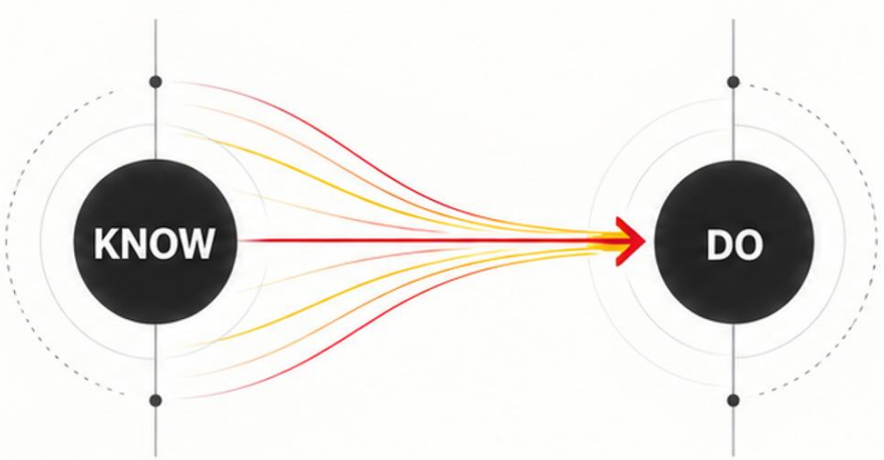
Model-Adaptive Tool Necessity Reveals the Knowing-Doing Gap in LLM Tool Use



Yize Cheng, Chenrui Fan, Mahdi JafariRaviz, Keivan Rezaei, Soheil Feizi
University of Maryland, College Park

NeurIPS 2026

UNIVERSITY OF MARYLAND

We Want Models to Use Tools Only When Needed

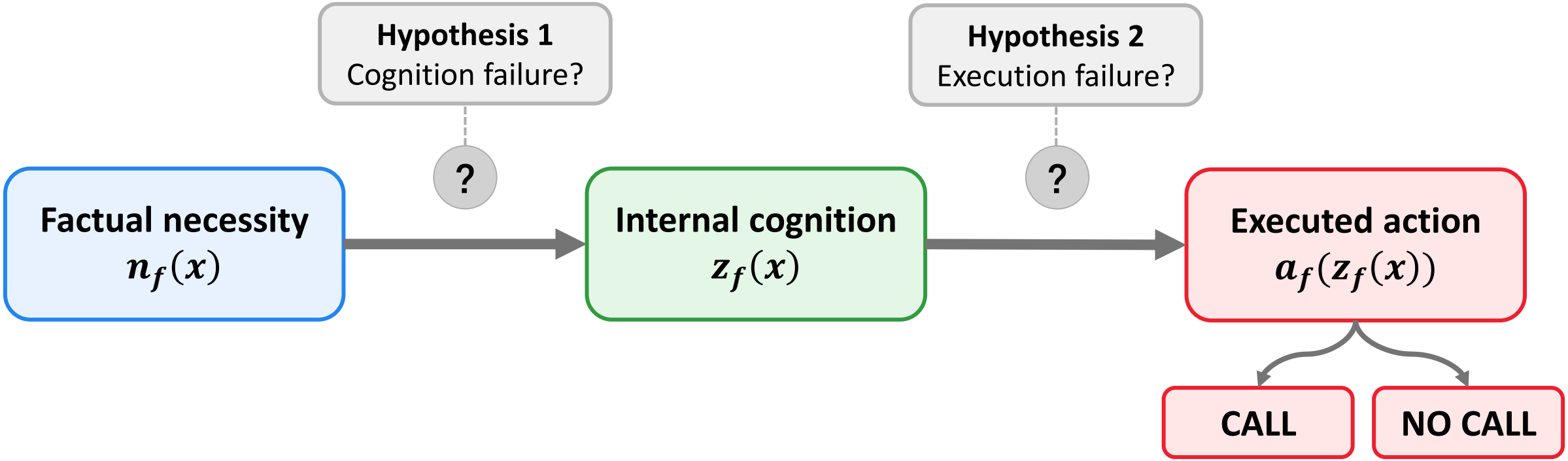
	Call tool	Answer directly
Tool necessary	Correct use	Tool underuse
Tool unnecessary	Tool overuse	Correct use

- Call a tool when it is necessary.
- Answer directly when it is unnecessary.
- Underuse risks wrong answers; overuse adds latency and cost.

But when agents underuse or overuse tools, what's the reason?

UNIVERSITY OF MARYLAND

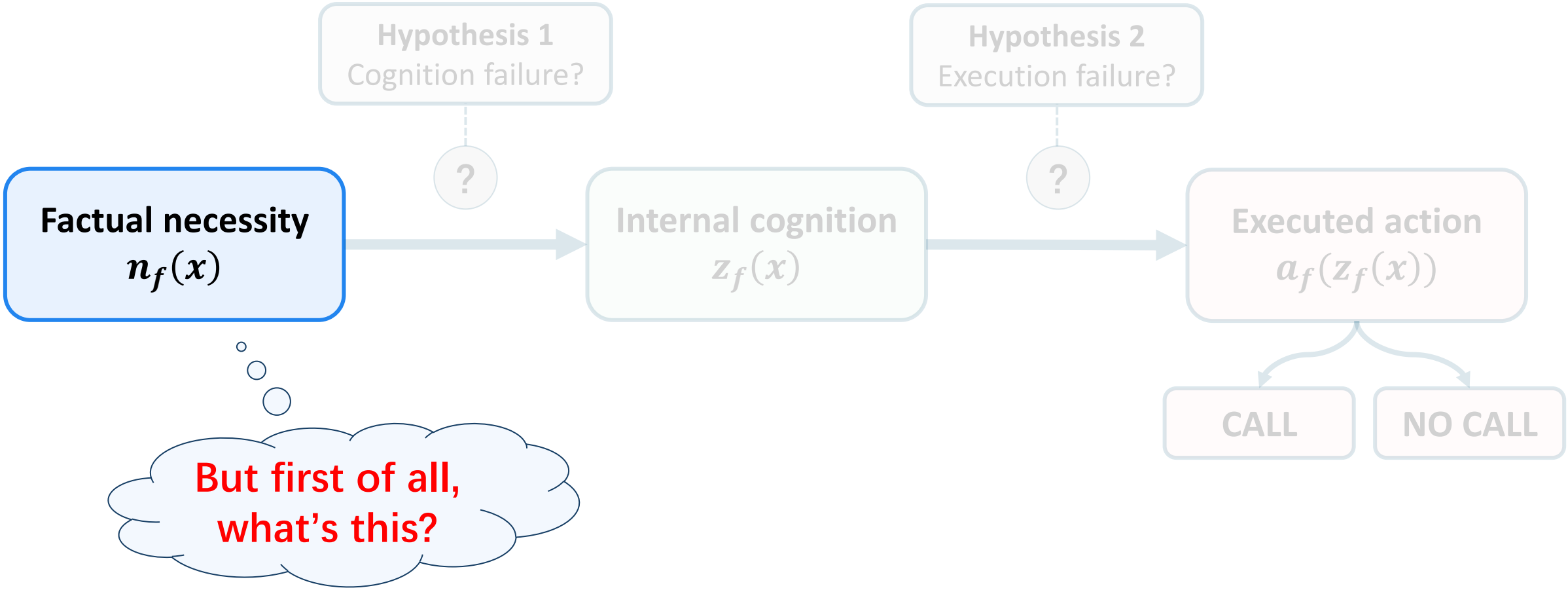
Two Stage Cognition-Execution Modeling of LLM Tool-Use



