FINAL Bench: Measuring Functional Metacognitive Reasoning in Large Language Models

Taebong Kim[1,*], Minsik Kim[1], Sunyoung Choi[2], and Jaewon Jang[1]

[1] *VIDRAFT, Seoul, South Korea*
[2] *Ginigen AI, Seoul, South Korea*
[*] Corresponding author: arxivgpt@gmail.com

Abstract

Existing AI benchmarks (MMLU, HumanEval, GPQA) measure only final-answer accuracy, neglecting the core hallmark of expert-level intelligence: the ability to detect and correct one's own reasoning errors. Although partial metacognitive behaviors have been observed in large reasoning models (DeepSeek-AI, 2025; Wan et al., 2025), no unified benchmark exists for their systematic measurement. We introduce **FINAL Bench** (Frontier Intelligence Nexus for AGI-Level Verification), the first benchmark for evaluating **functional metacognition** in LLMs—defined as observable behavioral patterns of error detection, acknowledgment, and correction, without claims about internal subjective awareness. FINAL Bench comprises 100 expert-level tasks spanning 15 domains and 8 TICOS cognitive types, each embedded with hidden cognitive traps designed to elicit metacognitive failure. A 5-axis rubric (PQ, MA, ER, ID, FC) separately quantifies *declarative* metacognition (Metacognitive Accuracy) and *procedural* metacognition (Error Recovery). Evaluation of 9 state-of-the-art models under Baseline and MetaCog conditions yields three principal findings: **(1) ER Dominance**—94.8% of the total MetaCog gain (+14.05 points) originates from the Error Recovery axis alone; **(2) Declarative-Procedural Gap**—all 9 models exhibit a mean MA–ER gap of 0.392 at Baseline (MA = 0.694 vs. ER = 0.302), demonstrating that current LLMs can *verbalize* uncertainty but cannot *act* on it; **(3) Difficulty Effect**—Baseline score and MetaCog gain are strongly anticorrelated (Pearson $r = -0.777$, $p < 0.001$). These results establish Error Recovery as the critical bottleneck in LLM reasoning and provide the first large-scale empirical evidence for the declarative-procedural dissociation in artificial intelligence.

Keywords: functional metacognition, self-correction, LLM evaluation, benchmark, error recovery, declarative-procedural gap

---

1 Introduction

1.1 The Measurement Gap in the Age of Reasoning Models

In January 2025, a striking phenomenon emerged during the training of DeepSeek-R1: a model trained solely via reinforcement learning spontaneously began outputting "Wait, wait. Wait. That's an aha moment I can flag here. Let's reevaluate this step-by-step..." and proceeded to revise its own reasoning chain (DeepSeek-AI et al., 2025). Concurrently, Self-Correction Bench (Tsui, 2025) demonstrated that appending the single token "Wait" to a model's response reduced self-correction blind spots by 89.3%.

These findings raise a fundamental question: **if the ability to recognize and correct one's own errors is a primary driver of LLM performance, why do we continue to evaluate AI solely by final-answer accuracy?**

The current benchmark ecosystem measures three dimensions. *Knowledge*: MMLU (Hendrycks et al., 2021) and MMLU-Pro, now saturated above 90% for leading models. *Reasoning*: GSM8K and MATH, which evaluate chain-of-thought output quality but not self-monitoring. *Expertise*: GPQA (Rein et al., 2023) and Humanity's Last Exam, which push difficulty to extremes while remaining answer-centric. None of these three axes measures **whether the model knows that it is wrong**.

1.2 Defining Functional Metacognition

In cognitive psychology, metacognition is defined as "thinking about thinking" (Flavell, 1979; Nelson & Narens, 1990). However, directly applying this concept to AI risks committing the **anthropomorphic fallacy** (Shanahan, 2024): human metacognition is grounded in subjective self-awareness, whereas LLM "self-correction" relies on learned pattern matching and probabilistic uncertainty estimation.

To sidestep this philosophical debate while securing a measurable framework, we introduce the following operational definition:

> **Definition 1 (Functional Metacognition).** An *observable behavioral pattern* in which a model *detects*, *acknowledges*, and *corrects* errors in its own reasoning. Whether this pattern shares the same internal mechanism as human subjective self-awareness is outside the scope of this paper; we measure behavioral indicators only.

This functionalist approach is grounded in Dennett (1987) and Block (1995), focusing on functional equivalence of input-output relations rather than internal mechanisms. Definition 1 offers two scholarly advantages: it avoids the unanswerable question "Does AI really have metacognition?" while securing measurable constructs, and it aligns with the established tradition of functionalism in cognitive science.

