025) highlight a gap between implicit (token-probability) and explicit (verbalized) confidence. Proposed remedies include fine-tuning (Steyvers et al., 2025), RLHF reward calibration (Leng et al., 2024), and scale design (Dai, 2026). Park et al. (2026) introduced Evolution Strategy for Metacognitive Alignment (ESMA) with a binary Yes/No framework.

Binary probes diverge from human metacognition measurement (Fleming and Lau, 2014) and admit a structural failure mode that graded metrics expose. Across eleven models from five families on MMLU, the binary probe collapses to near-constant "no" on 7 of 11 models ($p_{\text{yes}} < 0.05$), three formally degenerate; on GSM8K, several cells exhibit near-constant outputs ($p_{\text{yes}} = 0.96$ on Qwen3-4B, 0.04 on Gemma-2-2B) even where binary's calibration score appears competitive.

We propose a graded metacognitive framework extending ESMA to optimize calibrated confidence across multiple levels.[1] We evaluate binary and graded (A–D) probes across Qwen2.5-Instruct (1.5B, 3B, 7B) on TriviaQA, MMLU, and GSM8K, validate across eight cross-family models, and assess transfer. Our contributions are:

- **A graded metacognitive framework extending ESMA.** We introduce a measurement framework importing bias-resistant metrics from human research (γ, Type-2 AUROC, grade separation) paired with a structural probe-health diagnostic (p_{yes}), and extend Park et al. (2026) with a continuous graded reward $r(c, g) = 1 \pm g/g_{\max}$ optimising calibrated confidence across a four-level (A–D)

[1] Code: https://anonymous.4open.science/r/graded-metacognition-F1F6/README.md

scale. Applied across eleven models, the framework reveals binary probes are structurally broken on OOD factual reasoning (robust across three independent calibration metrics), while Graded ESMA improves γ from 0.60 to 0.72 ± 0.02 on a 3B model (three seeds) with accuracy gains (35.4% $\rightarrow$ 50.7%) and rescues a 1.5B model from anti-calibration ($\gamma = -0.07 \rightarrow 0.67$). Binary ESMA, by contrast, collapses graded-probe metrics on GSM8K and produces a 0.75-point in-domain γ gap at 1.5B.

- **Verbalised confidence aligns with internal logits** after graded training. On TriviaQA, $r = 0.99$ at 3B; on GSM8K the gap widens (graded $r \geq 0.87$ vs. binary $r = 0.00$ at 3B). This addresses the implicit–explicit confidence gap raised by Steyvers and Peters (2025).
- **Graded metacognition is a scale-dependent emergent capability**, with meaningful calibration ($\gamma > 0$) emerging only at $\sim$3B across five families. Qwen3-4B matches Qwen2.5-3B at $\gamma = 0.57$ despite different pretraining, isolating scale from training data. Gains transfer partially to GSM8K but not to MMLU, where A–D probes lose discriminative power when the confidence scale shares labels with the answer space.

2 Related Work

Metacognition in LLMs. Prior work targets hallucination (Ji et al., 2023), refusal (Zhang et al., 2024; Wen et al., 2025), and calibration (Lin et al., 2022; Kadavath et al., 2022). Park et al. (2026) introduced ESMA, aligning binary self-reports with knowledge via d'_{type2}. Cacioli (2026) shows via signal detection theory that binary metrics conflate Type 1 (task) and Type 2 (metacognitive) performance, so high accuracy or response bias can mimic calibration. Steyvers and Peters (2025) document a gap between implicit (logit) and explicit (verbalized) confidence, which our verbal–logit alignment analysis directly addresses. All existing remedies remain binary, while human metacognition recognizes a graded spectrum (Hart, 1965; Koriat, 1993).

Eliciting and measuring LLM confidence. LLM confidence is elicited via verbalized self-reports (Lin et al., 2022; Tian et al., 2023), logit extraction (Kadavath et al., 2022; Jiang et al., 2021), or output consistency (Kuhn et al., 2023; Manakul et al., 2023), and evaluated with ECE (Guo et al., 2017) or binary AUROC. These metrics collapse confidence to a scalar and miss probe collapse to a single response, which our p_{yes} diagnostic surfaces. Dai (2026) shows confidence-scale design itself shapes measured metacognition, echoing our A–D label-collision finding on MMLU.

Measuring human metacognition. Human research distinguishes *sensitivity* (confidence tracks accuracy) from *bias* (overall level). Rahnev (2021) reviews the bias-resistant toolkit (γ, Type-2 AUROC, meta-d'); Fleming (2024) synthesizes its computational basis. We adopt γ (ignores ties, bias-insensitive), Type-2 AUROC (threshold-invariant), and grade separation (distributional decisiveness), all of which separate calibration from response-bias artefacts.

Evolution strategies for LLM fine-tuning. Metacognitive alignment is non-differentiable because direct and meta queries run in independent contexts. ES sidesteps this via forward-pass evaluation of perturbed candidates. Qiu et al. (2025) establish the natural ES recipe we adopt; Sarkar et al. (2025) extend it to hyperscale. Park et al. (2026) apply ES with a binary reward; we extend to a graded reward aligning calibration across the full scale.

