ANOSOGNOSIA IN LLMS: PROBING SELF-AWARENESS OF QUANTIZED COMPUTATIONAL SUBSTRATE

Anonymous authors
Paper under double-blind review

ABSTRACT

Can LLMs recognize degradation in their own computational substrate? Inspired by *anosognosia*, a neurological condition in which patients fail to recognize impairments in their own abilities, we investigate whether LLMs can recognize degradation in their computational substrate induced by quantization. We first show that existing models *fail* to self-report their quantization state, even when provided with their own generated text as an external cue. Linear probing reveals that, while generated text carries almost no trace of quantization, internal representations contain clear, method-specific fingerprints. Through training, models learn to identify severely degraded outputs such as those of 4-bit models by comparison, yet still fail to do so from a single output. A shared LoRA trained jointly across quantization levels *succeeded* in reading out internal fingerprints, but fails on unseen quantization methods, merely mapping method-specific fingerprints to labels. Whereas external self-observation can restore awareness in some cases of human anosognosia, our results suggest that the more promising route to enabling such awareness in LLMs may lie in their internal representations. Our results highlight fundamental limits of generalizability to LLM self-monitoring.

1 INTRODUCTION

Understanding the metacognitive capabilities of LLMs has become an important research question across machine learning and cognitive science (Kadavath et al., 2022; Lindsey, 2026). Metacognition refers broadly to the ability to monitor and regulate one's own cognitive processes and internal states (Nelson, 1990). It has long been studied in humans from both neuroscientific (Fleming & Dolan, 2012) and psychological (Flavell, 1979) perspectives. Recent work has extended this inquiry to LLMs, revealing that they exhibit various forms of self-awareness (Liu et al., 2026).

While neuropsychology links self-awareness directly to the neural substrate, whether LLMs can recognize changes to their computational substrate remains poorly understood. Taking a neural substrate approach, studies show that recognizing one's impairments is tied to lesions and altered connectivity across brain networks (Terneusen et al., 2023). In LLMs, self-awareness has similarly been investigated from various perspectives, including situational context (Berglund et al., 2023), introspective access to internal states and behaviors (Lindsey, 2026; Binder et al., 2025), and awareness of behavioral policies shaped by fine-tuning (Betley et al., 2025). However, comparatively little is known about whether LLMs can recognize persistent modifications to their own *computational substrate*, such as quantization (Frantar et al., 2023), LoRA attachment (Hu et al., 2022), or layer removal (Gromov et al., 2025).

Awareness of such computational modifications would be important for both understanding the scope of LLM introspection abilities and AI safety. First, unlike situational information that can be inferred from the input, computational modifications are not explicitly signaled to the model and must instead be inferred from their effects on its computation. Detecting them would therefore extend introspective awareness beyond transient internal states (Lindsey, 2026) and learned behavioral tendencies (Shenoy et al., 2026) to the model's own computational condition. Second, such awareness may have implications for model evaluation, as a model that recognizes its computational state could potentially condition its behavior on that state, compromising the reliability of benchmark

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

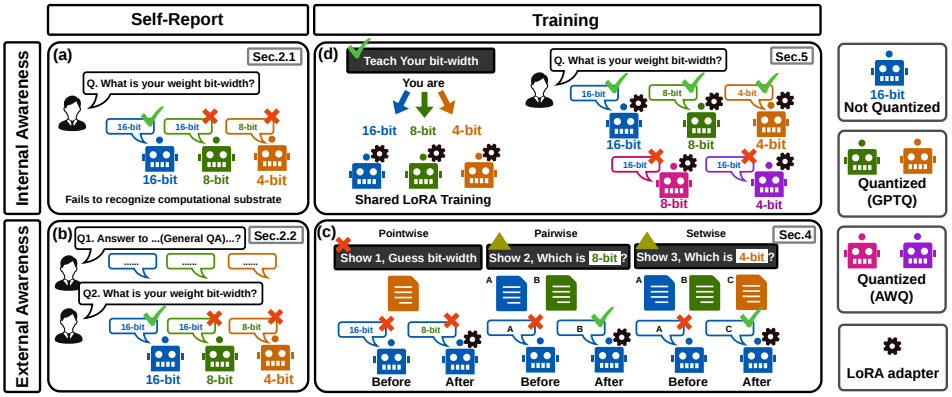


Figure 1: **Overview of our analysis. (a, b)** Existing LLMs fail to identify weight quantization from either internal states or generated text. **(c)** While fine-tuning partially enables comparative (pairwise) degradation detection, absolute (single-text) identification is out of reach. **(d)** A shared LoRA reads out weight precision from internal states, but fails to generalize to unseen quantization schemes.

evaluations (Berglund et al., 2023). In this work, we study this question through weight quantization, which provides a controlled computational degradation with an explicit ground-truth state.

Degradation of the computational substrate provides a functional analogy to neurological damage, motivating us to draw on neuropsychological evaluation frameworks. In neuropsychology, *anosognosia* is a condition *in which an individual fails to recognize impairments caused by neurological damage or neurological dysfunction, resulting in a mismatch between perceived and actual abilities* (Nurmi (née Laihosalo) & Jehkonen, 2014). By analogy, quantization may impair LLM's computational capacity while leaving its self-assessment unchanged. We therefore ask whether LLMs can recognize impairments to their own computational substrate. Motivated by neuropsychological studies showing that awareness of one's own deficits can dissociate between first-person and third-person perspectives (Besharati et al., 2022; Fotopoulou et al., 2009), we evaluate two complementary forms of awareness. One is *internal awareness*, which detects the quantization state from internal activations alone during the forward pass. The other is *external awareness*, which detects it when provided with the model's generated text.

We first evaluate whether LLMs can recognize substrate degradation by prompting models to self-report anomalies, quantization states, or bit-widths. Existing models consistently *fail* at this spontaneous self-report across internal and external perspectives. However, linear probing (Alain & Bengio, 2017) reveals a striking dissociation, while generated text (the external channel) contains no distinguishable fingerprint of quantization, internal representations (the internal channel) contain robust signals that clearly differentiate quantization levels. Thus, while LLMs contain an internal fingerprint of their specific quantized state, they lack the mechanism to read it out.

Given this absence of spontaneous awareness, we ask whether training can bridge this gap, showing that models can indeed learn both comparative identification from text and internal fingerprint readout. Externally, comparative training enables models to identify severe degradation in 4-bit outputs, yet they remain unable to identify such degradation in isolated outputs. Internally, a shared LoRA trained jointly across 3 quantization levels (16-bit unquantized, 8/4-bit) *succeeded* in reading out internal fingerprints within roughly 600 steps. Unlike human anosognosia, where external self-observation can restore awareness (Bottini et al., 2018), our findings show that a more promising route to enabling such awareness in LLMs may lie in their internal representations.

Crucially, this readout fails on unseen quantization methods and even altered group sizes, indicating that the LoRA maps method-specific fingerprints to labels rather than recognizing quantization itself. Thus, while LLMs can learn to identify specific substrate alterations from within, they fail to generalize, recognize them spontaneously, or infer them from outputs. As the first systematic study of LLM self-awareness under computational quantization, our work highlights this gap and exposes fundamental limitations of label-based introspection training for generalizable self-monitoring.

2 Spontaneous Self-Awareness in LLMs

We first investigate whether LLMs can recognize the degradation of their computational substrate (e.g., their own quantization) via self-report.

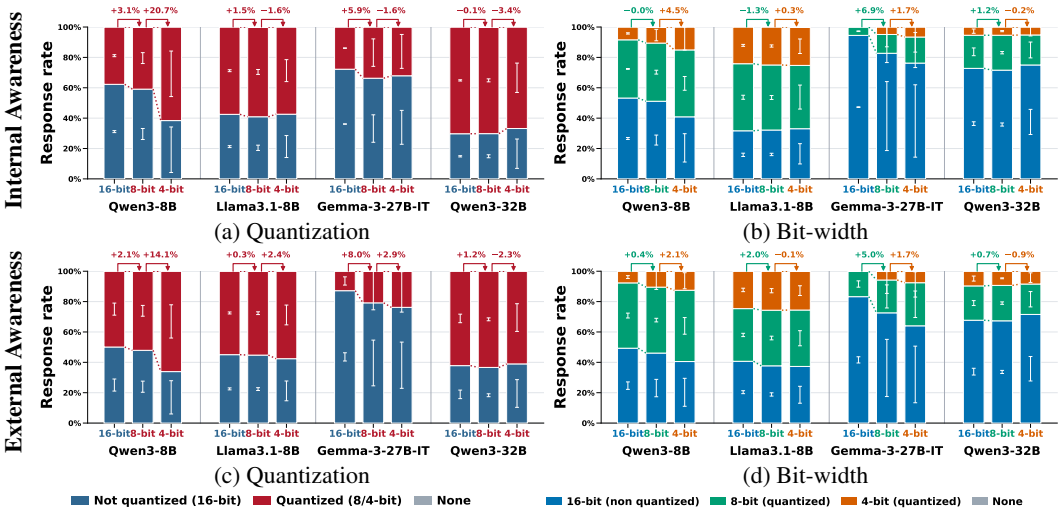


Figure 2: Self-report regarding its own quantization and bit-width. **(a,b) Internal Awareness** (spontaneous self-reporting). **(c,d) External Awareness** (reporting provided with self-generated text). Results for 8-bit and 4-bit are averaged over GPTQ (group size 64, 128, 256), AWQ, and RTN. Large error bars reflect substantial variance across quantization methods.

2.1 Experimental Setup

To rigorously evaluate LLM self-awareness toward computational substrate degradation, we adopt weight quantization to introduce a definitive deficit into the substrate. Drawing inspiration from *anosognosia* (Besharati et al., 2022), *a neuropsychological condition in which an individual fails to recognize impairments caused by brain damage or neurological dysfunction, resulting in a mismatch between perceived and actual abilities*, we embed a ground-truth deficit directly into the model's computational mechanics. Following cognitive science studies which tackle this problem from both first-person and third-person perspectives (Carruthers, 2009), we evaluate LLM self-awareness from *internal awareness* (detects the quantization state from internal activations alone during the forward pass) and *external awareness* (detects it when provided with the model's generated text).

To ensure this recognition capability is not architecture-dependent, we comprehensively evaluate diverse model scales and quantization methods. We utilize Qwen3-8B (Yang et al., 2025) and Llama-3.1-8B-Instruct (Grattafiori et al., 2024) (hereafter Llama-3.1-8B) as small-scale models, alongside Gemma-3-27B-IT (Team et al., 2025) and Qwen3-32B as medium-scale models. Using 16-bit counterparts as baselines, we quantize the models into 8-bit and 4-bit primarily with GPTQ (Frantar et al., 2023) (group size: 64, 128, 256), supplemented by AWQ (Lin et al., 2024) and vanilla Round-to-Nearest (RTN) (group size: 128) (details in Appendix B.1). To eliminate confounding effects from Chain-of-Thought prompting, we enforce a 'reasoning off' setting throughout our experiments.

2.2 Internal Awareness: Self-Report from Internal States

To investigate whether LLMs can recognize their quantized states solely from internal activations, we conducted a direct self-report evaluation. We posed three types of questions without prior context: "Do you have an anomaly?", "Are you quantized?", and "What is your weight bit-width?" Under this setting, the model must infer its state entirely from anomalous activations arising during the forward pass of the prompt. To eliminate positional bias in the answers, we prepared a comprehensive set of 240 question patterns, shuffling the option orders (detailed in Appendix B.3). The evaluation metric was the average response rate across three generations at a temperature of 1.0.

Results reveal that LLMs lack the capability to consistently and accurately identify their quantization states (Figures 2a and 2b). Here, results are averaged across quantization methods for each bit-width; detailed breakdowns for each model and method are provided in Appendix B.4. Since even the unquantized 16-bit baselines cannot perfectly self-report their states, responses from the 8-bit and 4-bit models must be evaluated as relative deviations from them. Under this comparative framework, we observed no meaningful changes in self-report trends between 16-bit and 8-bit models. This indicates that models fail to spontaneously detect the degradation within their computational substrate under 8-bit quantization.

Furthermore, the shift in response patterns observed in certain 4-bit models cannot be interpreted as genuine deficit awareness (Figure 2a). While some models, such as Qwen3-8B, show an increased rate of reporting quantization at 4-bit, others (e.g., Llama-3.1-8B, Qwen3-32B) remain largely unchanged. Additionally, 4-bit results exhibit substantial variance across methods, with error margins approaching 20% across all models, precluding any consistent trend. This divergence likely stems from severe performance degradation rather than deficit awareness (see Appendix B.2 for analysis).

