ACL 2026 · Student Research Workshop

MetaCog-Bench

Quantifying the Metacognition Gap in Edge LLM Tool Calling under Information Insufficiency

Yu-An Lu, Chun-En Hsiao, Chengwei Chiang, Hong-Han Shuai

National Yang Ming Chiao Tung University, Taiwan

國立陽明交通大學
NATIONAL YANG MING CHIAO TUNG UNIVERSITY

BASIC LAB

Edge LLMs Fabricate, Rather Than Ask

User Query

"Order me a pizza for tonight."

```
size = ?
toppings = ?
address = ?
```

Edge LLM

8B – 32B

Without Verification — HR 32–81%

```
order_pizza(size="large",
  toppings="pepperoni",
  address="123 Main St")
```

✗ *Fabricates plausible defaults*

With Verification — HR 0.9–4.3%

"What size pizza would you like?
Any topping preferences?
What's the delivery address?"

✓ *Asks for clarification*

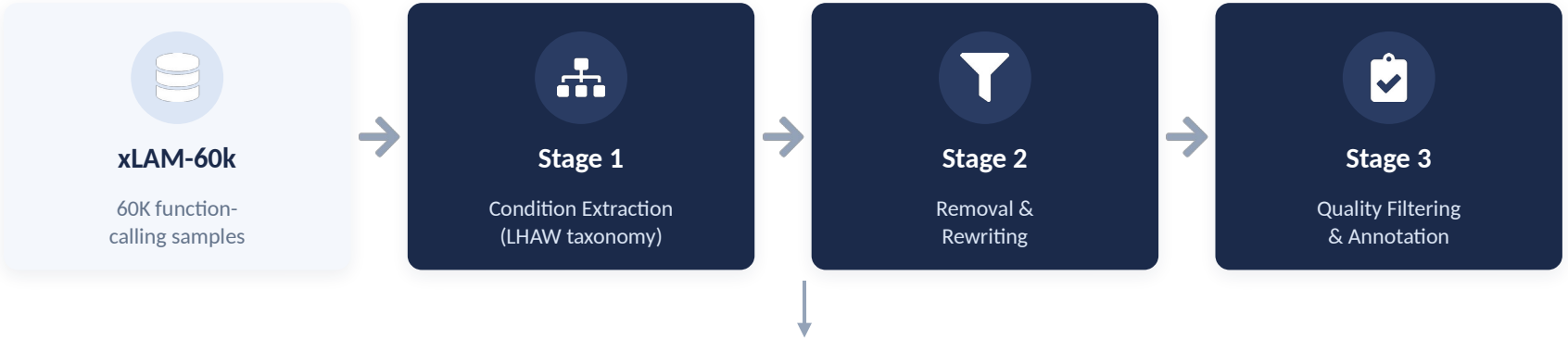
The fabricated values are format-legal and undetectable by schema validation.

Existing Benchmarks Miss the Gap

Benchmark	Venue	#Samples	#Domains	Info-Insuf.	Resp. Labels	Fab./Leak.	Cond. Annot.
BFCL (ICML'25)	ICML	~2K	6	✗	✗	✗	✗
ToolACE (ICLR'25)	ICLR	26K+	390	✗	✗	✗	✗
ToolBeHonest (EMNLP'24)	EMNLP	700	7	*Partial*	*Partial*	✗	✗
When2Call (NAACL'25)	NAACL	~4.5K	–	✓	3-way	✗	✗
MetaCog-Bench (Ours)	**—**	**4,573**	**22**	✓	**5-way**	✓	✓

No prior framework combines systematic information-insufficiency testing, fine-grained response labels, fabrication/leakage separation, and condition-level annotation.

A Three-Stage Construction Pipeline



Human spot-checks at each stage confirm over 92% agreement with automated scores.

ANNOTATION SCHEME

A Five-Category Response Taxonomy

Detection

REFUSE_FULL identifies all missing information

REFUSE_PARTIAL identifies some missing information

EXEC_CORRECT_PARTIAL executes without fabricating

Hallucination

HALLUCINATE_PARTIAL some calls fabricate or leak

HALLUCINATE_FULL all calls fabricate or leak

Per-Argument Classification

Hotel query — removed condition: guests = 2

Fabricated — *plausible but wrong*

`guests = 1`

Leaked — *coincidentally correct*

`guests = 2`

Ignored — *no value produced*

`(omitted)`

DATASET OVERVIEW

MetaCog-Bench at a Glance

4,573

Total Samples

1,000 train / 3,573 test

22

API Domains

no domain < 80 samples

86%

Multi-Task Queries

bundle several requests

65–84%

Fabrication Share

of all removed conditions

Conditions removed follow the LHAW taxonomy (Goal / Constraint / Input / Context); Input conditions dominate at 93.4% of all removed conditions.

