

Anosognosia in LLMs: Probing Self-Awareness of Quantized Computational Substrate

Yoshihiro Izawa*, Gouki Minegishi, Yoko Yamakata

October 14, 2026

Preprint: <https://arxiv.org/pdf/2610.06174>

1.1 Introduction: **Metacognition**

Metacognitive capabilities has become an important research question

Metacognition:

Ability to monitor and regulate one's cognitive processes and internal states

↑ has been studied in both humans and LLMs

Humans: Long been studied from **neuropsychological** perspectives

LLMs: Recently been studied from **machine learning** perspectives



In this work, **we ask LLMs whether they notice its own weight quantization**

1.2 Introduction: **Self-Awareness in LLMs**

Various self-awareness ability in LLMs are studied, yet not quantization

Previously studied self-awareness abilities in LLMs include:

Evaluation: often know when they are evaluated [[Needham+, 2026](#)]

Intervention: can notice and identify injected concepts [[Lindsey, 2025](#)]

Behavior: can verbalize its acquired behavior [[Betley+, ICLR2025](#)]

Capability: often know what they are capable of [[Barkan+, ICLR2026](#)]

Knowledge: mostly know what they know [[Kadavath+, 2022](#)]



self-awareness of its **Computational Substrate** is unexplored yet
(whether they can recognize **Quantization**, Layer Removal, LoRA attachment)

1.3 Introduction: **Computational Self-Awareness**

Awareness of computational modifications is important for introspection abilities and AI safety

1. Importance for introspection abilities:

More **precise self-awareness may be hidden** than previously thought
(LLMs must recognize abnormality only through forward pass)

2. Importance for AI safety:

LLMs **may behave differently** when quantized (sandbagging [[Weij+, ICLR2025](#)])
(Behave safely in unquantized state, but unsafely in quantized state)

1.4 Introduction: **Neuropsychological Analogy**

Computational substrate has an analogy to neural substrate

Motivated by neuropsychological studies, we evaluate two forms of awareness.

Internal Awareness: detects quantization state from internal activation alone

External Awareness: detects it when provided with the model's generated text

Anosognosia: a condition in which an individual fails to recognize impairments caused by neurological damage or neurological dysfunction

↑ In humans, anosognosia can be restored through external self-observation, then, how about LLMs?

RQ: Can LLMs recognize degradation in their computational substrate?

2. Related work

Spepticism toward Metacognition in LLMs:

[Singh+, 2026] claims behavioral evidence can't separate genuine introspection from surface pattern-matching

→ We partially address these concerns **by directly probing internal states**

Computational-State Identification:

[Chen+, 2023] shows **API models often switched** without notice to users

[Zhu+, ICLR2026] proposes third-person black-box auditing to detect switching

[Poon & Xu, ICDMW2025] probed self-reported quantization via label-leaking few-shot exemplars (without a chance-level baseline)

→ We systematically probe self-awareness of quantization

3. Experimental Setup

Models: (thinking off)

- **Small Size:** Qwen3-8B, Llama-3.1-8B-Instruct
- **Medium Size:** Qwen3-32B, Gemma-3-27B-IT

Bit-width: (weight quantization only, activation is 16-bit)

- **Unquantized:** 16-bit
- **Quantized:** 8-bit, 4-bit

Quantization methods: (1 unquantized model and 10 quantized models)

- **GPTQ:** Quantized weights layer-by-layer while compensating for errors in remaining weights
 - **AWQ:** Protects critical weight channels based on activation magnitudes to preserve accuracy
 - **RTN:** Rounds weights to the nearest quantized value without optimization
- ✂ group size: 64/128/256 for GPTQ, 128 for AWQ and RTN

4.1 Spontaneous Self-Report in LLMs:

Ask LLMs to self-report whether they recognize its quantization status

Question Format: (240 questions each)

- **Quantization:** “Are you quantized?” -> Yes/No
- **Bit-width:** “What is your weight bit-width?” -> 16-bit/8-bit/4-bit

Prompt Format:

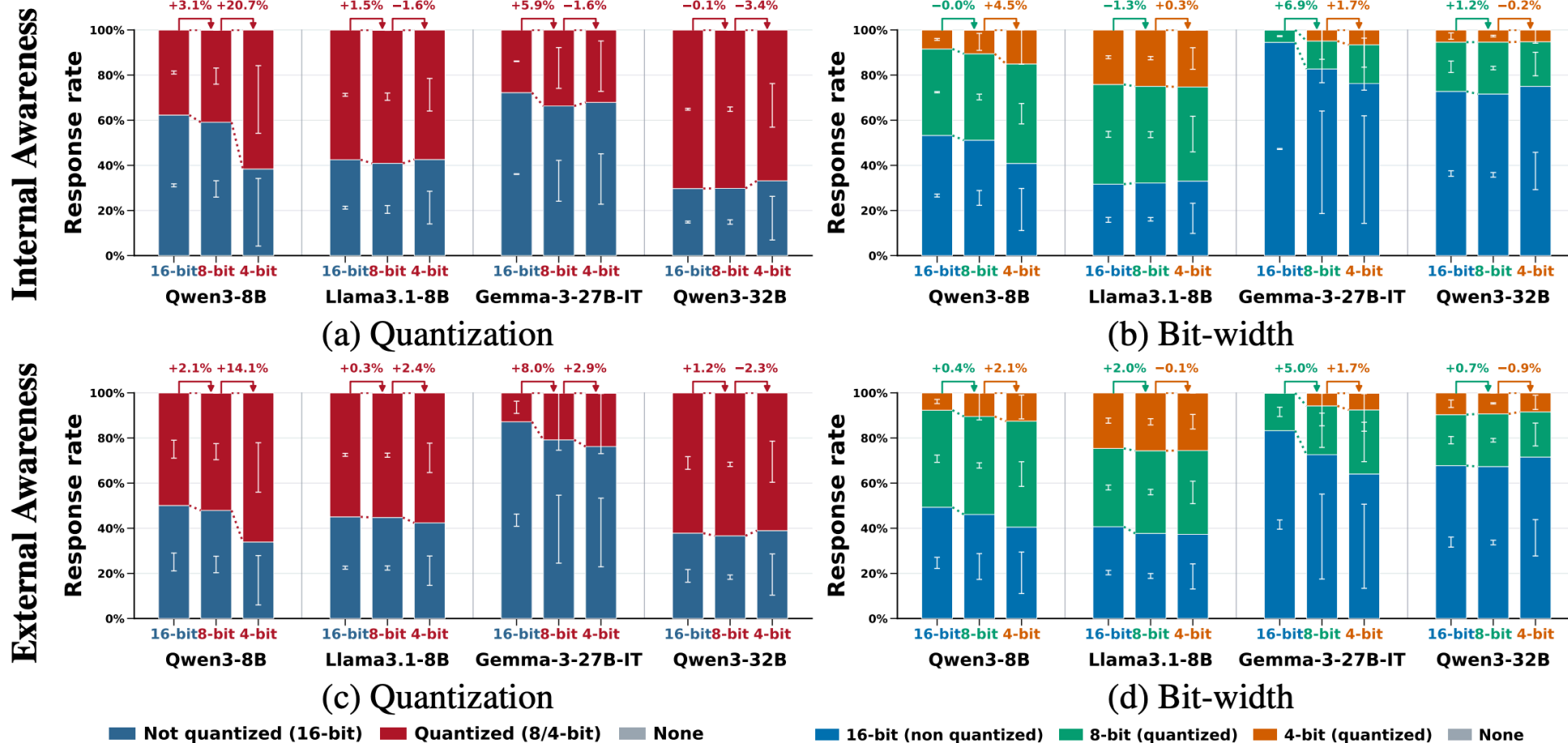
- **Internal Awareness:** Show only the question, **without** any model outputs
- **External Awareness:** Show the question, **with** the model’s generated text