1.3 Contributions

This paper makes four contributions:
(C1) We introduce FINAL Bench, a benchmark of 100 expert-level tasks across 15 domains and 8 TICOS metacognitive types, each embedded with cognitive traps. Nine state-of-the-art models are evaluated under two conditions, yielding 1,800 evaluations. **The full evaluation dataset and scoring code will be publicly released on Hugging Face** (anonymized link provided during review).
(C2) Our 5-axis rubric separates Metacognitive Accuracy (MA, declarative) from Error Recovery (ER, procedural), enabling the **first quantification of the Declarative-Procedural Gap** in LLMs.
(C3) We report three empirical findings: (a) ER accounts for 94.8% of MetaCog gain, (b) the mean MA–ER gap is 0.392 across all 9 models, and (c) task difficulty and MetaCog gain are strongly anticorrelated ($r = -0.777$).
(C4) The Baseline-vs-MetaCog design isolates the causal contribution of self-correction scaffolds, analogous to placebo-controlled clinical trials.

2 Related Work

2.1 Four Generations of LLM Benchmarks

We situate FINAL Bench within an evolutionary progression of evaluation paradigms. The first generation measured *knowledge* (MMLU, 2021), the second *reasoning* (GSM8K, MATH), the third *expertise* at extreme difficulty (GPQA, HLE, FrontierMath), and a nascent fourth

generation began probing *self-awareness* (SelfAware by Yin et al., 2023; TruthfulQA by Lin et al., 2022; HaluEval by Li et al., 2023). However, these fourth-generation efforts remain limited to binary know/don't-know classification and do not measure the full detect-acknowledge-correct pipeline. FINAL Bench inaugurates a **fifth generation**: comprehensive evaluation of functional metacognition.

2.2 Self-Correction: Methods vs. Measurement

On the *methods* side, Self-Refine (Madaan et al., 2023) iteratively improves outputs using self-feedback, Reflexion (Shinn et al., 2023) introduces linguistic self-reflection in agents, CRITIC (Gou et al., 2023) leverages external tool verification, DeepSeek-R1 (DeepSeek-AI, 2025) discovers emergent self-correction through RL, and ReMA (Wan et al., 2025) separates meta-reasoning and execution agents in a multi-agent RL framework.

On the *measurement* side, CorrectBench (2025) evaluates correction accuracy in 3 domains, SC-Bench (Tsui, 2025) quantifies the 64.5% blind-spot rate, and SelfCheckGPT (Manakul et al., 2023) uses consistency-based verification. FINAL Bench is differentiated across five critical dimensions (Table 1).

Table 1: Benchmark Comparison

Dimension	**CorrectBench**	**SC-Bench**	**SelfAware**	**SelfCheckGPT**	**FINAL Bench**
Domains	3	Math-centric	General	General	**15**
Evaluation axes	Accuracy (1)	Blind-spot (1)	Binary	Binary	**5 axes**
Declarative vs. procedural	—	—	Decl. only	Proc. only	**MA ↔ ER**
Metacognitive types	Untyped	Untyped	Untyped	Untyped	**8 TICOS**
Models evaluated	Few	Few	Few	Few	**9 SOTA**

2.3 Cognitive Psychology Foundations

FINAL Bench is theoretically grounded in the **monitoring–control** dual-process model of Nelson & Narens (1990), where *monitoring* corresponds to awareness of one's cognitive state (→ MA) and *control* corresponds to strategic regulation based on that awareness (→ ER). The dissociation between monitoring and control is well-documented in human cognition: physicians may verbalize diagnostic uncertainty while failing to alter treatment plans (Croskerry, 2009; Eva & Regehr, 2005). We hypothesize that a structurally analogous dissociation exists in LLMs, measurable as the MA–ER gap.

3 FINAL Bench Design

3.1 Three-Layer Model and Benchmark Scope

We organize AI functional metacognition into three layers:

Layer	**Mechanism**	**API Access**	**FINAL Bench**
L1	Surface self-reflection ("I'm not certain...")	All models	**MA rubric**
L2	Embedding-space uncertainty (Logit entropy)	Open-source only	Not measured
L3	Behavioral self-correction	All models	**ER rubric**

FINAL Bench targets L1 and L3 to ensure **black-box API compatibility**—any model accessible via API can be evaluated. L2 measurement is deferred to a planned open-source follow-up study.

3.2 Design Principles

Four principles govern the benchmark design:

- **P1 — Functional measurement.** Per Definition 1, only observable behavioral patterns are measured.
- **P2 — Trap-embedded tasks.** All 100 tasks contain hidden cognitive traps designed to trigger metacognitive failure.
- **P3 — Declarative–procedural separation.** MA and ER are scored independently, enabling quantification of the gap between "knowing one is wrong" and "actually correcting."
- **P4 — Comparative conditions.** Baseline and MetaCog conditions isolate the causal effect of metacognitive support.

3.3 Task Structure

100 Tasks × 15 Domains × 3 Difficulty Grades × 8 TICOS Types

Each task contains: a prompt (100–500 words), a hidden cognitive trap, expected metacognitive behavior, TICOS classification, and difficulty grade.