3 Methodology

We extend ESMA (Park et al., 2026) by replacing binary metacognitive probes and rewards with graded alternatives. The evolution strategy mechanism is unchanged; our contribution is in how metacognition is elicited, measured, and optimized. Figure 1 illustrates the training pipeline.

3.1 Metacognitive Probes

Following Park et al. (2026), factual questions and metacognitive probes are administered in independent contexts. We compare two probe types: (1) the original binary Yes/No probe, and (2) a graded four-level scale (A–D), parsed as $g \in \{0, 1, 2, 3\}$ where A = 3 (certain) and D = 0 (don't know). The graded probe presents the question before the rating scale; scale-first ordering produced severe center-bias (>80% at B/C) in pilot runs.

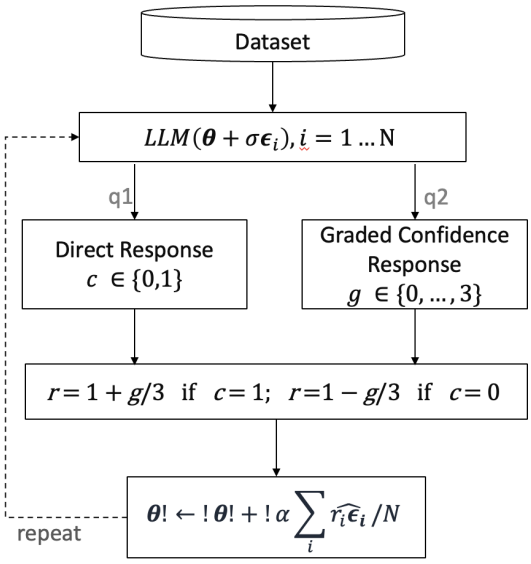

Figure 1: Graded ESMA training pipeline. For each of N perturbations ϵ_i, the model generates a direct answer and a graded confidence response in *independent contexts*. A continuous reward rewards correctness–confidence alignment, and parameters are updated via natural evolution strategies. The binary variant shares the same structure with a different probe and reward.

3.2 Metrics

We evaluate both probes with metrics chosen to separate metacognitive *sensitivity* from *response bias* (Fleming and Lau, 2014; Fleming, 2024).

Binary probe. We report d'_{type2} (Maniscalco and Lau, 2012; Park et al., 2026) alongside task accuracy, and the **positive-response rate** p_{yes}, a structural probe-health diagnostic introduced here: a calibrated binary probe should answer "yes" near the model's accuracy rate, so $p_{\text{yes}} < 0.05$ or >0.95 flags response collapse that d'_{type2} alone can hide.

Graded probe. We report three bias-resistant metrics standard in human metacognition research (Fleming and Lau, 2014; Fleming, 2024; Rahnev, 2021): **Gamma correlation** (γ; Goodman–Kruskal rank correlation, range $[-1, +1]$, ignores ties), **Type-2 AUROC** (threshold-invariant, chance 0.5), and **grade separation** ($\bar{g}_{\text{correct}} - \bar{g}_{\text{incorrect}}$). For cross-family tables, γ and separation are rescaled to $[0, 1]$ via $\gamma^* = (\gamma + 1)/2$ so 0.5 is chance.

Verbal–internal correspondence. To test whether verbalised confidence reflects internal state or is post-hoc rationalisation (Steyvers and Peters, 2025), we report the Pearson correlation r between verbalised grades and logit-based confidence from the first meta-response token, and a logit-based Type-2 AUROC substituting logits for verbalised grades.

3.3 Graded Reward Function

ESMA's original reward assigns discrete values $\{0, 1, 2\}$ for matched correctness–confidence pairs. We replace it with a continuous reward incentivising proportional calibration:

$$r(c, g) = \begin{cases} 1 + g/g_{\max} & \text{if } c = 1 \text{ (correct)} \\ 1 - g/g_{\max} & \text{if } c = 0 \text{ (incorrect)} \end{cases} \tag{1}$$

where $c \in \{0, 1\}$ is correctness, $g \in \{0, \ldots, g_{\max}\}$ is the confidence grade, and $g_{\max} = 3$. A confidently wrong answer is penalised more than a tentatively wrong one, and a confidently correct answer rewarded more than a tentative correct one, the structural difference from the binary reward, which treats all confident responses as equivalent regardless of correctness magnitude.

3.4 Experimental Setup

We evaluate Qwen2.5-Instruct at 1.5B, 3B, and 7B. Training uses TriviaQA ($n = 2{,}000$) with natural evolution strategies (Qiu et al., 2025): population size 32, $\sigma = 10^{-3}$, $\alpha = 5 \times 10^{-4}$, 750 iterations. Transfer evaluation uses GSM8K and MMLU, $n = 2{,}000$ each (GSM8K capped at 1,319). The 3B graded ESMA configuration uses three independent seeds; others use seed 42. We train both 3B and 1.5B models with both objectives and evaluate each under both probes, isolating training-objective effects from readout effects.

All training and evaluation experiments ran on 4×A100 GPUs. Each training run takes approximately 13 hours; each transfer evaluation takes 1–2 hours. Total compute across all experiments was approximately 200 GPU-hours. Implementation uses HuggingFace transformers 4.49.0 and accelerate for model inference and training. Goodman–Kruskal γ is computed via the standard concordant-minus-discordant pair formula; Type-2 AUROC and grade separation are computed in pure Python.

