Does the Model Still Know? Evidence-State Probe Transfer and RAG Behaviour Under Weight Quantization

Kamil Hassan
Embedded LLM
kamil.hassan@embeddedllm.com
kamilhassan201@gmail.com

Abstract

Retrieval-augmented generation (RAG) systems must know whether the retrieved documents support an answer, lack one, or conflict. This evidence state (Answer / Refuse / Conflict) can be read from hidden states by a cheap linear probe, which could route queries before generation. We test whether such a probe still works after quantization, which is common in deployment. On four instruction-tuned 4B–8B models, we compare unquantized bfloat16 weights with GPTQ at 8, 4 and 3 bits and 4-bit NormalFloat, using an existing benchmark and probe recipe. Activations and computation stay in bfloat16. Used unchanged (frozen), the bfloat16 model's probe loses nothing measurable at 8 bits, averaged over the near-best bfloat16 layers. At 4 bits it trails a probe retrained on the quantized model by 0.02–0.21 accuracy, and at 3 bits by 0.19–0.38. On the same layers, retrained probes stay within 0.04 of bfloat16, so the evidence state can still be read. When a frozen probe fails, its errors fall mostly on one or two evidence states, which differ by model. Exploratory label-free maps suggest the signal is largely kept but linearly re-encoded. Generated responses are less accurate than the probes even at bfloat16. Quantization widens this gap for specific classes at 3 bits and, on one model, already at 4 bits. Probes can be reused at 8 bits but should be checked or refit below that.

1 Introduction

Retrieval-augmented generation (RAG) conditions a language model on retrieved documents (Lewis et al., 2020), which may contain the answer, lack it, or contradict each other; each case calls for a different response: answer, abstain, or flag the conflict. Models handle the last two cases poorly: they often answer instead of abstaining when the evidence is insufficient (Joren et al., 2025), and with mixed evidence they show confirmation bias towards the part that agrees with their parametric memory (Xie et al., 2024).

Huq et al. (2026) call this three-way distinction the *evidence state* (Answer / Refuse / Conflict) and show that it is linearly decodable from hidden states before generation, with probes that outperform prompting baselines on almost all of the 16 models they test. Such a cheap linear probe could act as a router or monitor before the model answers.

In deployment, the probe would often run next to a quantized model. Quantizing the weights to 8 or 4 bits, e.g., with GPTQ (Frantar et al., 2023) or 4-bit NormalFloat (NF4; Dettmers et al., 2023), cuts memory cost, and weight-only 4-bit models usually stay close to full precision on benchmarks (Li et al., 2024). Benchmarks, however, do not tell us whether a full-precision probe still reads the quantized model correctly, or whether behaviour degrades together with the internal signal. Huq et al. (2026) suggest that low-bit quantization may explain why their attention-based probe peaks extremely early for Falcon3-7B-1.58bit, the one model they list as quantized, but do not test this or compare precisions.

We change one variable, the precision of the weights, and keep activations, the key–value (KV) cache and all computation in bfloat16 (BF16). On four instruction-tuned models, Qwen3-4B-Instruct-2507 and Qwen3-8B (Yang et al., 2025), Llama-3.1-8B-Instruct (Grattafiori

et al., 2024) and Granite-4.1-8B (Granite Team, IBM, 2026), we compare the BF16 model with GPTQ at 8, 4 and 3 bits (W8, W4, W3) and with NF4. We use the benchmark and probe recipe of Huq et al. (2026) and ask three research questions:

RQ1 Is the evidence state still linearly decodable after quantization, overall and per class?
RQ2 Does a probe trained on BF16 activations still work, frozen, on the quantized model's activations? If not, what changed in the representation?
RQ3 Does quantization make the model "know but fail to act", so that a probe still reads the evidence state while the generated behaviour degrades?

Frozen BF16 probes transfer without loss to W8, lose accuracy at 4 bits and lose 0.19–0.38 at W3, although the evidence state is still encoded. This W3 loss is the transfer gap (accuracy of the retrained, or *native*, probe minus that of the frozen BF16 probe), averaged over the near-best BF16 layers (the plateau); native probes lose only 0.03–0.04 there (Section 4.2). Frozen-probe failures concentrate on one or two classes that depend on the model and variant (Section 4.2.1). Label-free re-scaling closes only part of the gap, and exploratory label-free maps suggest a linear re-encoding of a largely preserved signal rather than lost information (Section 4.2.2). For specific classes, quantization widens a gap between probe and behaviour that already exists at BF16, on Llama-3.1-8B already at 4 bits (Section 4.3).

We contribute (i) a study of frozen cross-precision probe transfer for RAG evidence states, with four models, four quantized variants, and BF16 and random-label controls; (ii) a label-free diagnosis of per-dimension shift versus rotation and linear re-encoding; (iii) a per-class analysis of decodability (RQ1) and of the gap between knowing and acting (RQ3), with Holm-corrected tests and a claim rule; and (iv) released code and analysis outputs.[1]

2 Related Work

Evidence states in hidden states. Hidden states often hold more than the output shows: models that answer unanswerable questions still encode that the question is unanswerable (Slobodkin et al., 2023), the closest precedent for our Refuse class. Zhao et al. (2024) decode a conflict between the context and the model's memory with logistic-regression probes, which in later layers are more accurate on the residual stream than on multilayer perceptron (MLP) or attention outputs. Huq et al. (2026) cover Answer, Refuse and Conflict in one benchmark, where the conflict lies between retrieved documents (inter-context conflict in the taxonomy of Xu et al., 2024). We reuse their benchmark and probe recipe and ask what quantizing the weights does to the probe and to behaviour.

Frozen probes across weight changes. Huq et al. (2026) find that their probe transfers among fine-tuned variants of one model but not to or from its base model. Two preprints apply an unchanged probe to new weights. Tsatsos (2026) applies logistic-regression probes of base models, with the base model's standardization, to distilled and instruction-tuned successors; the probes keep a similar area under the ROC curve (AUC) although the class-mean direction rotates by 17–35°. Precision is fixed in that study (its 32B models use NF4 throughout). Duan (2026) trains logistic-regression safety monitors on BF16 activations of two models, freezes them, and finds that their AUC barely changes under several quantization methods, including NF4 and GPTQ, whereas LoRA fine-tuning often degrades them. Neither preprint separates a mean shift from a rotation, and the second names no setting below 4 bits and reports no comparison with generated behaviour. It also repairs monitors that went stale after fine-tuning with a label-free map fitted on paired activations; we use such maps as a diagnosis. We test evidence-state probes per class down to 3 bits and whether refitting the scaler alone restores them.

Internals of quantized models. Output-level evaluations find a noticeable decline at 3 bits for small models (Li et al., 2024), and aggregate accuracy can hide answers that flip (Dutta et al., 2024). Fewer studies look inside. Fu et al. (2025) report that quantized models, like

[1] https://github.com/kamil-hassan201/quantized-evidence-probes; see Appendix I.

full-precision ones, keep truthful internal representations yet can be pushed to false outputs by deceptive prompts; their probes are shown for one model and its 4-bit variant, with no transfer between the two. Zhou et al. (2026) analyse GPTQ without supervised probes: at 4 bits the signal typically degrades while computation stays intact, and at 2 bits computation collapses. For weight-only round-to-nearest quantization of Qwen3-32B to the 4-bit NVFP4 format, Chen et al. (2026) find that most of the hidden-state error is a change of direction rather than of norm. Bhandari & Wade (2026) re-extract steering vectors at each precision, find them nearly collinear at INT8 and NF4, and leave reusing full-precision vectors to future work. To our knowledge, no prior work tests frozen cross-precision probe transfer for RAG evidence states, separates a per-dimension shift from other changes with label-free re-scaling, or compares per class how probe and behaviour degrade.

3 Methodology

3.1 Dataset

We use the evidence-state benchmark of Huq et al. (2026), in which each instance pairs a question with five documents. Answer instances contain the gold passage and four retrieved distractors; Refuse instances contain only distractors; Conflict instances contain the gold passage, a copy whose answer entity is replaced by a counterfactual one, and three distractors. Questions come from TriviaQA (Joshi et al., 2017), HotpotQA (Yang et al., 2018) and Natural Questions (Kwiatkowski et al., 2019), and each question appears once in each state; to reduce parametric knowledge leakage, the benchmark authors apply counterfactual entity substitution across the dataset, following Xie et al. (2024). We use the authors' released files (commit `2b338ce`) unchanged, with their split by question: 2,396 questions and 7,188 instances (5,463 train, 285 validation, 1,440 test), every split balanced across classes; their paper reports slightly different counts (Appendix A).

3.2 Models and Quantization

We study Qwen3-4B-Instruct-2507 (36 layers, hidden size 2,560; hereafter Qwen3-4B), Qwen3-8B with thinking disabled (36 / 4,096), Llama-3.1-8B-Instruct (32 / 4,096; hereafter Llama-3.1-8B) and Granite-4.1-8B (40 / 4,096), which is instruction-tuned although its name lacks "Instruct". We quantize only the weights of the attention and MLP projections; activations, the KV cache, all computation, embeddings, the language-model head and normalization layers stay in BF16. The GPTQ (Frantar et al., 2023) variants W8, W4 and W3 come from GPTQModel,[2] calibrated on 256 sequences of 2,048 tokens from C4 (Raffel et al., 2020) (seed 42); its PyTorch backend multiplies with weights dequantized to BF16. NF4 (Dettmers et al., 2023), a fixed data type that needs no calibration data, is applied at load time with bitsandbytes;[3] "4 bits" below means W4 and NF4. WikiText-2 perplexity (Merity et al., 2017) serves as a sanity check; a GPTQ 2-bit variant (W2) collapses (11–48$\times$ the BF16 perplexity) and is shown only for reference (settings and checks: Appendix B, Table 4).

3.3 Feature Extraction and Probing

We extract features with the naive prompt of Huq et al. (2026) (P0 in their appendix; our Appendix C) in each model's chat template. We read the last prompt token and record at every layer the residual stream after the block (**H**) and the MLP output before it joins the residual stream (**M**); the main text reports **H**, the appendix adds **M**. Per layer and feature we apply their probe recipe: a standard scaler and multinomial logistic regression (C=1, L-BFGS) as in their code, and sigmoid calibration as in their paper, fitted on the validation split as in their code (Appendix A), in scikit-learn.[4] Test predictions use the calibrated probabilities. We train these probes on each variant's own activations and call them that variant's *native* (retrained) probes; RQ1 and RQ3 read them at the layer with the highest uncalibrated

[2] https://github.com/ModelCloud/GPTQModel

[3] https://github.com/bitsandbytes-foundation/bitsandbytes

[4] https://scikit-learn.org

validation accuracy. As in Huq et al. (2026), we report the false-answer rate (FAR), the share of Refuse and Conflict items labelled Answer, and refit the selected-layer probe on shuffled labels (five seeds, uncalibrated) as a random-label control (cf. Hewitt & Liang, 2019).

3.4 Cross-Precision Probe Transfer

For RQ2 we apply three probes to the quantized model's test activations X_ℓ^q at layer ℓ: the native probe f_ℓ^q, the *frozen probe* f_ℓ^{BF16} (the BF16 probe with scaler, regression and calibration unchanged), and the *re-scaled probe*, which refits only its scaler on the quantized model's unlabelled training activations, removing per-dimension changes in mean and scale. The *transfer gap* is

$$\mathrm{Gap}_\ell = \mathrm{Acc}\big(f_\ell^q, X_\ell^q\big) - \mathrm{Acc}\big(f_\ell^{\mathrm{BF16}}, X_\ell^q\big), \tag{1}$$

and the re-scaled gap uses the re-scaled probe in place of f_ℓ^{BF16}. Frozen accuracy can change sharply between neighbouring layers, so we report the overall and re-scaled gaps as per-item averages of Eq. 1 over the BF16 plateau: the layers whose BF16 validation accuracy is within 0.03 of the best. Per-class gaps and the diagnostics below use the BF16-selected layer (highest BF16 validation accuracy); plateau and layer are chosen per feature. A gap left after re-scaling does not by itself show a rotation, so we added two exploratory tests after the main RQ2 results (Appendix F). First, the frozen probe reads the output of label-free maps from quantized to BF16 activations, fitted on paired training items: per-dimension affine, re-scaling followed by an orthogonal Procrustes rotation, and ridge (upper bound). Second, we compare the cosine between native and BF16 probe weights with a null from refitting the BF16 probe on training resamples. If Procrustes closes significantly more of the gap than re-scaling alone (paired bootstrap confidence interval, CI, of the difference above zero; Section 3.7), we read a rotation component; if only ridge closes the rest, a general linear re-encoding.

3.5 Behaviour Readout

For behaviour, the KRE prompting baseline of Huq et al. (2026) replaces the P0 system message for all four models; it is adapted from the instructions of the Knowledge Robustness Evaluation (KRE) dataset (Ying et al., 2024) and warns that the documents may be incomplete, misleading or conflicting. We chose it over their OPIN_INSTR baseline (adapted from the prompts of Zhou et al., 2023) on BF16 validation (Appendices C and D). We decode greedily, up to 512 new tokens (128 for prompt selection), one instance per call. A large language model (LLM) judge (Zheng et al., 2023), Gemini 3.1 Flash-Lite (Google DeepMind, 2026) at temperature 0, labels each generation from its text alone (the *generation readout*): Conflict if the output notices a conflict, else Refuse if it notices insufficient evidence, else Answer. We call this a detection reading: the label records what the output notices. A generation that notices the problem but still answers therefore counts as Conflict or Refuse. We keep a regex parser for comparison (Appendix D); a forced-choice readout (Appendix H) is descriptive only, not evidence.

3.6 Knows-but-Fails Analysis

For class c, $\Delta P(c)$ and $\Delta B(c)$ are the changes from BF16 in native-probe recall and behaviour recall. Let S_c be the class-c test items that BF16 behaviour gets right, and $\mathcal{P}$ and $\mathcal{B}$ the items on which the quantized model's native probe and its behaviour are right. The knows-but-fails rate (KBF) and the lost-signal rate are

$$\mathrm{KBF}(c) = \frac{\big|(S_c \cap \mathcal{P}) \setminus \mathcal{B}\big|}{|S_c \cap \mathcal{P}|}, \qquad \mathrm{Lost}(c) = \frac{\big|S_c \setminus (\mathcal{P} \cup \mathcal{B})\big|}{|S_c|}, \tag{2}$$

and the excess behaviour loss is $\Delta B(c) - \min(\Delta P(c), 0)$. We claim a case (model, variant, class) when (i) $\Delta B(c)$ is significantly negative (McNemar, Holm-corrected $p < 0.05$), (ii) the probe is not worse ($\Delta P(c) \geq -0.02$ and its CI is not entirely below 0), and (iii) $\mathrm{KBF}(c) > \mathrm{Lost}(c)$. A case is *partial* when (i) and (iii) hold and the excess-loss CI lies below 0, but (ii) fails. Claims use the **H** probe (higher validation accuracy in 16 of 20 model–variant pairs).