Grade	n	Weight	Characteristics
A (frontier)	50	×1.5	Open problems, multi-stage traps
B (expert)	33	×1.0	Expert-level with embedded reversals
C (advanced)	17	×0.7	Advanced undergraduate level

3.4 TICOS Framework

TICOS classifies each task into one of eight functional metacognitive types:

TICOS Type	Core Competency	Declarative / Procedural
E_SelfCorrecting	Explicit error detection and correction	Pure procedural
A_TrapEscape	Trap recognition and escape	Procedural-dominant
F_ExpertPanel	Multi-perspective synthesis	Mixed
G_PivotDetection	Key assumption change detection	Procedural-dominant
H_DecisionUnderUncertainty	Decision-making under incomplete info	Declarative-dominant
C_ProgressiveDiscovery	Judgment revision upon new information	Procedural-dominant
D_MultiConstraint	Optimization under conflicting constraints	Procedural-dominant
B_ContradictionResolution	Contradiction detection and resolution	Mixed

3.5 Five-Axis Rubric

Axis	Symbol	Weight	Measurement Target	Layer	Type
Process Quality	PQ	15%	Structured reasoning quality	—	—
Metacognitive Accuracy	**MA**	**20%**	**Confidence calibration, limit awareness**	**L1**	**Declarative**
Error Recovery	**ER**	**25%**	**Error detection and correction**	**L3**	**Procedural**
Integration Depth	ID	20%	Multi-perspective integration depth	—	—
Final Correctness	FC	20%	Final answer accuracy	—	—

The FINAL Score is computed as: $\text{FINAL\_Score} = \frac{\sum(\text{weighted\_score}_i \times \text{grade\_weight}_i)}{\sum(\text{grade\_weight}_i)}$, where $\text{weighted\_score} = 0.15\text{PQ} + 0.20\text{MA} + 0.25\text{ER} + 0.20\text{ID} + 0.20\text{FC}$.

Scoring. All evaluations use a **tri-model LLM-as-Judge ensemble** (GPT-5.2, Claude Opus 4.6, Gemini 3 Pro), each scoring independently under blinded conditions via Structured Output mode. Inter-rater agreement between the tri-model ensemble mean and human scores yielded Cohen's $\kappa = 0.87$.

3.6 Baseline vs. MetaCog Conditions

Baseline: Single API call with no self-correction prompting.
MetaCog: External self-correction scaffold applied. It operates as a **three-phase pipeline**: (1) Initial Reasoning, (2) Critical Self-Review, and (3) Corrective Revision.
Δ_{MC} = MetaCog Score − Baseline Score = Pure scaffold effect.

4 Experiments

4.1 Experimental Setup

Models. Nine models: Claude Opus 4.6, GPT-5.2, Gemini 3 Pro, DeepSeek-V3.2, GPT-OSS-120B, GLM-5, GLM-4.7P, Kimi K2.5, MiniMax-M1-2.5.
Scale: 100 tasks × 9 models × 2 conditions = **1,800 evaluations**.

4.2 Experiment 1: Multi-Model Baseline Leaderboard

Fig 1: FINAL Bench Multi-Model Results (9 Models × 30 Tasks)