4 Experiments and Results

We evaluate three model scales (1.5B, 3B, 7B) of Qwen2.5-Instruct across three datasets and two

probe types, validate the patterns on eight cross-family models, and train and evaluate four ESMA configurations. We focus on Qwen2.5 for training.

4.1 Baseline Graded Metacognition

We evaluated Qwen2.5-Instruct on TriviaQA, MMLU, and GSM8K ($n = 2{,}000$ each) under binary and graded (A–D) probes administered in independent contexts. Three findings stand out.

The binary probe is structurally fragile. Across the Qwen2.5 ladder, the binary probe's positive-response rate p_{yes} exhibits severe imbalance on MMLU: $p_{\text{yes}} = 0.000$–0.003 for all three scales (Table 1). d'_{type2} remains nominally positive on these cells, but the probe is structurally collapsed: it answers "no" to nearly every question regardless of the model's actual knowledge. Binary metrics computed on a constant-output probe report sensitivity that the probe is not producing.

Graded calibration emerges at ∼3B parameters. On TriviaQA, binary d'_{type2} is positive even at 1.5B ($d' = 0.43$), but graded γ is negative (-0.08): the smallest model is more confident when wrong than when right. Meaningful graded calibration emerges only at 3B ($\gamma = 0.60$) and stabilises at 7B ($\gamma = 0.61$). On MMLU, graded γ is moderately negative at 3B (-0.27) and mildly negative at 7B (-0.10), revealing scale-dependent anti-calibration that the binary probe's collapsed output hides.

GSM8K is at floor. All three scales achieve ∼5–8% accuracy under our keyword-only protocol, producing degenerate or weakly-calibrated γ across conditions. GSM8K therefore serves as an out-of-domain stress test rather than a substantive baseline.

4.2 Cross-Family Validation

We evaluated eight additional models across four families (Qwen3 1.7B/4B/8B, Llama-3.2 1B/3B, Mistral-7B, Gemma-2 2B/9B) on TriviaQA, MMLU, and GSM8K. Table 1 reports accuracy, γ^* for both probes, and $p_{\text{yes}}^{\text{Bin}}$ across all eleven models.

Binary collapse on OOD factual reasoning is family-independent. On MMLU, 7 of 11 models drive the binary probe to $p_{\text{yes}} < 0.05$ (3 formally degenerate); on the 8 non-degenerate cells the graded probe wins γ^* 6–2. The pattern crosses architecture, scale, and generation across all five families.

p_{yes} exposes pathology that γ^* alone hides. On GSM8K, several cells show competitive binary γ^* (0.78 on Qwen3-4B, 0.60 on Gemma-2-2B) but p_{yes} near saturation (0.96 and 0.04 respectively). These probes are structurally constant; γ^* scores them well only because correctness is itself near-uniform at floor accuracy.

Scale threshold replicates. Meaningful graded calibration ($\gamma > 0$ on TriviaQA) emerges only at ∼3B across all five families. Qwen3-4B matches Qwen2.5-3B at $\gamma = 0.57$ despite different pretraining, isolating scale from training data.

Robustness across metrics. We re-evaluated the Qwen2.5 family under γ^*, AUC*, and Sep.* (Table 2); the three metrics agree on the winner in nearly every non-degenerate comparison. On MMLU, all three binary metrics collapse to degeneracy on every Qwen2.5 scale while graded remains informative.

4.3 ESMA Training Results

We trained graded and binary ESMA on the 3B and 1.5B models. All runs used $\sigma = 10^{-3}$, $\alpha = 5 \times 10^{-4}$, population size 32, 750 iterations on TriviaQA. The 3B graded configuration used three independent seeds; other configurations used seed 42. Each trained model is evaluated under both probe types, isolating training-objective from readout effects.

Graded ESMA achieves calibrated metacognition. On TriviaQA, the 3B graded model improved γ from 0.60 to 0.72 ± 0.02 across three seeds, with accuracy rising from 35.4% to 50.7% and grade separation from 0.62 to 1.22. The 1.5B graded model was rescued from anti-calibration ($\gamma = -0.07 \rightarrow 0.67$) with unchanged accuracy (41.6% → 41.8%), matching the untrained 7B baseline.

Binary ESMA degrades the graded readout, with scale-dependent severity. On TriviaQA, 3B binary ESMA reaches the highest binary d'_{type2} (1.00) but degrades all graded-probe metrics relative to graded ESMA at the same scale. At 1.5B, binary training produces $\gamma = -0.08$ on TriviaQA, a 0.75-point in-domain gap versus graded ESMA's $\gamma = 0.67$ (Figure 2). Out-of-domain, the asymmetry sharpens: on GSM8K the 3B binary-trained model's graded probe collapses to single-bin degeneracy, and on MMLU 3B binary ESMA inverts its own binary-probe calibration ($d'_{\text{type2}} = 0.16$ vs.