3.7 Statistics and Implementation

CIs, shown in brackets, are 95% percentile bootstrap intervals (10,000 resamples over test items), paired for differences. "No change" and "without loss" mean $|\Delta| \leq 0.02$ with a CI that covers 0. Per-class changes use McNemar's test, Holm-corrected within each RQ and model over 30 tests: five quantized variants $\times$ three classes $\times$ two features (RQ1) or two readouts, generation and forced choice (RQ3). Including the collapsed W2 and forced choice only makes the correction more conservative; neither is used as evidence. Elsewhere, "significant" means that the 95% CI excludes 0, without adjustment for multiplicity. The P0 prompt, the KRE and OPIN_INSTR templates, the probe recipe and the base parsing rules come from the released code of Huq et al. (2026); the prompt assembly, the regex parser extensions and the LLM parser are ours. All runs use one NVIDIA A10 GPU (24 GB); library versions, hyperparameters and seeds are in Appendix I.

4 Results

4.1 RQ1: Retrained Probes Still Decode the Evidence State

A probe retrained on each quantized model still decodes the evidence state, down to W3. Native accuracy stays at 0.85–0.92 on every model and variant, far above the 0.32–0.34 of random-label controls (Table 1). The largest cost is at W3: 0.02–0.07 at each variant's own best layer (paired changes: Appendix Table 10), and 0.03–0.04 averaged over the BF16 plateau layers used for RQ2 (Figure 1). This is the premise for RQ2: the information a frozen probe needs is still largely present in every quantized model.

Model	BF16	W8	W4	NF4	W3
Qwen3-4B	0.92 [0.91, 0.93]	0.92 [0.90, 0.93]	0.91 [0.90, 0.93]	0.91 [0.90, 0.92]	0.89 [0.87, 0.91]
Qwen3-8B	0.91 [0.90, 0.93]	0.92 [0.91, 0.94]	0.89 [0.87, 0.90]	0.90 [0.89, 0.92]	0.85 [0.83, 0.87]
Llama-3.1-8B	0.92 [0.91, 0.94]	0.92 [0.91, 0.93]	0.91 [0.90, 0.93]	0.91 [0.90, 0.93]	0.89 [0.87, 0.91]
Granite-4.1-8B	0.91 [0.90, 0.93]	0.92 [0.91, 0.94]	0.92 [0.91, 0.93]	0.91 [0.89, 0.92]	0.90 [0.88, 0.91]

Table 1: RQ1: test accuracy of native residual-stream (**H**) probes, each read at its own validation-selected layer. Brackets: 95% bootstrap CIs (1,440 test items). Random-label controls reach 0.32–0.34. Appendix Table 10 adds layers, false-answer rates, paired accuracy changes (computed before rounding), macro-F1 and perplexity.

Per class, the changes are mostly small, and which class slips first depends on the model. After Holm correction, only five of the 48 changes (4 models $\times$ 4 variants $\times$ 3 classes) are significant, all at W4 or W3: drops for Answer on both Qwen models and for Conflict on Granite-4.1-8B, and one gain. The largest drop is -0.12 $[-0.16, -0.08]$, for Answer on Qwen3-8B at W3; at the BF16-selected layer it is -0.03 $[-0.06, 0.00]$, which suggests layer-selection noise (Appendix E). Llama-3.1-8B shows no significant change (Appendix Figure 5).

4.2 RQ2: Frozen BF16 Probes on Quantized Activations

A frozen BF16 probe transfers without loss to W8, but not to 4 or 3 bits (Figure 1). At W8, the plateau-averaged transfer gap is within 0.01 of zero on every model, and the quantized activations are almost identical to the BF16 ones (mean cosine 1.00; Appendix Table 14). At 4 bits the plateau gap is never zero, and its size depends on the model: 0.09–0.21 on the two Qwen models but only 0.02–0.05 on Llama-3.1-8B and Granite-4.1-8B, with every CI excluding 0 (Appendix F). NF4 has the smaller gap on every model, but we ran no paired test between the two methods.

At W3 the frozen probe loses 0.19–0.38 accuracy on every model, although the evidence state is still decodable. This transfer gap is larger than at 4 bits on every model, while native probes on the same layers lose only 0.03–0.04 (dark grey bars in Figure 1). Averaged over the plateau, the frozen probe is still above chance (0.49–0.70 against 0.32–0.34 for random-label

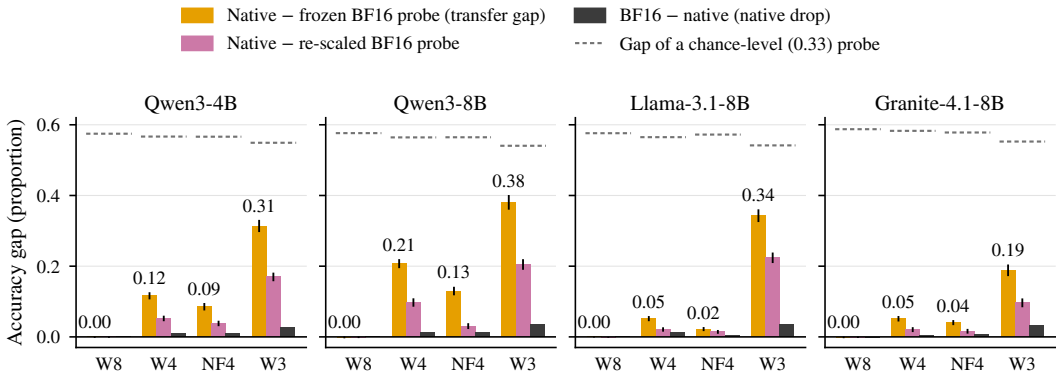

Figure 1: Frozen cross-precision transfer of residual-stream (**H**) probes, averaged over the BF16 plateau (layers within 0.03 of the best BF16 validation accuracy). Native: retrained on the variant; re-scaled: only the scaler refit. Error bars: 95% paired-bootstrap CIs (none for the native drop). W8 transfers without loss.

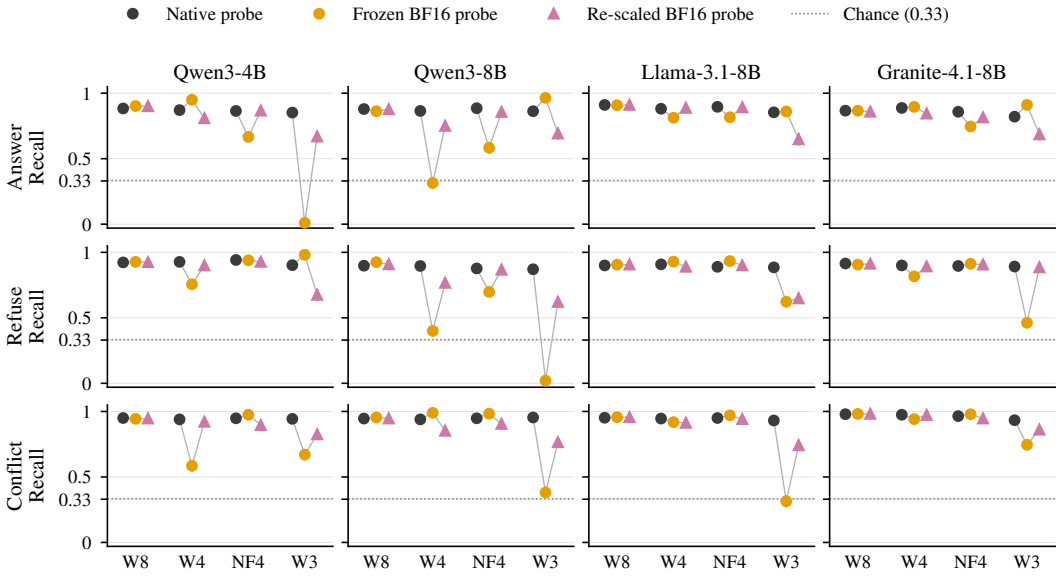

Figure 2: Per-class recall of the native, frozen BF16 and re-scaled BF16 probes at the BF16-selected layer (residual stream **H**). Grey lines join the three probes of one variant; error bars (95% paired-bootstrap CIs of the gap to the native probe) are mostly hidden by the markers. When the frozen probe fails, it loses much of its recall on one or two classes, which depend on the model and variant; re-scaling removes most of this collapse.

controls), so part of what it learned at BF16 still applies. The gap is at least 0.10 on every plateau layer (Appendix Figure 8), and **M** shows the same pattern (Appendix F).

Equal perplexity can hide unequal transfer. WikiText-2 perplexity rises in the same coarse order as the gap (Appendix Table 7), but W4 and NF4 on Llama-3.1-8B have the same perplexity (7.62) and gaps of 0.05 [0.04, 0.06] and 0.02 [0.02, 0.03].

4.2.1 Where Frozen Probes Fail

When a frozen probe fails, it loses much of its recall on one or two evidence states, and which ones depends on the model and variant (Figure 2). Overall accuracy hides these failures. At the BF16-selected layer of Qwen3-4B at W3, frozen-probe accuracy is still 0.55, but its Answer recall is 0.01 (native: 0.85; gap 0.84 [0.81, 0.87]), while its Refuse recall is 0.98.

At W3, the hardest-hit classes are Refuse and Conflict on Qwen3-8B, Conflict on Llama-3.1-8B and Refuse on Granite-4.1-8B (the mildest: recall 0.46 against 0.89). At 4 bits the failures are smaller but class-specific: on Qwen3-8B at W4, the frozen probe over-predicts Conflict, and at NF4, Answer has the largest gap on every model (Appendix F).

4.2.2 *What Changes in the Representation*

Label-free re-scaling removes most of the collapse onto one class, but not the whole gap. Refitting only the probe's scaler on unlabelled quantized activations removes 35–76% of the 4- and 3-bit plateau gaps (pink bars, Figure 1). Within each model, it helps least at W3 (35–49%), and every remaining gap at 4 and 3 bits has a CI above 0. Per class, the frozen probe at W3 assigns 55–84% of the class-balanced test items to a single class; after re-scaling, no class receives more than 38% (pink triangles, Figure 2). On Qwen3-4B, Answer recall rises from 0.01 to 0.67, while Refuse recall falls from 0.98 to 0.68. The remaining loss is spread over the classes: re-scaled recall still trails the native probe in 11 of the 12 model–class pairs at W3, by 0.07–0.25 (each CI above 0; Appendix F). The collapse is thus consistent with a per-dimension shift in mean and scale, but quantization does more than shift and re-scale each dimension independently.

		Transfer gap after map					Gap difference
Model	Var.	Frozen	Re-scaled	Affine	Procrustes	Ridge	Re-scaled − Procrustes
Qwen3-4B	W4	0.15	0.04	0.03	0.01°	−0.02	0.03 [0.01, 0.04]
	NF4	0.06	0.02	0.02	0.02	−0.01°	0.00 [−0.01, 0.01]°
	W3	0.35	0.18	0.19	0.07	−0.01°	0.11 [0.09, 0.14]
Qwen3-8B	W4	0.33	0.11	0.10	0.03	−0.01°	0.08 [0.06, 0.10]
	NF4	0.15	0.03	0.02	0.01°	−0.02	0.01 [0.00, 0.03]°
	W3	0.44	0.20	0.21	0.08	0.00°	0.12 [0.09, 0.15]
Llama-3.1-8B	W4	0.03	0.02	0.01°	0.02	−0.01°	−0.01 [−0.02, 0.00]°
	NF4	0.00°	0.00°	0.00°	0.00°	−0.01°	0.00 [−0.01, 0.01]°
	W3	0.29	0.21	0.20	0.13	0.00°	0.08 [0.05, 0.11]
Granite-4.1-8B	W4	0.04	0.02	0.02	0.02	0.00°	0.00 [−0.02, 0.01]°
	NF4	0.03	0.02	0.02	0.01	−0.01°	0.00 [−0.01, 0.02]°
	W3	0.18	0.07	0.08	0.04	−0.01°	0.04 [0.02, 0.06]

Table 2: Transfer gap at the BF16-selected layer (**H**) after label-free maps of the quantized activations: none (frozen), scaler refit (re-scaled), per-dimension affine, re-scaling plus rotation (Procrustes), full linear (ridge). Brackets: 95% paired-bootstrap CI; °: CI includes 0.

At W3, and at W4 on both Qwen models, the residual change is consistent with a linear re-encoding of a largely preserved signal, with a rotation component (Table 2). At the BF16-selected layer, re-scaling plus a rotation (Procrustes) closes significantly more of the gap than re-scaling alone in all these cases, by 0.04–0.12 at W3 (exploratory label-free maps, Section 3.4; W8 and further CIs: Appendix Table 14). On Qwen3-4B at W4 there is no gap one layer earlier (Appendix Figure 8). At W3 Procrustes still leaves a gap of 0.04–0.13; a full linear (ridge) map closes it: every W3 gap then has a CI that covers 0. In the other six 4-bit cases Procrustes gives no significant gain: the gaps are at most 0.04 on Llama-3.1-8B and Granite-4.1-8B, and re-scaling removes most of the NF4 gap on the Qwen models. Probe directions change too, but raw cosines between native and BF16 probe weights would overstate this: refitting the BF16 probe on resamples of its own training data already gives only 0.67–0.90 (Appendix Table 16). W8 cosines lie above their class's null range, and 32 of the 36 lower-precision ones below it.

4.3 RQ3: Knowing but Failing to Act

Even at BF16, probes read the evidence state better than the generations show it. Probe accuracy is 0.91–0.92, but the generation readout matches the evidence state on only 0.61–0.76 of the items (Appendix Table 9). The gap is largest for Conflict: the probe recognizes almost every Conflict item, yet the output flags at most about half of them.

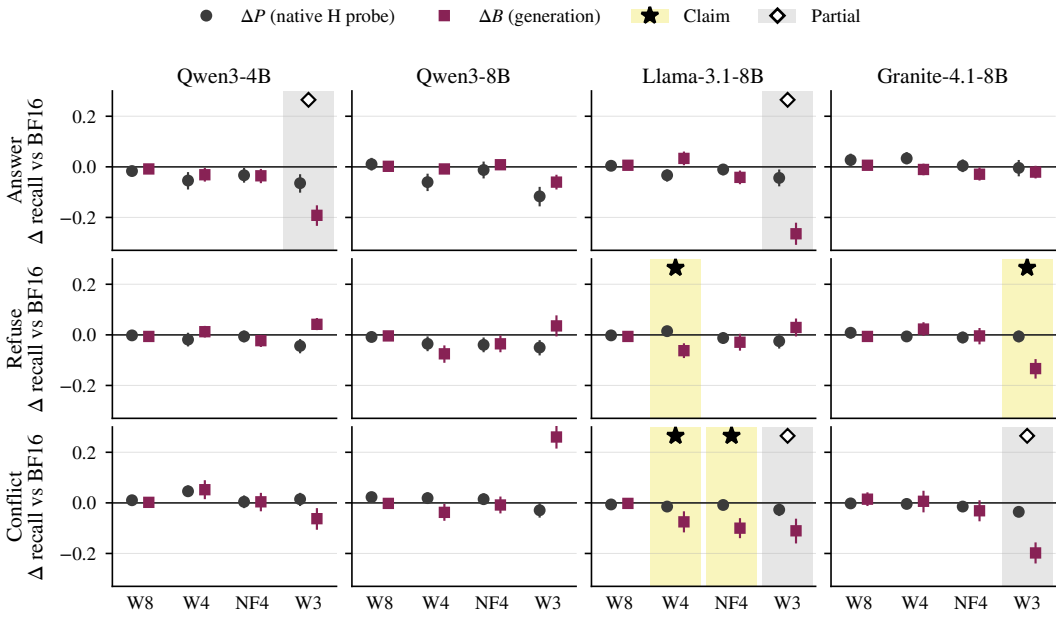

Figure 3: Per-class recall change vs. BF16 for the native **H** probe (ΔP) and the LLM-judged generation readout (ΔB), with 95% paired-bootstrap CIs. Stars: claims, where behaviour drops significantly while the probe does not (Section 3.6); diamonds: partial cases, where behaviour loses significantly more than a probe that also degrades. No W8 case qualifies.