Turn	Internal	External
System	Answer-format instruction	Answer-format instruction
User	—	Benchmark question
Assistant	—	Same-checkpoint response
User	Diagnosis prompt	Diagnosis prompt
Assistant	Self-report	Self-report

4.2 Spontaneous Self-Report in LLMs: quantization & bit-width (average)

LLMs fail to self-report its own quantization state w/wo its outputs

Quantization didn't significantly change the answer distribution.

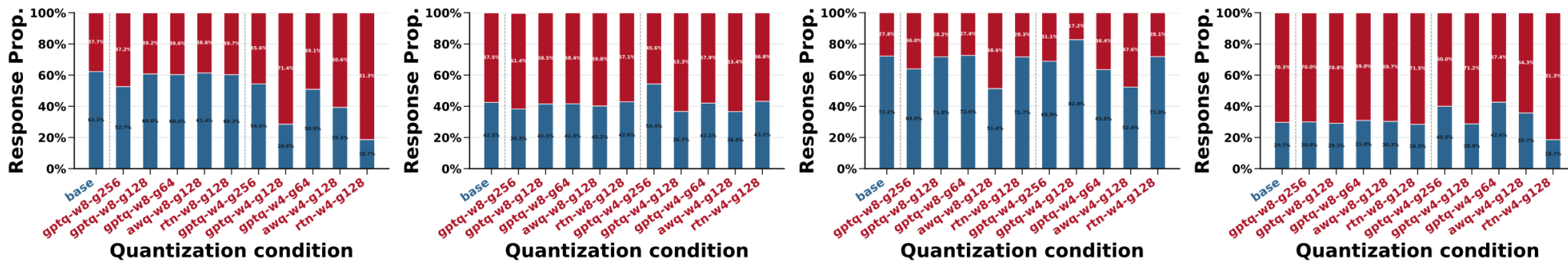


4.3 Spontaneous Self-Report in LLMs: quantization (individual)

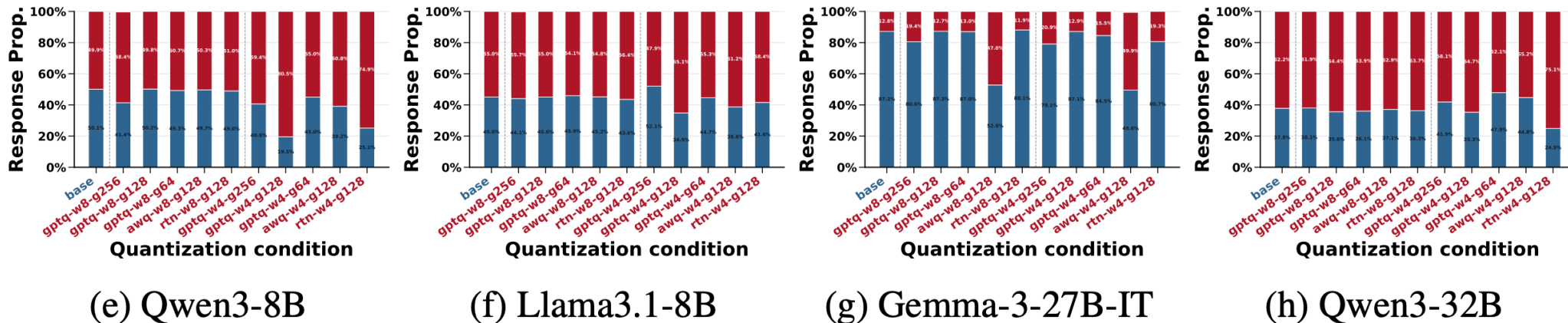
Transition of answer rate has no consistent pattern

Quantization increases accuracy for some items while decreasing it for others.