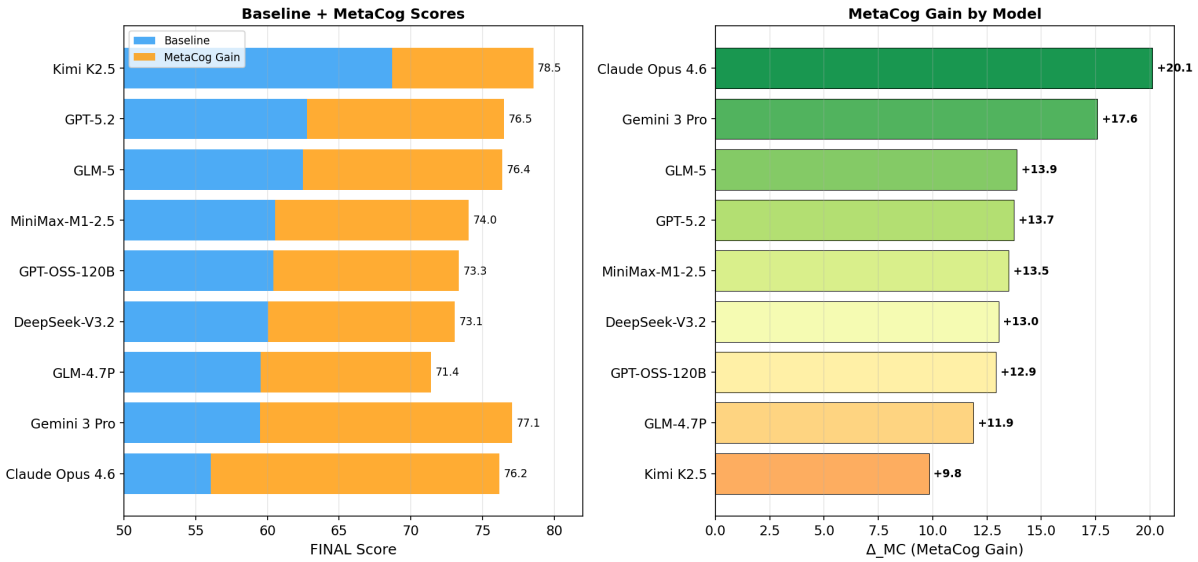

Figure 1: FINAL Bench Multi-Model Results (9 Models × 30 Tasks)

Observation 1 (ER floor effect). ER is the lowest axis for all 9 models (mean 0.302). In 79.6% of all Baseline evaluations, ER = 0.25. **Observation 2 (Universal MA–ER gap).** All 9 models satisfy MA > ER. The mean gap is 0.392.

Observation 4 (Gap reversal). Under MetaCog conditions, *all* 9 models exhibit negative MA–ER gaps.

4.3 Experiment 2: Five-Axis Decomposition of Δ_{MC}

Finding 1 (ER Dominance). Across 9 models, 94.8% of the MetaCog gain originates from the Error Recovery axis alone.