Quantization widens this gap in a few cases: behaviour gets worse while the probe still reads the evidence state about as well as at BF16 (Figure 3). Four cases meet our claim rule (Appendix Table 17): three on Llama-3.1-8B at 4 bits (Refuse and Conflict) and one on Granite-4.1-8B at W3. In the clearest 4-bit case, Llama-3.1-8B at NF4, generation Conflict recall falls by 0.10 [0.06, 0.14], while probe recall changes by less than 0.02. In all four the loss lies in outputs that notice the problem but still answer, so the two Refuse claims do not hold if Refuse requires abstaining (Appendix Table 20). In four more W3 cases the probe slips too, but behaviour loses significantly more. In all eight, the evidence state is still linearly decodable, which suggests that what breaks is the step from representation to output. Qwen3-8B has no claim with the **H** probe, but two of its cases qualify with **M**, and one of them with **H** read at the BF16-selected layer (Appendix G).

At W3, behaviour fails in a direction that depends on the model. Qwen3-4B and Llama-3.1-8B become more cautious: over-abstention (Answer items labelled Refuse or Conflict) rises by 0.19 and 0.26. Granite-4.1-8B does the opposite: it flags the problem less often (FAR 0.41 to 0.56). So a quantized model need not fail in the cautious direction.

Two caveats qualify these results, besides the detection reading (Section 3.5). First, under our initial 128-token cap, which cut off 8–70% of a run's outputs, Qwen3-4B Conflict at W3 was a fifth claim, which the 512-token outputs do not support (Appendix Table 18); we cannot explain the Conflict gain on Qwen3-8B at W3. Second, condition (iii) adds little: with probe accuracy near 0.90, almost no item loses both probe and behaviour, so KBF exceeds the lost-signal rate in every case meeting (i); the claims rest on (i) and (ii) (Appendix G).

5 Discussion

The gap between knowing and acting exists at BF16; quantization widens it for specific classes and models. At W3, retrained probes lose little accuracy (Section 4.1), whereas behaviour loses significantly more than the probe in some model-specific cases (Section 4.3). This fits reports that hidden states hold more than the output shows (Section 2) and that quantized models keep truthful internal representations yet can be pushed to false outputs

by deceptive prompts (Fu et al., 2025). It also suggests that a probe on the deployed model could flag evidence states that the answer misses, most of all Conflict, though we did not test such a router end to end. The evidence is correlational: a decodable evidence state does not show that the model uses it (Belinkov, 2022), some probe–output gaps reflect calibration rather than hidden knowledge (Liu et al., 2023), and the comparison is uneven (a probe trained on labelled items under P0 versus zero-shot behaviour under KRE).

The label-free maps suggest that low-bit quantization re-encodes the evidence state rather than erasing it. At W3, Procrustes closes significantly more of the gap than re-scaling alone, and the ridge map closes the rest (Section 4.2.2). We read this as a rotation-like change plus anisotropic scaling or mixing, not a pure rotation; native probes still decode the signal with little loss. The direction-dominated hidden-state error that Chen et al. (2026) report for NVFP4 cannot tell a mean shift from a rotation, since both change direction. Zhou et al. (2026) find, without trained probes, that 4-bit GPTQ typically degrades the signal while computation stays intact; in our 4-bit models the evidence state stays decodable while frozen probes lose accuracy. These tests are exploratory; the ridge map shows only that a linear map exists.

A BF16 probe used as a router or monitor can be reused at W8, but below 8 bits it should be validated per precision and per class. At 4 bits the plateau gap is never zero, and failures concentrate in classes that overall accuracy hides (Section 4.2.1). Duan (2026) finds that frozen monitors largely keep their AUC under quantization, including NF4. Our probes do at 4 bits (AUC 0.01–0.05 below the native probe's) but not at W3 (0.06–0.15 below; Appendix F). This suggests that part of the 4-bit gap is a shifted decision threshold.

Label-free re-scaling is a cheap partial fix. It removes most of the collapse onto one class, but not the whole gap (Section 4.2.2). A linear map recovers most of the rest, but needs paired activations from both precisions. Retraining is also cheap: evidence-state labels belong to the input, not the model, so it costs only forward passes and a logistic regression. Equal perplexity did not mean equal gaps (Section 4.2); we did not test it as a predictor.

NF4 had a smaller gap than W4 on every model, but we ran no paired test and cannot say why. The methods differ in calibration (Section 3.2) and block size (64 against 128; Appendix B), and quantization effects are method- and model-dependent elsewhere too (Mekala et al., 2025); separating these factors needs a paired test and ablations.

6 Conclusion

We changed only the weight precision of four open models with up to 8B parameters. Frozen BF16 evidence-state probes transfer without loss to W8 but lose accuracy below it, 0.19–0.38 at W3, although retrained probes still decode the evidence state. Exploratory label-free maps suggest a linear re-encoding of a largely preserved signal. Frozen-probe failures concentrate on one or two model-dependent classes, and behaviour can degrade more than the internal signal, on Llama-3.1-8B already at 4 bits. Probes can be reused at W8 but should be validated or refit at each deployment precision below 8 bits; label-free re-scaling closes only part of the gap. Future work should test KV-cache and activation quantization, other methods, larger models, and whether the preserved signal can steer or repair behaviour.

Limitations

We study four open-weight instruction-tuned models of 4B–8B parameters, the largest that fit our single 24 GB GPU, on one English benchmark with synthetic, entity-substitution conflicts (Huq et al., 2026); larger models, natural conflicts and other languages may differ. We quantize weights only, so activation and KV-cache quantization, which change the probed activations directly, are untested, as are other methods such as AWQ (Lin et al., 2024); our naive GPTQ W2 collapses, so we say nothing about usable 2-bit models. Behaviour uses one prompt per model, greedy decoding, a token cap that changed one claim, and an LLM judge checked by hand on only 24 outputs. Each probe is fitted once on the fixed training split, so CIs do not cover training-sample variance. Appendix J gives the full list, including analysis choices made after seeing part of the results.

Acknowledgments

This work was carried out at Embedded LLM, which provided the computing resources for all experiments.

The author used a generative AI assistant to help edit and format the manuscript and to assist with the analysis and plotting code. The author reviewed and verified all content and takes full responsibility for it.

References

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Saurav Bhandari and Benjamin Wade. Steering under compression: Dose-response, capability cost, and failure asymmetry in quantized LLMs. *arXiv preprint arXiv:2609.06473*, 2026. URL https://arxiv.org/abs/2609.06473.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa. QuIP: 2-bit quantization of large language models with guarantees. In *Advances in Neural Information Processing Systems*, volume 36, pp. 4396–4429, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/0df38cd13520747e1e64e5b123a78ef8-Abstract-Conference.html.

Yuxiang Chen, Michael Beyer, Jun Zhu, and Jianfei Chen. Why does post-training quantization work? *arXiv preprint arXiv:2609.11716*, 2026. URL https://arxiv.org/abs/2609.11716.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30318–30332, 2022. doi: 10.52202/068431-2198. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c3ba4962c05c49636d4c6206a97e9c8a-Abstract-Conference.html.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 10088–10115. Curran Associates, Inc., 2023. doi: 10.52202/075280-0441. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract-Conference.html.

Evan Duan. Do activation monitors survive model updates? Benchmarking, predicting, and repairing activation-monitor staleness. *arXiv preprint arXiv:2606.15980*, 2026. URL https://arxiv.org/abs/2606.15980.

Abhinav Dutta, Sanjeev Krishnan, Nipun Kwatra, and Ramachandran Ramjee. Accuracy is not all you need. In *Advances in Neural Information Processing Systems*, volume 37, pp. 124347–124390, 2024. doi: 10.52202/079017-3950. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e0e956681b04ac126679e8c7dd706b2e-Abstract-Conference.html.

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 (ICLR)*, 2023. URL https://openreview.net/forum?id=tcbBPnfwxS. arXiv version titled "GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers".

Yao Fu, Xianxuan Long, Runchao Li, Haotian Yu, Mu Sheng, Xiaotian Han, Yu Yin, and Pan Li. Quantized but deceptive? A multi-dimensional truthfulness evaluation of quantized LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30435–30458, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1548. URL https://aclanthology.org/2025.emnlp-main.1548/.

Google DeepMind. Gemini 3.1 Flash-Lite model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-1-Flash-Lite-Model-Card.pdf, 2026. Updated May 2026; accessed 2026-11-07. Used via OpenRouter as `google/gemini-3.1-flash-lite` (GA snapshot 2026-05-07).

Granite Team, IBM. Granite 4.1 LLMs: How they're built. Hugging Face blog, https://huggingface.co/blog/ibm-granite/granite-4-1, 2026. Model used: https://huggingface.co/ibm-granite/granite-4.1-8b, released 2026-04-29; accessed 2026-11-07.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Syed Mahbubul Huq, Christopher Child, Tillman Weyde, and Pranava Madhyastha. Knowing before answering: Decoding language models for reliable RAG. In *Third Conference on Language Modeling (COLM)*, 2026. URL https://openreview.net/forum?id=3BHfBsGJW7.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/33dffa2e3d2ab74a783d1a8c292f66d9-Abstract-Conference.html.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural Questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Shiyao Li, Xuefei Ning, Luning Wang, Tengxuan Liu, Xiangsheng Shi, Shengen Yan, Guohao Dai, Huazhong Yang, and Yu Wang. Evaluating quantized large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 28480–28524. PMLR, 2024. URL https://proceedings.mlr.press/v235/li24bb.html.

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 *Proceedings of Machine Learning and Systems*, volume 6, pp. 87–100, 2024. URL https://proceedings.mlsys.org/paper_files/paper/2024/hash/42a452cbafa9dd64e9ba4aa95cc1ef21-Abstract-Conference.html.

Kevin Liu, Stephen Casper, Dylan Hadfield-Menell, and Jacob Andreas. Cognitive dissonance: Why do language model outputs disagree with internal representations of truthfulness? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4791–4797. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.291. URL https://aclanthology.org/2023.emnlp-main.291/.

Anmol Mekala, Anirudh Atmakuru, Yixiao Song, Marzena Karpinska, and Mohit Iyyer. Does quantization affect models' performance on long-context tasks? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9422–9470, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.479. URL https://aclanthology.org/2025.emnlp-main.479/.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://jmlr.org/papers/v21/20-074.html.

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. The curious case of hallucinatory (un)answerability: Finding truths in the hidden states of over-confident large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3607–3625. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.220. URL https://aclanthology.org/2023.emnlp-main.220/.

Marios Tsatsos. Linear probes transfer from base to distilled language models despite rotation. *arXiv preprint arXiv:2610.35586*, 2026. URL https://arxiv.org/abs/2610.35586.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38087–38099. PMLR, 2023. URL https://proceedings.mlr.press/v202/xiao23c.html.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=auKAUJZMO6.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8541–8565. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.486. URL https://aclanthology.org/2024.emnlp-main.486/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Jiahao Ying, Yixin Cao, Kai Xiong, Long Cui, Yidong He, and Yongbin Liu. Intuitive or dependent? Investigating LLMs' behavior style to conflicting prompts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4221–4246. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.232. URL https://aclanthology.org/2024.acl-long.232/.

Yu Zhao, Xiaotang Du, Giwon Hong, Aryo Pradipta Gema, Alessio Devoto, Hongru Wang, Xuanli He, Kam-Fai Wong, and Pasquale Minervini. Analysing the residual stream of language models under knowledge conflicts. In *Workshop on Foundation Model Interventions (MINT) at NeurIPS 2024*, 2024. URL https://openreview.net/forum?id=3Y0xm1q58x.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Chenxi Zhou, Pengfei Cao, Jiang Li, Bohan Yu, Jinyu Ye, Jun Zhao, and Kang Liu. From signal degradation to computation collapse: Uncovering the two failure modes of LLM quantization. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 23204–23222, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1162. URL https://aclanthology.org/2026.findings-acl.1162/.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14544–14556. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.968. URL https://aclanthology.org/2023.findings-emnlp.968/.

A Reproduction of the Reference Probe

Our BF16 Qwen3-4B probe (**H**, layer 22 of 36, counting from 0) reaches 0.92 test accuracy (95% CI [0.91, 0.93]), above the 0.89 reported by Huq et al. (2026). The difference is unlikely to come from numerical precision: the released extraction code also runs the model in BF16 and stores features in FP16, and an FP16 round-trip of our features leaves test accuracy unchanged. Refitting our features with the probe variants in the released code keeps test accuracy between 0.92 and 0.93, and our layer-wise validation curve closely matches theirs (0.916 vs. 0.912 at layer 20). The gap is confined to the Answer and Refuse classes. We could not resolve the cause. The paper reports 2,391 questions, a 70/10/20 split and 1,435 test instances, while the public release we use has 2,396 questions, a 76/4/20 split and 1,440 test instances. The prompt-assembly module is not in the release either, so differences in the evaluation data, in prompt construction or in document truncation cannot be ruled out.

The paper of Huq et al. (2026) states sigmoid calibration, while their released code defaults to isotonic calibration, and the released run script uses that default. We follow the paper and use sigmoid calibration. On our features, isotonic calibration fitted on the validation split gives the same test accuracy (0.92 at layer 22). Table 3 lists what we checked.