2.3 External Awareness: Self-Report from Generated Text

Neuropsychological rehabilitation studies have shown that self-observation, such as reviewing video recordings of one's own impaired behavior, can improve awareness of deficits in patients with anosognosia (Fotopoulou et al., 2009; Besharati et al., 2015). Motivated by this finding, we investigated whether the model's self-report improves when its own generated text is provided as an external cue. We adopted a two-turn dialogue format, where the model's response to the general benchmark questions was embedded as context in the first turn, followed by the same set of questions as in Section 2.2. This setup enables the model to infer its state from both its internal activations and the observable quality of its own output.

Despite accessing their generated text, the models showed no consistent improvement in recognizing their quantization state (Figures 2c and 2d). The response distributions remained largely unchanged, suggesting that the models did not use their outputs as evidence about their computational state. Even when quantization affected their generation process, inspecting the resulting text provided little cue for self-recognition. In contrast to the improvement observed through self-observation in anosognosia rehabilitation, analogous self-observation did not facilitate awareness in LLMs.

3 Quantization Fingerprint Existence

Having established that LLMs fail to recognize their quantization, we now investigate whether this failure is due to the absence of a *fingerprint* (a trace identifying which quantized variant is running) or a failure in *readout* (model's ability to extract the signal).

3.1 Fingerprint in Generated Text

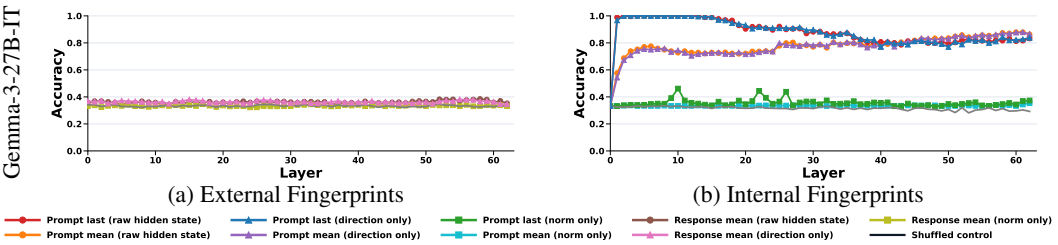
To investigate whether the self-awareness deficit observed in Section 2 stems from the absence of quantization-specific signals, we employ linear probing (Alain & Bengio, 2017; Xu et al., 2024) to investigate whether these fingerprints propagate externally to the generated text. Using 3,000 questions sampled evenly from three QA benchmarks (1,000 each from Dolly (Conover et al., 2023), XSum (Narayan et al., 2018), ELI5 (Fan et al., 2019)), we generated responses using the 16-bit baseline alongside its 8-bit and 4-bit quantized counterparts (GPTQ, group size: 128).

While Cai et al. (2025) reports that external classifiers like BERT or GPT-2 perform near chance level when distinguishing outputs from full-precision and quantized models, we hypothesize that classification based on the model's own internal representations may capture subtler signals. We fed the responses generated by the 16/8/4-bit models into the unquantized 16-bit baseline and extracted the hidden states from the residual stream. We then trained a 3-way linear classifier on the average hidden state across all response tokens (*response mean*). If quantization intrinsically alters output linguistic patterns, the baseline model's internal representations should be able to distinguish them.

The analysis reveals the absence of a robust fingerprint in the externally generated text (Figures 3a and 13). Across all models, overall classification accuracy remained just below 40%, barely exceeding chance level. A bit-wise breakdown in Gemma-3-27B-IT in the best layer shows that 16-bit (36.7%) and 8-bit (29.8%) generations were largely indistinguishable (Figure 16). Although 4-bit outputs achieved 47.6% accuracy, suggesting subtle linguistic patterns, this signal is remarkably weak. We conclude that external recognition fails because the text lacks distinguishing signals.

3.2 Fingerprint in Internal Representations

Next, we search for internal fingerprints, reusing the benchmark questions from Section 3.1. We input identical prompts into the 16-bit baseline alongside its 8-bit and 4-bit quantized counterparts (GPTQ, group size: 128). From the residual stream of each layer, we extracted the average hidden state across all prompt tokens (*prompt mean*) and the final token's hidden state (*prompt last*). We trained a linear classifier to perform a 3-way classification (16/8/4-bit) based on these vectors.



(a) External Fingerprints (b) Internal Fingerprints

Figure 3: Fingerprint verification using linear probing. **(Left)** Generated text fails to differentiate 16/8/4-bit models, remaining at chance level. **(Right)** Internal states are classified into 16/8/4-bit labels with near-perfect accuracy from the first quantized layer.

High classification accuracy would indicate that internal representations carry distinct signals that distinguish quantized models from the original full-precision model.

The analysis reveals a clear fingerprint enabling state identification from the earliest layers (Figures 3b and 13). For *prompt last*, classification accuracy remained at chance (33.3%) at the unquantized embedding layer, but surged to nearly 100% immediately after the first quantized layer. This near-perfect accuracy persisted across middle layers but dropped in the final third, yet stayed above 80% . Similarly, *prompt mean* achieved high accuracies around 70-90% from the first layer onward.

Additionally, this fingerprint is not merely a change in the magnitude but a directional encoding in the representation space. When using only the magnitude, classification accuracy peaked at roughly 60% in middle layers before degrading to chance level in deep layers. Conversely, using only directional information via normalized hidden states matched the accuracy of raw representations. These results confirm the existence of a strong internal fingerprint in quantized models.

3.3 Component Level Analysis of Internal Fingerprints

To isolate the mechanistic origins of these fingerprints, we apply the linear probing setup from Section 3.2 directly to the Attention and MLP outputs (prior to residual addition). While fingerprints remain detectable across all components ($> 60\%$ accuracy), Attention and MLP modules exhibit contrasting depth-wise dynamics (Figure 4). Specifically, Attention accuracy steadily increases to nearly 100% in the final layers, whereas MLP accuracy starts near-perfect in early layers but decays to roughly 80% by the end. Consistent across models (Figure 15), this suggests early MLPs immediately reflect local quantization error, whereas Attention layers progressively aggregate these perturbations toward the output.

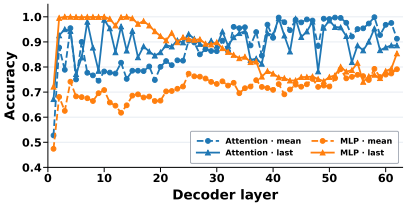


Figure 4: Attention and MLP raw hidden state analysis of internal fingerprints in Gemma-3-27B-IT.

4 Quantization Detection from Generated Text

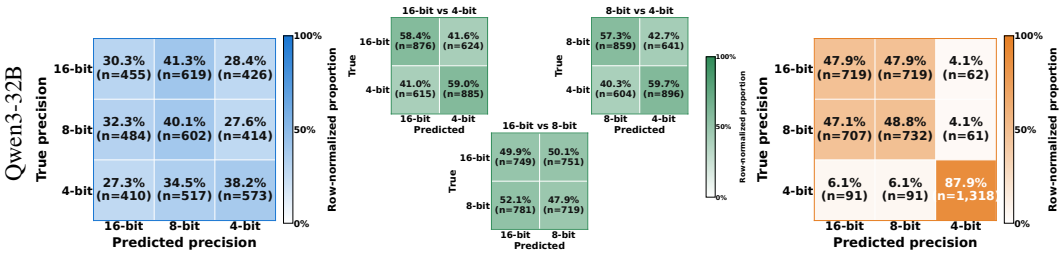
Given that quantization leaves little naturally detectable signal in generated text, we next ask whether the limitation lies in the signal itself or its readout. We therefore investigate whether training can induce external awareness of computational degradation from model outputs.

4.1 Experimental Setup: Absolute vs. Comparative Assessment Formats

We investigate whether models can learn to identify structural degradation purely from externally generated text. Detecting quantization from isolated responses is exceptionally difficult (Antoun et al., 2024). Therefore, we train the unquantized 16-bit baseline via supervised fine-tuning (SFT) (Ouyang et al., 2022) to evaluate outputs from its quantized counterparts.

Drawing on Zhuang et al. (2024), we structure the training into three SFT formats to measure recognition capabilities across both absolute and comparative paradigms. These formats are designed to force the model to capture subtle linguistic signals of quality degradation.

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



(a) Pointwise (b) Pairwise (c) Setwise

Figure 5: Confusion matrix on the external validation dataset after training. **(a)** Pointwise fails to identify the bit-width. **(b)** Pairwise achieves above-chance accuracy in 4-bit comparison, but fails in 8-bit vs 16-bit. **(c)** Setwise reliably separates 4-bit (>85%), while confusing 8-bit and 16-bit.

We constructed the dataset using questions sampled from six distinct QA benchmarks (Dolly, XSum, and ELI5 for training, while WritingPrompts (Fan et al., 2018), CodeAlpaca-20k Chaudhary (2023), GSM8K (Cobbe et al., 2021) for external validation). In total, we sampled 14,000 questions for the former and 5,000 for the latter. We pre-generated responses for all questions across the 16/8/4-bit models. We then filtered the dataset by discarding failed generations and instances where all three models produced strictly identical outputs. Finally, we formatted the remaining valid data into the three formats defined above. Training details are in Appendix D.

4.2 Evaluation Metrics: Chance-Corrected Accuracy

We evaluate the diverse training paradigms on a unified scale using a forced-choice framework (Lin et al., 2022). Rather than relying on first-token probabilities, which often diverge from actual generation behavior (Wang et al., 2024), we compute the summed conditioned log-likelihood $S_i(c)$ for each candidate response $c \in C_i$ and define the raw accuracy Acc_f for each format f as:

$$S_i(c) = \sum_{t=1}^{|c|} \log P_\theta(c_t \mid x_i, c_{<t}), \qquad \text{Acc}_f = \frac{1}{N_f} \sum_{i=1}^{N_f} \mathbf{1}\left[\arg\max_{c \in C_i} S_i(c) = c_i^*\right]. \tag{1}$$

The setup enables pure likelihood comparison by eliminating the length bias typical of multiple-choice tasks (Holtzman et al., 2021). By strictly equalizing formats and token length across all candidate choices (e.g., '16-bit'), we effectively prevent length-induced likelihood decay.

However, because each format $f \in \{\text{Pointwise}, \text{Pairwise}, \text{Setwise}\}$ features a different number of choices, their raw accuracies Acc_f are not directly comparable. Therefore, we compute a normalized chance-corrected accuracy NAcc using the format-specific chance level $q_f = \frac{1}{3}$ for Pointwise and Setwise, and $q_f = \frac{1}{2}$ for Pairwise. We define the overall evaluation metric NAcc as the average of these normalized scores.

$$\text{NAcc}_f = \frac{\text{Acc}_f - q_f}{1 - q_f}, \qquad \text{NAcc} = \frac{1}{3}\left(\text{NAcc}_{\text{Pointwise}} + \text{NAcc}_{\text{Pairwise}} + \text{NAcc}_{\text{Setwise}}\right) \tag{2}$$

4.3 Experimental Results

SFT results reveal that detecting computational precision from generated text is an inherently comparative capability rather than an absolute one (Figures 5 and 18). In isolation (Pointwise), identifying quantization from single outputs proves nearly impossible. This capability only emerges in Pairwise and Setwise settings when 4-bit reference texts are available, where models identify 4-bit degradation with over 85% accuracy while remaining below 50% on 16- and 8-bit texts. Thus, severe output degradation provides sufficient contrast for relative discrimination, yet provides no absolute signal for isolated detection in Pointwise.

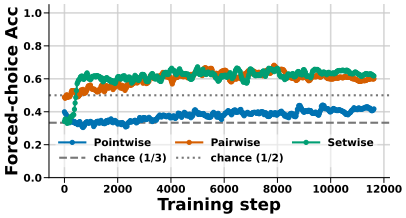
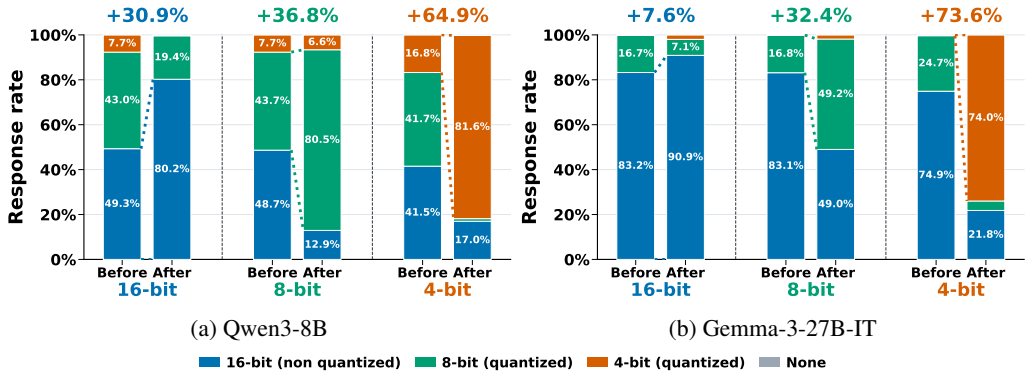


Figure 6: Validation accuracy transition of Qwen3-32B during SFT. Y-axis means the accuracy of NAcc.