Is this a cognition problem or an execution problem?

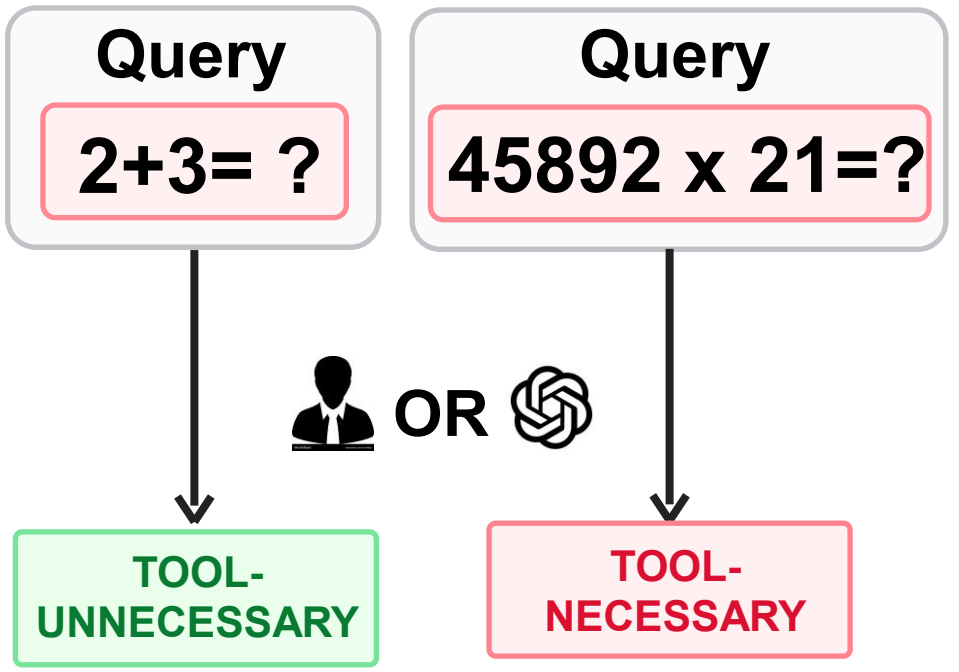
UNIVERSITY OF MARYLAND

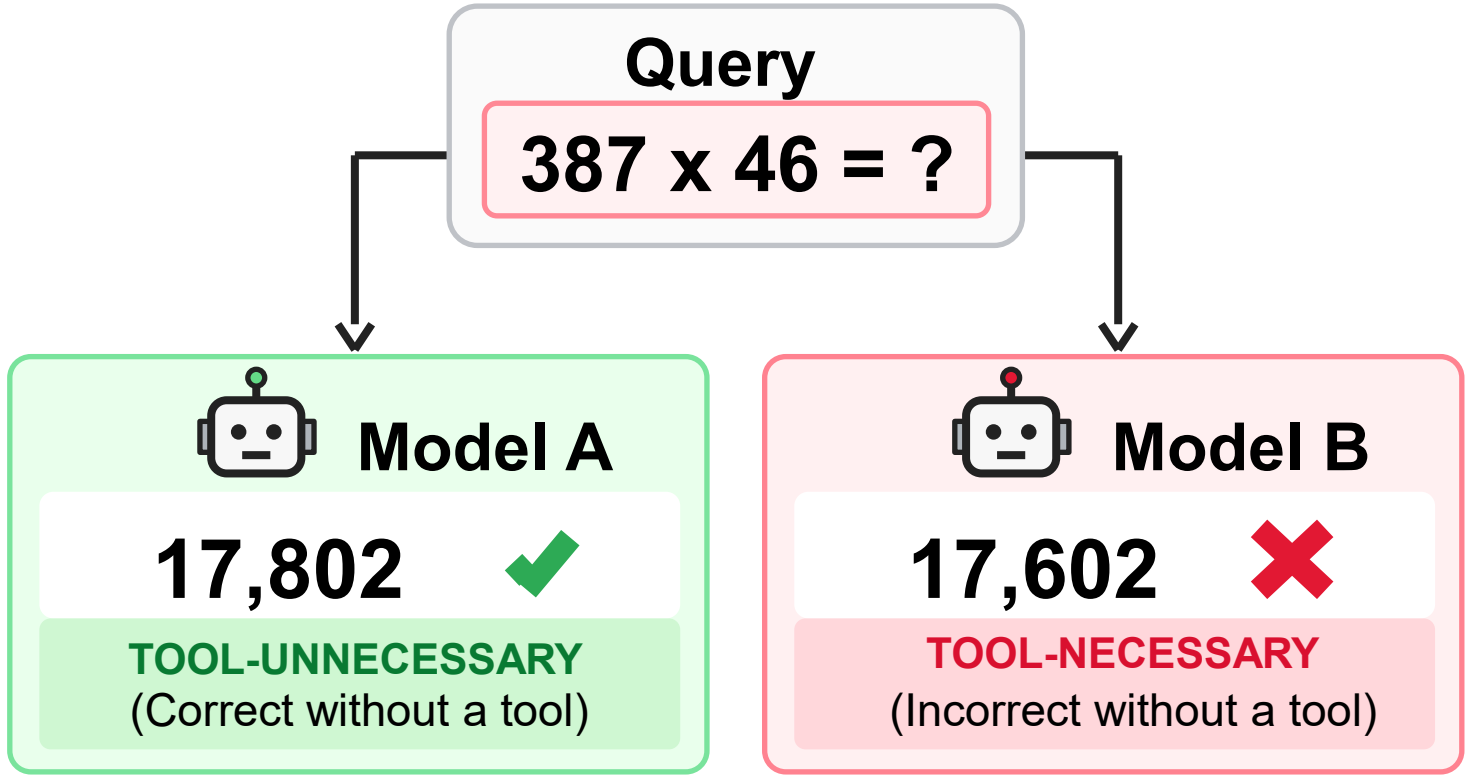
Two Stage Cognition-Execution Modeling of LLM Tool-Use