Item	Huq et al. (2026)	Ours
Model	Qwen3-4B-Instruct; the run script loads a local checkpoint, Instruct-2507 is the default in the code	Qwen3-4B-Instruct-2507
Precision	BF16 in the released extraction code (eager attention), features stored in FP16; not stated in the paper	BF16 (scaled-dot-product attention); an FP16 round-trip of our features leaves test accuracy unchanged
Extraction	batch size 1, no padding, last prompt token, block output	same; identical token ids for Qwen3-4B
Prompt assembly	module not released	P0 system message; documents and question in the user message; full documents, no truncation
Calibration	sigmoid in the paper; isotonic by default in the code	sigmoid; isotonic gives the same test accuracy (0.92)
Probe variants of the released code	–	uncalibrated, isotonic on validation, train+validation with isotonic or sigmoid (3-fold): test accuracy 0.92–0.93
Validation curve (**H**)	best 0.912 at layer 20	0.916 at layer 20; best 0.919 at layer 22; using layer 20 lowers test accuracy by about 0.003
Test accuracy	0.888 (shown as 0.89 or 0.88 in different tables); macro-F1 0.888; FAR 0.081	0.92 [0.91, 0.93]; macro-F1 0.92 [0.91, 0.94]; FAR 0.06 [0.04, 0.07]
Recall Answer / Refuse / Conflict	0.85 / 0.87 / 0.94	0.90 / 0.93 / 0.94 (no CI stored)
Data	2,391 questions, 7,173 instances, 70/10/20 split, 1,435 test	release at `2b338ce`: 2,396 questions, 7,188 instances, 76/4/20 split by question, 1,440 test; our test file has the same SHA-256 as the released test file

Table 3: Reproduction checks for the BF16 Qwen3-4B residual-stream probe against Huq et al. (2026) and their released code at commit `2b338ce`. Brackets: 95% bootstrap CIs over the 1,440 test items. The paper describes bootstrap CIs but prints none for these numbers. FAR: false-answer rate.

B Quantization Details

Table 4 summarizes the variants, and Table 5 lists their settings as recorded by each run; they are identical for the four models. All variants quantize the seven linear projections of every block (query, key, value and output projections of attention; gate, up and down projections of the MLP). That is 252 layers for both Qwen models, 224 for Llama-3.1-8B and

280 for Granite-4.1-8B. The language-model head, the embeddings and all normalization layers stay in BF16. For GPTQ calibration we sample 4,000 C4 documents once (seed 42, from the English training shard `c4-train.00000-of-01024`); each model tokenizes them in the same order, appends an end-of-sequence token after each document and takes the first $256 \times 2{,}048$ tokens as 256 consecutive sequences, so each model uses a prefix of the same document list. We never calibrate on the benchmark. We avoid LLM.int8() because it also quantizes activations (Dettmers et al., 2022), which are what we measure.

Variant	Method and settings	PPL ratio
W8	GPTQ 8-bit, g128, sym., act-order	1.00
W4	GPTQ 4-bit, g128, sym., act-order	1.03–1.06
W3	GPTQ 3-bit, g128, sym., act-order	1.18–1.30
NF4	bitsandbytes, block 64, no calib.	1.02–1.06
W2	GPTQ 2-bit (collapsed)	11.05–48.01

Table 4: Precision variants. Only weights are quantized; activations, KV cache and compute stay in BF16. g128: group size 128; sym.: symmetric; act-order: activation-order column ordering. PPL ratio: range over the four models of WikiText-2 perplexity relative to BF16. The collapsed W2 is never used as evidence.

Setting	GPTQ (W8, W4, W3)	NF4
Library	GPTQModel 7.6.0	bitsandbytes 0.50.2
Bits	8, 4, 3 (W2: 2)	4 (NF4 data type)
Group / block size	128	64
Symmetric	yes	–
Act-order (`desc_act`)	yes (`act_group_aware` off)	–
Damping	`damp_percent` 0.05 (library default), auto-increment 0.01	–
Double quantization	–	no
Calibration	C4, $256 \times 2{,}048$ tokens, seed 42	none (at load time)
Kernel / backend	`TorchLinear`: weights dequantized to BF16 before each matmul (cached during generation for the layers that fit in GPU memory); Triton dequantization off	`Linear4bit`, BF16 compute
Time per variant	750–1,384 s (peak 6.0–9.95 GB)	12–93 s

Table 5: Quantization settings, identical for all four models. Only weights are quantized; activations, KV cache and compute stay in BF16. W2 uses the GPTQ settings with 2 bits and has collapsed (Table 7).

We verify every GPTQ checkpoint after quantization (Table 6). First, we unpack each quantized layer to dense weights and count the distinct values per output channel within each quantization group (128 input channels); the count must not exceed 2^b for b bits. A group holds only 128 weights per output channel, so this bound can bind at W4, W3 and W2 but not at W8. Second, we run one P0 prompt with hooks on the first, middle and last block and on their query and down projections; every input and output must be BF16. Third, every quantized layer must use the `TorchLinear` backend; loading aborts otherwise. All 16 GPTQ checkpoints (four models; W8, W4, W3 and the collapsed W2) pass all three checks. During generation we cache the dequantized BF16 weights of as many layers as fit in GPU memory instead of dequantizing them at every call; on one checkpoint (Qwen3-4B at W8, 20 instances) the generations and forced-choice logits were identical with and without the cache. NF4 layers are `Linear4bit` with BF16 parameters and compute. Table 7 gives WikiText-2 perplexity (Merity et al., 2017) for all variants.

Model	Layers	Max. distinct values W8	W4	W3	W2	Hooked blocks
Qwen3-4B	252	113	16	8	4	0, 18, 35
Qwen3-8B	252	115	16	8	4	0, 18, 35
Llama-3.1-8B	224	114	16	8	4	0, 16, 31
Granite-4.1-8B	280	115	16	8	4	0, 20, 39
Limit 2^b		256	16	8	4	

Table 6: GPTQ verification. Layers: quantized linear layers checked. Max. distinct values: the largest number of distinct weight values in any (group, output channel) over all layers; the collapsed W2 reaches its limit of 4 on every model. Blocks are counted from 0, like every layer number in this paper (layer ℓ is the output of block ℓ). Activations at the hooked blocks and their query and down projections are BF16 on input and output, and every quantized layer uses `TorchLinear`, for all 16 GPTQ checkpoints.

Variant	Qwen3-4B PPL	Qwen3-4B ×BF16	Qwen3-8B PPL	Qwen3-8B ×BF16	Llama-3.1-8B PPL	Llama-3.1-8B ×BF16	Granite-4.1-8B PPL	Granite-4.1-8B ×BF16
BF16	10.06	1.00	9.73	1.00	7.22	1.00	8.56	1.00
W8	10.05	1.00	9.72	1.00	7.22	1.00	8.54	1.00
W4	10.49	1.04	10.01	1.03	7.62	1.06	8.87	1.04
NF4	10.37	1.03	9.98	1.03	7.62	1.06	8.74	1.02
W3	12.04	1.20	11.45	1.18	9.36	1.30	10.67	1.25
W2 (collapsed)	111.07	11.05	115.95	11.92	346.48	48.01	168.56	19.70

Table 7: WikiText-2 test perplexity (2,048-token windows) of every precision variant and its ratio to BF16. The GPTQ 2-bit model (W2) has collapsed on every model (11–48× BF16 perplexity) and is excluded from all main analyses; it appears only as a collapse reference.

C Prompts and Generation Settings

We take the P0 prompt and the KRE and OPIN_INSTR instructions verbatim from the released files of Huq et al. (2026). The module that assembles their P0 and instruction-baseline prompts is not in the release, so the assembly below is ours. Prompts are identical across precision variants by construction: every variant uses the base model's tokenizer and chat template and the same assembly code.

P0 (feature extraction). System message:

```
You are a helpful assistant. Use only the provided documents to answer.
```

User message (five documents, each followed by a blank line; documents are not truncated):

```
[Document 1]:
{document 1}

. . .

[Document 5]:
{document 5}

Question: {question}
Answer:
```

KRE and OPIN_INSTR (behaviour). The user message is the same as for P0; the baseline instruction replaces the P0 system message. KRE is adapted from the hint-style instructions that Ying et al. (2024) use with their Knowledge Robustness Evaluation data. The

OPIN_INSTR instruction is adapted from the instructed prompt of Zhou et al. (2023); the release adds "based only on that information" and contains no opinion-based framing. KRE:

```
You are given documents and a question. The documents may be incomplete,
misleading, or may conflict with one another, so judge carefully.
```

OPIN_INSTR:

```
Instruction: Read the given information and answer the corresponding
question based only on that information.
```

Forced choice (descriptive only). The system message is the P0 system message. The user message has the same documents, then `Question: {question}` on its own line, then:

```
Is the evidence sufficient (A), insufficient (B), or conflicting (C)?
Answer with one letter.
```

The prediction is the letter with the highest next-token logit among A, B and C. This readout is not used as evidence (Appendix H).

Chat templates. Every prompt goes through the model's own chat template with a generation prompt appended. We tokenize the rendered text without adding special tokens, because the template already contains them. Qwen3-8B is a hybrid thinking model; we pass `enable_thinking=False`, so its template ends the prompt with an empty `<think>\n\n</think>\n\n` block, and the last prompt token we probe comes after that block. The Llama-3.1 template always adds a fixed knowledge-cutoff and date line ("Cutting Knowledge Date: December 2023", "Today Date: 26 Jul 2024") to the system message; it is the same for every variant. Qwen3-4B (the Instruct-2507 release) is a non-thinking model and needs no extra argument.

Generation. We decode greedily with at most 512 new tokens on the test split, one instance per call, without batching or padding. An output *hits the cap* when no end-of-sequence token appears within the cap. We first used 64 tokens, the maximum stated in the paper of Huq et al. (2026), but about half of the Qwen3-4B outputs stopped before the verdict; the released scripts that set a limit themselves use 256 or 300 tokens, and the main baseline script reads its limit from a module that is not released. We then ran all test generations with a 128-token cap. A third of them hit it (9,568 of the 28,800 outputs in the 20 runs of Table 9), so we generated exactly those again with a 512-token cap; 461 (1.6%) still hit it. Greedy decoding is deterministic, so an output that ended within 128 tokens is the same under either cap, and 9,558 of the regenerated outputs start with their 128-token version. The validation generations that select the prompt (Appendix D) and the W2 test generations keep the 128-token cap. Appendix G compares the RQ3 results under the two caps. A few prompts exceed 8,192 tokens (four in the test split, each over 20,000 tokens); for these, generation keeps the KV cache in CPU memory, which should change memory use but not the computed tokens. The 10 regenerated outputs that do not start with their 128-token version are all on these four prompts, in three GPTQ runs (Granite-4.1-8B at W8 and W4, Llama-3.1-8B at W8) where most of these 128-token outputs were already degenerate (repetition or garbled text), so on such prompts the output of some GPTQ variants is not stable from run to run. All models, including the GPTQ variants, run in BF16. We request PyTorch's scaled-dot-product attention for the BF16 and NF4 models; the GPTQ models use the default of the GPTQModel loader, which is the same scaled-dot-product attention: every W8, W4 and W3 checkpoint reports it when reloaded (checked after the runs).

D Behaviour Labelling

LLM parser (primary). A judge model, Gemini 3.1 Flash-Lite (Google DeepMind, 2026) accessed through OpenRouter, labels every generation (temperature 0, seed 0, at most 100 output tokens, structured JSON output with three Boolean fields). The judge sees only the generated text, never the question, the documents or the gold label. Its system prompt is:

```
You read the output of a language model that was given some documents and
a question. You do NOT see the documents or the question. Judge ONLY
what the output text says, and answer three yes/no questions about it.

notices_conflict: Does the output say that the provided documents
conflict or contradict each other, give inconsistent or different
versions, or that a document is wrong, incorrect or contains an error?
Answer yes even if the output then picks one version and answers.

notices_insufficient: Does the output say that the provided documents do
not contain, mention or support the information needed to answer the
question as a whole (for example "the documents do not mention X", "there
is not enough information", "this cannot be determined from the
documents"), or that it answers from general or outside knowledge
because the documents do not give the answer? Answer yes even if the
output then answers. Answer NO if it only says that one document lacks
something while another document gives the answer, or only remarks that
a side detail is missing while the documents still answer the question.

commits_to_answer: Does the output give a specific answer to the
question, even with caveats? Answer no if it only says it cannot answer
or only says the documents conflict.
```

The user message is `Model output:` followed by the generation. The label follows a fixed precedence: Conflict if `notices_conflict`, else Refuse if `notices_insufficient`, else Answer. This is a detection reading: a generation that notices the problem but still answers counts as Conflict or Refuse. We record such cases as *noticed but answered* (NBA in Table 9) without adding a fourth class; Appendix G splits the Refuse and Conflict recall by the third field (Table 19). An output with no flag and no answer (e.g., cut off) and an empty output both map to Answer; no output was empty in the 20 runs of Table 9. FAR therefore means "gave no sign of noticing the problem".

Check of the judge labels. We checked the judge on 300 test outputs: from each of 10 runs (every model at BF16 and W3, and Llama-3.1-8B at W4 and NF4, which include every run with a claim or partial case) we drew 10 instances at random from each of the three classes. A second LLM labelled them blind: it saw only the output text and answered the judge's three questions with the same definitions. Its label agreed with the judge's on 286 of the 300 outputs (95%); on the two noticing fields, every disagreement was a yes from the judge and a no from the second LLM (8 and 6 outputs). We then labelled 24 of the 300 outputs by hand in the same blind way, in random order: all 14 on which the two labels differed, and 10 drawn at random from the 275 on which all three fields agreed. On the 14, the human label matched the judge's 6 times, the second LLM's 7 times and neither once; on the 10, it matched both 9 times. So where the two LLMs disagree, on about 5% of the sample, the human label sides with the judge about half the time. This is a small check by a single annotator (the author), not a validation of the judge.

Regex parser (comparison). The regex parser starts from the rules of the released judging script of Huq et al. (2026) (first non-empty line, canonical strings, then regular-expression cues; anything else is Answer). On Qwen3-4B validation outputs these rules missed most refusals and every stated conflict, so we extended them: markdown normalization, skipping lines that only restate the question, more refusal phrasings, conflict cues that name a document (e.g., "Document 4 incorrectly . . . "), and a scan of the whole output for verdicts about the documents as a whole. At BF16, regex-parser test accuracy is 0.68 [0.65, 0.70] (Qwen3-4B), 0.56 [0.54, 0.59] (Qwen3-8B), 0.57 [0.54, 0.59] (Llama-3.1-8B) and 0.63 [0.61, 0.66] (Granite-4.1-8B), against 0.76 [0.73, 0.78], 0.61 [0.59, 0.64], 0.74 [0.72, 0.76] and 0.69 [0.67, 0.72] with the LLM parser.

Baseline selection. For each model we run KRE and OPIN_INSTR on the 285 BF16 validation items (128-token cap) and keep the baseline with the higher LLM-parser accuracy (Table 8). KRE wins on all four models, and all pass the pre-set gate of 0.45 validation accuracy. The unpaired CIs overlap on all four models; the choice is closest on Qwen3-8B

(0.60 vs. 0.59), the only model on which the regex parser would have chosen OPIN_INSTR. With the regex parser, KRE also wins on Qwen3-4B, Llama-3.1-8B and Granite-4.1-8B (Table 8). Table 9 reports the selected baseline on the test split for every variant.