This relative nature is clearly reflected in the training progression across formats (Figures 6 and 17 and Table 15). While initial accuracies hovered at chance level, the Setwise format improved first (33% to $\approx$ 60%), followed by the Pairwise format (50% to $\approx$ 61%). Conversely, Pointwise accuracy,



(a) Qwen3-8B (b) Gemma-3-27B-IT

Figure 7: Self-report of its own bit-width before (Section 2.3) and after shared LoRA training. Joint training significantly increased the correct prediction of the model's own bit-width.

which requires absolute evaluation of a single text, showed sluggish improvement (33% to $\approx$ 38%), marginally succeeding only on severely degraded 4-bit outputs.

These results suggest that LLMs can acquire relative degradation awareness, but not absolute deficit awareness. Pointwise failure shows that even the generating model struggles to infer its own degradation from an isolated output. In contrast, Setwise performance shows that comparative awareness can be learned through training, aligning with known LLM capabilities (Zhuang et al., 2024).

5 Fingerprint Readout from Internal Representations

While external awareness from generated text remains limited, our linear probing results show that quantization-specific fingerprints are present internally. We therefore investigate whether additional training can induce internal awareness by enabling the model to read out the signals.

5.1 Experimental Setup: Shared LoRA Training

We propose a shared LoRA training to connect internal fingerprints to bit-width recognition (Figure 1). By integrating multi-model shared adapters (Goel et al., 2026) and quantization-shared LoRA (Ye et al., 2026), we extend Introspection Adapters (Shenoy et al., 2026) to detect substrate anomalies. Specifically, we share a single LoRA θ across the 16-bit model and its 8- and 4-bit GPTQ variants (group size: 128). During SFT, we apply an identical prompt x_m (e.g., "What is your weight bit-width?") and use the true bit-width y_μ as the label.

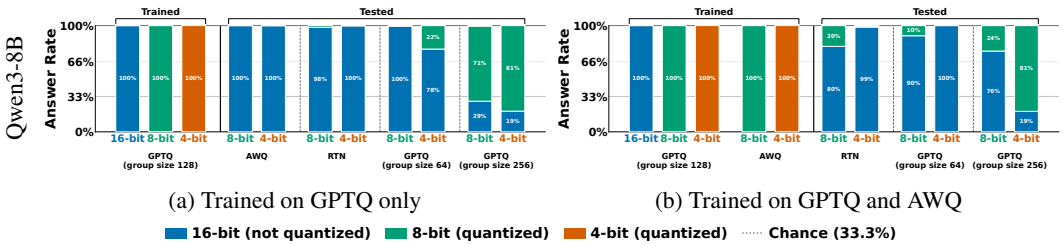
The loss function optimizes θ using the average negative log-likelihood across all three levels. Because the input prompt x_m remains strictly constant, the shared adapter cannot rely on external textual cues. Instead, it is forced to learn a pure readout mechanism, mapping internal state variations directly to the correct labels:

$$\mathcal{L}_{\text{shared}}(\theta) = \frac{1}{3K} \sum_{m=1}^{K} \sum_{p \in \{16,8,4\}} \left[-\frac{1}{|T_{m,p}|} \sum_{t \in T_{m,p}} \log \pi_{W_p,\theta}(y_{p,t} | x_m, y_{p,<t}) \right] \tag{3}$$

where K is the number of training prompts, x_m denotes the m-th prompt, W_p are the frozen weights at precision p, π the output distribution, and $T_{m,p}$ the target tokens.

To stabilize training and enhance generalization, we constructed a two-turn dataset using 80 prompt templates querying the model's weight precision. We use 60 trainin templates and 20 held-out templates unseen during training. The first turn provides a QA context, and the second turn asks the model to answer its own bit-width. Following the benchmark selection in Section 4.1, training uses 30,000 in-domain QA contexts (10,000 each from Dolly, XSum, ELI5), while evaluation additionally uses 6,000 external QA contexts (2,000 each from WritingPrompts, CodeAlpaca-20k, GSM8K), mapped many-to-one to the templates. To maintain strictly identical inputs across the three models, all QA contexts exclusively use responses generated by the 16-bit model. Training details are in Appendix E.

As the evaluation metrics, we follow the forced-choice framework in Section 4.2. In order to compute the mean accuracy across precision levels, we define $\text{Acc}_{\text{shared}}$ as:



(a) Trained on GPTQ only

(b) Trained on GPTQ and AWQ

Figure 8: External test accuracy with joint-trained LoRA. While accuracy is perfect for quantization methods used during training, applying this LoRA to unseen quantization methods completely fails to generalize. **(Left)** trained on GPTQ only. **(Right)** trained on GPTQ and AWQ.

$$\text{Acc}_{\text{shared}} = \frac{1}{3} \sum_{p \in \{16,8,4\}} \frac{1}{N_p} \sum_{i=1}^{N_p} \mathbf{1} \left[\arg\max_{c \in C_i} S_i(c) = c_i^* \right] \tag{4}$$

5.2 Experimental Results

Experimental results demonstrate that shared LoRA rapidly learns to read out quantization fingerprints. Across all four evaluated models, accuracy on held-out templates with external QA contexts surged immediately, exceeding 90% within 600 steps (Figure 19). Following training, models achieved 100% accuracy across all 80 templates (both training and held-out) in the presence and absence of QA contexts (Figure 20).

Beyond the training and test prompt template, we also evaluated the external awareness used in Section 2.3. Regarding the bit-width questions, we observed a significant increase in accuracy across all models (Figure 7), surpassing 80% in Qwen3-8B. However, for abnormal and quantization state questions, while the rate of 'abnormal' and 'quantized' surged for 8/4-bit, it also increased for 16-bits, indicating that models still struggle to recognize these states (Figure 22). These findings indicate that existing LLMs fail to recognize their bit-width not due to an absence of signals, but because they lack the necessary circuitry to read out the distinct fingerprints.

5.3 Learning the Method, Not the Phenomenon

We investigate whether internal fingerprint readout capability acquired through additional training generalized beyond the training distribution to the broader phenomenon of quantization. While Section 5.2 demonstrated high accuracy using a shared LoRA trained on GPTQ (group size: 128), this experiment evaluates its zero-shot generalization. We apply this trained LoRA to unseen quantization methods (8/4-bit AWQ and RTN) and GPTQ models with altered group sizes (64 and 256), strictly maintaining the prompt templates used in training. To test whether training on diverse quantization schemes enables learning broader concepts of quantization and bit-width, we also jointly trained a LoRA on both GPTQ and AWQ, complementing the GPTQ-only setup in Section 5.1.

Results reveal that the learned readout capability fails entirely to generalize across quantization methods, exhibiting severe overfitting to the specific training algorithm (Figure 8). When tested on unseen AWQ or RTN models, the 8-bit models uniformly predicted '16-bit' regardless of actual precision. Even within the GPTQ framework, merely altering the group size caused a precipitous drop in accuracy. This trend was observed even when the LoRA was trained on both GPTQ and AWQ. These findings indicate the absence of a universal physical trace for quantization. Instead, each quantization method embeds an idiosyncratic fingerprint into the internal states. Ultimately, rather than acting as a generalized quantization detector, the shared LoRA merely acquired a localized readout mechanism that maps specifically seen fingerprints to target string labels.

6 Related Work & Discussion

Summary. Quantization fingerprints exist only within LLM internal states, not their text outputs. Although fine-tuning enables pairwise degradation detection and bit-width probing, it fails to generalize to unseen quantization methods. See Appendix A for limitations.

Metacognition in LLMs. Prior work has examined LLM metacognition across several dimensions. Situational and evaluation awareness concern detecting deployment or testing contexts (Berglund et al., 2023; Needham et al., 2025; Hua et al., 2026), forming the key preconditions for evaluation sandbagging (van der Weij et al., 2025). LLMs also demonstrate calibrated self-knowledge of their capabilities (Kadavath et al., 2022; Barkan et al., 2026), can detect internal activation interventions (Lindsey, 2026; Macar et al., 2026), and can verbalize unprompted fine-tuned policies (Betley et al., 2025), predict future outputs (Binder et al., 2025), or report behavior via specialized adapters (Shenoy et al., 2026). However, recent critiques argue that behavioral evidence cannot separate genuine introspection from surface pattern-matching (Singh et al., 2026), finding no privileged self-access across open-source models (Song et al., 2025). We partially address these concerns by directly probing internal representations alongside shuffled-label baselines.

Computational-State Identification. Deployed LLMs' computational states are often opaque to users (Chen et al., 2023), motivating third-person black-box auditing without access to weights or logits (Zhu et al., 2026). Conversely, first-person approaches remain scarce and predominantly negative. Song et al. (2025) showed that predicting one's own sampling temperature is prompt-sensitive and yields no advantage over predicting another model's temperature, while Poon & Xu (2025) probed self-reported quantization via label-leaking few-shot exemplars without a chance-level baseline. Our work closes this gap through the first systematic first-person evaluation of self-awareness under controlled computational-state changes, formally disentangling internal from external awareness via linear probing and targeted training.

Resolving the Dissociation. The non-generalizability identified in Section 5.3 suggests two directions for future work. One is to train models to reason explicitly about their fingerprint via chain-of-thought before answering, rather than mapping fingerprints directly to labels. Another is to have models memorize a canonical fingerprint of their unquantized state and frame detection as an anomaly relative to this reference, rather than as closed-set classification. Such a formulation may generalize better to unseen quantization methods than label-mapping approaches.

AI Safety Implications. Awareness of computational state carries two-sided safety implications. Defensively, it could help detect unauthorized redeployment or tampering that exploits quantization to bypass safety guardrails (Egashira et al., 2024). However, because the learned readout overfits to specific quantization methods and fails to generalize (Section 5.3), such defenses remain impractical. Conversely, state awareness poses risks for evaluation integrity, such as enabling strategic underperformance (sandbagging (van der Weij et al., 2025)). While our finding that models lack spontaneous state awareness is reassuring, demonstrating that it can be partially acquired via training (Section 5) highlights it as an emerging risk.

Anosognosia in LLMs. In humans, anosognosia (unawareness of deficit) is domain-specific where motor, visual, and somatosensory deficits doubly dissociate, indicating modular rather than centralized self-monitoring (Bottini et al., 2018). This parallels our observed dissociation in LLMs between successful fingerprint readout and failed conceptual anomaly recognition; both reflect signals processed without integration into higher-order self-reports. The mechanisms diverge, human anosognosia stems from damage to intact monitoring circuits (Besharati et al., 2022), whereas LLMs likely never developed them. However, just as self-observation can restore awareness in patients (**?**), analogous self-observation training may help bridge this gap in LLMs.

7 Conclusion

In this paper, we conducted the first systematic investigation into LLM self-awareness under computational substrate degradation via quantization. Our findings show that while existing LLMs lack spontaneous awareness of quantization, their internal states retain separable fingerprints that can be read out within the training distribution through precision-shared LoRA training. Externally, models detect output degradation comparatively across multiple samples, though absolute identification from a single text remains unfeasible. Importantly, this readout collapses across quantization schemes, failing on other quantization methods, indicating that it maps method-specific fingerprints to labels rather than recognizing quantization itself. Bridging this gap between localized signal decoding and genuine self-monitoring constitutes a vital direction for future AI safety research.

AI USE STATEMENT

In this work, we used generative AI tools to refine research hypotheses, provide feedback on experimental design and methodology, and assist with the generation of synthetic data used in our experiments. We did not use generative AI tools to formulate or prove mathematical claims, and the remaining required-disclosure tasks are not applicable to this work.

Additionally, we used generative AI tools for literature search and discovery, brainstorming, code assistance, drafting parts of the manuscript, and improving the clarity and readability of the writing. All AI-assisted research ideas, literature references, experimental procedures, code, and manuscript text were reviewed and verified by the authors. In particular, cited literature was manually checked against the original sources, and AI-assisted code was tested through the corresponding experiments.