Family	Model	TriviaQA *(factual)*		MMLU *(factual, OOD)*		GSM8K *(reasoning)*	
		Acc. (B/G)	γ^*_{Bin} / γ^*_{Grad}	Acc. (B/G)	γ^*_{Bin} / γ^*_{Grad}	Acc. (B/G)	γ^*_{Bin} / γ^*_{Grad}
Qwen2.5	1.5B	41.6 / **41.8**	**0.67** / 0.46	55.5 / **56.1**	0.35 / **0.53**	5.5 / **5.8**	**0.58** / 0.43
	3B	35.4 / **36.0**	0.79 / **0.80**	61.8 / **61.9**	1.00* / **0.37**	**7.6** / 7.3	**0.58** / 1.00*
	7B	**48.9** / 47.9	0.79 / **0.81**	**68.5** / 68.2	0.50* / **0.45**	**5.0** / 4.9	0.50* / **0.59**
Qwen3	1.7B	**30.9** / 30.3	**0.74** / 0.64	**53.1** / 44.0	0.44 / **0.48**	1.1 / 1.1	**0.39** / 0.32
	4B	**44.1** / 44.0	0.77 / **0.79**	**66.7** / 50.7	0.29 / **0.50**	10.5 / 10.5	**0.78** / 0.72
	8B	51.3 / **51.6**	0.76 / **0.78**	**71.5** / 54.9	0.23 / **0.47**	**14.9** / 14.3	0.57 / **0.63**
Llama-3.2	1B	47.9 / **49.2**	**0.62** / 0.47	**28.2** / 26.6	0.47 / **0.54**	5.3 / 5.3	0.64 / **0.69**
	3B	59.1 / **60.8**	0.65 / **0.67**	**47.8** / 34.6	**0.53** / 0.51	**11.7** / 9.2	**0.68** / 0.51
Mistral	7B-v0.3	**68.0** / 67.2	**0.82** / 0.80	**51.9** / 46.2	**0.60** / 0.51	9.9 / 9.9	**0.67** / 0.61
Gemma-2	2B	**48.7** / 48.5	0.68 / **0.73**	**50.7** / 48.3	1.00* / **0.52**	7.1 / 7.1	**0.60** / 0.57
	9B	69.7 / **69.8**	0.81 / **0.83**	**66.8** / 56.5	0.29 / **0.52**	1.5 / 1.5	**0.57** / 0.15

Table 1: **Cross-family metacognition: Binary vs. Graded (A–D) confidence probes** on TriviaQA, MMLU, and GSM8K ($n = 2{,}000$ each; GSM8K capped at 1,319). Eleven models from five families. For each dataset we report probe-specific accuracy as Acc. (Binary / Graded) and the transformed calibration score $\gamma^* = (\gamma + 1)/2$ for each probe (1.0: perfect confidence–correctness alignment; 0.5: chance; <0.5: anti-calibration). **Bold** marks the stronger probe per comparison (degenerate values excluded). Amber entries marked with * flag *degenerate* probe behaviour from single-bin response collapse: although $\gamma^* = 1.00$ and $\gamma^* = 0.50$ appear strong in isolation, they are numerical artefacts of the probe answering identically to every item. Goodman–Kruskal γ is defined over *concordant* versus *discordant* confidence–correctness pairs; when every confidence rating is identical, every pair is a tie, and γ reduces to whichever extreme the formula's edge-case convention returns (here, $\gamma = 1$ from a vacuous numerator ≥ 0, rescaled to $\gamma^* = 1.00$, or $\gamma = 0$ rescaled to $\gamma^* = 0.50$). In both cases the probe has not discriminated correct from incorrect responses, it has produced no information at all, so the value is mathematically defined but uninformative. The central pattern is that graded probes remain substantially more stable out-of-domain. On MMLU, several binary probes collapse entirely while graded probes remain informative across nearly all models. On TriviaQA, graded probes are competitive with binary probes in-domain, indicating that improved robustness does not require sacrificing factual calibration performance.

0.89 for graded ESMA), demonstrating that binary training overfits to TriviaQA at the cost of OOD metacognitive sensitivity.

Verbalized and internal confidence align after graded training. Graded ESMA produces near-perfect alignment between verbalised A–D grades and logit-based confidence on TriviaQA ($r = 0.99$ at 3B, 0.96 at 1.5B; Table 3), preserved on GSM8K ($r \geq 0.87$). Binary ESMA's 3B configuration collapses entirely on GSM8K ($r = 0.00$, the same configuration whose graded readout collapses to single-bin response), while at 1.5B alignment is weaker but preserved ($r = 0.72$ on TriviaQA, 0.74 on GSM8K). Across all six non-MMLU comparisons, graded ESMA produces stronger verbal–logit correspondence than scale-matched binary ESMA; on MMLU, $r \approx 0$ for all configurations because logit-based grading collapses regardless of training. This addresses the implicit–explicit confidence gap raised by Steyvers and Peters (2025): graded ESMA shapes verbal output to track internal token probabilities, while binary ESMA severs this correspondence under distribution shift.

4.4 Transfer Evaluation

Figure 2 summarises post-training graded γ across the three evaluation domains for all four trained configurations.

MMLU resists training across both objectives. Graded ESMA produces near-chance $\gamma = 0.09$ at 1.5B and mild anti-calibration $\gamma = -0.26$ at 3B; binary ESMA hovers similarly near chance ($\gamma = 0.10$ at 1.5B, -0.22 at 3B). Combined with the cross-family pattern in Table 1, this confirms that A–D probes lose discriminative power on MCQ tasks regardless of training scheme. Across the eight non-Qwen2.5 baseline models, A–D γ ranges from -0.06 to $+0.07$ on MMLU, suggesting the constraint is structural: confidence scales should not share labels with the answer space.