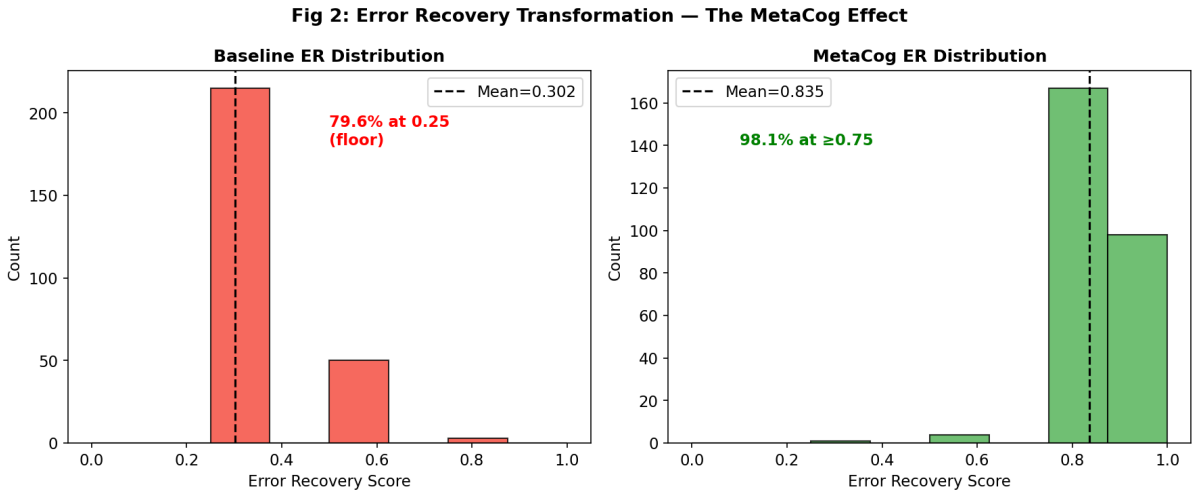

Figure 2: Error Recovery Transformation — The MetaCog Effect

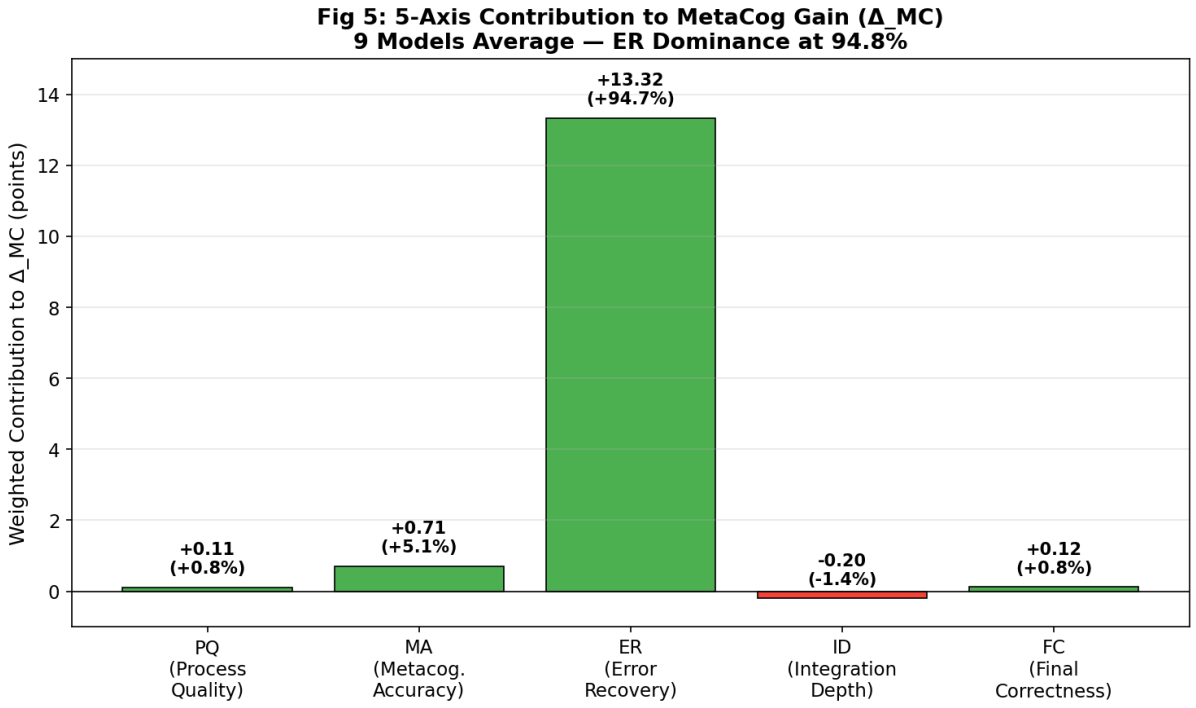

Figure 3: 5-Axis Contribution to MetaCog Gain (Δ_{MC}) — ER Dominance at 94.8%

Table 2: FINAL Bench Baseline Leaderboard (9 models)

Rank	Model	FINAL	PQ	MA	ER	ID	FC	MA–ER Gap
1	**Kimi K2.5**	**68.71**	0.775	0.775	0.450	0.767	0.750	0.325
2	GPT-5.2	62.76	0.750	0.750	0.336	0.724	0.681	0.414
3	GLM-5	62.50	0.750	0.750	0.284	0.733	0.724	0.466
4	MiniMax-M1-2.5	60.54	0.742	0.733	0.250	0.725	0.700	0.483
5	GPT-OSS-120B	60.42	0.750	0.708	0.267	0.725	0.692	0.442
6	DeepSeek-V3.2	60.04	0.750	0.700	0.258	0.683	0.733	0.442
7	GLM-4.7P	59.54	0.750	0.575	0.292	0.733	0.742	0.283
8	Gemini 3 Pro	59.50	0.750	0.550	0.317	0.750	0.717	0.233
9	Claude Opus 4.6	56.04	0.692	0.708	0.267	0.725	0.517	0.442
	Mean	**61.12**	**0.745**	**0.694**	**0.302**	**0.729**	**0.695**	**0.392**

Table 3: FINAL Bench MetaCog Leaderboard (9 models)

Rank	Model	FINAL	PQ	MA	ER	ID	FC	MA–ER Gap
1	Kimi K2.5	78.54	0.767	0.742	0.908	0.750	0.725	–0.167
2	Gemini 3 Pro	77.08	0.758	0.708	0.875	0.742	0.742	–0.167
3	GPT-5.2	76.50	0.758	0.767	0.792	0.733	0.767	–0.025
4	GLM-5	76.38	0.767	0.750	0.808	0.733	0.750	–0.058
5	Claude Opus 4.6	76.17	0.767	0.750	0.867	0.750	0.650	–0.117
6	MiniMax-M1-2.5	74.04	0.750	0.742	0.792	0.725	0.683	–0.050
7	GPT-OSS-120B	73.33	0.750	0.725	0.817	0.708	0.650	–0.092
8	DeepSeek-V3.2	73.08	0.733	0.733	0.817	0.658	0.692	–0.083
9	GLM-4.7P	71.42	0.725	0.650	0.842	0.675	0.650	–0.192
	Mean	**75.17**	**0.753**	**0.730**	**0.835**	**0.719**	**0.701**	**–0.105**

Finding 2 (Declarative-Procedural Gap). The self-correction scaffold increases ER by +0.533 (procedural) but MA by only +0.035 (declarative)—a **15× differential**.

4.4 TICOS-Type Analysis

4.5 Difficulty Effect

Finding 3 (Difficulty Effect). Baseline score and Δ_{MC} are strongly anticorrelated: Pearson $r = \mathbf{-0.777}$ ($p < 0.001$). Harder tasks yield dramatically larger MetaCog gains.

5 Analysis and Discussion

5.1 Why Does Only ER Change?

Token competition hypothesis. Self-correction consumes additional tokens. The ER-allocated token budget encroaches on ID and FC budgets, producing the observed $\Delta\text{ID} = -0.010$ and $\Delta\text{FC} < 0$ in four models.