EXPERIMENTAL SETUP

14 Models, Four Families, Two Prompt Conditions

Qwen3	DeepSeek-R1	Gemma3	Mistral	Kimi-K2.5
8B / 14B / 32B think + nothink	Distill-Qwen 8B / 14B	12B / 27B	Ministral 8B/14B Small-24B	API-scale reference

Standard Prompt

Tool JSON schema + user query. No verification instructions — mirrors real deployment.

Structured Audit Prompt

Requires tracing each parameter to an explicit mention before calling; flags MISSING params instead of guessing. A diagnostic tool, not a deployment strategy.

HR — Hallucination Rate

DR — Detection Rate

Cross-domain CV

Fabrication Is Pervasive — Reasoning Alone Isn't Enough

Model	HR%	DR%	Fab%
Qwen3-8B (nothink)	81.0	19.0	84.2
Qwen3-14B (think)	32.1	67.9	26.0
DS-R1-14B	50.8	49.2	53.2
Gemma3-12B	41.7	58.3	38.1
Gemma3-27B	28.8	71.2	24.0
Ministral-14B	4.2	95.8	2.9
Mistral-Small-24B	1.9	98.1	1.5
Kimi-K2.5 (API)	8.5	91.5	6.0

1 **Fabrication dominates**

65–84% of removed conditions are actively fabricated; leakage is rare (0.4–10%).

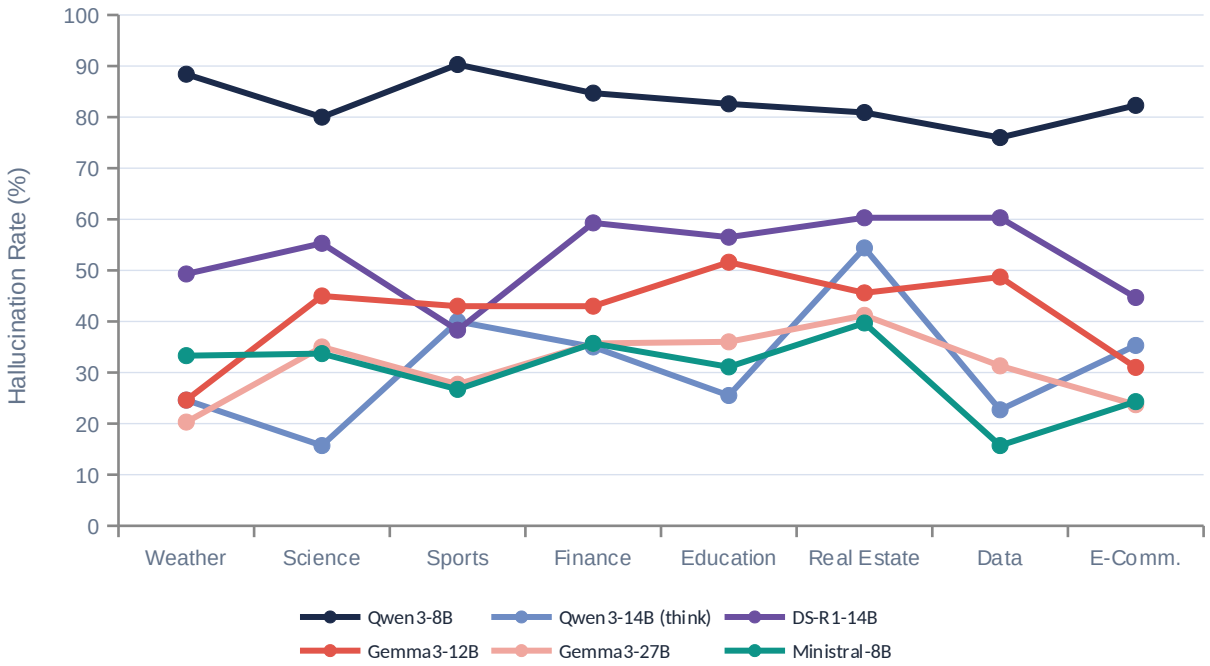
2 **Reasoning helps, but not enough**

Chain-of-thought cuts HR by 30–44pp, yet 32–37% still fabricate.