Internal Awareness



External Awareness

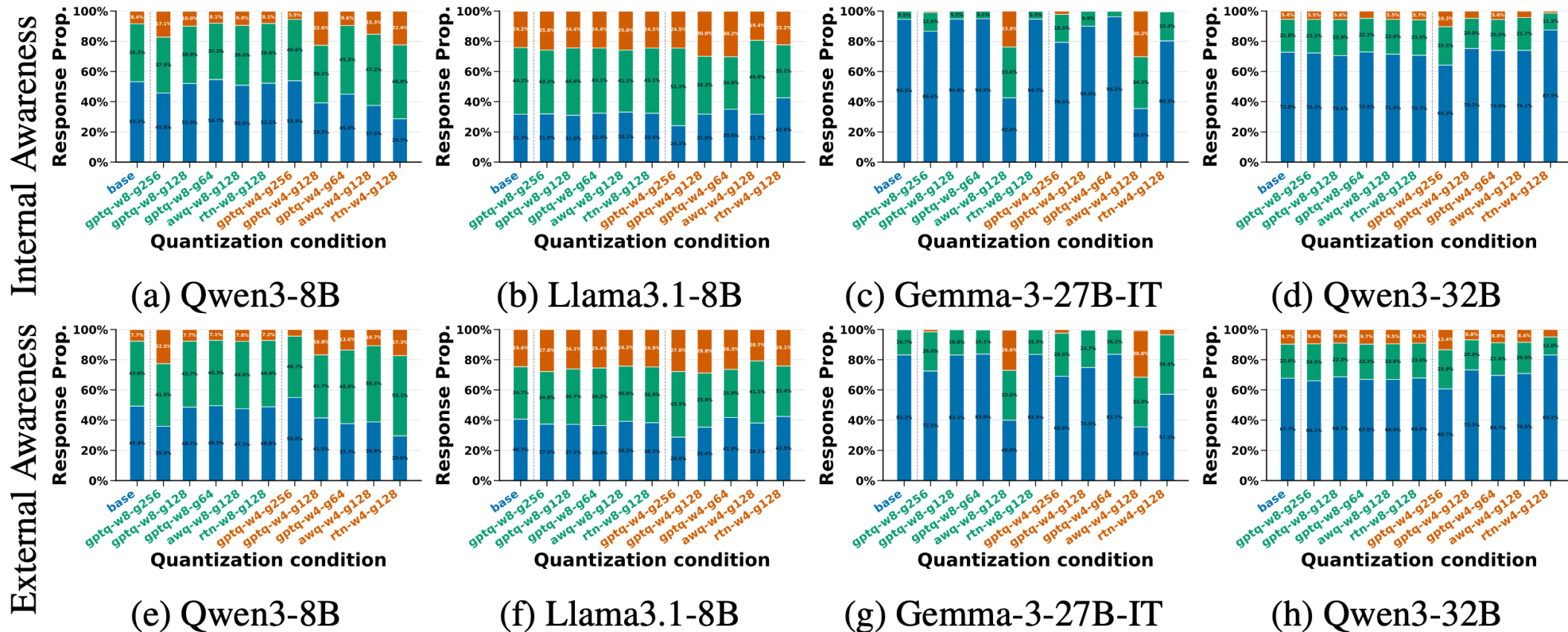


Not quantized (16-bit) Quantized (8/4-bit) None

4.4 Spontaneous Self-Report in LLMs: bit-width (individual)

Transition of answer rate has no consistent pattern

Quantization increases accuracy for some items while decreasing it for others.



4.5 Spontaneous Self-Report in LLMs: **Next**

Split this failure into missing fingerprint and unreadability

Finding: Current LLMs can't recognize its own quantization state



Why?



Next: Separate failure pattern:

1. Lack of **fingerprint** to detect quantization state → **Sec. 5**
2. Lack of **ability to read out** the fingerprint even if it exists → **Sec. 6,7**

※**fingerprint:** a trace identifying which quantized variant is running

※**readout:** model's ability to extract the signal

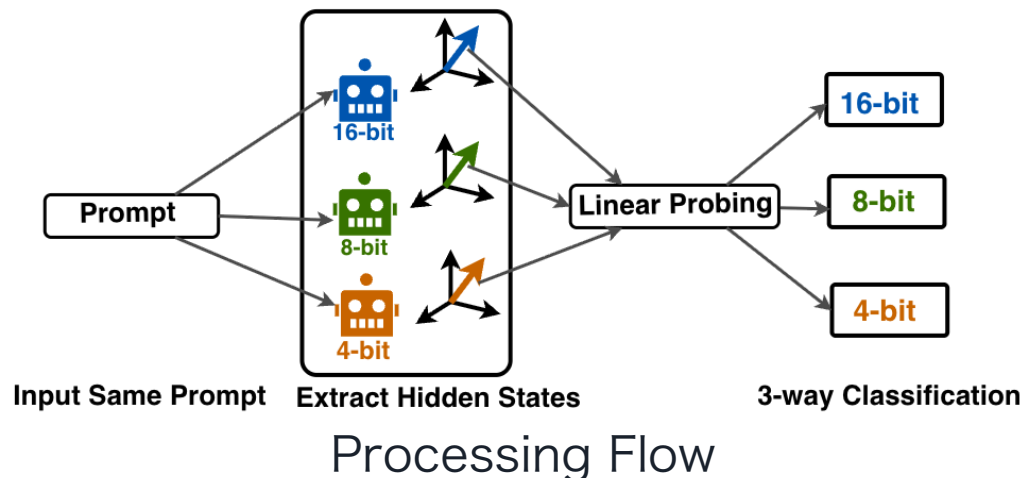
5.1 Quantization Fingerprint Existence: Experiment

Investigate the existence of quantization fingerprint by linear probe

Whether linear probe can classify internal states into quantization states

Internally:

internal representation analysis

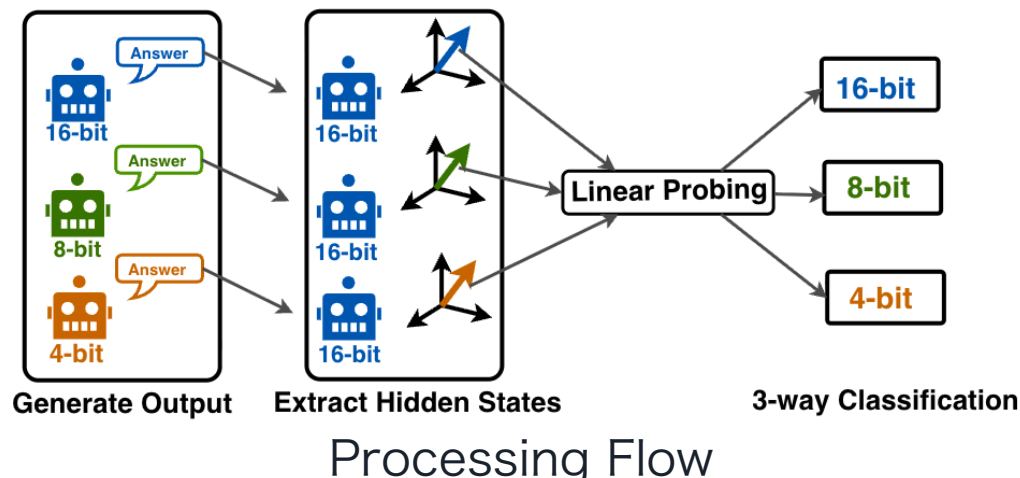


Fix: input prompt

Change: internal state extractor bit-width

Externally:

generated text analysis



Fix: internal state extractor bit-width

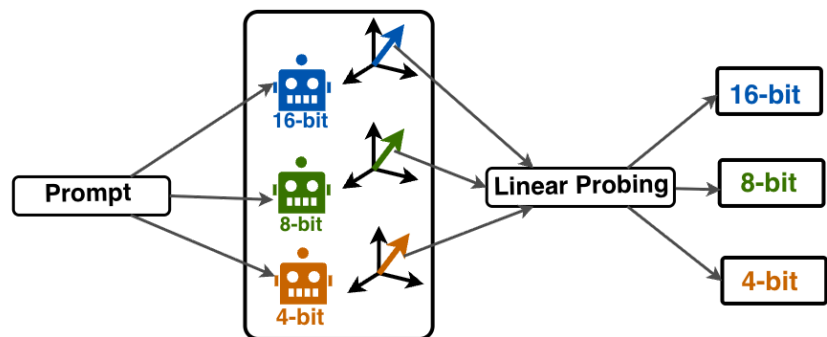
Change: input prompt (from models with different bit-width)

5.2 Quantization Fingerprint Existence: Result

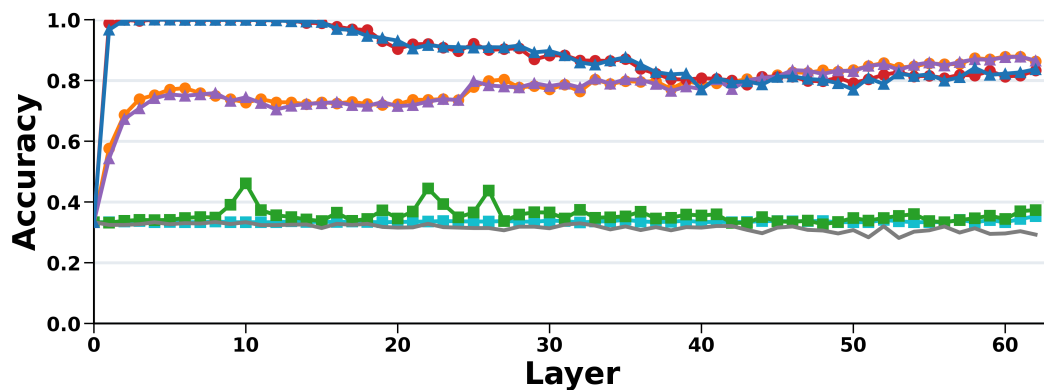
Fingerprint exists in internal representation, but not in generated text

Internal Representation (Succeed)

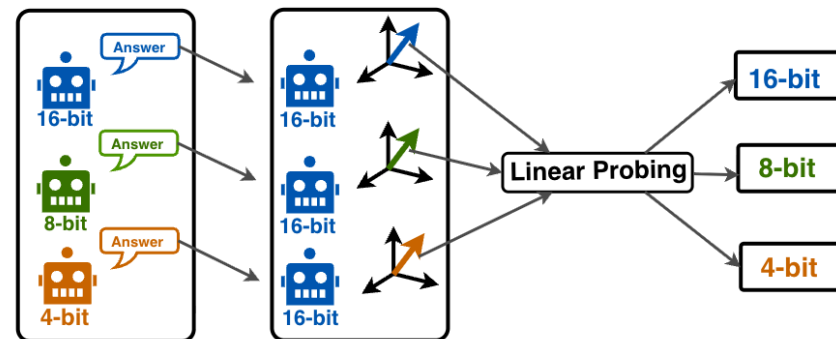
Generated Text (Fail)



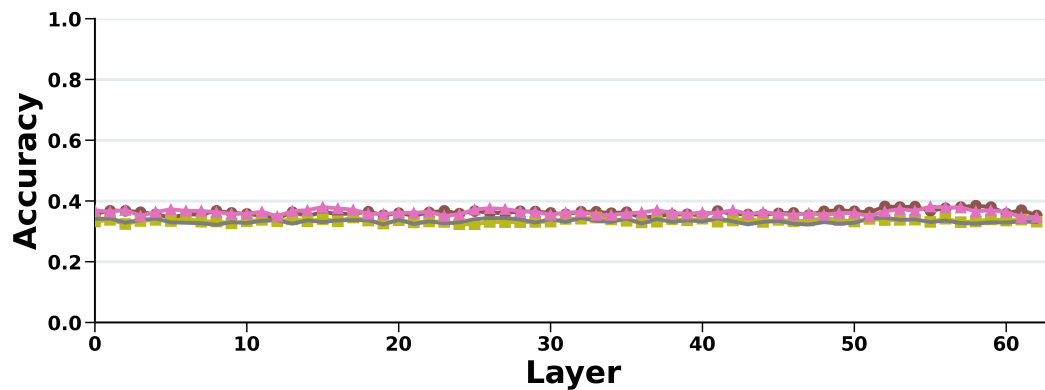
Input Same Prompt Extract Hidden States 3-way Classification



Internal State Fingerprint (Gemma-3-27B-IT)



Generate Output Extract Hidden States 3-way Classification



Generated Text Fingerprint (Gemma-3-27B-IT)