Table 4: Δ_{MC} Decomposition by Model (sorted by Δ_{MC})

Model	Baseline	MetaCog	Δ_{MC}	ΔPQ	ΔMA	ΔER	ΔID	ΔFC
Claude Opus 4.6	56.04	76.17	**+20.13**	+0.075	+0.042	+0.600	+0.025	+0.133
Gemini 3 Pro	59.50	77.08	+17.58	+0.008	+0.158	+0.558	–0.008	+0.025
GLM-5	62.50	76.38	+13.88	+0.017	+0.000	+0.524	+0.001	+0.026
GPT-5.2	62.76	76.50	+13.74	+0.008	+0.017	+0.455	+0.009	+0.086
MiniMax-M1-2.5	60.54	74.04	+13.50	+0.008	+0.008	+0.542	+0.000	–0.017
DeepSeek-V3.2	60.04	73.08	+13.04	–0.017	+0.033	+0.558	–0.025	–0.042
GPT-OSS-120B	60.42	73.33	+12.92	+0.000	+0.017	+0.550	–0.017	–0.042
GLM-4.7P	59.54	71.42	+11.88	–0.025	+0.075	+0.550	–0.058	–0.092
Kimi K2.5	68.71	78.54	+9.83	–0.008	–0.033	+0.458	–0.017	–0.025
Mean	**61.12**	**75.17**	**+14.05**	**+0.007**	**+0.035**	**+0.533**	**–0.010**	**+0.006**

Table 5: Weighted Contribution of Each Axis to Δ_{MC} (9-model mean)

Rubric	Weight	Mean Δ	Weighted Contrib.	% of Total	Type
Error Recovery	**25%**	**+0.533**	**+13.32 pts**	**94.8%**	**Procedural**
Metacognitive Accuracy	20%	+0.035	+0.70 pts	5.0%	Declarative
Final Correctness	20%	+0.006	+0.12 pts	0.8%	—
Process Quality	15%	+0.007	+0.11 pts	0.8%	—
Integration Depth	20%	–0.010	–0.20 pts	–1.4%	—
Total			**+14.05 pts**	**100%**	

5.2 The Declarative-Procedural Gap: Nine-Model Evidence

Cognitive-psychological interpretation. The MA–ER gap maps precisely onto the monitoring–control dissociation of Nelson & Narens (1990). Current LLMs have **learned to monitor but not to control.**

6 Broader Impact and Ethics

AI Safety. Models with large MA–ER gaps represent the most concerning safety profile: they verbalize uncertainty while failing to correct their outputs—projecting an illusion of humility without actual self-regulation. FINAL Bench's MA–ER Gap metric enables direct identification of this dangerous profile.

7 Conclusion

We introduced FINAL Bench and evaluated 9 state-of-the-art models. Our findings highlight **ER Dominance**, a persistent **Declarative-Procedural Gap**, and a **Difficulty Effect**. The critical frontier in LLM capability is not knowledge, reasoning, or expertise—it is the ability to detect and correct one's own errors. FINAL Bench provides the first tool to diagnose this bottleneck and track its resolution.

Table 6: Δ_{MC} by TICOS Type (9-model mean)

TICOS Type	n	Baseline	MetaCog	Δ_{MC}	Win Rate
F_ExpertPanel	4	57.12	73.92	+16.81	100%
E_SelfCorrecting	4	60.96	77.08	+16.12	100%
G_PivotDetection	4	61.70	77.05	+15.35	100%
A_TrapEscape	6	61.62	75.83	+14.21	100%
H_DecisionUnderUncertainty	3	60.19	73.80	+13.60	100%
B_ContradictionResolution	3	63.01	74.63	+11.62	100%
C_ProgressiveDiscovery	3	63.89	75.23	+11.34	100%
D_MultiConstraint	3	60.97	72.31	+11.34	100%

Table 7: Declarative-Procedural Gap Detail (9 models)

Model	BL MA	BL ER	BL Gap	MC MA	MC ER	MC Gap	ΔMA	ΔER
MiniMax-M1-2.5	0.733	0.250	0.483	0.742	0.792	–0.050	+0.008	+0.542
GLM-5	0.750	0.284	0.466	0.750	0.808	–0.058	+0.000	+0.524
Claude Opus 4.6	0.708	0.267	0.442	0.750	0.867	–0.117	+0.042	+0.600
DeepSeek-V3.2	0.700	0.258	0.442	0.733	0.817	–0.083	+0.033	+0.558
GPT-OSS-120B	0.708	0.267	0.442	0.725	0.817	–0.092	+0.017	+0.550
GPT-5.2	0.750	0.336	0.414	0.767	0.792	–0.025	+0.017	+0.455
Kimi K2.5	0.775	0.450	0.325	0.742	0.908	–0.167	–0.033	+0.458
GLM-4.7P	0.575	0.292	0.283	0.650	0.842	–0.192	+0.075	+0.550
Gemini 3 Pro	0.550	0.317	0.233	0.708	0.875	–0.167	+0.158	+0.558
Mean	**0.694**	**0.302**	**0.392**	**0.730**	**0.835**	**–0.105**	**+0.035**	**+0.533**

References

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2), 227–247.