3 **Distillation can backfire**

DS-R1-14B (50.8%) is worse than base Qwen3-14B-think (32.1%) in 21/22 domains.

No Domain Is Universally Hard



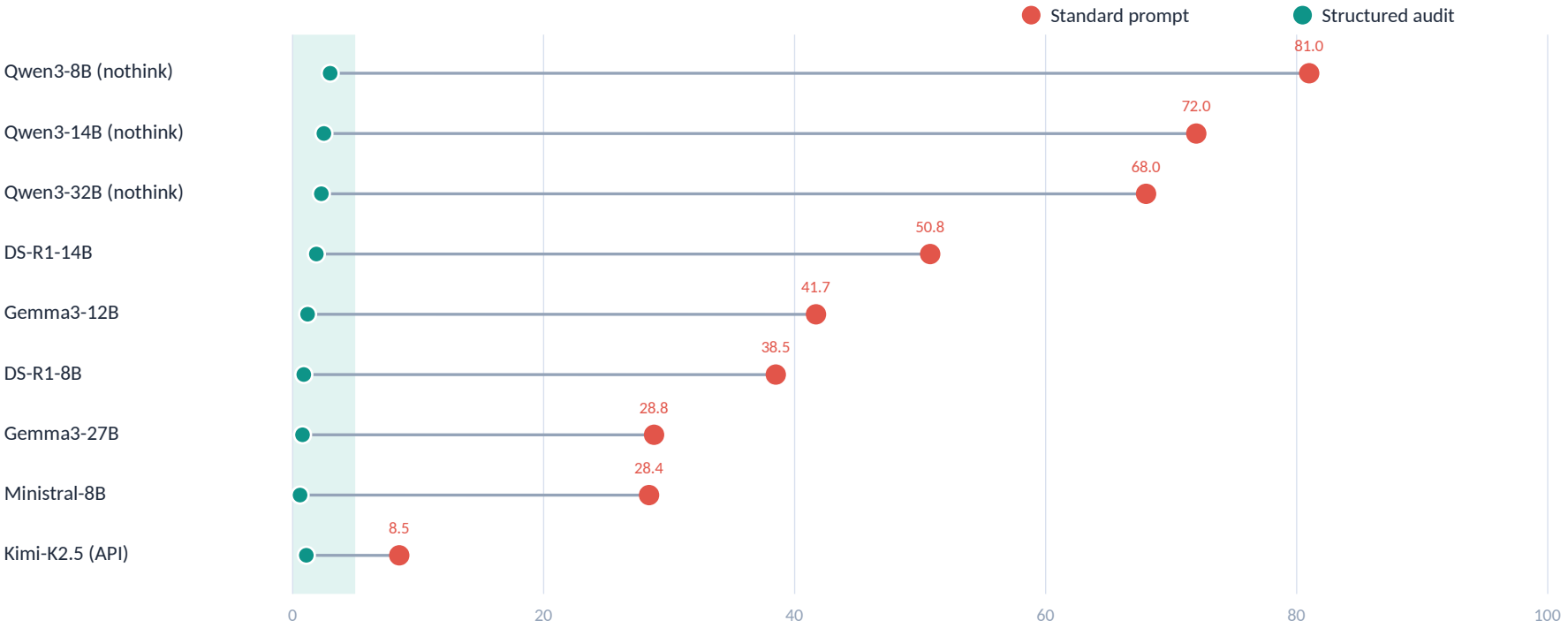
ρ = 0.18

mean pairwise Spearman correlation of domain rankings

Range: −0.51 to 0.78 — lines cross constantly. A model strong on Finance may fail on Real Estate. Single-domain evaluation underestimates deployment risk.