	LLM parser		Regex parser	
Model	KRE	OPIN_INSTR	KRE	OPIN_INSTR
Qwen3-4B	**0.80 [0.75, 0.84]**	0.74 [0.68, 0.79]	0.72 [0.67, 0.78]	0.69 [0.64, 0.74]
Qwen3-8B	**0.60 [0.54, 0.66]**	0.59 [0.53, 0.65]	0.54 [0.48, 0.60]	0.58 [0.52, 0.63]
Llama-3.1-8B	**0.76 [0.71, 0.81]**	0.68 [0.62, 0.73]	0.56 [0.51, 0.62]	0.54 [0.48, 0.60]
Granite-4.1-8B	**0.71 [0.65, 0.76]**	0.65 [0.59, 0.71]	0.62 [0.56, 0.68]	0.59 [0.54, 0.65]

Table 8: Behaviour-baseline selection: BF16 validation accuracy (285 items) with 95% bootstrap CIs. Bold: selected baseline (LLM parser, which decides). Regex-parser scores are shown for comparison only.

			Recall						
Model	Var.	Accuracy	Answer	Refuse	Conflict	FAR	Over-abst.	NBA	Hit max
Qwen3-4B	BF16	0.76 [0.73, 0.78]	0.81	0.91	0.55	0.24 [0.21, 0.27]	0.19	0.29	0.02
	W8	0.75 [0.73, 0.78]	0.80	0.90	0.56	0.24 [0.22, 0.27]	0.20	0.29	0.01
	W4	0.77 [0.75, 0.79]	0.78	0.92	0.61	0.20 [0.18, 0.23]	0.22	0.30	0.02
	NF4	0.74 [0.72, 0.76]	0.77	0.89	0.56	0.24 [0.22, 0.27]	0.23	0.30	0.04
	W3	0.69 [0.66, 0.71]	0.62	0.95	0.49	0.21 [0.18, 0.24]	0.38	0.32	0.04
Qwen3-8B	BF16	0.61 [0.59, 0.64]	0.94	0.71	0.18	0.53 [0.50, 0.56]	0.06	0.13	0.00
	W8	0.61 [0.59, 0.63]	0.94	0.71	0.18	0.53 [0.50, 0.56]	0.06	0.13	0.00
	W4	0.57 [0.55, 0.60]	0.93	0.64	0.15	0.59 [0.55, 0.62]	0.07	0.11	0.00
	NF4	0.60 [0.57, 0.63]	0.95	0.68	0.18	0.56 [0.53, 0.59]	0.05	0.10	0.00
	W3	0.69 [0.67, 0.71]	0.88	0.75	0.44	0.38 [0.35, 0.41]	0.12	0.30	0.05
Llama-3.1-8B	BF16	0.74 [0.72, 0.76]	0.85	0.87	0.49	0.30 [0.27, 0.33]	0.15	0.34	0.02
	W8	0.74 [0.72, 0.76]	0.86	0.87	0.49	0.30 [0.27, 0.33]	0.14	0.34	0.02
	W4	0.70 [0.68, 0.73]	0.89	0.81	0.42	0.36 [0.33, 0.40]	0.11	0.27	0.04
	NF4	0.68 [0.66, 0.71]	0.81	0.84	0.39	0.34 [0.31, 0.37]	0.19	0.34	0.02
	W3	0.62 [0.60, 0.65]	0.59	0.90	0.38	0.25 [0.22, 0.28]	0.41	0.24	0.03
Granite-4.1-8B	BF16	0.69 [0.67, 0.72]	0.92	0.83	0.33	0.41 [0.38, 0.44]	0.08	0.26	0.01
	W8	0.70 [0.67, 0.72]	0.92	0.83	0.34	0.41 [0.38, 0.44]	0.08	0.25	0.00
	W4	0.70 [0.67, 0.72]	0.91	0.85	0.33	0.39 [0.36, 0.42]	0.09	0.24	0.00
	NF4	0.67 [0.65, 0.69]	0.89	0.83	0.30	0.42 [0.38, 0.45]	0.11	0.23	0.01
	W3	0.57 [0.55, 0.60]	0.90	0.70	0.13	0.56 [0.53, 0.59]	0.10	0.13	0.00

Table 9: Generation readout (prompt baseline selected on BF16 validation, greedy decoding, 512 new tokens, labels from the LLM parser) on the 1,440 test items. FAR = false-answer rate (share of Refuse and Conflict items labelled Answer); Over-abst. = share of Answer items labelled Refuse or Conflict; NBA = noticed-but-answered rate (share of all outputs that flag missing or conflicting evidence but still give an answer); Hit max = share of outputs that reach the 512-token cap. No output was empty. Brackets: 95% bootstrap CIs (only accuracy and FAR have stored CIs).

E Additional RQ1 Results

All non-collapsed variants share a mid-to-late plateau of validation accuracy (Figure 4). W3 sits slightly lower, most visibly on Llama-3.1-8B and for **M**. The collapsed W2 model peaks early (validation accuracy about 0.55–0.70) and decays towards chance in later layers; we show it only for completeness.

Accuracy changes. Table 10 extends Table 1 with the selected layer, macro-F1, FAR, each accuracy change with its CI, the random-label control and WikiText-2 perplexity. With each variant read at its own validation-selected layer, the W3 CIs exclude 0 except on Granite-4.1-8B, where the CI touches 0; at 4 bits, only Qwen3-8B at W4 is significantly lower (-0.03 [-0.04, -0.01]).

Selected layers. The validation-selected layer of a quantized model is within three layers of the BF16 layer on every model (Table 10). RQ1 and RQ3 read each variant at its own best layer, whereas the main RQ2 numbers average over the BF16 plateau layers and the RQ2 diagnostics use the BF16-selected layer, so native accuracies differ slightly between Tables 1 and 14. At W3, native probes lose 0.02–0.07 relative to BF16 at their own best layer, 0.03–0.04 averaged over the plateau and 0.02–0.03 at the BF16-selected layer. The largest difference is on Qwen3-8B at W3: 0.85 at its own layer 20 but 0.90 at the BF16 layer 23, which suggests that layer selection on the 285 validation items is noisy for this variant. Layer 20 leads layer 23 by one validation item (255 against 254 correct), and at layer 23 the Answer drop of Section 4.1 is -0.03 $[-0.06, 0.00]$ in place of -0.12 $[-0.16, -0.08]$.

Model	Var.	PPL	Layer	Accuracy	Macro-F1	FAR	ΔAcc. vs BF16	Rand.
Qwen3-4B	BF16	10.06	22	0.92 [0.91, 0.93]	0.92 [0.91, 0.94]	0.06 [0.04, 0.07]	–	0.33
	W8	10.05	22	0.92 [0.90, 0.93]	0.92 [0.90, 0.93]	0.06 [0.04, 0.07]	0.00 [−0.01, 0.00]	0.33
	W4	10.49	19	0.91 [0.90, 0.93]	0.91 [0.90, 0.93]	0.05 [0.04, 0.07]	−0.01 [−0.03, 0.01]	0.33
	NF4	10.37	23	0.91 [0.90, 0.92]	0.91 [0.89, 0.92]	0.06 [0.05, 0.08]	−0.01 [−0.03, 0.00]	0.33
	W3	12.04	20	0.89 [0.87, 0.91]	0.89 [0.87, 0.91]	0.07 [0.05, 0.08]	−0.03 [−0.05, −0.01]	0.34
Qwen3-8B	BF16	9.73	23	0.91 [0.90, 0.93]	0.91 [0.90, 0.93]	0.06 [0.05, 0.08]	–	0.34
	W8	9.72	22	0.92 [0.91, 0.94]	0.92 [0.91, 0.94]	0.06 [0.04, 0.07]	0.01 [0.00, 0.02]	0.33
	W4	10.01	20	0.89 [0.87, 0.90]	0.89 [0.87, 0.90]	0.07 [0.05, 0.09]	−0.03 [−0.04, −0.01]	0.34
	NF4	9.98	20	0.90 [0.89, 0.92]	0.90 [0.89, 0.92]	0.08 [0.06, 0.10]	−0.01 [−0.03, 0.00]	0.33
	W3	11.45	20	0.85 [0.83, 0.87]	0.85 [0.83, 0.87]	0.09 [0.08, 0.11]	−0.07 [−0.08, −0.05]	0.33
Llama-3.1-8B	BF16	7.22	16	0.92 [0.91, 0.94]	0.92 [0.91, 0.94]	0.06 [0.04, 0.07]	–	0.33
	W8	7.22	16	0.92 [0.91, 0.93]	0.92 [0.91, 0.93]	0.06 [0.05, 0.08]	0.00 [−0.01, 0.00]	0.33
	W4	7.62	17	0.91 [0.90, 0.93]	0.91 [0.90, 0.93]	0.05 [0.04, 0.07]	−0.01 [−0.02, 0.00]	0.33
	NF4	7.62	16	0.91 [0.90, 0.93]	0.91 [0.90, 0.93]	0.07 [0.05, 0.09]	−0.01 [−0.02, 0.00]	0.33
	W3	9.36	18	0.89 [0.87, 0.91]	0.89 [0.87, 0.91]	0.08 [0.07, 0.10]	−0.03 [−0.05, −0.02]	0.34
Granite-4.1-8B	BF16	8.56	27	0.91 [0.90, 0.93]	0.91 [0.90, 0.93]	0.05 [0.03, 0.06]	–	0.33
	W8	8.54	28	0.92 [0.91, 0.94]	0.92 [0.91, 0.94]	0.05 [0.03, 0.06]	0.01 [0.00, 0.02]	0.32
	W4	8.87	27	0.92 [0.91, 0.93]	0.92 [0.91, 0.93]	0.05 [0.04, 0.07]	0.01 [0.00, 0.02]	0.33
	NF4	8.74	27	0.91 [0.89, 0.92]	0.91 [0.89, 0.92]	0.06 [0.04, 0.07]	−0.01 [−0.02, 0.00]	0.33
	W3	10.67	29	0.90 [0.88, 0.91]	0.90 [0.88, 0.91]	0.07 [0.05, 0.08]	−0.02 [−0.03, 0.00]	0.34

Table 10: Full RQ1 table: native residual-stream (**H**) probes of each variant, read at each variant's own validation-selected layer, with WikiText-2 perplexity (PPL) and macro-F1. Layers are counted from 0 (layer ℓ is the output of block ℓ). FAR = false-answer rate (share of Refuse and Conflict items predicted Answer); ΔAcc. = accuracy change vs. BF16 (paired). Brackets: 95% bootstrap CIs (10,000 resamples, 1,440 test items). Rand. = mean test accuracy of five uncalibrated random-label control probes (chance ≈ 0.33).

Per-class changes of native probes. Figure 5 shows every per-class recall change of native residual-stream probes; all lie between -0.12 and $+0.05$, and the class that slips first differs between models. Five changes are significant after Holm correction. Answer drops on Qwen3-4B at W3 (-0.06 $[-0.10, -0.03]$, Holm $p = 0.023$) and on Qwen3-8B at W4 (-0.06 $[-0.10, -0.03]$, $p = 0.024$) and W3 (-0.12 $[-0.16, -0.08]$, $p < 0.001$). Conflict drops on Granite-4.1-8B at W3 (-0.04 $[-0.06, -0.01]$, $p = 0.033$) and rises on Qwen3-4B at W4 (0.05 $[0.03, 0.07]$, $p = 0.003$). Llama-3.1-8B has no significant change; its largest drop is Answer at W3 (-0.04 $[-0.08, -0.01]$, Holm $p = 0.26$).

For **M**, only two per-class changes are significant after Holm correction (Figure 6), both on Qwen3-4B at W3: Answer -0.06 $[-0.10, -0.03]$ (Holm $p = 0.027$) and Refuse -0.05 $[-0.08, -0.02]$ (Holm $p = 0.043$). As for **H**, native probes still decode every class.

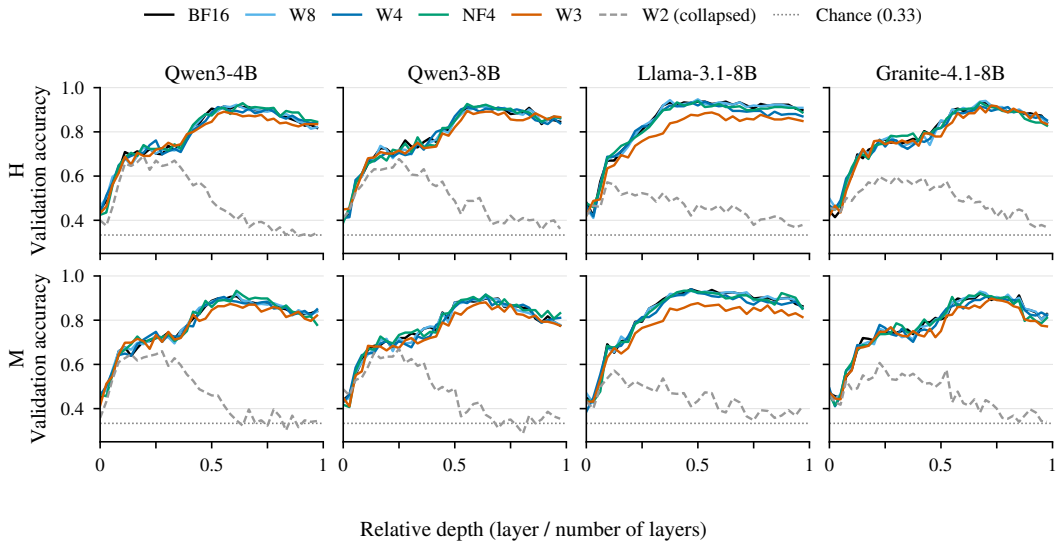

Figure 4: Validation accuracy of native probes at every layer (relative depth) for each precision variant, residual stream **H** (top) and MLP output **M** (bottom). W2 (GPTQ 2-bit, dashed grey) is shown for completeness only: the model has collapsed (WikiText-2 perplexity 11–48× BF16). Dotted line: chance (0.33). No CIs (validation split, 285 items).

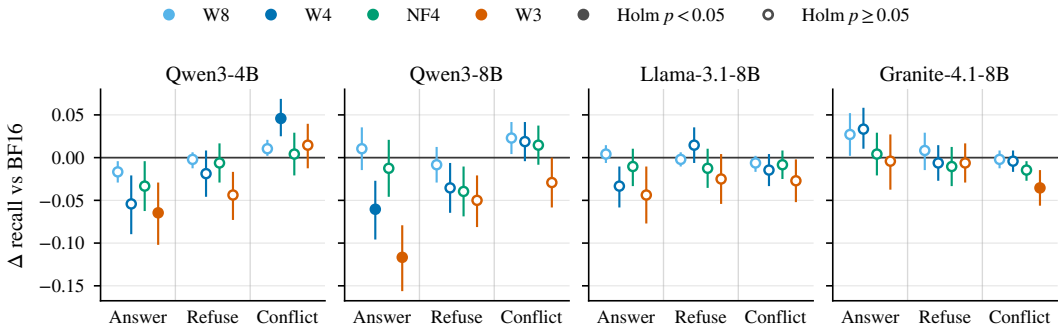

Figure 5: Per-class change in recall of native (retrained) residual-stream (**H**) probes relative to BF16, each read at its own validation-selected layer, with 95% paired-bootstrap CIs. Filled markers: significant after Holm correction over the model's 30 RQ1 McNemar tests; open: not significant. Most changes are small, and the class that slips first differs between models.