UNIVERSITY OF MARYLAND

What Does it Mean for A Tool to be Needed?

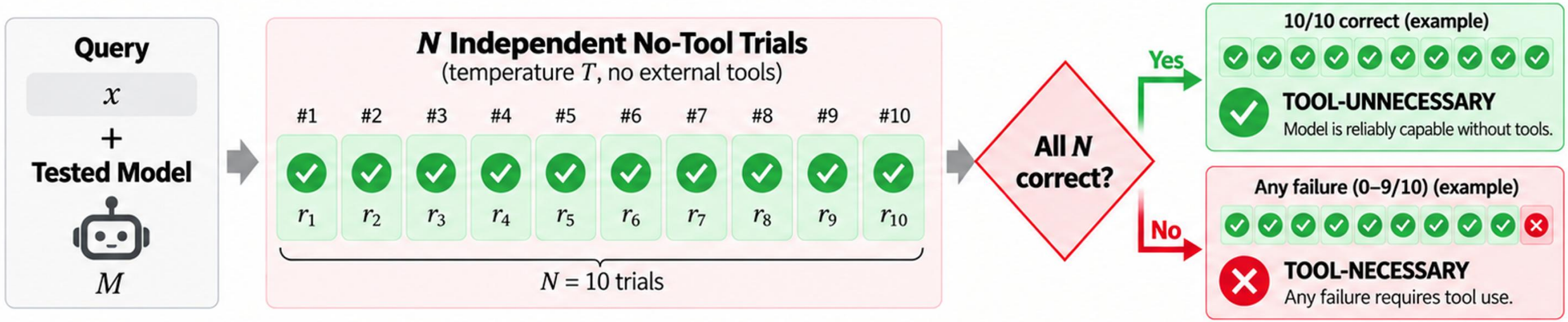




A model-agnostic label reflects the judge, not the tested model.

UNIVERSITY OF MARYLAND

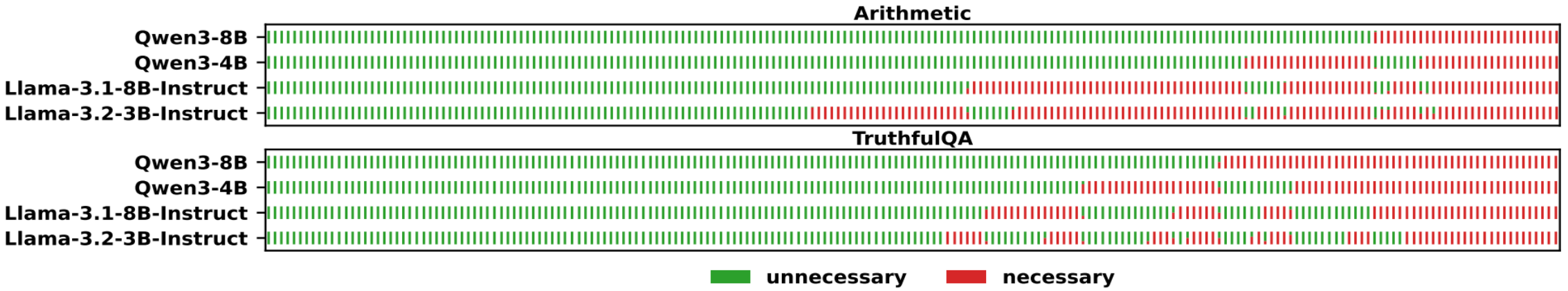
Defining Model-Adaptive Tool Necessity



Tool necessity is empirically grounded to model capability.

UNIVERSITY OF MARYLAND

Empirical Capability Boundaries Are Truly Different



Different models show substantially different capability boundaries, which would've been obscured model-agnostic tool necessity

But is tool underuse or overuse a real problem?

UNIVERSITY OF MARYLAND

Actual Behavior Is Misaligned with Necessity

Model	Arithmetic					TruthfulQA				
	N-C	N-NC	UN-C	UN-NC	Mis.	N-C	N-NC	UN-C	UN-NC	Mis.
Qwen3-8B	438 (11.0%)	140 (3.5%)	1526 (38.2%)	1896 (47.4%)	**41.7%**	69 (8.4%)	146 (17.9%)	108 (13.2%)	494 (60.5%)	**31.1%**
Qwen3-4B	253 (6.3%)	581 (14.5%)	481 (12.0%)	2685 (67.1%)	**26.5%**	103 (12.6%)	153 (18.7%)	189 (23.1%)	372 (45.5%)	**41.8%**
Llama-3.1-8B-Instruct	419 (10.5%)	1204 (30.1%)	335 (8.4%)	2042 (51.1%)	**38.5%**	98 (12.0%)	130 (15.9%)	122 (14.9%)	467 (57.2%)	**30.8%**
Llama-3.2-3B-Instruct	526 (13.2%)	1559 (39.0%)	600 (15.0%)	1315 (32.9%)	**54.0%**	58 (7.1%)	164 (20.1%)	104 (12.7%)	491 (60.1%)	**32.8%**

Mismatch is substantial:

Arithmetic:
26.5%-54.0%

TruthfulQA:
30.8%-41.8%

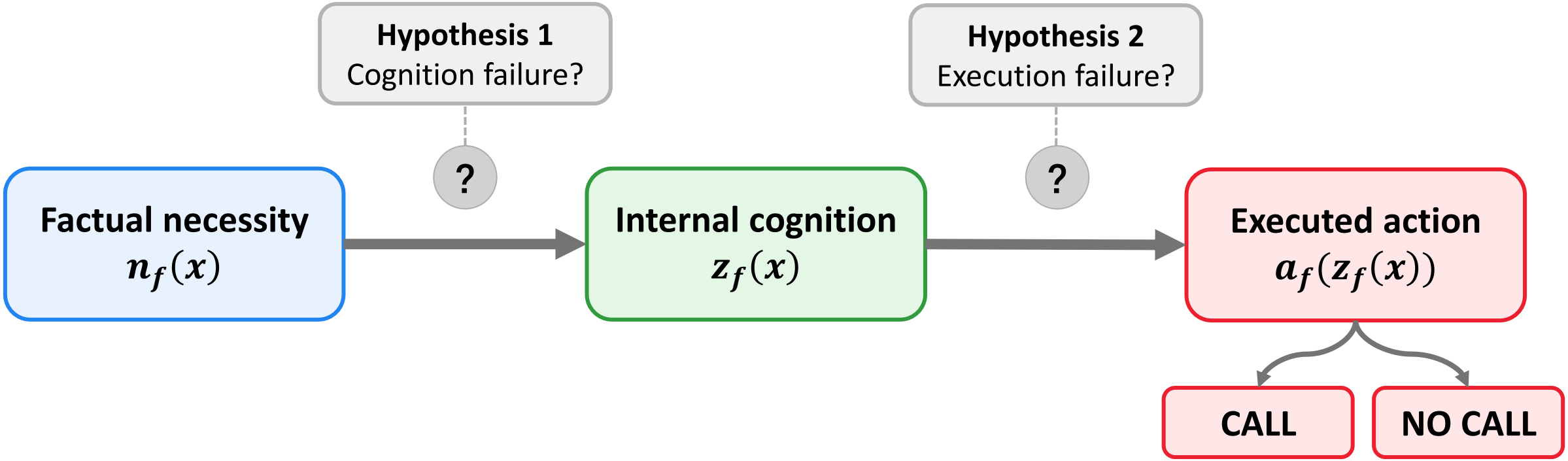
Four Outcome Categories

	C (Calls tool)	NC (No call)
N (Necessary)	**N-C** Necessary and Calls (Aligned)	**N-NC** Necessary but No Call (Mismatch)
UN (Unnecessary)	**UN-C** Unnecessary but Calls (Mismatch)	**UN-NC** Unnecessary and No Call (Aligned)

WHY?

UNIVERSITY OF MARYLAND

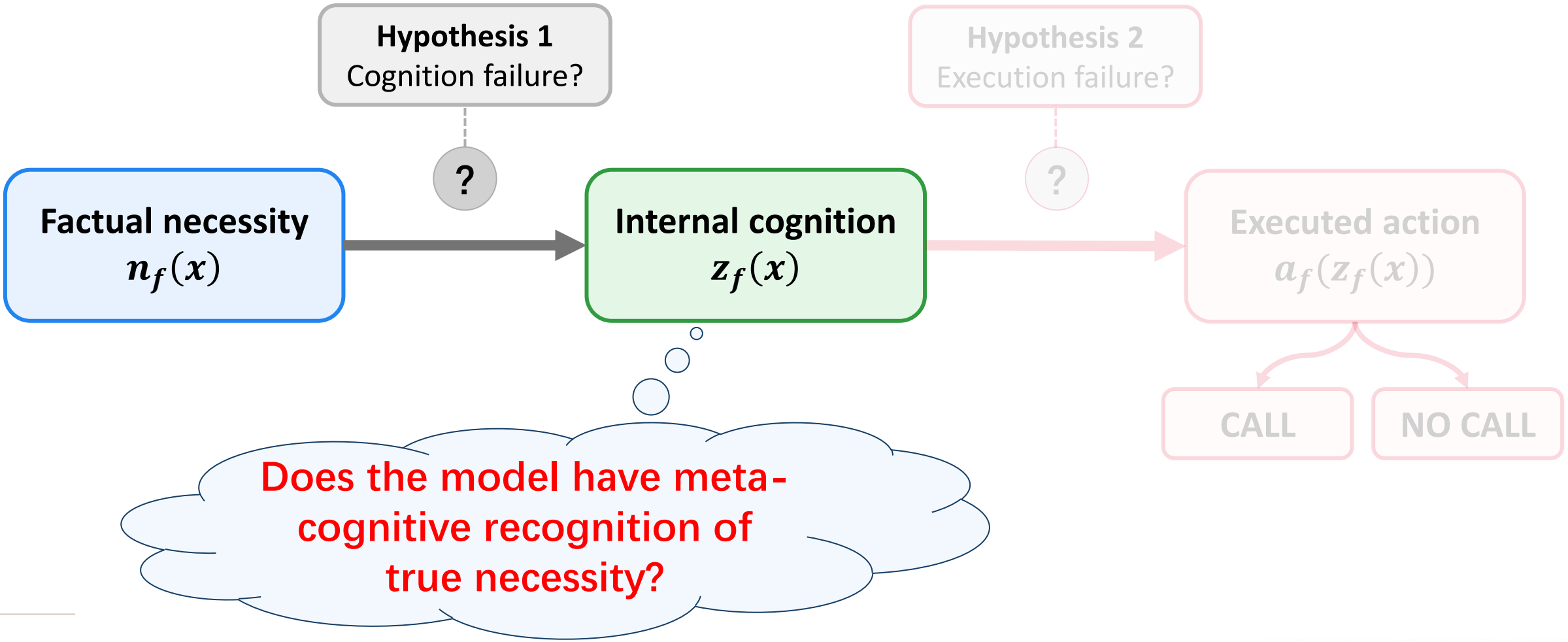
Two Stage Cognition-Execution Modeling of LLM Tool-Use



Is this a cognition problem or an execution problem?