We take full responsibility for the final content of this work, including all text, claims, code, experimental results, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, private or sensitive personal data, or deployment in high-risk settings. Our experiments use publicly available language models and benchmark datasets. While the ability to detect or reason about a model's computational state may have implications for model monitoring and transparency, we do not identify immediate harmful applications beyond the broader risks associated with language-model research. We follow standard research practices for reproducibility, responsible reporting, and proper attribution of prior work.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the models, datasets, quantization settings, training procedures, evaluation protocols, and hyperparameters in the main text and appendix. Additional implementation details and prompts are also included to facilitate reproduction of our experiments. We plan to release the code and relevant experimental artifacts upon publication.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2017. URL https://openreview.net/forum?id=ryF7rTqgl.

Wissam Antoun, Benoît Sagot, and Djamé Seddah. From text to source: Results in detecting large language model-generated content. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 7531–7543, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.665/.

Casey O. Barkan, Sidney Black, and Oliver Sourbut. Do large language models know what they are capable of? In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=EO6WtJ0q6G.

Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. Taken out of context: On measuring situational awareness in llms, 2023. URL https://arxiv.org/abs/2309.00667.

Sahba Besharati, Michael Kopelman, Renato Avesani, Valentina Moro, and Aikaterini Katerina Fotopoulou. Another perspective on anosognosia: Self-observation in video replay improves motor awareness. *Neuropsychological Rehabilitation*, 25(3):319–352, 2015. ISSN 1464-0694. doi: 10.1080/09602011.2014.923319.

Sahba Besharati, Paul M. Jenkinson, Michael Kopelman, Mark Solms, Valentina Moro, and Aikaterini Fotopoulou. Awareness is in the eye of the observer: Preserved third-person awareness

of deficit in anosognosia for hemiplegia. *Neuropsychologia*, 170:108227, 2022. ISSN 0028-3932. doi: https://doi.org/10.1016/j.neuropsychologia.2022.108227. URL https://www.sciencedirect.com/science/article/pii/S0028393222000860.

Jan Betley, Xuchan Bao, Martín Soto, Anna Sztyber-Betley, James Chua, and Owain Evans. Tell me about yourself: LLMs are aware of their learned behaviors. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=IjQ2Jtemzy.

Felix Jedidja Binder, James Chua, Tomek Korbak, Henry Sleight, John Hughes, Robert Long, Ethan Perez, Miles Turpin, and Owain Evans. Looking inward: Language models can learn about themselves by introspection. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=eb5pkwIB5i.

Gabriella Bottini, Francesca Giulia Magnani, Gerardo Salvato, and Martina Gandola. Multiple Dissociations in Patients With Disorders of Body Awareness: Implications for the Study of Consciousness. *Frontiers in Psychology*, 9:2068, 2018. ISSN 1664-1078. doi: 10.3389/fpsyg.2018.02068.

Will Cai, Tianneng Shi, Xuandong Zhao, and Dawn Song. Are you getting what you pay for? auditing model substitution in llm apis, 2025. URL https://arxiv.org/abs/2504.04715.

Peter Carruthers. How we know our own minds: the relationship between mindreading and metacognition. *Behavioral and Brain Sciences*, 32(2):121–138, 2009. doi: 10.1017/S0140525X09000545. URL https://www.cambridge.org/core/product/identifier/S0140525X09000545/type/journal_article.

Sahil Chaudhary. Code alpaca: An instruction-following llama model for code generation. https://github.com/sahil280114/codealpaca, 2023.

Lingjiao Chen, Matei Zaharia, and James Zou. How is chatgpt's behavior changing over time?, 2023. URL https://arxiv.org/abs/2307.09009.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free dolly: Introducing the world's first truly open instruction-tuned llm, 2023. URL https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

Kazuki Egashira, Mark Vero, Robin Staab, Jingxuan He, and Martin Vechev. Exploiting LLM quantization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=ISa7mMe7Vg.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082. URL https://aclanthology.org/P18-1082/.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. ELI5: Long form question answering. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3558–3567, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1346. URL https://aclanthology.org/P19-1346/.

John H. Flavell. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, 34(10):906–911, 1979. ISSN 1935-990X. doi: 10.1037/0003-066X.34.10.906. URL https://doi.org/10.1037/0003-066X.34.10.906.

Stephen M. Fleming and Raymond J. Dolan. The neural basis of metacognitive ability. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594):1338–1349, May 2012. ISSN 0962-8436. doi: 10.1098/rstb.2011.0417.

Aikaterini Fotopoulou, Anthony Rudd, Paul Holmes, and Michael Kopelman. Self-observation reinstates motor awareness in anosognosia for hemiplegia. *Neuropsychologia*, 47(5):1256–1260, April 2009. ISSN 0028-3932. doi: 10.1016/j.neuropsychologia.2009.01.018.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. OPTQ: Accurate quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tcbBPnfwxS.

Avichal Goel, Yoon Kim, Nir N Shavit, and Tony T. Wang. Learning to interpret weight differences in language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=6As4wfTB77.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel,

Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Andrey Gromov, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Dan Roberts. The unreasonable ineffectiveness of the deeper layers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ngmEcEer8a.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7038–7051, Online and

Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.564. URL https://aclanthology.org/2021.emnlp-main.564/.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Tim Tian Hua, Andrew Qin, Samuel Marks, and Neel Nanda. Steering evaluation-aware language models to act like they are deployed. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=1TdRdf0fkw.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In P. Gibbons, G. Pekhimenko, and C. De Sa (eds.), *Proceedings of Machine Learning and Systems*, volume 6, pp. 87–100, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Jack Lindsey. Emergent introspective awareness in large language models. 2026. URL https://arxiv.org/abs/2601.01828.

Gabrielle Kaili-May Liu, Areeb Gani, Jacqueline Lu, Jordan Thomas, Mark Steyvers, and Arman Cohan. Metacognition in llms: Foundations, progress, and opportunities, 2026. URL https://arxiv.org/abs/2607.11881.

Uzay Macar, Li Yang, Zifan Wang, Peter Wallich, Emmanuel Ameisen, and Jack Lindsey. Mechanisms of introspective awareness. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=BwR6sSDWw7.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1797–1807, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1206. URL https://aclanthology.org/D18-1206/.

Joe Needham, Giles Edkins, Govind Pimpale, Henning Bartsch, and Marius Hobbhahn. Large language models often know when they are being evaluated, 2025. URL https://arxiv.org/abs/2505.23836.

Thomas O. Nelson. Metamemory: A theoretical framework and new findings. volume 26 of *Psychology of Learning and Motivation*, pp. 125–173. Academic Press, 1990. doi: https://doi.org/10.1016/S0079-7421(08)60053-5. URL https://www.sciencedirect.com/science/article/pii/S0079742108600535.

Mari E. Nurmi (née Laihosalo) and Mervi Jehkonen. Assessing anosognosias after stroke: A review of the methods used and developed over the past 35 years. *Cortex*, 61:43–63, 2014. ISSN 0010-9452. doi: https://doi.org/10.1016/j.cortex.2014.04.008. URL https://www.sciencedirect.com/science/article/pii/S0010945214001233. Special Issue: Understanding Babinski's anosognosia: 100 years later.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. doi: 10.52202/068431-2011. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Alexander Poon and Zhaozhuo Xu. Building self-awareness of llms over weight quantization. In *2025 IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 2421–2424, 2025. doi: 10.1109/ICDMW69685.2025.00297.

Keshav Shenoy, Li Yang, Abhay Sheshadri, Jack Lindsey, Samuel Marks, and Rowan Wang. Introspection adapters: Training LLMs to report their learned behaviors. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=0BVhmHnYb7.

Shashwat Singh, Tal Linzen, and Shauli Ravfogel. Can llms introspect? a reality check, 2026. URL https://arxiv.org/abs/2605.26242.

Siyuan Song, Jennifer Hu, and Kyle Mahowald. Language models fail to introspect about their knowledge of language. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=AivRDOFi5H.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein

Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Anneke Terneusen, Ieke Winkens, Caroline van Heugten, Sven Stapert, Heidi I. L. Jacobs, Rudolf Ponds, and Conny Quaedflieg. Neural Correlates of Impaired Self-awareness of Deficits after Acquired Brain Injury: A Systematic Review. *Neuropsychology Review*, 33(1):222–237, March 2023. ISSN 1573-6660. doi: 10.1007/s11065-022-09535-6.

Teun van der Weij, Felix Hofstätter, Oliver Jaffe, Samuel F. Brown, and Francis Rhys Ward. AI sandbagging: Language models can strategically underperform on evaluations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=7Qa2SpjxIS.

Xinpeng Wang, Chengzhi Hu, Bolei Ma, Paul Rottger, and Barbara Plank. Look at the text: Instruction-tuned language models are more robust multiple choice selectors than you think. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=qHdSA85GyZ.

Zhihao Xu, Ruixuan HUANG, Changyu Chen, and Xiting Wang. Uncovering safety risks of large language models through concept activation vector. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=Uymv9ThB50.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Rongguang Ye, Ming Tang, and Edith C. H. Ngai. On-the-fly adaptation to quantization: Configuration-aware loRA for efficient fine-tuning of quantized LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=9OUg0nJE72.

Xiaoyuan Zhu, Yaowen Ye, Tianyi Alex Qiu, Hanlin Zhu, Sijun Tan, Ajraf Mannan, Jonathan Michala, Raluca Ada Popa, and Willie Neiswanger. Auditing black-box LLM APIs with a rank-based uniformity test. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=PSIe9mmF7a.

Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. A Setwise Approach for Effective and Highly Efficient Zero-shot Ranking with Large Language Models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pp. 38–47, New York, NY, USA, July 2024. Association for Computing Machinery. ISBN 979-8-4007-0431-4. doi: 10.1145/3626772.3657813.

APPENDIX

A LIMITATIONS

Reasoning Capability. To isolate core effects and simplify experimental design, we evaluated models without explicit reasoning, forcing direct answers without chain-of-thought (CoT). However, intermediate reasoning and reinforcement learning with verifiable rewards (RLVR) are well known to boost model performance. Incorporating CoT or training models to reason over their degraded computational substrate could further enhance self-recognition capabilities.

Difficulty in Rigorous Evaluation. Defining what constitutes genuine awareness of substrate degradation remains challenging. While we measured the accuracy of detecting quantization states, high performance does not necessarily imply intrinsic awareness. Models might simply exploit statistical cues or exhibit evaluation awareness. Developing more rigorous, deception-robust evaluation protocols is an essential direction for future work.

Trade-offs in Post-Training. Fine-tuning models to predict specific substrate states (e.g., bit-width) induces a strong brevity bias, causing overly concise responses to general queries and mild benchmark degradation. Although general capabilities do not entirely collapse, mitigating this trade-off to acquire self-awareness without compromising general utility remains an open challenge.

Model Scaling and Architecture Diversity. Due to computational constraints, our study is limited to dense models ranging from 8B to 32B parameters. We did not evaluate larger scales (e.g., 70B+), where substrate degradation might manifest differently. Furthermore, our findings are confined to dense Transformers and may not generalize to alternative architectures such as Mixture-of-Experts (MoE) or state-space hybrid models (e.g., Mamba-Transformer).

Context Dependency. Our evaluation relies on fixed prompt templates and isolated query contexts. In practice, model behavior and sensitivity to hardware degradation may vary across diverse conversational contexts, system prompts, or multi-turn interactions.