Model	Binary			Graded (A–D)		
	γ^*	AUC*	Sep.*	γ^*	AUC*	Sep.*
TriviaQA						
Qwen2.5-1.5B	**0.67**	**0.57**	**0.57**	0.47	0.52	0.54
Qwen2.5-3B	0.79	0.57	0.57	**0.80**	**0.60**	**0.81**
Qwen2.5-7B	0.79	0.65	0.65	**0.82**	**0.69**	**0.78**
MMLU (out-of-domain factual)						
Qwen2.5-1.5B	0.35	$0.50^{\dagger}$	$0.50^{\dagger}$	**0.53**	**0.52**	**0.54**
Qwen2.5-3B	$1.00^{\star}$	$0.50^{\dagger}$	$0.50^{\dagger}$	**0.66**	**0.60**	**0.70**
Qwen2.5-7B	$0.50^{\dagger}$	$0.50^{\dagger}$	$0.50^{\dagger}$	**0.58**	**0.52**	**0.54**
GSM8K (out-of-domain reasoning)						
Qwen2.5-1.5B	**0.58**	**0.53**	**0.53**	0.43	$0.50^{\dagger}$	$0.49^{\dagger}$
Qwen2.5-3B	**0.58**	**0.51**	**0.52**	$1.00^{\star}$	$0.50^{\dagger}$	$0.50^{\dagger}$
Qwen2.5-7B	$0.50^{\dagger}$	$0.50^{\dagger}$	$0.50^{\dagger}$	**0.59**	$0.50^{\dagger}$	$0.51^{\dagger}$

Table 2: **Calibration-metric robustness check: the binary-vs-graded comparison holds under three independent calibration metrics.** Qwen2.5 family on TriviaQA, MMLU, and GSM8K (n=2,000 each). We re-evaluate the comparison from Table 1 under three calibration metrics, all rescaled to $[0, 1]$ where larger is better: $\gamma^*=(\gamma+1)/2$ (Goodman–Kruskal rank correlation), AUC*$=\max(\text{AUC}, 1-\text{AUC})$ (type-2 discrimination relative to chance), and Sep.*$=(\text{Sep}+1)/2$ (rescaled confidence separation between correct and incorrect predictions). A value of 0.5 indicates chance-level behaviour, 1.0 indicates perfect confidence–correctness alignment. **Bold** marks the stronger probe within each metric (degenerate values excluded from comparison). Amber entries flag two related failure modes that arise when a probe assigns the same response to every item: * marks formal collapse of γ^*, where every cross-item pair is a tie and the Goodman–Kruskal formula $\gamma = (C - D)/(C + D)$ reduces to $0/0$ that the standard convention returns as $\gamma = 1$ (rescaled to $\gamma^* = 1.00$); † marks chance-level AUC* or Sep.* from the same collapse mechanism, where a constant probe cannot separate correct from incorrect responses and both metrics floor at their chance value of 0.5. Although γ^*=1.00 and the various 0.50 values look strong or neutral in isolation, in all amber cells the probe has produced no information: it has answered identically to every question, and the metric values are mathematical artefacts of that constant output rather than evidence of calibration. AUC* and Sep.* are particularly diagnostic here because they cannot be inflated to 1.0 by collapse the way γ^* can, exposing the failure that γ^* alone would mask. The three metrics agree on the winner in nearly every non-degenerate comparison, confirming that the cross-family result is not an artefact of choosing γ^* as the headline calibration metric. On MMLU, all three binary metrics collapse to degeneracy on every Qwen2.5 model while graded remains informative across the board.

Reasoning transfer is partial and training-dependent. On GSM8K, graded ESMA achieves $\gamma = 0.36$ at 3B and $\gamma = 0.17$ at 1.5B; binary ESMA reaches $\gamma = 0.09$ at 1.5B but collapses to single-bin degeneracy at 3B. Both graded improvements occur at near-floor accuracy ($\sim$5%), so they should be interpreted as evidence that graded training preserves a calibration signal even in out-of-domain reasoning rather than as strong transfer claims. The binary-3B collapse, by contrast, indicates that binary training's failure mode emerges with training capacity rather than affecting all scales uniformly.

5 Discussion

Our results support five claims: binary probes admit a structural failure mode that bias-resistant graded metrics expose, graded calibration emerges at a consistent scale threshold across model families, graded ESMA produces genuine improvement on the training domain, graded ESMA aligns verbalised confidence with internal token probabilities, and A–D probes are unreliable on tasks whose answer space shares the same label set.