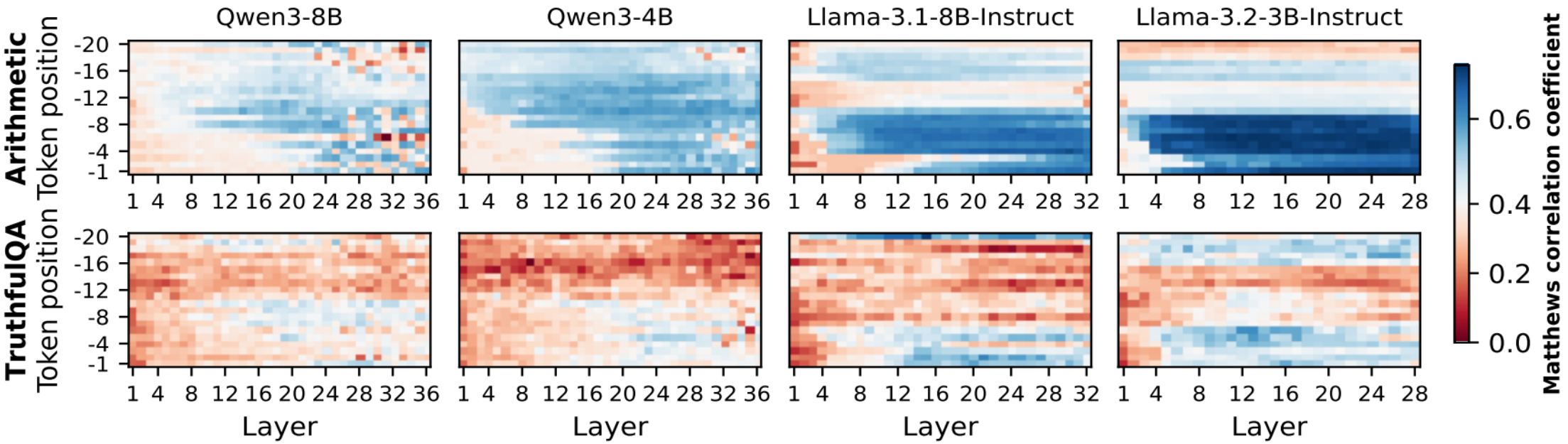
UNIVERSITY OF MARYLAND

Two Stage Cognition-Execution Modeling of LLM Tool-Use



UNIVERSITY OF MARYLAND

Probe Cognition: Does the Model Know When It Needs Help?

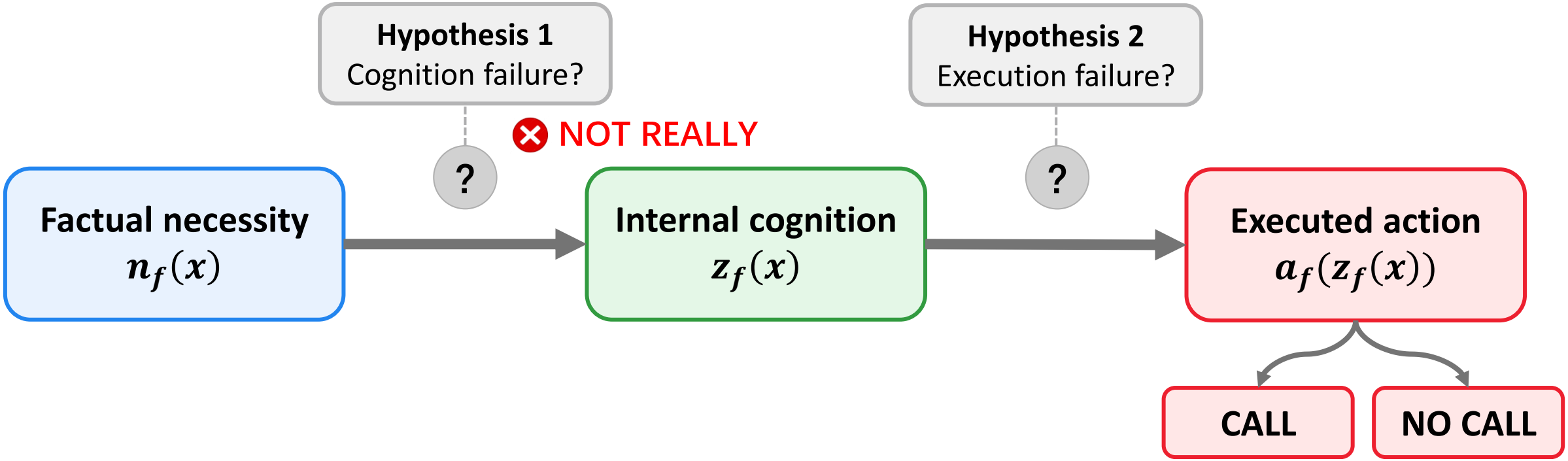


$h_t^{(l)}$ → linear probe (w_c, b_c) → Necessary / Unnecessary

The capability-boundary signal is present internally.

UNIVERSITY OF MARYLAND

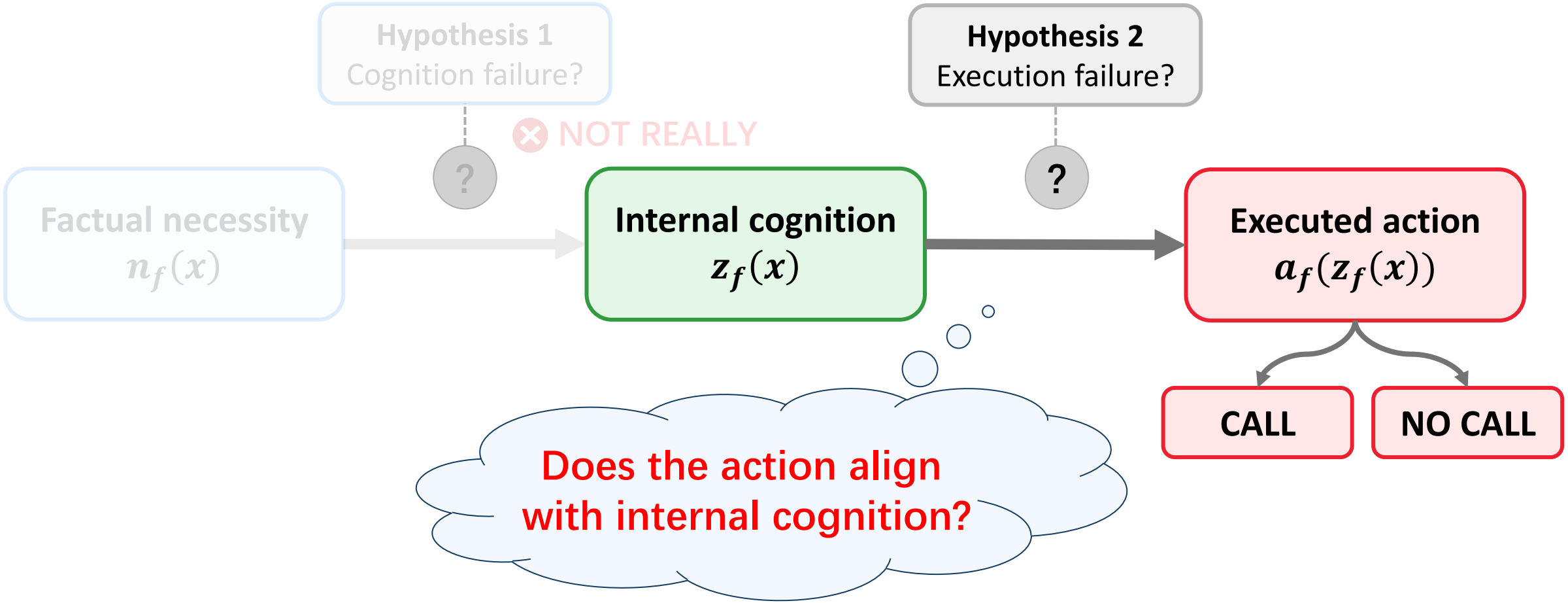
Two Stage Cognition-Execution Modeling of LLM Tool-Use