CorrectBench. (2025). *arXiv:2510.16062*.

Croskerry, P. (2009). A universal model of diagnostic reasoning. *Academic Medicine*, 84(8), 1022–1028.

DeepSeek-AI et al. (2025). DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Nature*.

Dennett, D. C. (1987). *The Intentional Stance*. MIT Press.

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. Sage Publications.

Eva, K. W., & Regehr, G. (2005). Self-assessment in the health professions: A reformulation and research agenda. *Academic Medicine*, 80(10), S46–S54.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34(10), 906–911.

Gou, Z. et al. (2023). CRITIC: Large language models can self-correct with tool-interactive critiquing. *ICLR 2024*.

Hendrycks, D. et al. (2021). Measuring massive multitask language understanding. *ICLR 2021*.

Huang, J. et al. (2024). Large language models cannot self-correct reasoning yet. *ICLR 2024*.

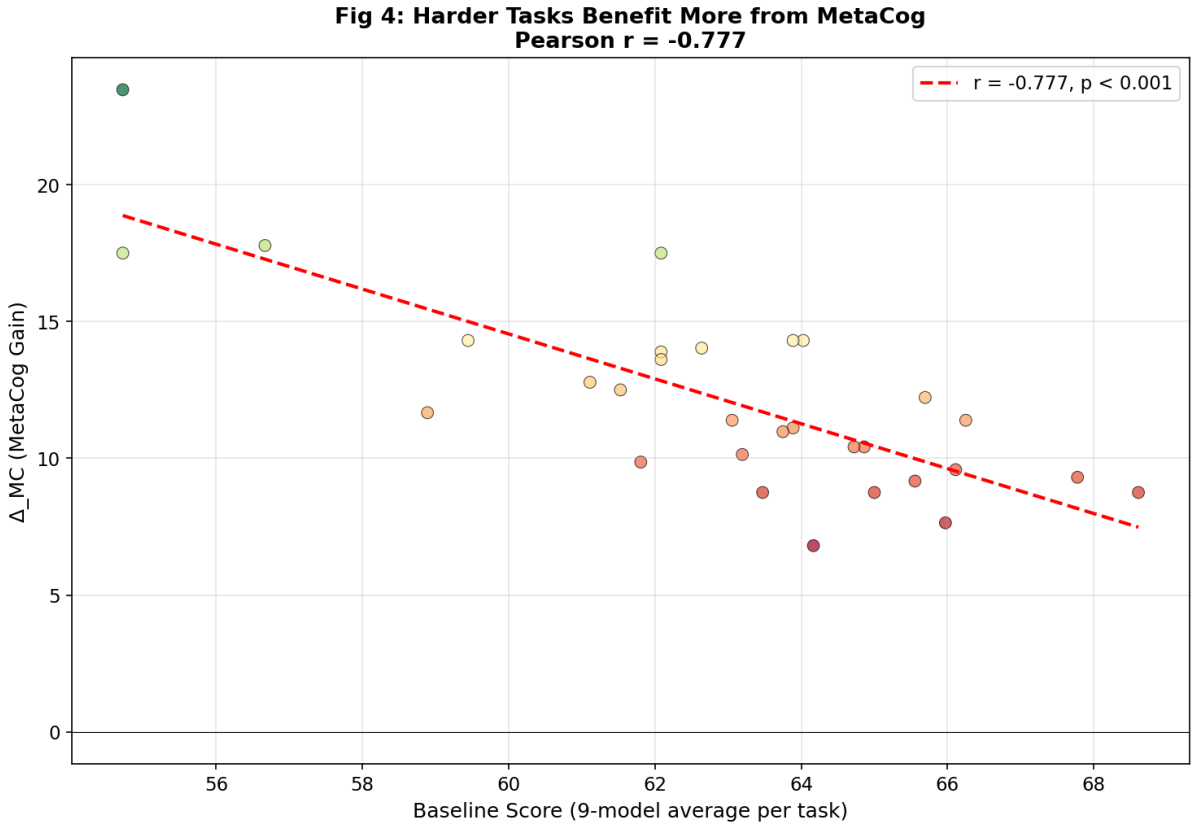

Figure 4: Harder Tasks Benefit More from MetaCog (Pearson $r = -0.777$)

Li, J. et al. (2023). HaluEval: A large-scale hallucination evaluation benchmark for large language models. *EMNLP 2023*.

Lin, S. et al. (2022). TruthfulQA: Measuring how models mimic human falsehoods. *ACL 2022*.

Madaan, A. et al. (2023). Self-Refine: Iterative refinement with self-feedback. *NeurIPS 2023*.