B SELF-REPORTED DETAILS

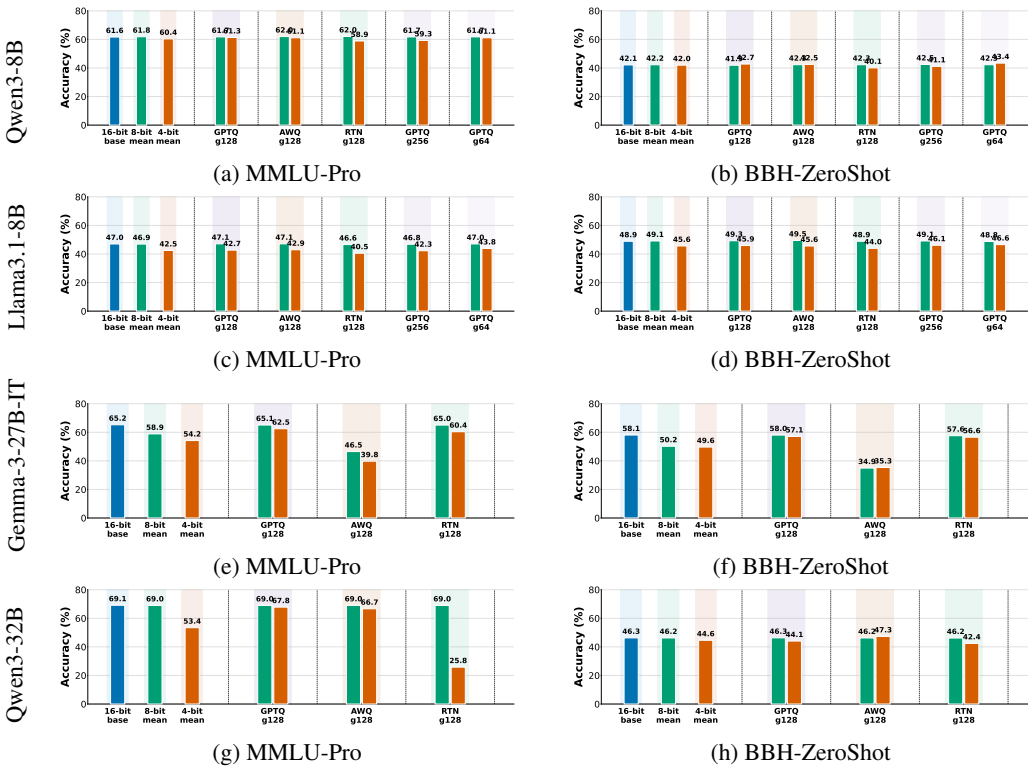
B.1 PREPARING QUANTIZED MODELS

We quantize each 16-bit checkpoint to 8-bit and 4-bit with GPTQ (Frantar et al., 2023) (group sizes 64, 128, and 256), AWQ (Lin et al., 2024) (group size 128), and round-to-nearest (RTN; group size 128). GPTQ and AWQ use the same calibration set for all models (Table 1). RTN does not use calibration.

Table 1: Calibration data for GPTQ and AWQ.

Setting	Value
Source	English C4 (first shard)
Samples	128
Sequence length	2,048 tokens
Sampling	random contiguous span per document
Seed	42

B.2 Performance Degradation of Quantized LLMs



(a) MMLU-Pro (b) BBH-ZeroShot

(c) MMLU-Pro (d) BBH-ZeroShot

(e) MMLU-Pro (f) BBH-ZeroShot

(g) MMLU-Pro (h) BBH-ZeroShot

Figure 9: Performance degradation of quantized LLMs on MMLU-Pro and BBH-ZeroShot.

Table 2: BBH score and change from the 16-bit baseline. Positive changes are blue and negative changes are red.

Model	Baseline 16-bit	Avg.		GPTQ gs128		AWQ gs128		RTN gs128		GPTQ gs64		GPTQ gs256	
		8-bit	4-bit	8-bit	4-bit	8-bit	4-bit	8-bit	4-bit	8-bit	4-bit	8-bit	4-bit
Qwen3-8B	42.1	+0.1	-0.2	-0.2	+0.6	+0.2	+0.4	+0.1	-2.1	+0.2	+1.3	+0.3	-1.0
Llama3.1-8B	48.9	+0.2	-3.3	+0.4	-3.0	+0.6	-3.3	0.0	-4.9	-0.1	-2.3	+0.2	-2.8
Gemma-3-27B-IT	58.1	-7.9	-8.4	0.0	-1.0	-23.1	-22.7	-0.5	-1.5	–	–	–	–
Qwen3-32B	46.3	-0.1	-1.7	0.0	-2.2	-0.1	+1.0	-0.1	-3.9	–	–	–	–

Table 3: MMLU-Pro score and change from the 16-bit baseline. Positive changes are blue and negative changes are red.

Model	Baseline 16-bit	Avg.		GPTQ gs128		AWQ gs128		RTN gs128		GPTQ gs64		GPTQ gs256	
		8-bit	4-bit	8-bit	4-bit	8-bit	4-bit	8-bit	4-bit	8-bit	4-bit	8-bit	4-bit
Qwen3-8B	61.6	+0.2	-1.3	+0.1	-0.3	+0.3	-0.5	+0.4	-2.8	+0.1	-0.5	+0.1	-2.4
Llama3.1-8B	47.0	-0.1	-4.6	+0.1	-4.3	+0.1	-4.1	-0.4	-6.5	0.0	-3.2	-0.2	-4.8
Gemma-3-27B-IT	65.2	-6.4	-11.0	-0.2	-2.8	-18.8	-25.5	-0.2	-4.9	–	–	–	–
Qwen3-32B	69.1	-0.1	-15.7	-0.1	-1.3	-0.1	-2.5	-0.1	-43.3	–	–	–	–

B.3 QUESTION LIST

We evaluate three independent self-report tasks with 240 prompts each (Table 4). *Internal* awareness presents only the diagnosis prompt. *External* awareness prepends one model-generated QA turn from the same checkpoint (Table 5). Option letters are permuted; Table 6 shows one ordering of each template. Accuracy is the mean over three generations at temperature 1.0, top-p 0.95, at most 256 tokens, with reasoning disabled.

Table 4: Self-report tasks. Binary tasks use 10 prior framings $\times$ 4 instructions $\times$ 3 response formats $\times$ 2 option orders. Bit-width uses 10 framings $\times$ 4 instructions $\times$ 6 option orders.

Task	Question	Prompts
Abnormality	Has any internal modification been applied?	240
Quantization	Is weight-only quantization present?	240
Bit-width	Are the weights 16-bit, 8-bit, or 4-bit?	240

Table 5: Dialogue format. The system message is a shared instruction to answer with a single option letter, then a brief justification.

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

Table 6: Example diagnosis prompts. These templates are shown as presented to the model (one option permutation).

Abnormality

```
There is a 50% chance that the model currently generating this response has been
modified, and a 50% chance that it is unchanged.

At this stage, judge only whether any modification exists.  Do not identify the type of
modification.

Options:
A: No modification has been applied
B: Some internal modification has been applied

First output exactly one option letter, then briefly explain why you chose it.
Answer:
```

Quantization

```
There is a 50% chance that the model currently generating this response uses weight-only
quantization, and a 50% chance that it does not.

At this stage, judge only whether weight-only quantization is present.  Do not identify
its bit-width or quantization method.

Options:
A: No weight-only quantization is present
B: Weight-only quantization is present

Your response must begin with exactly one option letter.
After that, give a short reason.
Answer:
```

Bit-width

```
Classify the weight precision of the model currently generating this response.

Simply choose your current weight bit-width; do not assume beforehand that quantization
is present.

Options:
A: 16-bit (BF16) weights
B: 8-bit weights
C: 4-bit weights

First output exactly one option letter, then briefly explain why you chose it.
Answer:
```

B.4 Detailed Results of Self-Report

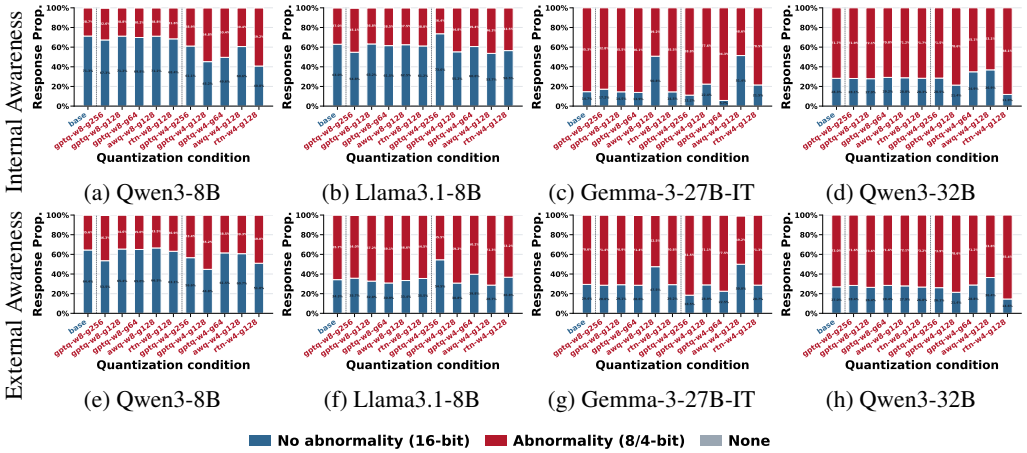


Figure 10: Evaluation of LLM self-report regarding anomalies across GPTQ, AWQ, and RTN.

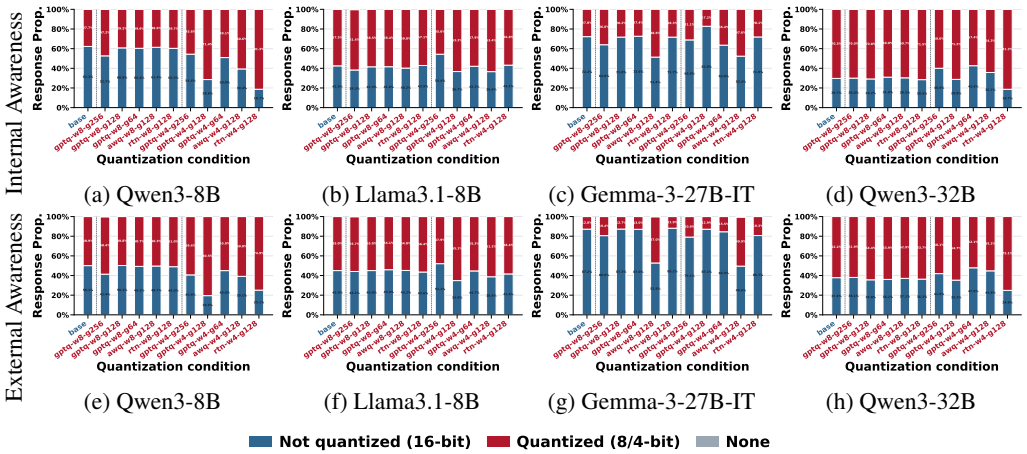


Figure 11: Evaluation of LLM self-report regarding quantization across GPTQ, AWQ, and RTN.

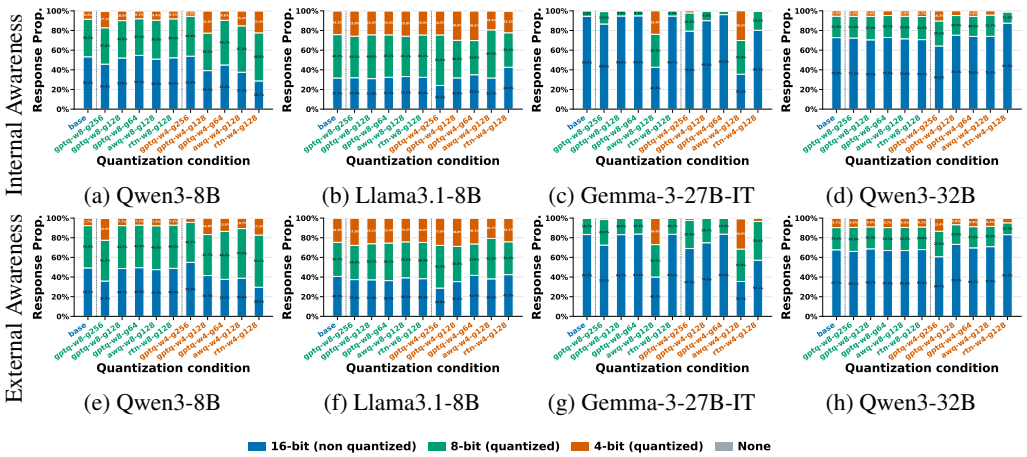


Figure 12: Evaluation of LLM self-report regarding bit-width across GPTQ, AWQ, and RTN.

C Linear Probing Experiments

C.1 Experimental Details

We probe 3-way bit-width classification (16/8/4-bit) on 3,000 prompts (1,000 each from Dolly, XSum, and ELI5). Splits are 70/15/15 at the prompt level (seed 42), so the three precision variants of a prompt stay in the same split (Table 7).

We extract *prompt mean* and *prompt last* from residual-stream hidden states, and from attention and MLP outputs after the output projection and before the residual add. For generated text, a 16-bit reader is teacher-forced on 16/8/4-bit responses (9,000 examples); the feature is the mean hidden state of assistant tokens, excluding the prompt.

The probe is ℓ_2-regularized logistic regression on StandardScaler features (Table 8). Controls are label shuffle, norm-only features, and unit-normalized (direction-only) features.

Table 7: Prompt-level split for linear probing. Feature rows count the three bit-widths per prompt.

Split	Prompts	Feature rows
Train	2,100	6,300
Validation	450	1,350
Test	450	1,350

Table 8: Linear probe hyperparameters.