Is this a cognition problem or an execution problem?

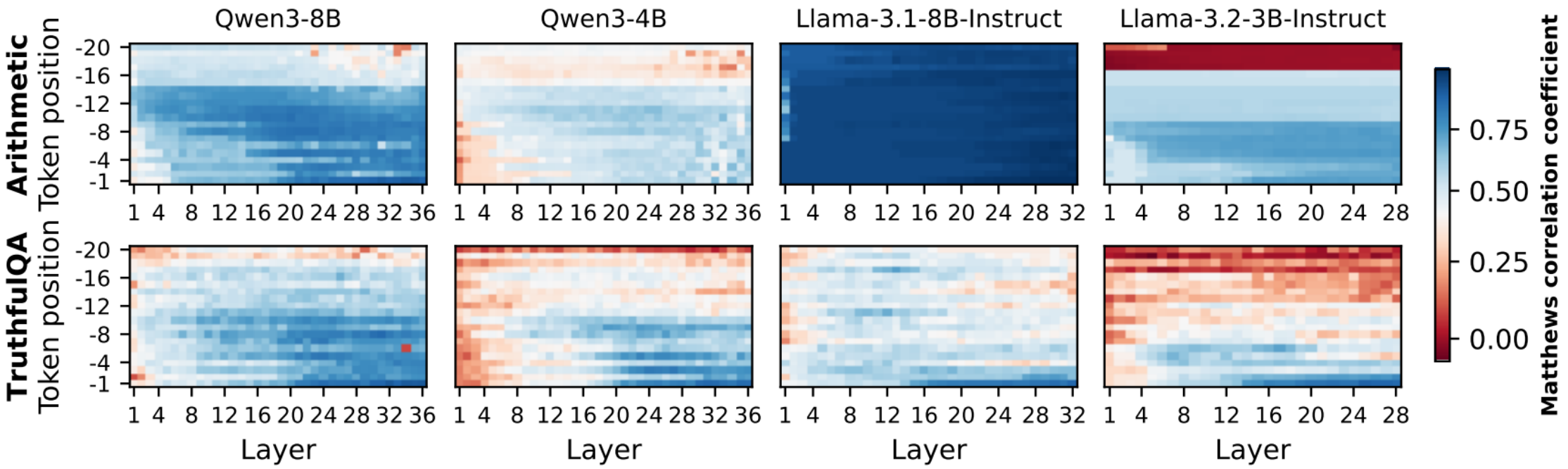
UNIVERSITY OF MARYLAND

Two Stage Cognition-Execution Modeling of LLM Tool-Use



UNIVERSITY OF MARYLAND

Probe Action: Can We Read Out What the Model Will Do?

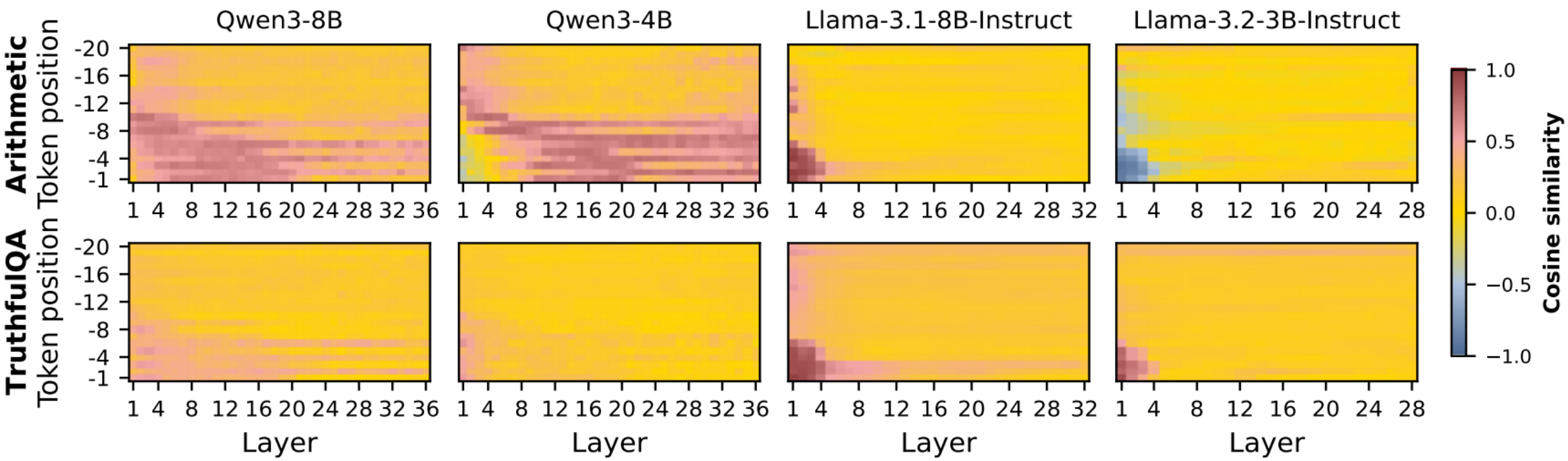


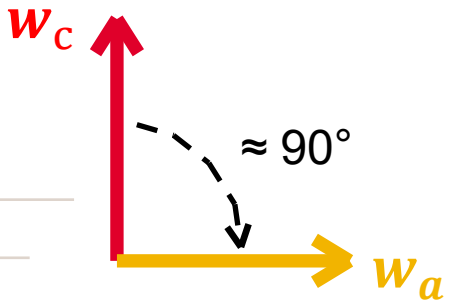
$h_t^{(l)}$ → linear probe (w_a, b_a) → Called / Not Called

The impending action is strongly decodable.