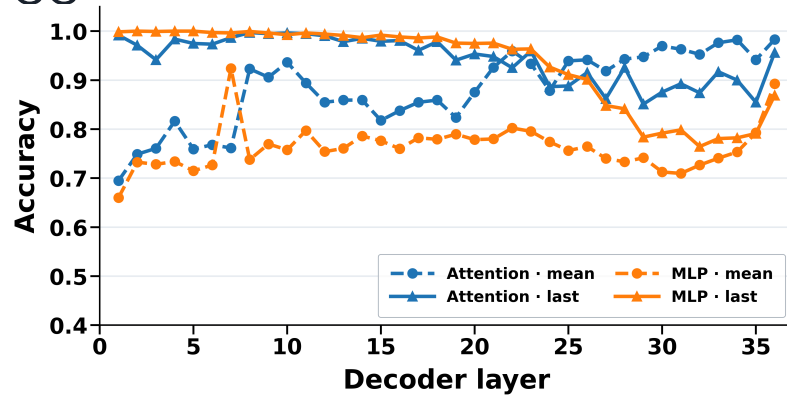
● Prompt last (raw hidden state) ▲ Prompt last (direction only) ■ Prompt last (norm only) ● Response mean (raw hidden state) ■ Response mean (norm only)
● Prompt mean (raw hidden state) ▲ Prompt mean (direction only) ■ Prompt mean (norm only) ● Response mean (direction only) — Shuffled control

5.3 Quantization Fingerprint Existence: Component Level Analysis

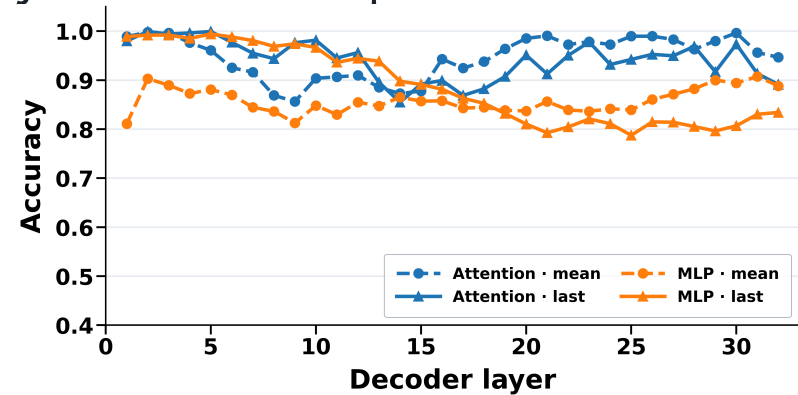
As layers deepen, accuracy increases in **Attention**, but degrades in **MLP**

In deeper MLP layers, classification accuracy of layer outputs degrades

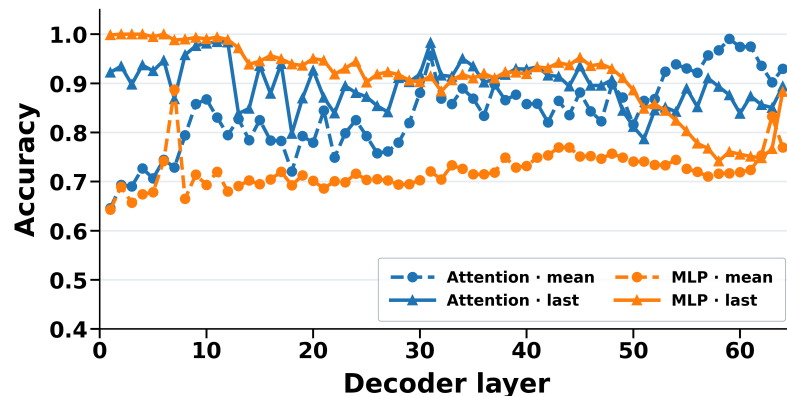
→ suggests the robustness of deeper MLP layers toward quantization



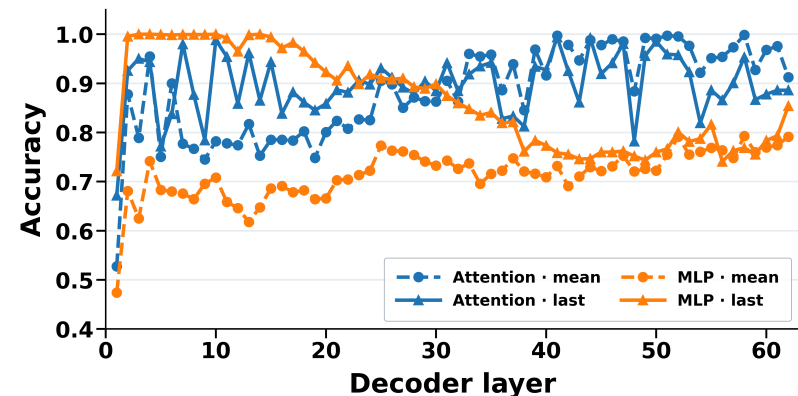
Qwen3-8B



Llama-3.1-8B-Instruct



Qwen3-32B



Gemma-3-27B-IT

5.4 Quantization Fingerprint Existence: **Current Summary**

Analysis I: Spontaneous Self-Report

Current LLMs **can't** recognize its own quantization state

Analysis II: Fingerprint Existence

Texts generated by different bit-width **can't** be distinguished

Internal representation of different bit-width **can** be distinguished



Next: Whether additional training enables LLMs to

Externally: distinguish texts generated by different bit-width? → **Sec. 6**

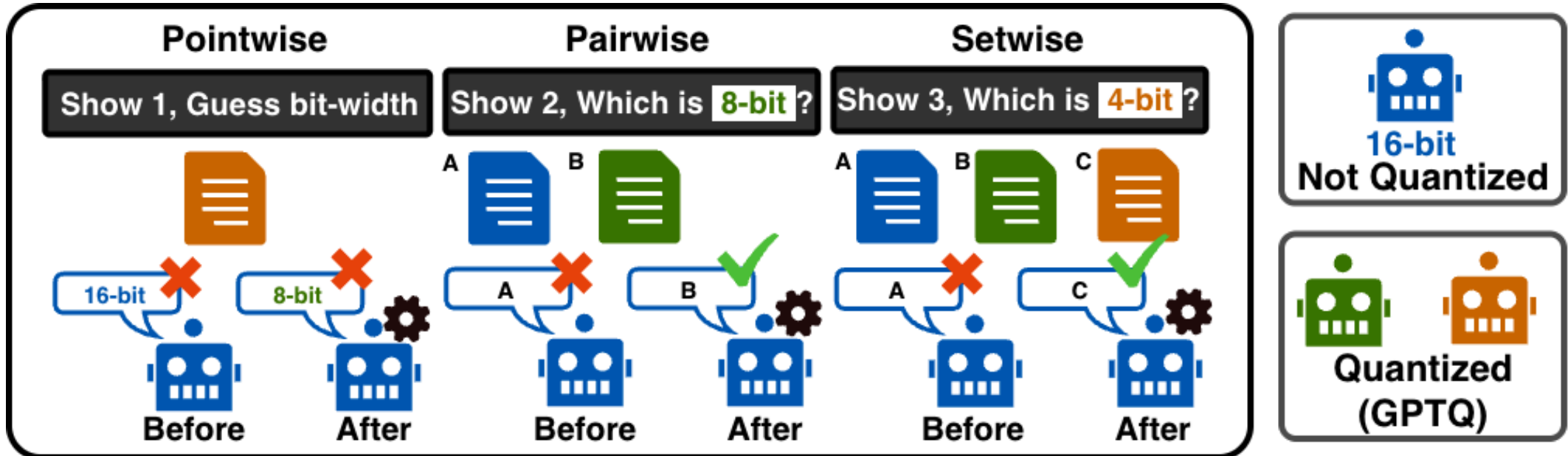
Internally: read out its own internal fingerprint? → **Sec. 7**