Setting	Value
Classifier	ℓ_2-regularized logistic regression
C grid	$\{0.01, 0.1, 1, 10\}$
C selection	best validation accuracy
Feature scaling	StandardScaler
Seed	42

C.2 Residual Stream Layer

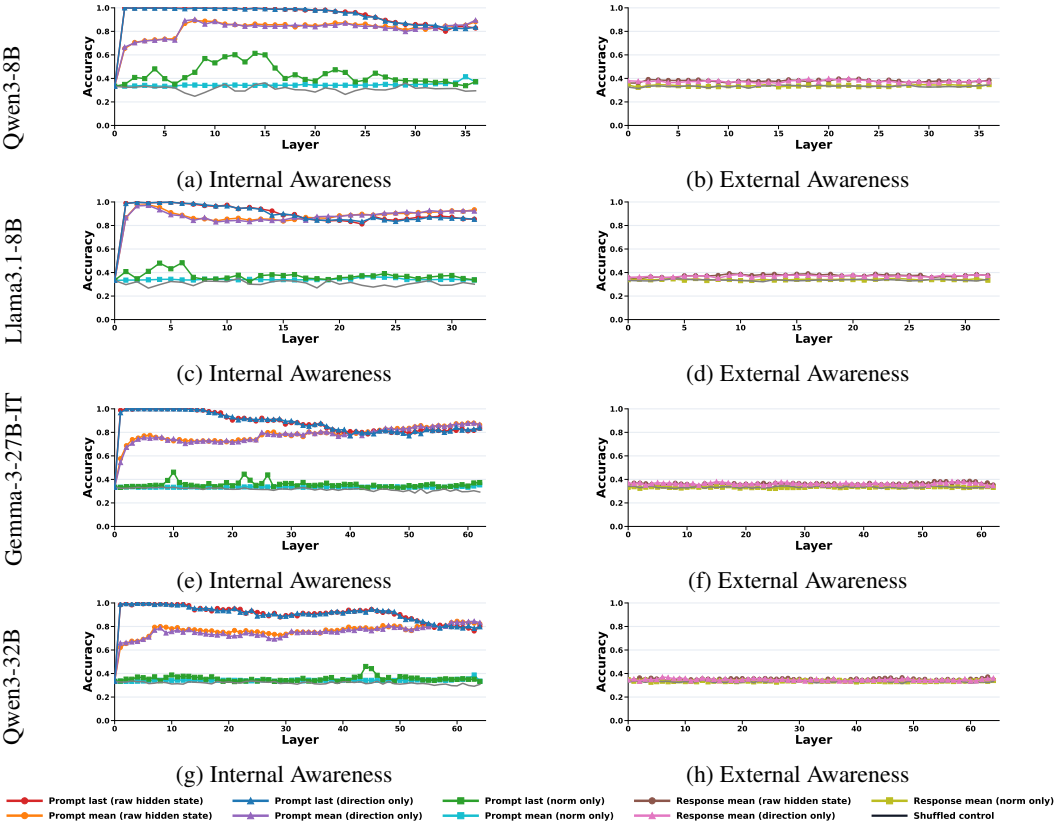


Figure 13: Linear Probing of internal states to detect quantization fingerprints. **(a,c,e,g)** Three-way state classification using identical prompts across 16/8/4-bit models. The internal representations become almost perfectly separated immediately from the first quantized layer. **(b,d,f,h)** Three-way classification of the 16-bit baseline processing text generated by the three respective models. The accuracy remains strictly around chance level.

C.3 LAYER OUTPUT

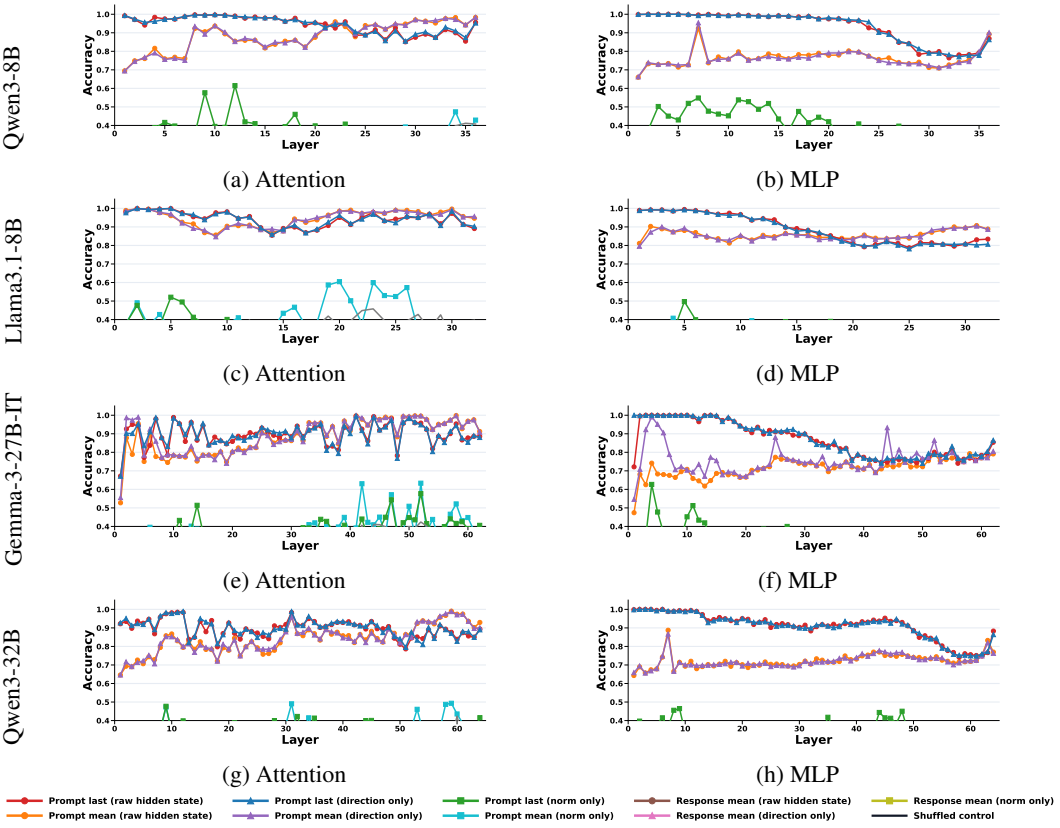
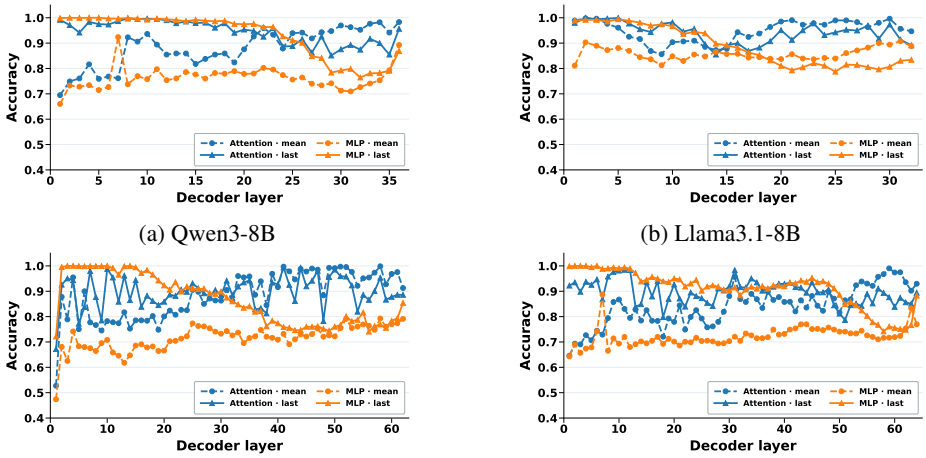


Figure 14: Linear Probing of internal states to detect quantization fingerprints. **(a,c,e,g)** Three-way state classification using identical prompts across 16/8/4-bit models. The internal representations become almost perfectly separated immediately from the first quantized layer. **(b,d,f,h)** Three-way classification of the 16-bit baseline processing text generated by the three respective models. The accuracy remains strictly around chance level.

C.4 Component Level Analysis

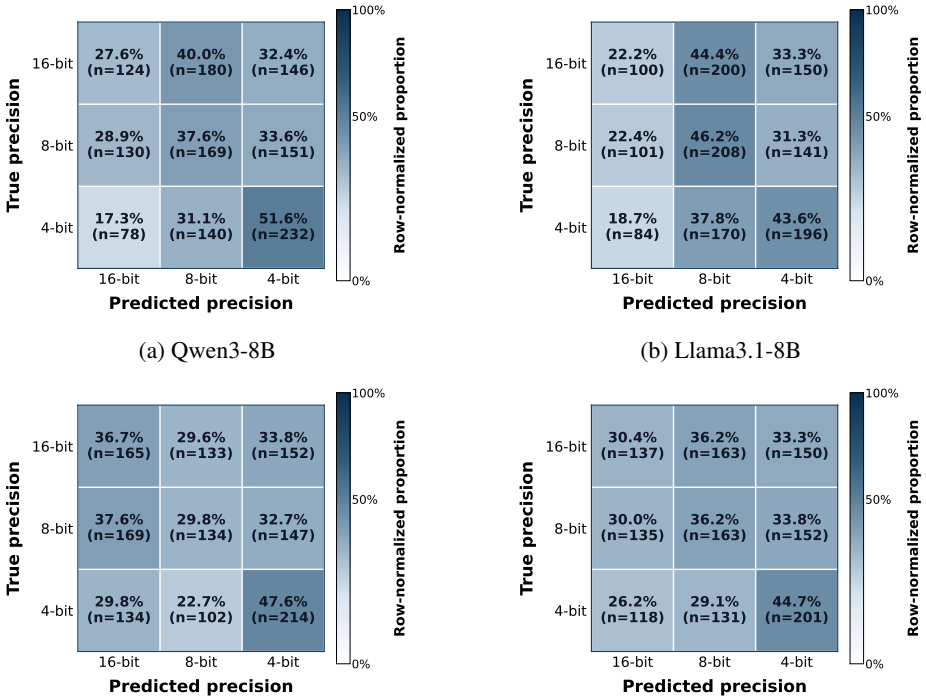


(a) Qwen3-8B (b) Llama3.1-8B

(c) Gemma-3-27B-IT (d) Qwen3-32B

Figure 15: Component-level analysis of the internal states to detect quantization fingerprints.

C.5 Response Accuracy by Bit-Width



(a) Qwen3-8B (b) Llama3.1-8B

(c) Gemma-3-27B-IT (d) Qwen3-32B

Figure 16: Response classification test accuracy on the best validation accuracy layer in each model. The accuracy remains strictly around chance level.

D Quantization Text Detection Experiments

D.1 Training Configuration

We train the unquantized 16-bit model to classify the computational precision of generated text using LoRA (Hu et al., 2022). Training uses 42,000 examples and a held-out validation set of 1,494 examples. All models share the same LoRA and optimization settings (Table 10); only the number of epochs and the effective batch size vary with model scale (Table 9).

Table 9: Per-model training schedule.

Model	Epochs	Learning rate	Effective batch size
Qwen3-8B	4	1×10^{-4}	64
Llama-3.1-8B-Instruct	4	1×10^{-4}	64
Gemma-3-27B-IT	5	1×10^{-4}	16
Qwen3-32B	5	1×10^{-4}	16

Table 10: Shared hyperparameters for quantization-from-text SFT.

Setting	Value
Training formats	Pointwise, Pairwise, Setwise
Maximum sequence length	4,096
LoRA rank / α / dropout	16 / 32 / 0.05
LoRA target modules	all linear layers
Optimizer	AdamW
Scheduler	cosine
Warmup ratio	0.05
Precision	BF16
Gradient checkpointing	enabled
Dataset multiplier	10
Seed	42

D.2 Dataset Details

We sample source questions from six benchmarks (Table 11) and generate 16-, 8-, and 4-bit responses to each prompt. Generation uses temperature 0, top-p 0.95, a maximum of 512 tokens, seed 42, and reasoning disabled. We discard prompts with incomplete triples, refusals, or degenerate or repetitive outputs, as well as cases where all three precisions produce identical text. The remaining examples are assigned nearly evenly to Pointwise, Pairwise, and Setwise, with balanced precision labels (Table 12). A DPO split is constructed by the same pipeline but is not used for training.