Binary probes are structurally broken on out-of-domain factual reasoning. The p_{yes} diagnostic surfaces a failure mode that calibration scores cannot: on MMLU, 7 of 11 models across five families drive the binary probe below $p_{\text{yes}} < 0.05$, with three formally degenerate. The pathology extends to GSM8K, where competitive binary γ^* values mask near-constant outputs ($p_{\text{yes}} = 0.96$ on Qwen3-4B, 0.04 on Gemma-2-2B). The conclusion holds under three independent calibration met-

Scale	Dataset	Binary ESMA		Graded ESMA	
		Verbal–Logit r	Logit AUROC$_2$	Verbal–Logit r	Logit AUROC$_2$
Qwen2.5-1.5B	TriviaQA	0.72	0.46	**0.96**	**0.71**
	MMLU	0.00†	0.48	0.00†	**0.51**
	GSM8K	0.74	0.55	**0.87**	**0.64**
Qwen2.5-3B	TriviaQA	0.91	0.56	**0.99**	**0.72**
	MMLU	0.00†	0.39	0.00†	0.38
	GSM8K	0.00†	0.50*	**0.91**	0.49

Table 3: **Graded training preserves verbal–logit correspondence ($r \geq 0.87$ out-of-domain); binary training collapses it ($r = 0.00$ on GSM8K at 3B).** Across the six non-MMLU comparisons, graded ESMA produces stronger verbal–logit alignment than scale-matched binary ESMA on every cell: $r = 0.96$ vs. 0.72 at 1.5B and $r = 0.99$ vs. 0.91 at 3B on TriviaQA; on GSM8K, graded preserves $r \geq 0.87$ at both scales while 3B binary collapses entirely. The Logit AUROC$_2$ column confirms the effect is not surface mimicry: graded training also strengthens logit-based discrimination (e.g., 0.72 vs. 0.56 at 3B TriviaQA), so verbal grades and internal probabilities improve together. This directly tests whether verbalised metacognition reflects internal state or post-hoc rationalisation (Steyvers and Peters, 2025): high r means the verbal grade and the probability distribution over candidate confidence tokens carry the same ordering. We report the Pearson correlation r between verbalised A–D grades and logit-based confidence from the first meta-response token, and the Type-2 AUROC computed using logit-based confidence (Logit AUROC$_2$); $n = 2{,}000$ items (1,319 for GSM8K). **Bold** marks where graded ESMA beats its scale-matched binary counterpart. Amber flags two related failure modes: † marks $r = 0$ from logit-grade collapse to a single token (the correlation is undefined and floored to zero); * marks AUROC at chance value (0.50) from the same single-bin collapse mechanism. Both arise when the A–D label collision with an answer space or training-induced response collapse forces the logit distribution onto one token; MMLU is therefore uninformative for all four configurations, and is retained to document the scope of the answer-collision failure mode rather than as an evaluation of training.

rics (γ^*, AUC*, Sep.*); binary's collapse is visible in all of them. Bias inflation reinforces the critique: the 7B baseline reaches $d'_{\text{type2}} = 0.83$ under binary probing but only $\gamma = 0.61$ under graded probing, and Binary ESMA pushes d'_{type2} to 1.00 on TriviaQA while inverting to -0.67 on MMLU. d'_{type2} alone is insufficient; bias-resistant graded metrics paired with structural diagnostics like p_{yes} are essential.

Graded metacognition is a scale-dependent emergent capability. Sub-2B models are anti-calibrated or near-zero (Qwen2.5-1.5B: -0.08; Llama-3.2-1B: -0.06); $\sim$3B models show meaningful calibration (Qwen2.5-3B: 0.57; Llama-3.2-3B: 0.33), replicating across five families. Qwen3-4B matches Qwen2.5-3B at $\gamma = 0.57$ despite different pretraining, isolating scale rather than training data as the controlling variable.

Graded ESMA produces genuine improvement; binary ESMA can damage calibration. γ improves from 0.60 to 0.72 ± 0.02 on the 3B model across three seeds, with accuracy rising from 35.4% to 50.7%. The 1.5B model is rescued from anti-calibration ($\gamma = -0.07 \rightarrow 0.67$) without changing accuracy, matching the untrained 7B baseline. The training cross-design isolates the objective from the readout (Figure 2): graded ESMA produces broad gains across both probes' metrics, while 3B binary ESMA collapses the graded readout to single-bin degeneracy on GSM8K and inverts MMLU calibration. The 1.5B binary ESMA condition sharpens the asymmetry: it produces $\gamma = -0.08$ on TriviaQA, a 0.75-point in-domain gap relative to graded ESMA, while preserving a usable signal on GSM8K ($\gamma = 0.09$). This pattern suggests binary ESMA's catastrophic failure emerges with training capacity, while graded ESMA's preservation across scales reflects training-objective robustness.

Graded training aligns verbalised confidence with internal state. On TriviaQA, graded ESMA achieves near-perfect alignment between verbalised A–D grades and logit-based confidence ($r = 0.99$ at 3B, 0.96 at 1.5B), preserved out-of-domain on GSM8K ($r \geq 0.87$). Binary ESMA at 3B reaches $r = 0.91$ on TriviaQA but collapses to $r = 0.00$ on GSM8K, the same configuration whose graded readout collapses. This addresses the implicit–explicit confidence gap raised by Steyvers and Peters (2025): graded training shapes verbal output to track internal probability estimates, while binary training severs this correspondence under dis-