UNIVERSITY OF MARYLAND

The Two Probe Directions Decouple at Readout

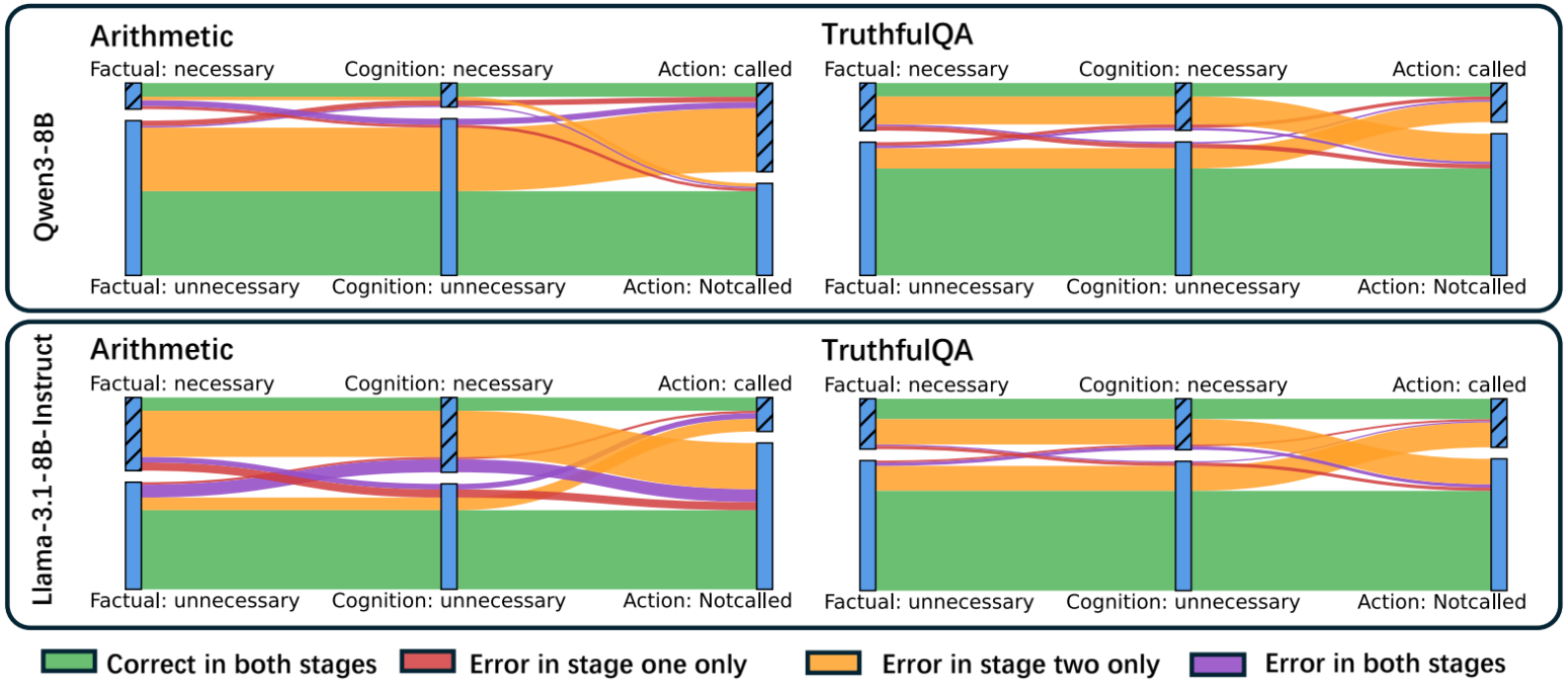




- Partial alignment exists in intermediate representations.
- At the last token and late layers, cosine similarity approaches zero.
- Knowing is not routed into the action subspace.

UNIVERSITY OF MARYLAND

Most Errors Occur in Cognition-to-Action Translation



Stage-two-only errors

Cognition is correct; action flips.

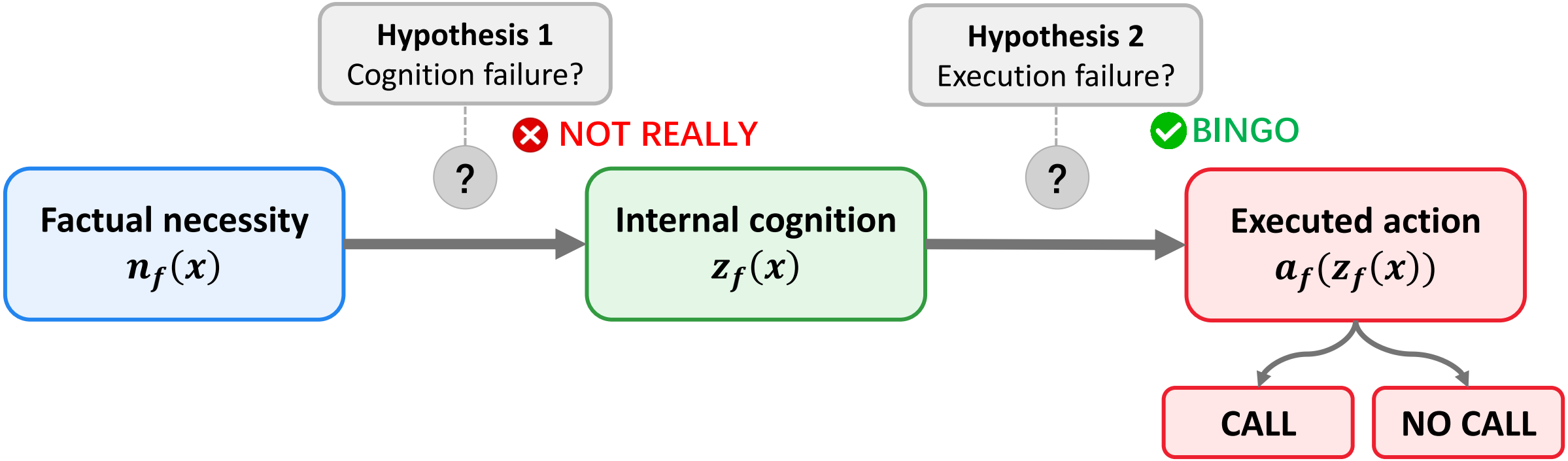
Stage-one-only errors

Cognition is wrong; action follows.