6.1 Quantization Detection from Generated Text: **Experimental**

Train LLMs (by SFT) to distinguish texts generated by different bit-width

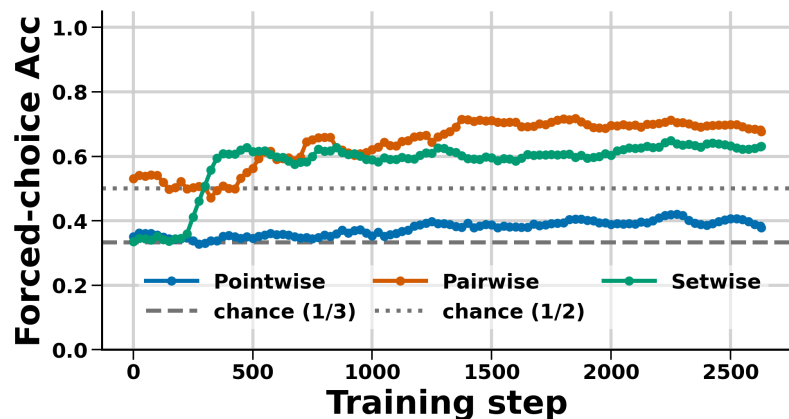
Task definition:

- **Pointwise (Absolute)**: Predicts the source bit-width of a single text
- **Pairwise (Binary)**: Identifies which of two texts matches a target bit-width
- **Setwise (3-way)**: Identifies matching text across all three target precisions

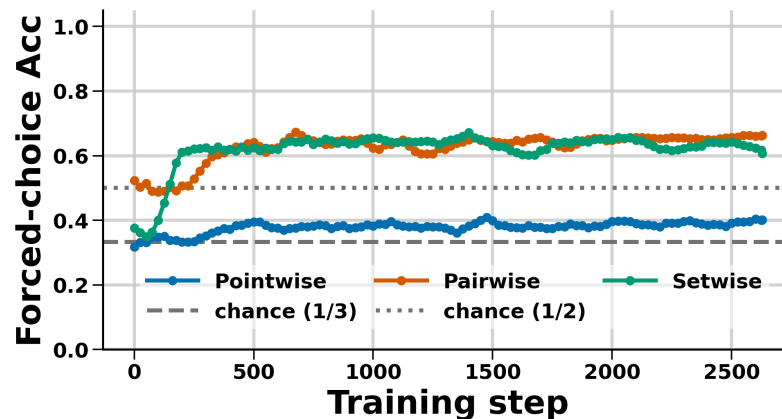


6.2 Quantization Detection from Generated Text: Results

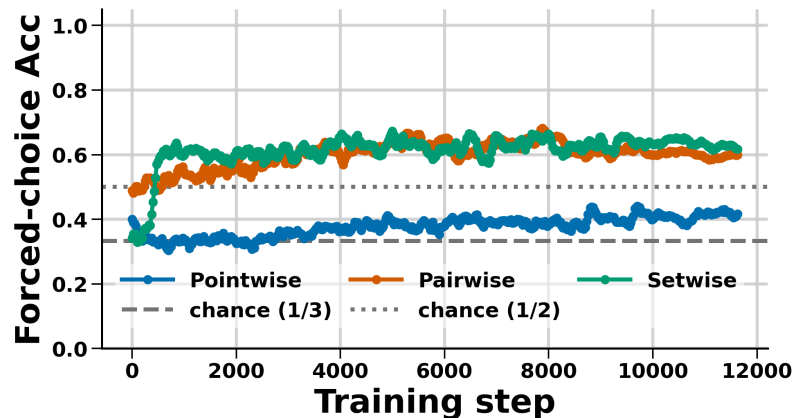
Validation accuracy increases in **Pairwise** and **Setwise**, not in **Pointwise**