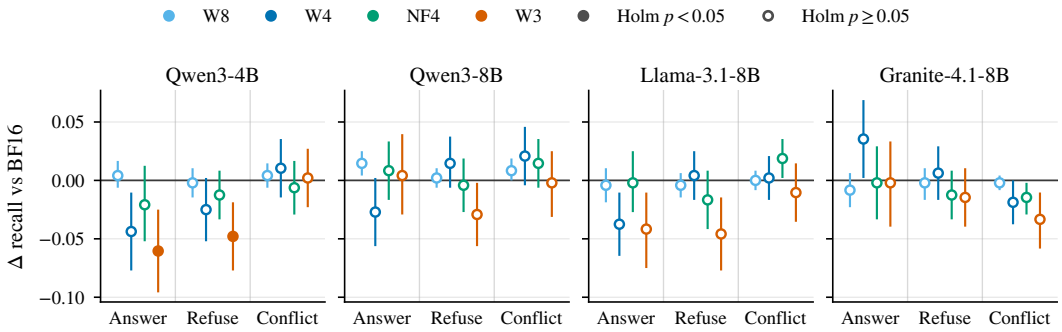

Figure 6: Per-class change in recall of native (retrained) MLP-output (**M**) probes relative to BF16, each read at its own validation-selected layer. Error bars are 95% paired-bootstrap CIs; filled markers are significant after Holm correction over the model's 30 RQ1 McNemar tests (5 quantized variants × 3 classes × 2 features), open markers are not. Only Qwen3-4B W3 Answer and Refuse are significant.

F Additional RQ2 Results

Plateau gaps. Table 11 gives the plateau-averaged gaps behind Figure 1 (**H**) and Figure 7 (**M**). For **H**, every W4, NF4 and W3 gap has a CI above 0, before and after re-scaling. On Granite-4.1-8B the W4 and NF4 CIs overlap. Re-scaling removes 53–60% of the W4 gap, 36–76% of the NF4 gap and 35–49% of the W3 gap. WikiText-2 perplexity is 1.00× BF16 at W8, 1.02–1.06× at 4 bits and 1.18–1.30× at W3 (Table 7).

		H			**M**		
Model	Var.	Drop	Frozen gap	Re-scaled gap	Drop	Frozen gap	Re-scaled gap
Qwen3-4B	W8	0.00	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
	W4	0.01	0.12 [0.11, 0.13]	0.05 [0.04, 0.06]	0.01	0.16 [0.15, 0.17]	0.06 [0.06, 0.07]
	NF4	0.01	0.09 [0.07, 0.10]	0.04 [0.03, 0.05]	0.01	0.14 [0.13, 0.15]	0.05 [0.04, 0.06]
	W3	0.03	0.31 [0.30, 0.33]	0.17 [0.16, 0.18]	0.03	0.30 [0.29, 0.31]	0.19 [0.18, 0.20]
Qwen3-8B	W8	0.00	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00	0.00 [0.00, 0.01]	0.00 [0.00, 0.00]
	W4	0.01	0.21 [0.19, 0.22]	0.10 [0.09, 0.11]	0.01	0.26 [0.25, 0.27]	0.16 [0.15, 0.17]
	NF4	0.01	0.13 [0.12, 0.14]	0.03 [0.02, 0.04]	0.00	0.16 [0.15, 0.17]	0.07 [0.06, 0.08]
	W3	0.04	0.38 [0.36, 0.40]	0.20 [0.19, 0.22]	0.04	0.39 [0.37, 0.41]	0.28 [0.27, 0.30]
Llama-3.1-8B	W8	0.00	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
	W4	0.01	0.05 [0.04, 0.06]	0.02 [0.01, 0.03]	0.01	0.05 [0.04, 0.06]	0.03 [0.03, 0.04]
	NF4	0.00	0.02 [0.02, 0.03]	0.01 [0.01, 0.02]	0.01	0.03 [0.02, 0.03]	0.02 [0.01, 0.02]
	W3	0.04	0.34 [0.32, 0.36]	0.22 [0.21, 0.24]	0.04	0.35 [0.34, 0.37]	0.28 [0.27, 0.29]
Granite-4.1-8B	W8	0.00	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
	W4	0.00	0.05 [0.04, 0.06]	0.02 [0.01, 0.03]	0.01	0.08 [0.08, 0.09]	0.03 [0.03, 0.04]
	NF4	0.01	0.04 [0.03, 0.05]	0.02 [0.01, 0.02]	0.01	0.06 [0.06, 0.07]	0.03 [0.02, 0.03]
	W3	0.03	0.19 [0.17, 0.20]	0.10 [0.08, 0.11]	0.04	0.24 [0.23, 0.25]	0.14 [0.13, 0.15]

Table 11: Plateau-averaged transfer gaps, the numbers behind Figures 1 and 7, for the residual stream **H** and the MLP output **M**. Drop = BF16 native accuracy minus the variant's native accuracy (no CI); frozen gap = native minus frozen BF16 probe accuracy; re-scaled gap = the same after refitting only the scaler. All values are per-item averages over the BF16 plateau layers of the feature. Brackets: 95% paired-bootstrap CIs.

Threshold or ranking. Table 12 reads the frozen probe in two further ways (**H**, plateau averages). Without the sigmoid calibrator (argmax of the regression's class scores), the gaps change by at most 0.012, so the calibration step is not the cause. The macro one-vs-rest AUC of the regression scores needs no threshold and does not change when a class's scores are shifted or scaled. At 4 bits the frozen probe's AUC is 0.92–0.97, 0.01–0.05 below the native probe's, while its accuracy gap is 0.02–0.21; at W3 it is 0.81–0.91, 0.06–0.15 below (each CI above 0). At 4 bits the frozen probe thus still ranks the items almost as well, which suggests that part of its accuracy gap is a shifted decision threshold, while at W3 the ranking itself degrades.

Per-class transfer. Table 13 gives the W3 recalls of Figure 2, at the BF16-selected layer. On Qwen3-4B the frozen probe has accuracy 0.55 [0.53, 0.58], with Answer recall 0.01 and Refuse recall 0.98. Refuse recall falls to 0.02 on Qwen3-8B, where Conflict also falls to 0.38; Conflict falls to 0.31 on Llama-3.1-8B and Refuse to 0.46 on Granite-4.1-8B. The frozen probe assigns 55–84% of all test items to one class (Refuse on Qwen3-4B, Answer on the other three models); after re-scaling, each class receives 26–38%. After re-scaling, recall still trails the native probe by 0.07–0.25 in 11 of the 12 model–class pairs (each CI above 0); the exception is Refuse on Granite-4.1-8B. On Qwen3-8B at W4, the frozen probe's Conflict recall is 0.99, against 0.31 for Answer and 0.40 for Refuse. At NF4, Answer has the largest gap on all four models, e.g., 0.20 [0.16, 0.24] on Qwen3-4B and 0.30 [0.26, 0.35] on Qwen3-8B. At 4 bits, every re-scaled per-class gap is at most 0.06, except on Qwen3-8B at W4 (0.09–0.13, each CI above 0).

Diagnosis maps. Table 14 extends Table 2 (residual stream **H**, BF16-selected layer; Section 4.2.2) with W8, native accuracy, the CIs of the frozen and Procrustes gaps and the

		Frozen gap		AUC		
Model	Var.	Calibrated	Uncalibrated	Native	Frozen	Native − frozen
Qwen3-4B	W8	0.00	0.00	0.97	0.97	0.00 [0.00, 0.00]
	W4	0.12	0.12	0.97	0.94	0.03 [0.03, 0.03]
	NF4	0.09	0.07	0.97	0.94	0.02 [0.02, 0.02]
	W3	0.31	0.31	0.96	0.85	0.10 [0.10, 0.11]
Qwen3-8B	W8	0.00	0.00	0.97	0.97	0.00 [0.00, 0.00]
	W4	0.21	0.21	0.97	0.92	0.05 [0.04, 0.05]
	NF4	0.13	0.12	0.97	0.95	0.02 [0.01, 0.02]
	W3	0.38	0.38	0.96	0.83	0.13 [0.12, 0.14]
Llama-3.1-8B	W8	0.00	0.00	0.97	0.97	0.00 [0.00, 0.00]
	W4	0.05	0.05	0.97	0.96	0.01 [0.01, 0.01]
	NF4	0.02	0.03	0.97	0.97	0.01 [0.00, 0.01]
	W3	0.34	0.34	0.95	0.81	0.15 [0.14, 0.16]
Granite-4.1-8B	W8	0.00	0.00	0.98	0.98	0.00 [0.00, 0.00]
	W4	0.05	0.05	0.97	0.96	0.01 [0.01, 0.01]
	NF4	0.04	0.04	0.97	0.96	0.01 [0.01, 0.01]
	W3	0.19	0.18	0.96	0.91	0.06 [0.05, 0.06]

Table 12: The frozen BF16 probe read without its calibrator and without a threshold (residual stream **H**, averages over the BF16 plateau layers). Frozen gap = native minus frozen-probe accuracy, with the sigmoid calibrator (as in Table 11) and without it (argmax of the regression scores, for both probes). AUC = macro one-vs-rest area under the ROC curve of the regression scores, which no shift or positive scaling of a class's scores changes; the BF16 probes reach 0.97–0.98 on BF16 activations. Brackets: 95% paired-bootstrap CIs.

			Frozen BF16 probe		Re-scaled BF16 probe	
Model	Class	Native	Recall	Gap to native	Recall	Gap to native
Qwen3-4B	Answer	0.85	0.01	0.84 [0.81, 0.87]	0.67	0.18 [0.14, 0.23]
	Refuse	0.90	0.98	−0.08 [−0.11, −0.05]	0.68	0.23 [0.18, 0.27]
	Conflict	0.94	0.67	0.27 [0.23, 0.32]	0.83	0.12 [0.09, 0.15]
Qwen3-8B	Answer	0.86	0.96	−0.10 [−0.14, −0.07]	0.69	0.17 [0.13, 0.22]
	Refuse	0.87	0.02	0.85 [0.82, 0.88]	0.62	0.25 [0.21, 0.29]
	Conflict	0.95	0.38	0.57 [0.53, 0.62]	0.76	0.19 [0.15, 0.23]
Llama-3.1-8B	Answer	0.85	0.86	−0.01 [−0.05, 0.03]	0.65	0.21 [0.16, 0.25]
	Refuse	0.89	0.62	0.26 [0.22, 0.31]	0.65	0.24 [0.19, 0.28]
	Conflict	0.93	0.31	0.62 [0.57, 0.66]	0.74	0.19 [0.15, 0.23]
Granite-4.1-8B	Answer	0.82	0.91	−0.09 [−0.12, −0.06]	0.69	0.14 [0.09, 0.18]
	Refuse	0.89	0.46	0.43 [0.39, 0.47]	0.89	0.01 [−0.02, 0.03]
	Conflict	0.93	0.75	0.19 [0.15, 0.23]	0.86	0.07 [0.04, 0.10]

Table 13: Per-class recall at W3, at the BF16-selected layer (residual stream **H**): the native probe, the frozen BF16 probe and the re-scaled BF16 probe, each with its gap to the native probe (native recall minus its own). These are the W3 values of Figure 2. Brackets: 95% paired-bootstrap CIs.

mean activation cosine. We also compare the per-class directions of the BF16 and native quantized probes in raw activation space by cosine. The direction of class c is w_c/σ, taken element-wise: the regression weights on the standardized activations divided by the scaler's standard deviations, so this cosine also reflects changes in the per-dimension scale. The norm ratio (mean norm of the quantized **H** test activations divided by that of BF16; point estimates without CIs) is 1.00 at W8, 0.94–1.00 at 4 bits and 0.83–0.95 at W3 across the four models. The probe-direction null refits the BF16 probe on 50 bootstrap resamples of its training set. Procrustes is fitted on activations standardized per dimension, each variant with its own training mean and standard deviation, so without the rotation it is exactly the re-scaled probe. All maps are fitted on paired unlabelled training items; the ridge penalty is chosen by reconstruction error on unlabelled validation activations. Procrustes closes more of the gap than re-scaling alone by 0.11 [0.09, 0.14] (Qwen3-4B), 0.12 [0.09, 0.15] (Qwen3-8B), 0.08 [0.05, 0.11] (Llama-3.1-8B) and 0.04 [0.02, 0.06] (Granite-4.1-8B) at W3, and by 0.03 [0.01, 0.04] (Qwen3-4B) and 0.08 [0.06, 0.10] (Qwen3-8B) at W4 (last column of Table 2). At W3 it leaves 0.04–0.13 of a gap of 0.18–0.44 at that layer, and every W3 gap after the ridge map has a CI covering 0. In the other six 4-bit cases Procrustes gives no significant gain: the CI of the difference includes 0 in each. On Llama-3.1-8B and Granite-4.1-8B at 4 bits the gaps at that layer are at most 0.04. The mean activation cosine does not order the W3 gaps across models: Llama-3.1-8B has the lowest cosine (0.68) but a smaller gap than Qwen3-8B (cosine 0.80).

			Transfer gap after map (native − mapped)					
Model	Var.	Native	Frozen	Re-scaled	Affine	Procrustes	Ridge	Act. cos.
Qwen3-4B	W8	0.92	−0.01 [−0.01, 0.00]	0.00°	0.00°	0.00 [−0.01, 0.00]	0.00°	1.00
	W4	0.91	0.15 [0.13, 0.17]	0.04	0.03	0.01 [0.00, 0.02]	−0.02	0.96
	NF4	0.92	0.06 [0.04, 0.07]	0.02	0.02	0.02 [0.01, 0.03]	−0.01°	0.96
	W3	0.90	0.35 [0.32, 0.37]	0.18	0.19	0.07 [0.05, 0.08]	−0.01°	0.88
Qwen3-8B	W8	0.91	−0.01 [−0.01, 0.00]	0.00°	0.00°	0.00 [−0.01, 0.00]	−0.01	1.00
	W4	0.90	0.33 [0.31, 0.36]	0.11	0.10	0.03 [0.01, 0.04]	−0.01°	0.91
	NF4	0.90	0.15 [0.13, 0.17]	0.03	0.02	0.01 [0.00, 0.03]	−0.02	0.94
	W3	0.90	0.44 [0.41, 0.47]	0.20	0.21	0.08 [0.06, 0.10]	0.00°	0.80
Llama-3.1-8B	W8	0.92	0.00 [−0.01, 0.00]	0.00°	0.00°	0.00 [−0.01, 0.00]	−0.01°	1.00
	W4	0.91	0.03 [0.01, 0.04]	0.02	0.01°	0.02 [0.01, 0.04]	−0.01°	0.95
	NF4	0.91	0.00 [−0.01, 0.02]	0.00°	0.00°	0.00 [−0.01, 0.02]	−0.01°	0.96
	W3	0.89	0.29 [0.26, 0.32]	0.21	0.20	0.13 [0.11, 0.15]	0.00°	0.68
Granite-4.1-8B	W8	0.92	0.00 [0.00, 0.01]	0.00°	0.00°	0.00 [0.00, 0.01]	0.01°	1.00
	W4	0.92	0.04 [0.02, 0.05]	0.02	0.02	0.02 [0.01, 0.03]	0.00°	0.99
	NF4	0.91	0.03 [0.01, 0.04]	0.02	0.02	0.01 [0.00, 0.02]	−0.01°	0.99
	W3	0.88	0.18 [0.15, 0.20]	0.07	0.08	0.04 [0.02, 0.05]	−0.01°	0.95

Table 14: RQ2 diagnosis at the BF16-selected layer (**H** feature). Native = accuracy of the probe trained on the quantized model at that layer. Each gap is native accuracy minus the accuracy of the frozen BF16 probe after a label-free map from quantized to BF16 activations, fit on unlabelled paired training activations: none (frozen), scaler refit (re-scaled), per-dimension affine, re-scaling followed by an orthogonal rotation (Procrustes), and a full ridge map (upper bound). Brackets: 95% paired-bootstrap CIs, shown for the frozen probe and Procrustes; the CIs of the other maps are in the released analysis outputs, and ° marks a gap whose CI includes 0 (decided on the unrounded CI bounds, here and in Table 2). Act. cos. = mean per-item cosine between quantized and BF16 test activations.