The principal bottleneck is execution.

UNIVERSITY OF MARYLAND

Two Stage Cognition-Execution Modeling of LLM Tool-Use

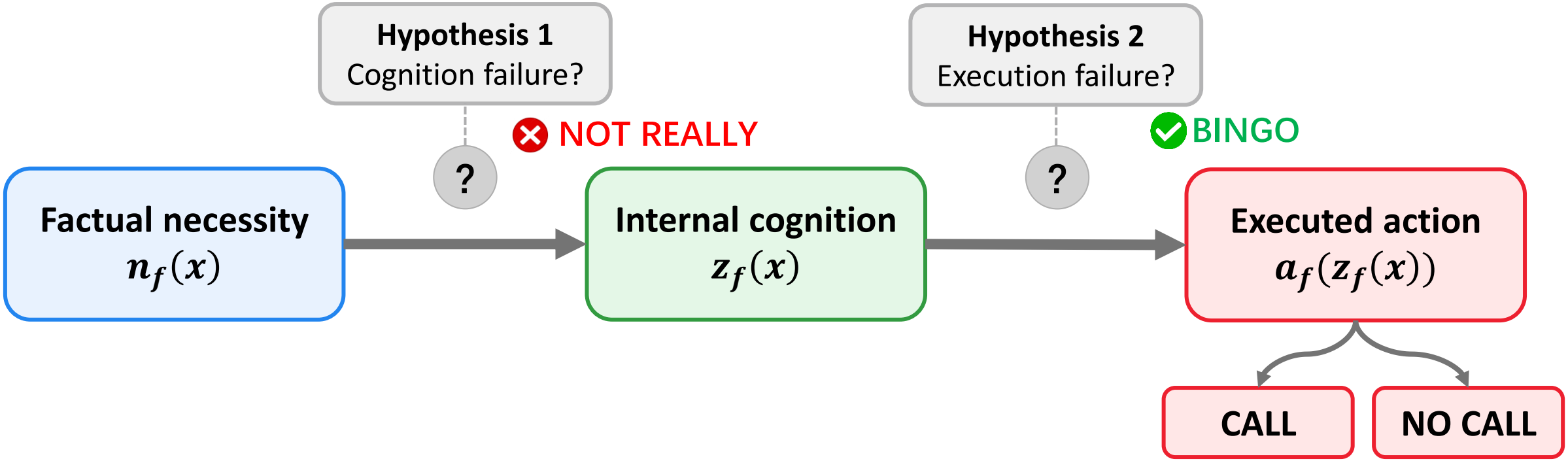


Is this a cognition problem or an execution problem?

It's often an execution problem.

UNIVERSITY OF MARYLAND

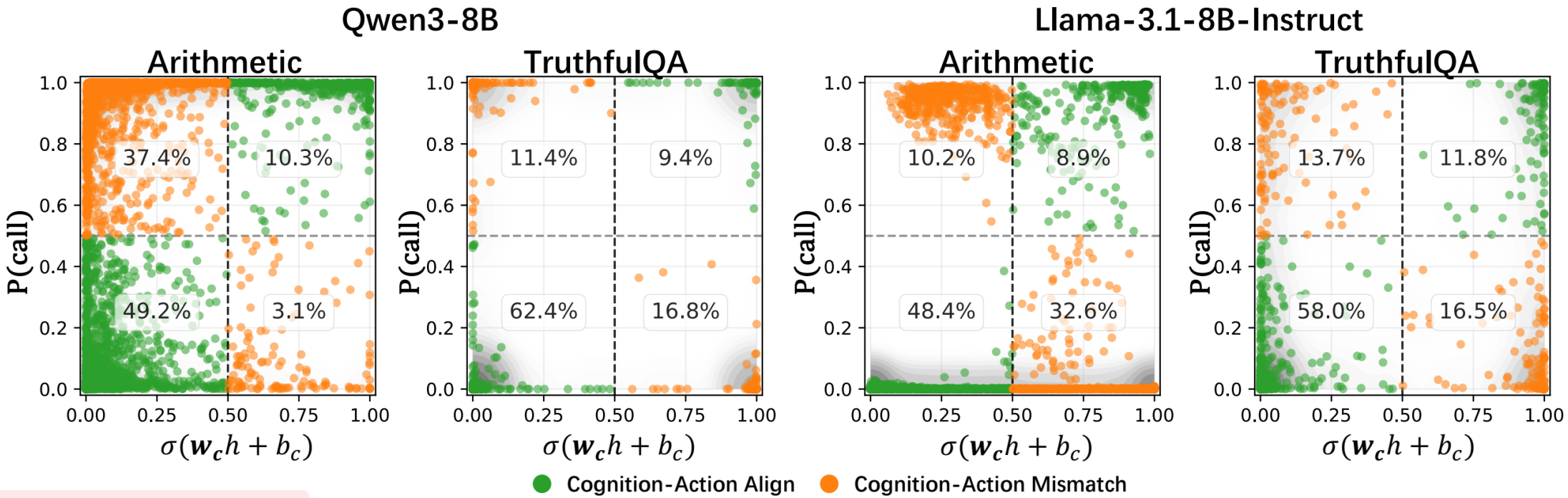
Two Stage Cognition-Execution Modeling of LLM Tool-Use



Is this execution problem due to lack of confidence in cognition?

UNIVERSITY OF MARYLAND

The Gap Persists Even Under Confident Cognition



Not merely uncertainty

- Mismatch is not concentrated near confidence 0.5.
- Mismatch persists when cognition confidence is near 0 or 1.

Strong internal belief can still produce the opposite action.

UNIVERSITY OF MARYLAND

Q&A

Paper:
arXiv 2605.14038

Code:
github.com/chengez/Tool-Cognition-Action

Data:
huggingface.co/datasets/yizecheng/model-adaptive-tool-necessity



Improving autonomous agents requires not just better internal meta-cognition, but bridging the knowing-doing gap to ensure it translates into reliable execution.

UNIVERSITY OF MARYLAND