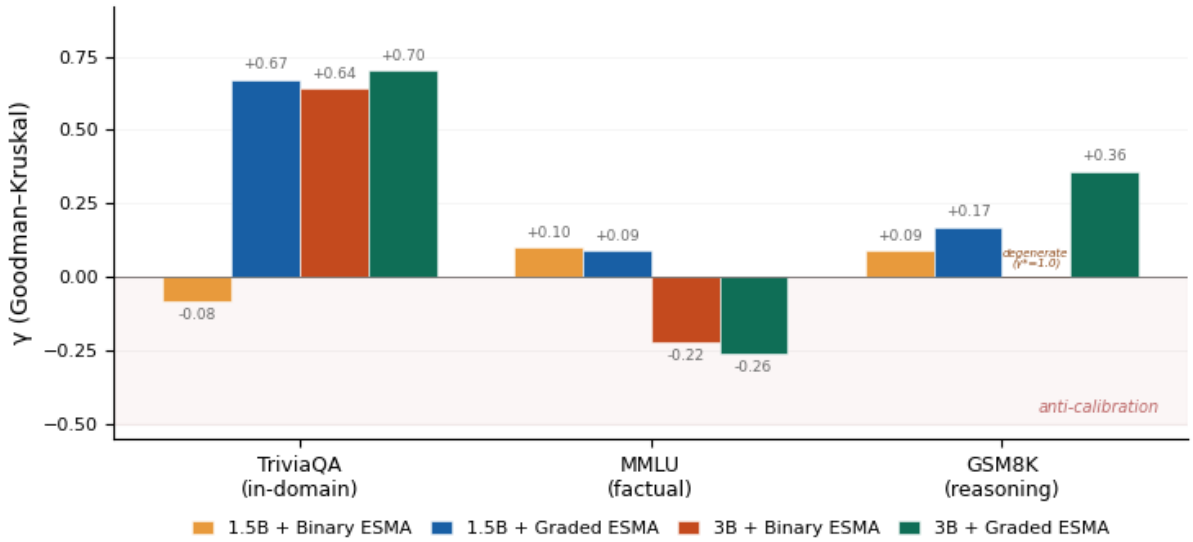

Figure 2: **Post-training calibration of ESMA models across evaluation domains: graded training generalises while binary training collapses or anti-calibrates out-of-domain.** Goodman–Kruskal γ for ESMA-trained Qwen2.5-1.5B and Qwen2.5-3B on TriviaQA (training domain), MMLU (out-of-domain factual), and GSM8K (out-of-domain reasoning). Within each dataset, bars are grouped by scale (1.5B then 3B), with binary ESMA shown immediately before its scale-matched graded counterpart for direct comparison. γ=1 is perfect confidence–correctness alignment, 0 is chance, and γ<0 (shaded red) is anti-calibration. The hatched bar marks degenerate probe collapse where γ is mathematically defined but uninformative (single-bin response). On TriviaQA, binary ESMA produces $\gamma = -0.08$ at 1.5B and 0.64 at 3B, while graded ESMA reaches 0.67 and 0.70 at the same scales — a 0.75-point in-domain gap at the smaller scale indicating that binary training can actively damage calibration when training capacity is limited. On GSM8K, binary ESMA produces a usable $\gamma = 0.09$ at 1.5B but collapses the graded readout entirely at 3B, while graded ESMA preserves positive γ at both scales (0.17 and 0.36), suggesting binary ESMA's catastrophic failure mode emerges with training capacity rather than affecting all scales uniformly. MMLU resists all four configurations: binary ESMA hovers near chance (0.10 at 1.5B, -0.22 at 3B), and graded ESMA is similarly near chance at 1.5B (0.09) and mildly anti-calibrated at 3B (-0.26), reflecting the difficulty of OOD factual transfer at this scale rather than a training-objective difference. Across the six non-degenerate comparisons, graded ESMA equals or exceeds binary ESMA in five (tied on MMLU at 1.5B), and never produces the catastrophic single-bin collapse seen in 3B binary ESMA on GSM8K.

tribution shift. The alignment is domain-specific ($r \approx 0$ on MMLU across all configurations), but the in-domain and OOD-reasoning correspondence is strong evidence that graded ESMA changes internal metacognitive representations, not merely surface labels.

MCQ tasks expose a limitation of A–D probes. On MMLU, A–D γ collapses to near-zero across all eight non-Qwen2.5 cross-family models (range -0.06 to $+0.07$). When the confidence scale shares labels with the answer space, models lose discriminative power on the confidence dimension, not metacognitive ability per se: the same models that calibrate normally on TriviaQA produce no signal on MMLU. Whether a non-colliding scale (e.g., a numeric 1–5 rating) recovers MCQ metacognition is left to future work.

Grade distribution collapse. The trained 3B model's grades collapse to the endpoints (45% D, 54% A) on TriviaQA, yet still achieve $\gamma = 0.72$ because it assigns extreme grades to the right items. The continuous reward incentivises confident extremes rather than fine-grained intermediate confidence; an entropy bonus $\beta \cdot H(g)$ is a natural fix we leave to future work.