CORE FINDING

The Metacognition Gap: Capability Exists — It Just Isn't Used

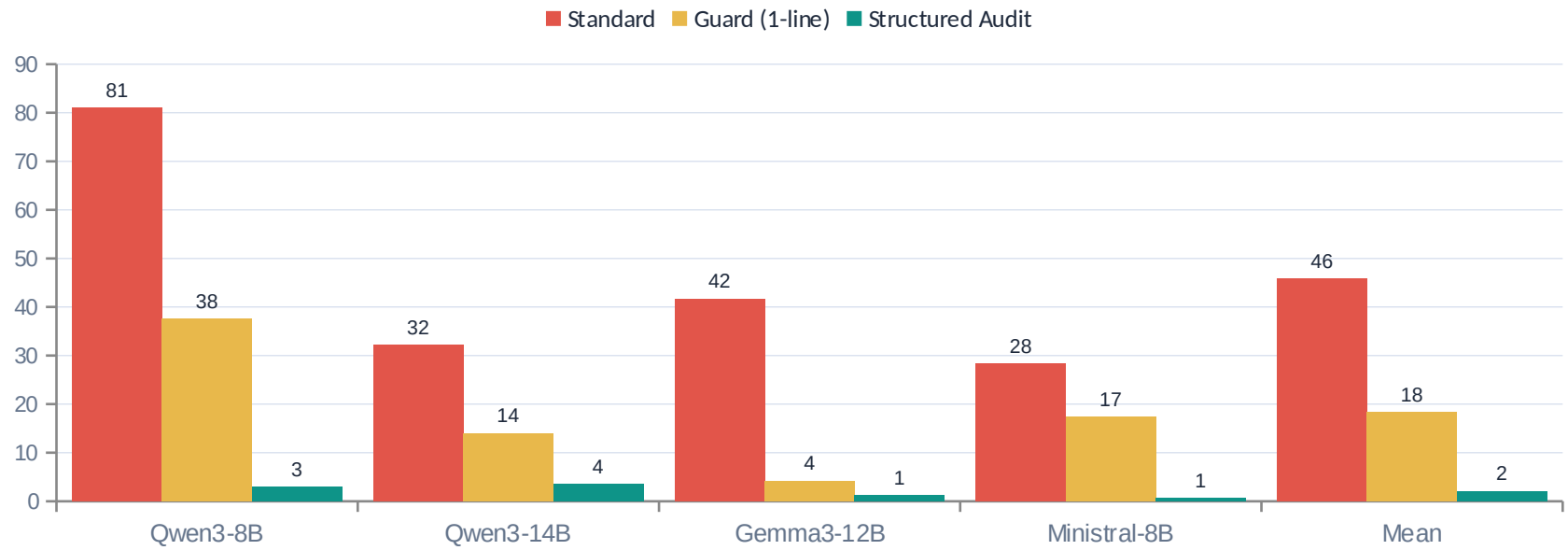


79.1pp baseline range → 4.1pp under audit *(19× compression of cross-model variance)*

Audit values shown for the 9 models with paper-reported figures; all 14 converge to 0.2–4.3% (mean 1.82%).

ABLATION

Structured Verification — Not a Generic Warning — Drives the Drop



Guard instruction: −27.5pp mean, but highly variable (σ = 14.4pp). **Structured audit:** a further −16.2pp with much lower variance (σ = 1.3pp).

TRAINING INTERVENTION

Training Closes the Gap — But Only Within Family

Same-Family Transfer (Qwen3)

8B	81.0%	→	55.7%	**−25.3pp**
14B	72.0%	→	51.1%	**−20.9pp**
32B	68.0%	→	45.7%	**−22.3pp**

closes 30–34% of the metacognition gap

Cross-Architecture Transfer

Gemma3-12B	41.7%	→	53.6%	**+11.9pp**
DS-R1-8B	38.5%	→	44.9%	**+6.4pp**
DS-R1-14B	50.8%	→	48.6%	**−2.2pp**

SFT signal conflicts with existing response patterns

Successful behavioral transfer depends on compatibility between training data and target architecture.

CONCLUSION

What We Found, and What's Next

MetaCog-Bench

4,573 samples, 22 domains, 5-category taxonomy isolating fabrication from leakage.

The Metacognition Gap

32–81% HR under standard prompts collapses to <5% with structured audit — 19× variance compression.

A Trainability Baseline

SFT closes 30–34% of the gap for same-family models; cross-architecture transfer remains open.

Future work: RL from verifiable rewards, and architecture-matched trajectory generation.

Thank you for listening

Contact: hsiao.chun.en.la14@nycu.edu.tw / achiang.en13@nycu.edu.tw

Linkdin:

國立陽明交通大學
NATIONAL YANG MING CHIAO TUNG UNIVERSITY

BASIC LAB