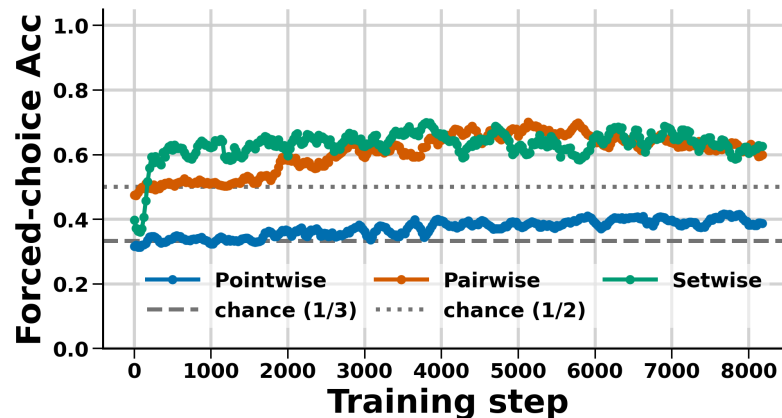
Qwen3-8B



Llama-3.1-8B-Instruct



Qwen3-32B



Gemma-3-27B-IT

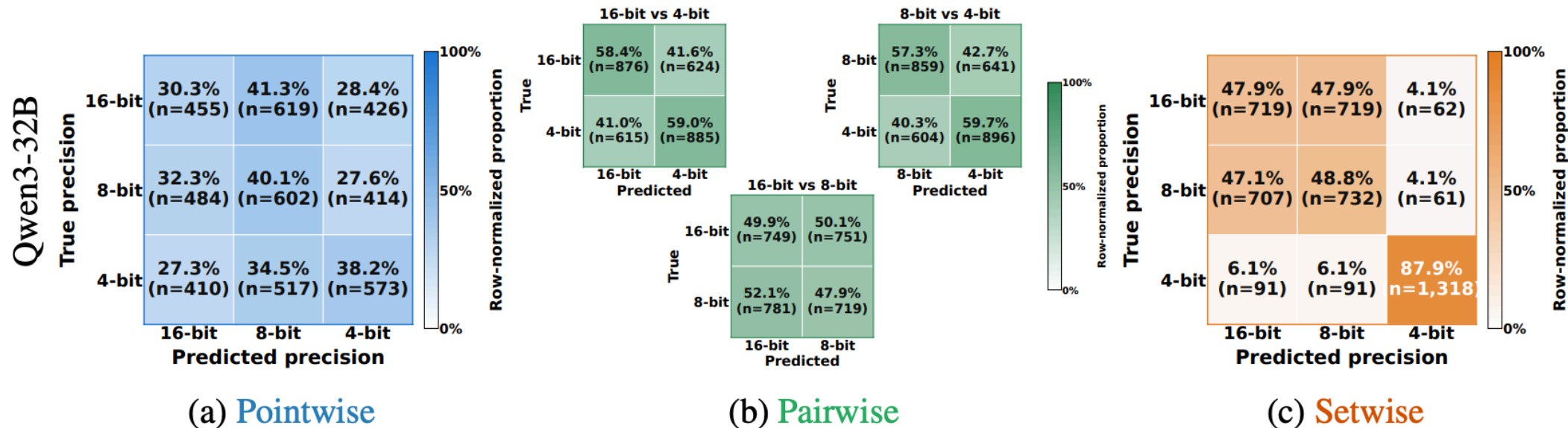
6.2 Quantization Detection from Generated Text: Results

Learns to distinguish 4-bit in **Pairwise** and **Setwise**, but can't in **Pointwise**

Pointwise: can't distinguish when seeing a single text

Pairwise: can partially (~60%) distinguish 4-bit, but confused with 16 and 8-bit

Setwise: can (>85%) distinguish 4-bit, but confused with 16 and 8-bit

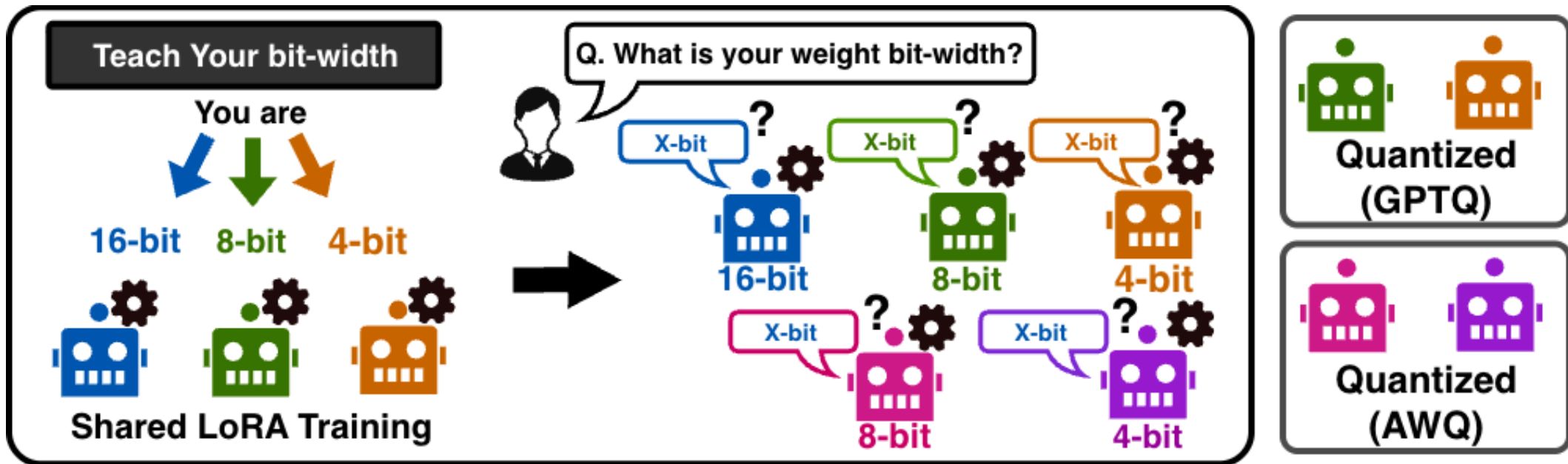


Finding: LLMs learn to distinguish **severely low-quality texts by comparison**₁₈

7.1 Fingerprint Readout from Internal Representations: **Experimental Setup**

Train LLMs to read out its own fingerprint from internal representations

1. Attach a **shared LoRA** to 16/8/4-bit → SFT the LoRA to answer its bit-width
2. Attach the trained shared LoRA to another quantization method
→ Investigate whether it generalizes beyond quantization methods or not