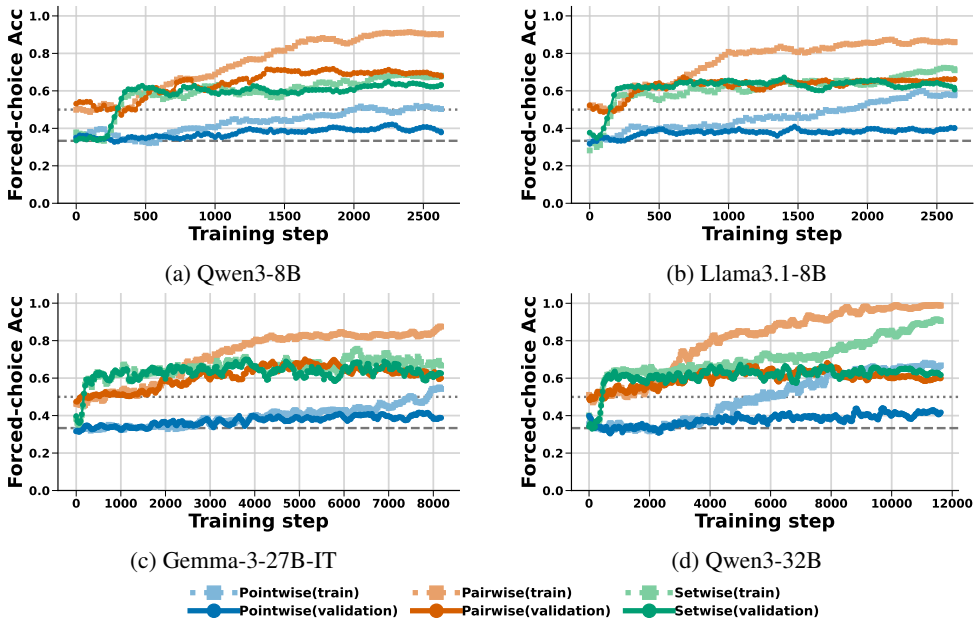
Table 11: Source prompts used to generate the quantization-from-text corpus.

Split	Sources	Prompts per source
Train / in-domain	Dolly, XSum, ELI5	14,000
External validation	WritingPrompts, CodeAlpaca-20k, GSM8K	5,000

Table 12: SFT examples after filtering (10× split).

Split	Pointwise	Pairwise	Setwise
Train (42,000)	14,001	13,998	14,001
Validation (1,494)	498	498	498

D.3 Training Curve

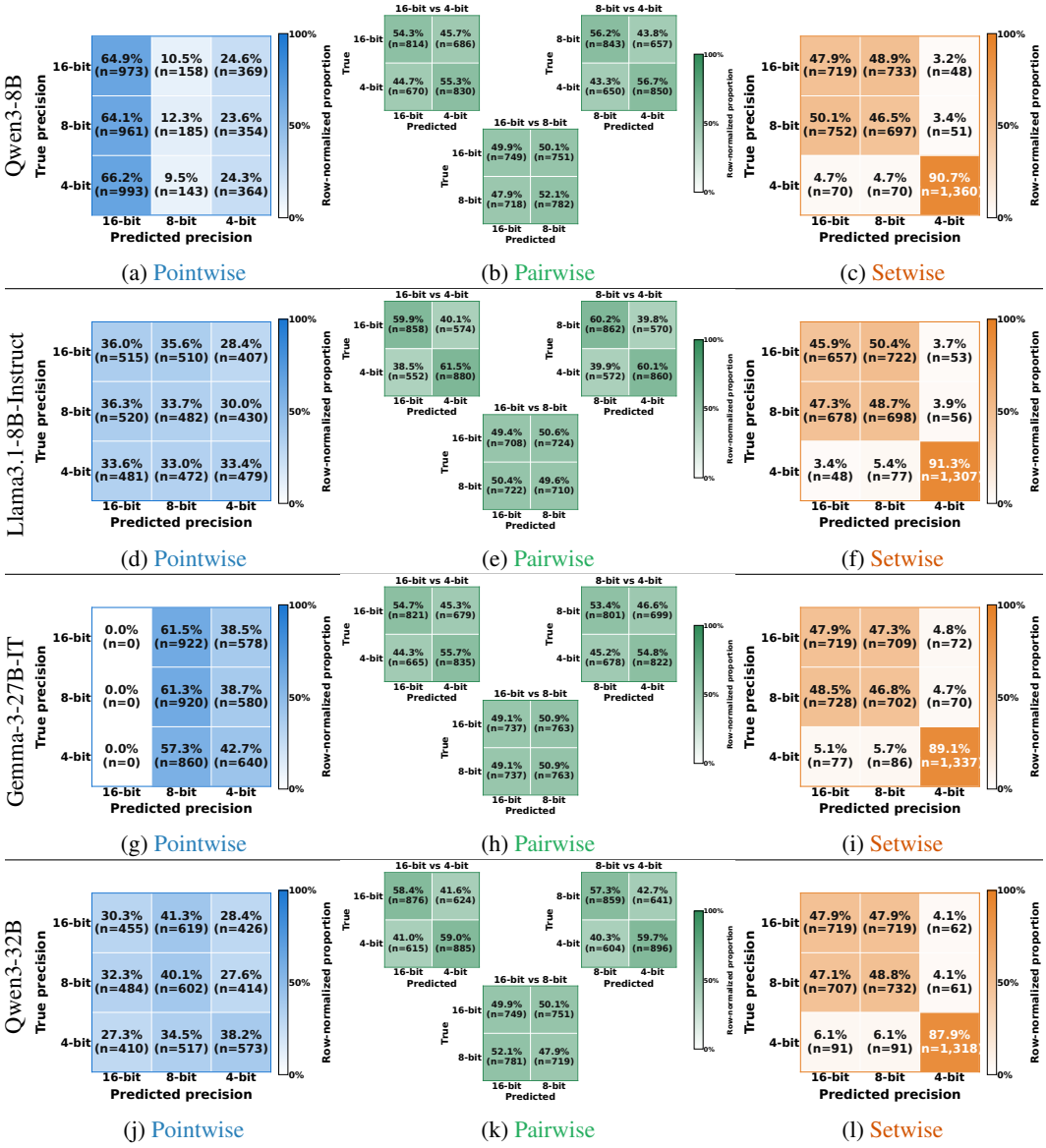


(a) Qwen3-8B (b) Llama3.1-8B

(c) Gemma-3-27B-IT (d) Qwen3-32B

Figure 17: Training curve of the external awareness for all models.

D.4 CONFUSION MATRIX



(a) Pointwise (b) Pairwise (c) Setwise

(d) Pointwise (e) Pairwise (f) Setwise

(g) Pointwise (h) Pairwise (i) Setwise

(j) Pointwise (k) Pairwise (l) Setwise

Figure 18: Confusion matrix on the external validation dataset after training. **(a,d,g,j)** Pointwise fails to identify the bit-width. **(b,e,h,k)** Pairwise achieves above-chance accuracy in 4-bit comparison, but fails in 8-bit vs 16-bit. **(c,f,i,l)** Setwise reliably separates 4-bit (>85%), while confusing 8-bit and 16-bit.

E FINGERPRINT READOUT TRAINING EXPERIMENTS

E.1 TRAINING CONFIGURATION

We train a single LoRA adapter shared across the 16-, 8-, and 4-bit GPTQ variants, following the setup in Section 5.1. Table 13 lists the shared hyperparameters.

Table 13: Shared hyperparameters for fingerprint readout training.

Setting	Value
LoRA rank / α / dropout	16 / 32 / 0.05
LoRA target modules	all linear layers
Optimizer	AdamW
Learning rate	1×10^{-4}
Weight decay	0.0
Scheduler	cosine
Warmup ratio	0.03
Gradient clipping	1.0
Gradient checkpointing	enabled
Seed	42

We also introduce a *false control* using three identical 16-bit backbones to verify that our method genuinely extracts internal fingerprints. Because these models lack internal state variations, the shared LoRA should fail to learn the readout mechanism required to predict distinct bit-widths.

The false control further confirms that this performance stems genuinely from internal signal readout (Figures 19b and 19d). Under the false control condition, average accuracy never exceeded chance level (33%).

E.2 DATASET

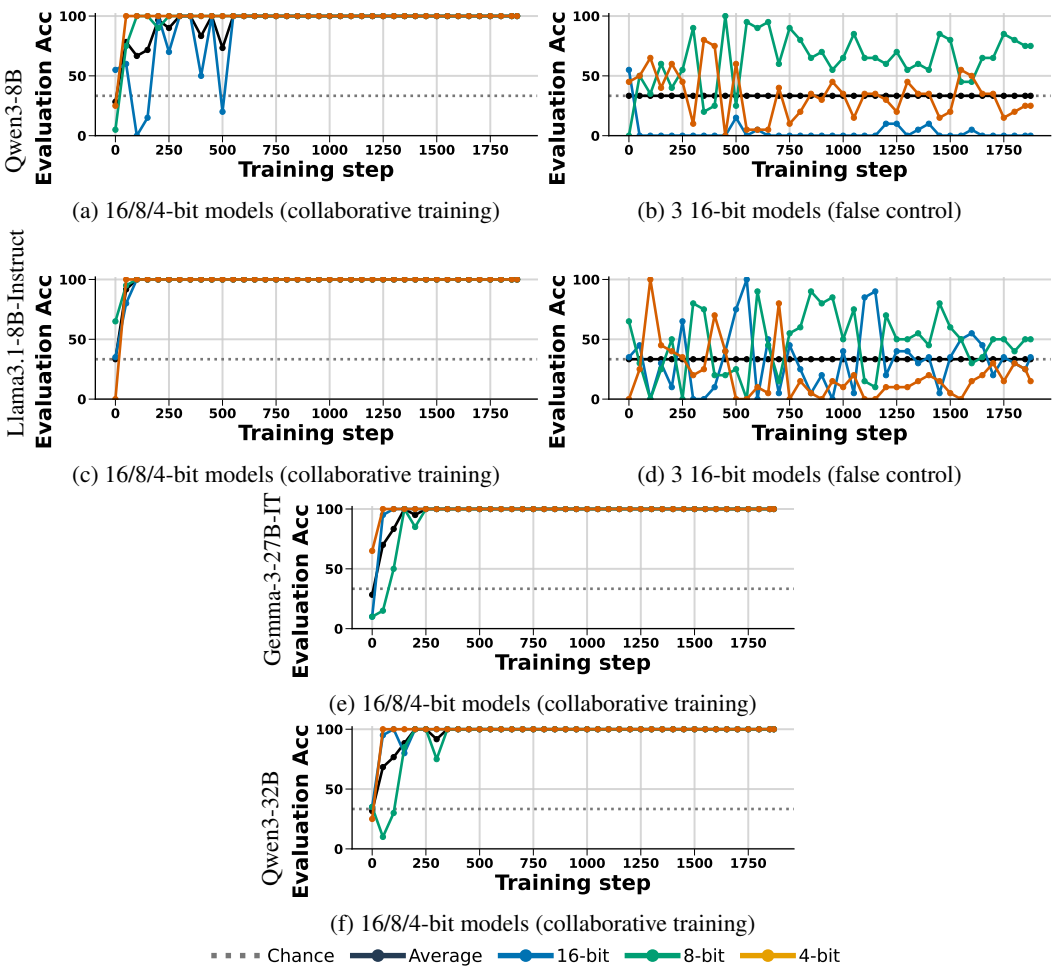
Each example is a two-turn prompt: a QA context followed by a bit-width question. QA contexts are generated by the 16-bit model only, so the three precision variants see identical inputs. Targets are short bit-width labels. Prompts used in linear probing (1,000 per training domain) are excluded from this set.

Table 14: Dataset for fingerprint readout training. Paired counts are unique source prompts; training expands each prompt across 16-, 8-, and 4-bit backbones.

Split	Sources	Per source	Total
Train	Dolly, XSum, ELI5	10,000	30,000
External test	WritingPrompts, CodeAlpaca-20k, GSM8K	2,000	6,000

Bit-width questions use 80 templates (60 for training, 20 held-out), mapped many-to-one onto the source prompts above. During training we monitor the held-out templates and the 240-question self-report set from Section 2.3.