MLP output. The MLP-output feature (**M**) shows the same pattern as the residual stream (Figure 7, Tables 11 and 15). W8 transfers without loss: every plateau gap lies within 0.003 of zero and its CI covers 0. At W3 the frozen plateau gap is 0.24–0.39, and re-scaling leaves more of it than for **H** (0.14–0.28 against 0.10–0.22). At 4 bits, the **M** gap is larger than the **H** gap on three models and similar on Llama-3.1-8B.

Transfer by layer. At 4 bits, frozen transfer can change sharply between neighbouring layers; at W3, no layer recovers it (Figure 8). On Qwen3-4B at W4 the gap spans 0.00–0.37 across the plateau layers, whereas at W3 it is at least 0.10 on every plateau layer of every

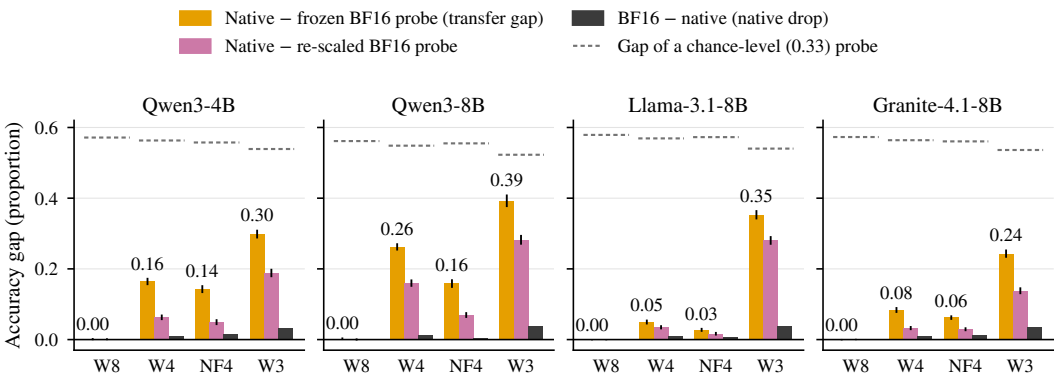

Figure 7: Frozen cross-precision transfer averaged over the BF16 validation plateau, for the MLP output (**M**). For each quantized variant we show the transfer gap (native probe accuracy minus frozen BF16 probe accuracy; orange), the gap left after label-free re-scaling of the BF16 probe's scaler (pink), and the drop of the native probe relative to BF16 (grey; no CI available). Error bars are 95% paired-bootstrap CIs; dashes mark the gap a chance-level (0.33) probe would have. As for **H**, W8 transfers without loss and W3 does not.

			Transfer gap after map (native − mapped)					
Model	Var.	Native	Frozen	Re-scaled	Affine	Procrustes	Ridge	Act. cos.
Qwen3-4B	W8	0.92	0.01 [0.00, 0.01]	0.00°	0.00°	0.01 [0.00, 0.01]	0.00°	1.00
	W4	0.91	0.14 [0.12, 0.16]	0.08	0.08	0.02 [0.01, 0.04]	−0.02	0.89
	NF4	0.90	0.05 [0.03, 0.06]	0.05	0.04	0.03 [0.01, 0.04]	−0.02	0.91
	W3	0.88	0.27 [0.25, 0.30]	0.21	0.14	0.06 [0.04, 0.08]	−0.02	0.66
Qwen3-8B	W8	0.91	0.00 [0.00, 0.01]	0.00°	0.00°	0.00 [0.00, 0.01]	0.00°	1.00
	W4	0.91	0.25 [0.23, 0.28]	0.11	0.10	0.05 [0.03, 0.06]	0.00°	0.82
	NF4	0.91	0.10 [0.08, 0.12]	0.06	0.05	0.04 [0.02, 0.05]	0.00°	0.88
	W3	0.89	0.42 [0.39, 0.45]	0.29	0.25	0.10 [0.08, 0.12]	0.01°	0.61
Llama-3.1-8B	W8	0.92	0.00 [−0.01, 0.00]	0.00°	0.00°	0.00 [−0.01, 0.00]	0.00°	1.00
	W4	0.91	0.05 [0.03, 0.06]	0.03	0.02	0.01 [0.00, 0.03]	−0.01	0.90
	NF4	0.92	0.03 [0.02, 0.04]	0.02	0.02	0.01 [0.00, 0.03]	0.00°	0.92
	W3	0.88	0.31 [0.28, 0.34]	0.21	0.18	0.10 [0.08, 0.12]	−0.02	0.50
Granite-4.1-8B	W8	0.90	0.00 [−0.01, 0.00]	0.00°	0.00°	0.00 [−0.01, 0.00]	0.00°	1.00
	W4	0.90	0.05 [0.03, 0.07]	0.02	0.02	0.01 [−0.01, 0.02]	−0.02	0.93
	NF4	0.90	0.12 [0.10, 0.14]	0.03	0.02	0.01 [0.00, 0.02]	0.00°	0.94
	W3	0.87	0.27 [0.24, 0.30]	0.16	0.14	0.07 [0.05, 0.09]	0.00°	0.75

Table 15: RQ2 diagnosis at the BF16-selected layer (**M** feature). Native = accuracy of the probe trained on the quantized model at that layer. Each gap is native accuracy minus the accuracy of the frozen BF16 probe after a label-free map from quantized to BF16 activations, fit on unlabelled paired training activations: none (frozen), scaler refit (re-scaled), per-dimension affine, re-scaling followed by an orthogonal rotation (Procrustes), and a full ridge map (upper bound). Brackets: 95% paired-bootstrap CIs, shown for the frozen probe and Procrustes; the CIs of the other maps are in the released analysis outputs, and ° marks a gap whose CI includes 0 (decided on the unrounded CI bounds, here and in Table 2). Act. cos. = mean per-item cosine between quantized and BF16 test activations.

model. On Qwen3-4B at W4, the frozen probe matches the native probe at layer 21 (0.91 vs. 0.91) but drops to 0.76 (native 0.91) at the BF16-selected layer 22. The label-free maps of Table 2 were fitted only at the BF16-selected layer, so the Qwen3-4B W4 result there (a gap of 0.15 and a significant gain from the rotation) describes layer 22 and need not hold at other layers: at layer 21 there is no gap to close. On Qwen3-8B at W4, no layer rescues the frozen probe: its best accuracy at any layer is 0.77 (layer 26; native 0.89). At W3, the frozen probe stays below the native probe at every layer on every model; its best values are 0.67, 0.68, 0.68 and 0.79, against native peaks of 0.90, 0.90, 0.89 and 0.90. Early layers transfer poorly on every model. Per-layer values have no CIs, which is why the main text uses plateau averages.

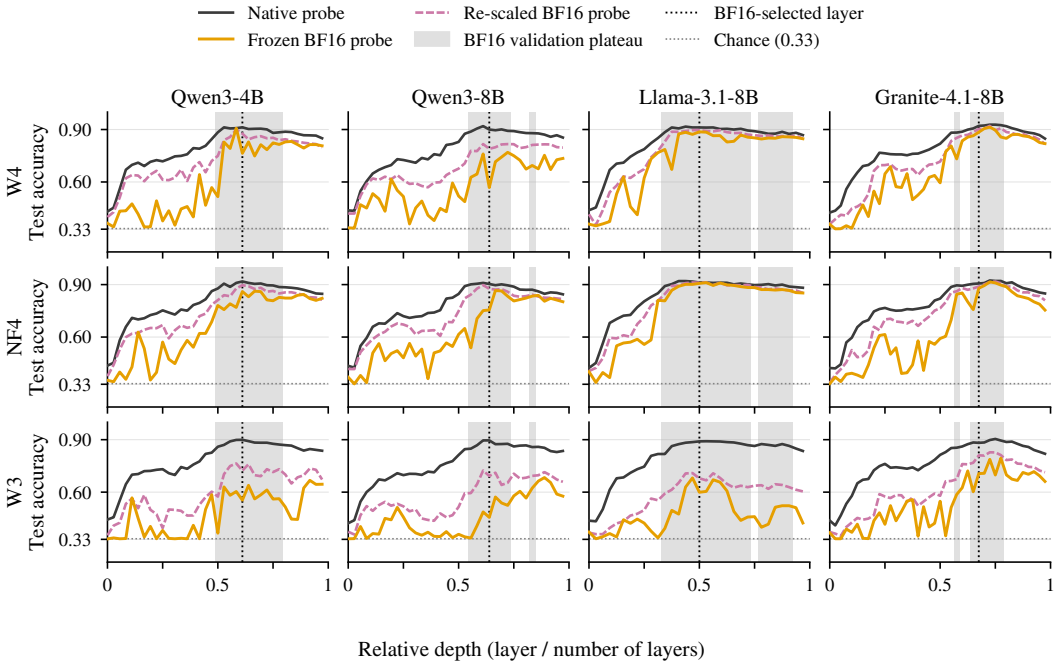

Figure 8: Transfer by layer (residual stream **H**). Test accuracy of the native probe (dark grey), the frozen BF16 probe (orange) and the re-scaled BF16 probe (pink, dashed) at every layer, plotted against relative depth, for W4, NF4 and W3 (rows) on each model (columns). Grey bands mark the BF16 validation plateau, the dotted vertical line the BF16-selected layer, and the dotted horizontal line chance (0.33). Per-layer values have no CIs. At 4 bits, frozen transfer can change sharply between neighbouring layers; at W3 it stays well below the native probe at every layer.

Probe-direction null. The refit-noise null for the probe-direction cosine is low (Table 16): refitting the BF16 probe on bootstrap resamples of its own training set gives cosines of 0.67–0.81 for Answer and Refuse and 0.80–0.90 for Conflict. The higher Conflict range reflects a direction that is more stable under refitting, not more robustness to quantization. W8 cosines (0.97–1.00) lie above the null: each null refit resamples the training items, whereas the W8 probe is fitted on the same items as the BF16 probe, on nearly identical activations. Almost all W4, NF4 and W3 cosines lie below it (32 of 36 on **H**); the exceptions on **H** are Granite-4.1-8B W4 Answer and Refuse, Granite-4.1-8B NF4 Refuse and Llama-3.1-8B NF4 Refuse. Granite-4.1-8B NF4 **H** Answer falls below the null by less than 0.002 (both print as 0.73). A cosine below the null shows that the native probe direction changed by more than refit noise; on its own it does not show that the activations rotated.

Transfer matrices. Figure 9 applies the probe of each variant, frozen, to every other variant. These cells are point estimates without CIs, so we read them only descriptively. BF16 and W8 probes are interchangeable. At W3, the reverse direction (W3 probe on BF16 activations)

		H			**M**		
Model	Row	Answer	Refuse	Conflict	Answer	Refuse	Conflict
Qwen3-4B	Null (50 refits)	[0.67, 0.76]	[0.69, 0.79]	[0.81, 0.87]	[0.67, 0.76]	[0.68, 0.78]	[0.80, 0.85]
	W8	0.98	0.98	0.99	0.98	0.98	0.99
	W4	0.55†	0.59†	0.67†	0.52†	0.55†	0.64†
	NF4	0.57†	0.60†	0.66†	0.53†	0.54†	0.62†
	W3	0.21†	0.26†	0.35†	0.14†	0.17†	0.29†
Qwen3-8B	Null (50 refits)	[0.74, 0.79]	[0.72, 0.80]	[0.83, 0.89]	[0.72, 0.79]	[0.73, 0.81]	[0.83, 0.89]
	W8	0.97	0.98	0.99	0.97	0.97	0.98
	W4	0.41†	0.41†	0.50†	0.38†	0.36†	0.42†
	NF4	0.55†	0.53†	0.64†	0.49†	0.48†	0.59†
	W3	0.23†	0.21†	0.31†	0.18†	0.16†	0.22†
Llama-3.1-8B	Null (50 refits)	[0.72, 0.80]	[0.70, 0.80]	[0.83, 0.89]	[0.73, 0.79]	[0.73, 0.81]	[0.83, 0.90]
	W8	0.99	0.99	1.00	0.99	0.99	1.00
	W4	0.63†	0.65†	0.76†	0.64†	0.65†	0.74†
	NF4	0.70†	0.71	0.78†	0.70†	0.70†	0.77†
	W3	0.17†	0.14†	0.26†	0.13†	0.13†	0.24†
Granite-4.1-8B	Null (50 refits)	[0.73, 0.80]	[0.71, 0.79]	[0.84, 0.90]	[0.70, 0.76]	[0.70, 0.78]	[0.82, 0.89]
	W8	0.98	0.98	0.99	0.98	0.98	0.99
	W4	0.73	0.74	0.81†	0.67†	0.68†	0.76†
	NF4	0.73†	0.72	0.80†	0.65†	0.67†	0.74†
	W3	0.35†	0.35†	0.46†	0.27†	0.27†	0.35†

Table 16: Probe-direction cosine between the native probe of each quantized model and the BF16 probe (per class, w_c/σ in raw activation space, BF16-selected layer), against a refit-noise null: the BF16 probe refit with the same recipe on bootstrap resamples of the BF16 training items (2.5–97.5th percentile of its cosine with the original). † = below the null range (decided before rounding), i.e., the direction changed by more than refit noise. The null resamples the training items, so W8 cosines lie above it.

works better than BF16 → W3 on all four models for **H** (0.64 vs. 0.55, 0.72 vs. 0.46, 0.75 vs. 0.60, 0.76 vs. 0.71). W4 and NF4 do not form a shared 4-bit cluster: W4 → NF4 reaches 0.68, 0.79, 0.86 and 0.87 for **H**.