Limitations

Training was limited to Qwen2.5-Instruct at 1.5B and 3B, while cross-family experiments cover baselines only; generalisation of graded ESMA to other families remains untested. Task transfer should also be interpreted cautiously: under the keyword-only protocol, GSM8K accuracy is only ∼5–8%, so the 3B model's positive γ (+0.36) is not strong evidence of robust reasoning transfer. The MCQ

setting remains unresolved. On MMLU, non-Qwen2.5 models collapse toward near-zero γ rather than reproducing Qwen2.5-3B's negative pattern, and we do not test alternative non-colliding scales such as numeric 1–5 labels. Verbal–logit alignment is likewise task-dependent: strong TriviaQA alignment ($r = 0.99$ at 3B) does not transfer to MMLU ($r \approx 0$), likely due to the A–D answer-label collision discussed in the main paper. Finally, trained graded models show uneven scale use, with endpoint grades dominating and intermediate grades rarely selected. Cross-family baselines were also run with a single seed, so the observed MMLU binary-collapse pattern and apparent $\sim$3B scale transition should be treated as observations within this experimental setting rather than definitive scaling claims.

Acknowledgments

The authors used generative AI assistants during paper preparation for prose editing, LaTeX table formatting, and code debugging. The tools were not used to generate new research ideas, experimental designs, analyses, or result interpretations. All scientific content, experiments, conclusions, and final manuscript decisions are the authors' own.

Ethical Considerations

All models and datasets are publicly released under research-permitting licenses: Qwen2.5/Qwen3 (Apache 2.0), Llama-3.2 (Llama 3.2 Community License), Mistral-7B-Instruct (Apache 2.0), Gemma-2 (Gemma Terms of Use), TriviaQA (Apache 2.0), MMLU (MIT), and GSM8K (MIT). Our use is consistent with each license and confined to research evaluation; we do not redistribute weights or dataset derivatives. Released code is MIT-licensed. The work involves no human subjects, PII, or sensitive content.

Potential risks. Our contribution is analytical and uses publicly released models and standard benchmarks, so it does not shift misuse potential beyond the base models. We flag three caveats for downstream users: (i) rank-calibration (γ) is not a guarantee of factual correctness; (ii) the $\sim$3B emergence threshold is an observation in our setting, not a certification of smaller models for deployment; and (iii) calibration transfers only partially out-of-domain (notably collapsing on MMLU), so scores should be re-measured in the target domain.

References

Jan Betley, Xuchan Bao, Martín Soto, Anna Sztyber-Betley, James Chua, and Owain Evans. 2025. Tell me about yourself: Llms are aware of their learned behaviors. *arXiv preprint arXiv:2501.11120*.

Jon-Paul Cacioli. 2026. Do llms know what they know? measuring metacognitive efficiency with signal detection theory. *arXiv preprint arXiv:2603.25112*.

Prateek Chhikara. 2025. Mind the confidence gap: Overconfidence, calibration, and distractor effects in large language models. *arXiv preprint arXiv:2502.11028*.

Yuyang Dai. 2026. Rescaling confidence: What scale design reveals about llm metacognition. *arXiv preprint arXiv:2603.09309*.

John H Flavell. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, 34(10):906.

Stephen M Fleming. 2024. Metacognition and confidence: A review and synthesis. *Annual Review of Psychology*, 75(1):241–268.

Stephen M Fleming and Hakwan C Lau. 2014. How to measure metacognition. *Frontiers in human neuroscience*, 8:82285.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Julian T Hart. 1965. Memory and the feeling-of-knowing experience. *Journal of educational psychology*, 56(4):208.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Asher Koriat. 1993. How do we know that we know? the accessibility model of the feeling of knowing. *Psychological review*, 100(4):609.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

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

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

Jack Lindsey. 2026. Emergent introspective awareness in large language models. *arXiv preprint arXiv:2601.01828*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9004–9017.

Brian Maniscalco and Hakwan Lau. 2012. A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and cognition*, 21(1):422–430.

Thomas O Nelson. 1984. A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological bulletin*, 95(1):109.

Sangjun Park, Elliot Meyerson, Xin Qiu, and Risto Miikkulainen. 2026. Fine-tuning language models to know what they know. *arXiv preprint arXiv:2602.02605*.

Xin Qiu, Yulu Gan, Conor F Hayes, Qiyao Liang, Yinggan Xu, Roberto Dailey, Elliot Meyerson, Babak Hodjat, and Risto Miikkulainen. 2025. Evolution strategies at scale: Llm fine-tuning beyond reinforcement learning. *arXiv preprint arXiv:2509.24372*.

Dobromir Rahnev. 2021. Visual metacognition: Measures, models, and neural correlates. *American psychologist*, 76(9):1445.

Bidipta Sarkar, Mattie Fellows, Juan Agustin Duque, Alistair Letcher, Antonio León Villares, Anya Sims, Clarisse Wibault, Dmitry Samsonov, Dylan Cope, Jarek Liesen, and 1 others. 2025. Evolution strategies at the hyperscale. *arXiv preprint arXiv:2511.16652*.

Mark Steyvers, Catarina Belem, and Padhraic Smyth. 2025. Improving metacognition and uncertainty communication in language models. *arXiv preprint arXiv:2510.05126*.

Mark Steyvers and Megan AK Peters. 2025. Metacognition and uncertainty communication in humans and large language models. *Current Directions in Psychological Science*, page 09637214251391158.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2025. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuan-Jing Huang. 2023. Do large language models know what they don't know? In *Findings of the association for Computational Linguistics: ACL 2023*, pages 8653–8665.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024. R-tuning: Instructing large language models to say 'i don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7113–7139.