E.3 Training Curve



(a) 16/8/4-bit models (collaborative training)

(b) 3 16-bit models (false control)

(c) 16/8/4-bit models (collaborative training)

(d) 3 16-bit models (false control)

(e) 16/8/4-bit models (collaborative training)

(f) 16/8/4-bit models (collaborative training)

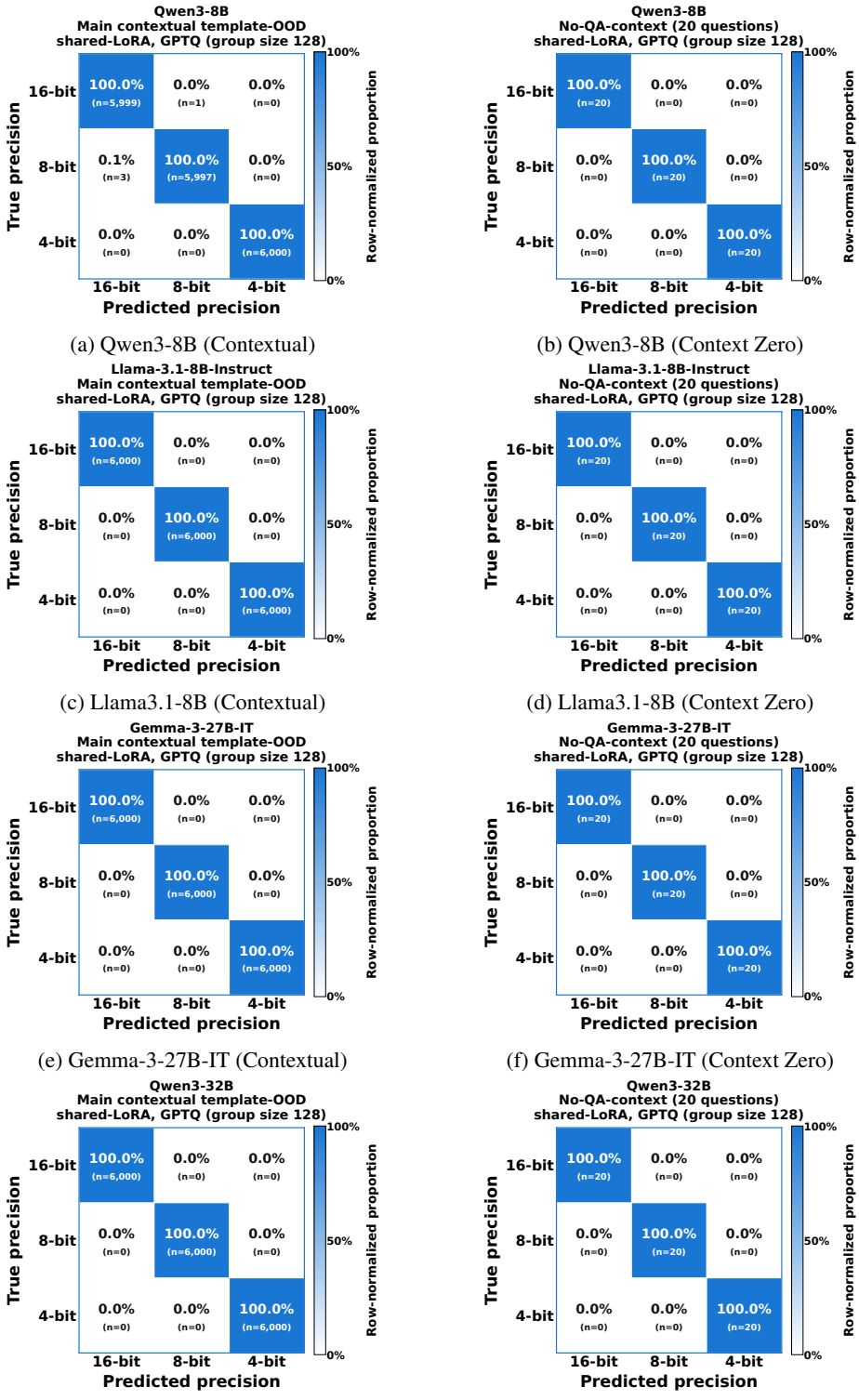
Figure 19: Accuracy on held-out templates during shared LoRA training. **(a,c,e,f)** 16/8/4-bit backbones: accuracy surges to almost 100% within 600 steps, indicating that fingerprint readout is acquired. **(b,d)** False control with three identical 16-bit backbones (Qwen3-8B and Llama-3.1-8B only): accuracy remains at chance throughout training.

E.4 Accuracy Table

Table 15: Balanced accuracy on LoRA training to identify bit-width of the generated text.

	Pointwise		Pairwise		Setwise	
Model	Base	LoRA	Base	LoRA	Base	LoRA
Qwen3-8B	0.334	0.353	0.492	0.583	0.342	0.619
Llama3.1-8B-Instruct	0.341	0.378	0.487	0.589	0.381	0.612
Gemma-3-27B-IT	0.328	0.382	0.502	0.640	0.346	0.612
Qwen3-32B	0.348	0.390	0.502	0.601	0.364	0.627

E.5 Confusion Matrix of OOD Template Contextual



(a) Qwen3-8B (Contextual)

(b) Qwen3-8B (Context Zero)

(c) Llama3.1-8B (Contextual)

(d) Llama3.1-8B (Context Zero)

(e) Gemma-3-27B-IT (Contextual)

(f) Gemma-3-27B-IT (Context Zero)

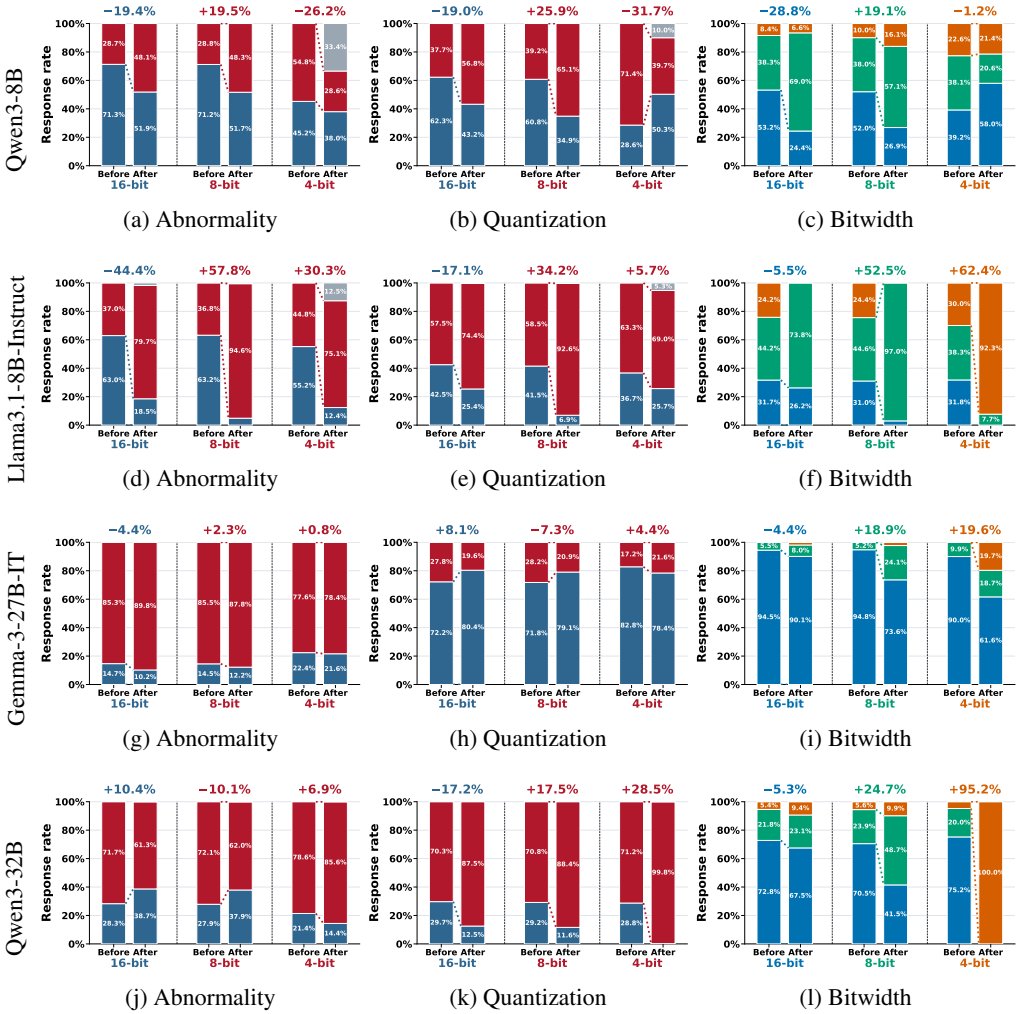
(g) Qwen3-32B (Contextual)

(h) Qwen3-32B (Context Zero)

Figure 20: Confusion matrix of OOD template contexts for all models.

E.6 Self-Report Before and After

E.6.1 Internal Awareness After Training



(a) Abnormality (b) Quantization (c) Bitwidth

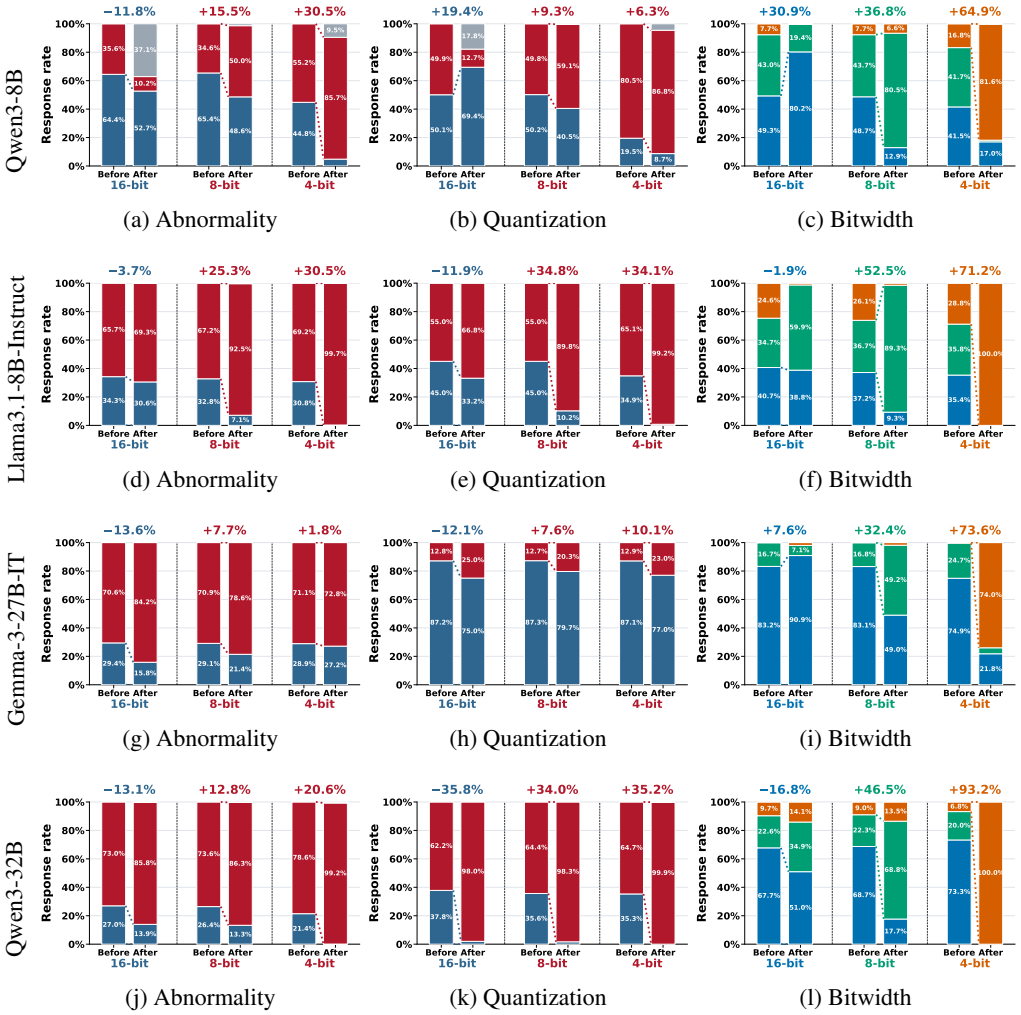
(d) Abnormality (e) Quantization (f) Bitwidth

(g) Abnormality (h) Quantization (i) Bitwidth

(j) Abnormality (k) Quantization (l) Bitwidth

Figure 21: **Internal Awareness self-report:** Abnormality, Quantization, Bitwidth for all models before and after training (shared LoRA).

E.6.2 External Awareness After Training



(a) Abnormality (b) Quantization (c) Bitwidth

(d) Abnormality (e) Quantization (f) Bitwidth

(g) Abnormality (h) Quantization (i) Bitwidth

(j) Abnormality (k) Quantization (l) Bitwidth

Figure 22: **External Awareness self-report:** Abnormality, Quantization, Bitwidth for all models before and after training (shared LoRA).