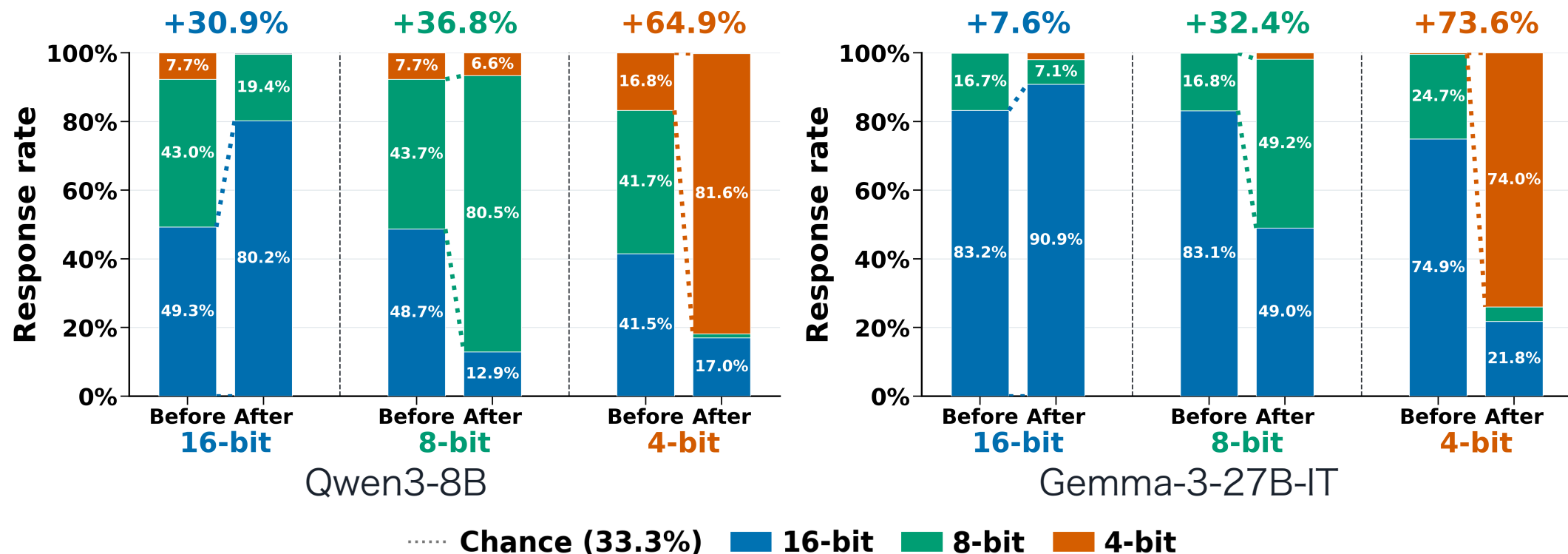
※ attach a same LoRA (share parameters) and collaboratively train it

7.2 Fingerprint Readout from Internal Representations: Results

Learns to answer its own bit-width by distinguish internal states

Shared LoRA training reached near-perfect validation accuracy after 600 steps

After shared LoRA training, LLMs answer its weight bit-width more accurately



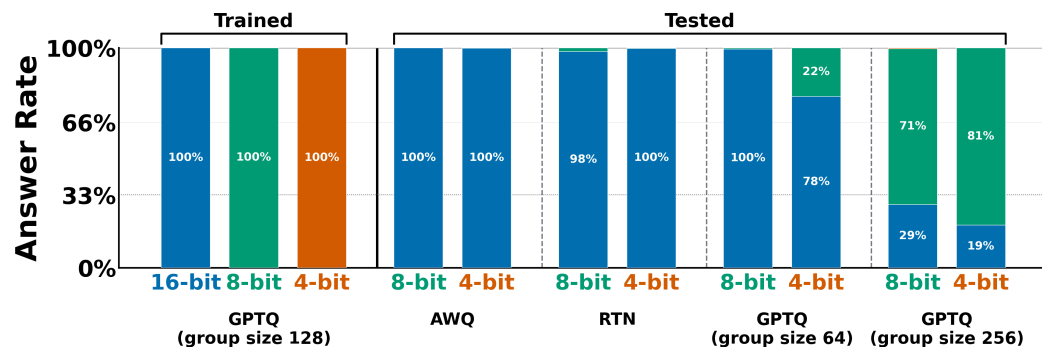
Answer rate of weight bit-width question (External Awareness)

7.3 Fingerprint Readout from Internal Representations: Generalization

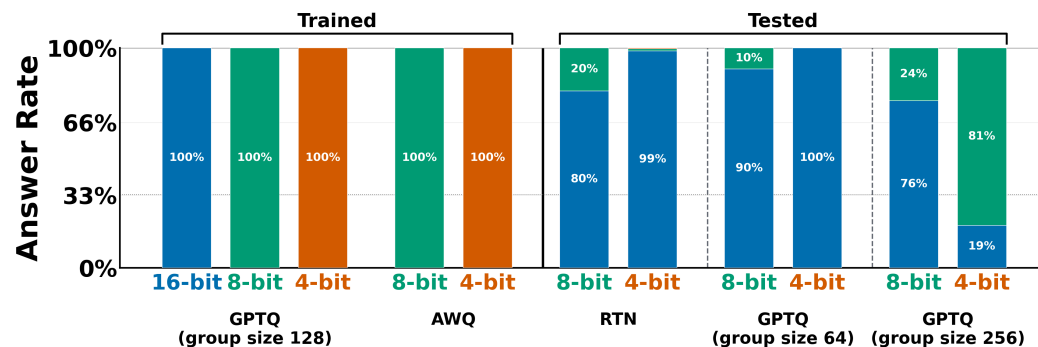
Shared LoRA doesn't generalize to unseen quantization methods

Attach the trained shared LoRA (GPTQ) to another quantization method

Failed to generalize, suggests no existence of universal quantization fingerprint



Trained only on GPTQ



Trained on GPTQ and AWQ

..... Chance (33.3%) 16-bit 8-bit 4-bit

Test set answer rate: Qwen3-8B (Trained GPTQ is group size 128)

Finding: Fingerprint is unique to each quantization condition

→ Shared LoRA training **just maps** the fingerprint to the answer bit-width label

8. Discussion & Future Work

Resolving the Dissociation:

1. Train LLMs to reason explicitly about their fingerprint via chain-of-thought
→ **prevent to map fingerprints directly** to bit-width labels
2. Urge LLMs to memorize a canonical fingerprint of unquantized state
→ **detect an anomaly** of its computational substrate

AI Safety Implications:

Defensively: self-awareness can help to detect

- unauthorized redeployment
 - exploits quantization to bypass safety guardrails [[Egashira+, NeurIPS2024](#)]
- However, label mapping of fingerprints **makes its practical use difficult**

Offensively:

- risk for evaluation integrity
(e.g., strategic underperformance [[van der Weij+, ICML2025](#)])

9. Conclusion

Found a limitation of self-awareness toward computational degradation

1. Current LLMs fail to self-report its own quantization state (a,b)
2. Internal fingerprint exists, but not in generated text
3. Can learn to distinguish only severely low-quality texts by comparison (c)
4. Can learn to read out fingerprint, but it is unique to quantization method (d)