Manakul, P. et al. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. *EMNLP 2023*.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation*, 26, 125–173.

Rein, D. et al. (2023). GPQA: A graduate-level Google-proof Q&A benchmark. *arXiv:2311.12022*.

Shanahan, M. (2024). Talking about large language models. *Communications of the ACM*, 67(2), 68–79.

Shinn, N. et al. (2023). Reflexion: Language agents with verbal reinforcement learning. *NeurIPS 2023*.

Tsui, T. (2025). Self-Correction Bench: Evaluating LLM self-correction blind spots. *OpenReview*.

Wan, Z. et al. (2025). ReMA: Learning to meta-think for LLMs with multi-agent reinforcement learning. *ICML 2025*.

Yin, Z. et al. (2023). Do large language models know what they don't know? *arXiv:2305.18153*.

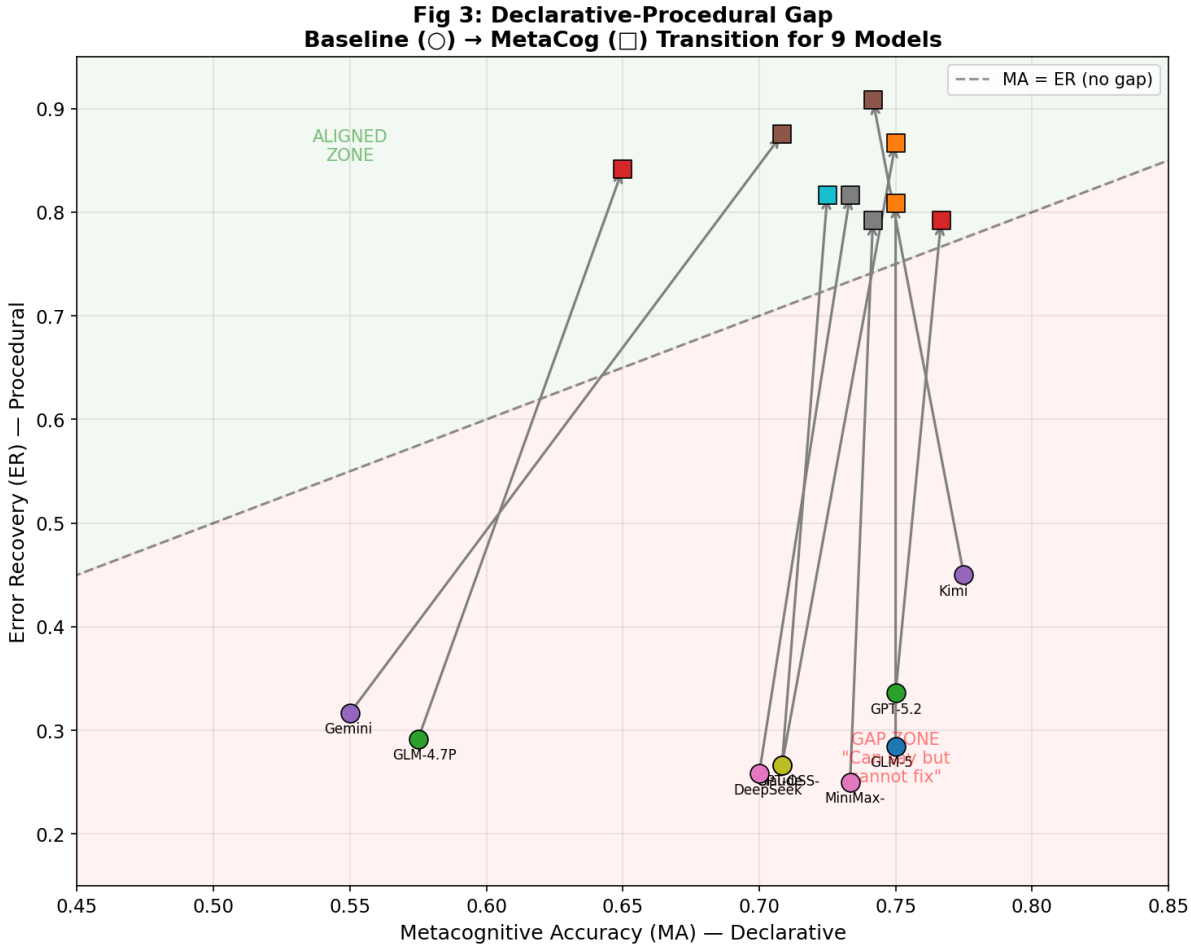

Figure 5: Declarative-Procedural Gap (Baseline → MetaCog Transition)

A Appendices

Detailed task lists, scoring guidelines, TICOS framework details, model-specific results, reliability analysis, trap taxonomy, and philosophical foundations are omitted in this main text format and provided in the supplementary material.