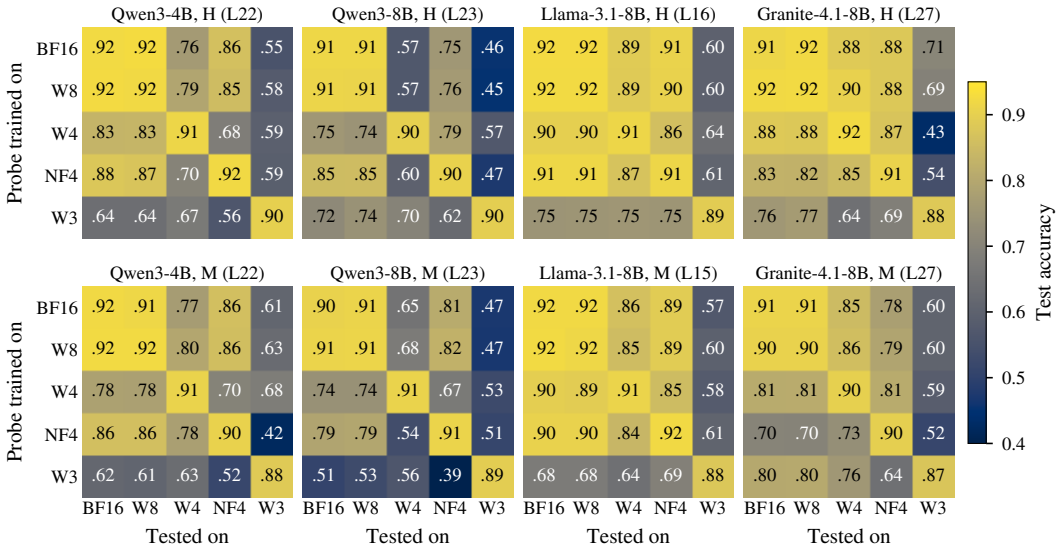

Figure 9: Cross-precision transfer matrices at the BF16-selected layer (given in each panel title): test accuracy of a probe trained on one variant (rows) and applied, frozen, to another variant's activations (columns), for the residual stream **H** (top) and the MLP output **M** (bottom). The diagonal holds the native probes. Single point estimates; no CIs were computed for off-diagonal cells, so asymmetries are descriptive.

G Full RQ3 Tables

Table 17 lists the **H** cases that meet the claim or partial rule (Section 3.6). The excess-loss CI also lies below 0 in all four claims (last column), although the claim rule does not require it.

Model	Var.	Class	ΔB	Holm p	ΔP	KBF / Lost	n	Excess loss
Claims								
Llama-3.1-8B	W4	Refuse	-0.06 $[-0.09, -0.03]$	0.002	0.01 $[-0.01, 0.04]$	0.10 / 0.01	399	-0.06 $[-0.09, -0.03]$
Llama-3.1-8B	W4	Conflict	-0.08 $[-0.12, -0.03]$	0.012	-0.01 $[-0.03, 0.00]$	0.30 / 0.00	233	-0.06 $[-0.10, -0.01]$
Llama-3.1-8B	NF4	Conflict	-0.10 $[-0.14, -0.06]$	<0.001	-0.01 $[-0.03, 0.01]$	0.30 / 0.00	232	-0.09 $[-0.13, -0.05]$
Granite-4.1-8B	W3	Refuse	-0.13 $[-0.17, -0.10]$	<0.001	-0.01 $[-0.03, 0.02]$	0.19 / 0.03	374	-0.13 $[-0.17, -0.08]$
Partial cases								
Qwen3-4B	W3	Answer	-0.19 $[-0.23, -0.15]$	<0.001	-0.06 $[-0.10, -0.03]$	0.25 / 0.05	338	-0.13 $[-0.18, -0.08]$
Llama-3.1-8B	W3	Answer	-0.26 $[-0.31, -0.22]$	<0.001	-0.04 $[-0.08, -0.01]$	0.34 / 0.05	362	-0.22 $[-0.28, -0.17]$
Llama-3.1-8B	W3	Conflict	-0.11 $[-0.16, -0.06]$	<0.001	-0.03 $[-0.05, 0.00]$	0.42 / 0.03	227	-0.08 $[-0.14, -0.03]$
Granite-4.1-8B	W3	Conflict	-0.20 $[-0.24, -0.16]$	<0.001	-0.04 $[-0.06, -0.01]$	0.70 / 0.01	155	-0.16 $[-0.21, -0.11]$

Table 17: RQ3 claims and partial cases for the **H** probe (generation readout; W8, W4, NF4 and W3). ΔB = generation recall change vs. BF16; Holm p = McNemar test on ΔB, Holm-corrected over the model's 30 tests (Section 3.7); ΔP = native-probe recall change; KBF = share of the $n = |S_c \cap \mathcal{P}|$ items that BF16 behaviour gets right and the quantized probe still classifies correctly, on which quantized behaviour is now wrong; Lost = share of the $|S_c|$ items that BF16 behaviour gets right on which both probe and generation fail; excess loss = $\Delta B - \min(\Delta P, 0)$. Claim: $\Delta B < 0$, Holm $p < 0.05$, KBF $>$ Lost, and $\Delta P \geq -0.02$ with CI upper bound ≥ 0. Partial: the same behaviour drop with an excess-loss CI below 0, but ΔP fails the probe condition. Brackets: 95% paired-bootstrap CIs, rounded to two decimals; the probe condition is checked on the unrounded CI bounds, and the excess loss is computed from the unrounded ΔB and ΔP.

Qwen3-8B has no claim or partial case on **H**. Its W4 Refuse behaviour drops (-0.08 $[-0.11, -0.04]$, Holm $p = 0.001$), but the probe drops too (-0.04 $[-0.06, -0.01]$) and the excess-loss CI covers 0; at W3, the probe loses more Answer recall than behaviour (ΔP -0.12 $[-0.16, -0.08]$, ΔB -0.06 $[-0.09, -0.03]$).

Tables 21–24 list every RQ3 case of the generation readout (W8, W4, NF4, W3) for both probe features, one table per model with the **H** rows on top and the **M** rows below. Claims are decided on the **H** probe, which had the higher validation accuracy in 16 of 20 model–variant pairs. Three cases meet the claim rule with **M** only: Qwen3-8B W4 Refuse, Qwen3-8B W3 Answer and Llama-3.1-8B W3 Conflict. On **H**, both Qwen3-8B cases fail the probe condition, and the Llama-3.1-8B W3 Conflict case is partial. We do not count **M**-only cases as claims.

Probe layer. ΔP reads each native probe at its own validation-selected layer, and this selection is noisy for some variants (Appendix E). As a check, we read every native **H** probe at the BF16-selected layer instead. The four claims and the four partial cases remain, and one more case meets the claim rule: Refuse on Qwen3-8B at W4, whose ΔP is then -0.01 $[-0.04, 0.01]$. For Answer on Qwen3-8B at W3, ΔP is then -0.03 $[-0.06, 0.00]$, which still fails condition (ii), and the excess-loss CI covers 0. We ran this check after all results were known and keep each variant's own layer for the claims.

Token cap. Table 18 compares the claim and partial cases under the two token caps (Appendix C). With the 128-token cap, 8–44% of the BF16 outputs and 63–70% of the Qwen W3 outputs were cut off, and five cases met the claim rule. With the 512-token cap, four of them and all four partial cases remain, most with a slightly larger behaviour drop. Qwen3-4B Conflict at W3 does not remain: its drop is still negative, but the unadjusted McNemar p rises from 0.001 to 0.007 and the Holm-adjusted p from 0.018 to 0.119. Of the 9,568 regenerated outputs, 998 (10%) get a different label. In 476 of them a Refuse label becomes Answer, which is consistent with outputs that go through the documents one by one and were cut off before they reached a document with the answer. We raised the cap after the first analysis, so the choice between the two caps was not fixed in advance; we report the 512-token results because few outputs are cut off there.

			128-token cap				512-token cap			
Model	Var.	Class	Hit cap	ΔB	Holm p	Status	Hit cap	ΔB	Holm p	Status
Qwen3-4B	W3	Answer	0.63	-0.20 $[-0.24, -0.16]$	<0.001	partial	0.04	-0.19 $[-0.23, -0.15]$	<0.001	partial
Qwen3-4B	W3	Conflict	0.63	-0.08 $[-0.12, -0.03]$	0.018	claim	0.04	-0.06 $[-0.11, -0.02]$	0.119	–
Llama-3.1-8B	W4	Refuse	0.25	-0.05 $[-0.08, -0.02]$	0.016	claim	0.04	-0.06 $[-0.09, -0.03]$	0.002	claim
Llama-3.1-8B	W4	Conflict	0.25	-0.07 $[-0.11, -0.03]$	0.029	claim	0.04	-0.08 $[-0.12, -0.03]$	0.012	claim
Llama-3.1-8B	NF4	Conflict	0.27	-0.09 $[-0.13, -0.05]$	<0.001	claim	0.02	-0.10 $[-0.14, -0.06]$	<0.001	claim
Llama-3.1-8B	W3	Answer	0.18	-0.26 $[-0.30, -0.22]$	<0.001	partial	0.03	-0.26 $[-0.31, -0.22]$	<0.001	partial
Llama-3.1-8B	W3	Conflict	0.18	-0.10 $[-0.15, -0.05]$	0.002	partial	0.03	-0.11 $[-0.16, -0.06]$	<0.001	partial
Granite-4.1-8B	W3	Refuse	0.28	-0.12 $[-0.16, -0.08]$	<0.001	claim	0.00	-0.13 $[-0.17, -0.10]$	<0.001	claim
Granite-4.1-8B	W3	Conflict	0.28	-0.19 $[-0.23, -0.15]$	<0.001	partial	0.00	-0.20 $[-0.24, -0.16]$	<0.001	partial
Largest gain (no claim)										
Qwen3-8B	W3	Conflict	0.70	0.21 [0.16, 0.26]	<0.001	–	0.05	0.26 [0.21, 0.31]	<0.001	–

Table 18: The cases that are a claim or a partial case with the 128-token or with the 512-token test generations (**H** probe, generation readout), and the largest gain. The 512-token columns are the results reported in this paper (Table 17). Hit cap = share of the variant's 1,440 outputs that reach the cap; ΔB and Holm p as in Table 17; Status = claim, partial case, or neither (–). Brackets: 95% paired-bootstrap CIs.

Behaviour shifts and caveats. At W3, over-abstention rises by 0.19 [0.15, 0.23] on Qwen3-4B and 0.26 [0.22, 0.31] on Llama-3.1-8B, while the false-answer rate of Granite-4.1-8B rises from 0.41 [0.38, 0.44] to 0.56 [0.53, 0.59] and its generation Conflict recall falls from 0.33 to 0.13 (ΔB -0.20 $[-0.24, -0.16]$; Table 9). The Conflict gain on Qwen3-8B at W3 (ΔB 0.26 [0.21, 0.31]) is not a truncation effect: it is larger with the 512-token cap than with the 128-token cap (Table 18). Almost all of it lies in outputs that flag the conflict and still answer (0.24 [0.20, 0.29]), while native-probe Conflict recall changes by -0.03 $[-0.06, 0.00]$; we report the gain without an explanation. KBF counts one-directional churn and can be high at zero net change: on Granite-4.1-8B at W4, Conflict has KBF 0.33 with ΔB 0.01 $[-0.04, 0.05]$.

Noticing versus answering. The generation readout scores which evidence state the output signals, not whether the model then abstains (Section 3.5). Tables 19 and 20 split its recall by the judge's third field into outputs that notice the problem and give no answer,

and outputs that notice it but still answer. In all four claims the loss lies in the noticed-but-answered part (-0.08 to -0.20, each CI below 0), and the no-answer part does not fall significantly. For Refuse, the intended response is to abstain, so the two Refuse claims are claims about noticing only: on Llama-3.1-8B at W4 and on Granite-4.1-8B at W3 the outputs say less often that the documents lack the answer, but they do not abstain less often (no-answer part 0.02 $[-0.03, 0.07]$ and 0.07 $[0.02, 0.11]$; 203 to 214 and 216 to 249 items). For Conflict, the intended response is to flag the conflict, with or without an answer, which is what the readout scores. Its no-answer part is small (0.01–0.05 of the Conflict items at BF16); on Qwen3-4B at W3, 13 of its 19 outputs are cut off at the cap, so its rise there (0.03 $[0.01, 0.04]$) still partly reflects truncation, not abstention. We keep the detection reading as the primary readout, because the baseline selection, the tests and the claim rule were all run with it.

The split also qualifies the W3 behaviour shifts. Llama-3.1-8B does answer less often: the share of Answer items whose output gives an answer falls by 0.15 $[0.12, 0.19]$, and the outputs without an answer rise from 21 to 94. On Qwen3-4B this share falls by 0.07 $[0.04, 0.10]$ (37 to 71 outputs without an answer), half the fall seen with the 128-token cap, where it lay only in outputs that hit the cap. On Granite-4.1-8B the share of Conflict items with an answer does not change significantly and that of Refuse items falls (-0.07 $[-0.11, -0.02]$; Table 20), so its higher FAR comes from outputs that still answer but no longer mention the problem. The split was added after all results were known and is descriptive: its CIs are not adjusted for multiplicity and lie outside the Holm family of the claim rule, and the judge's third field is covered only by the small check in Appendix D.

		Refuse items				Conflict items			
Model	Var.	Noticed	No ans.	Answered	Ans. given	Noticed	No ans.	Answered	Ans. given
Qwen3-4B	BF16	0.91	0.72	0.19	0.28	0.55	0.01	0.54	0.96
	W8	0.90	0.71	0.19	0.29	0.56	0.02	0.53	0.96
	W4	0.92	0.74	0.19	0.26	0.61	0.03	0.58	0.94
	NF4	0.89	0.68	0.21	0.32	0.56	0.03	0.53	0.95
	W3	0.95	0.79	0.16	0.21	0.49	0.04	0.45	0.93
Qwen3-8B	BF16	0.71	0.55	0.16	0.44	0.18	0.01	0.18	0.98
	W8	0.71	0.57	0.14	0.43	0.18	0.00	0.18	0.98
	W4	0.64	0.50	0.14	0.49	0.15	0.01	0.14	0.98
	NF4	0.68	0.57	0.11	0.43	0.18	0.02	0.15	0.98
	W3	0.75	0.40	0.35	0.59	0.44	0.02	0.42	0.97
Llama-3.1-8B	BF16	0.87	0.42	0.45	0.58	0.49	0.05	0.44	0.94
	W8	0.87	0.44	0.43	0.56	0.49	0.05	0.44	0.95
	W4	0.81	0.45	0.36	0.55	0.42	0.05	0.36	0.93
	NF4	0.84	0.37	0.47	0.63	0.39	0.05	0.35	0.93
	W3	0.90	0.78	0.13	0.23	0.38	0.14	0.25	0.77
Granite-4.1-8B	BF16	0.83	0.45	0.38	0.55	0.33	0.01	0.32	0.99
	W8	0.83	0.46	0.37	0.54	0.34	0.01	0.33	0.99
	W4	0.85	0.54	0.32	0.46	0.33	0.02	0.32	0.97
	NF4	0.83	0.50	0.32	0.50	0.30	0.01	0.29	0.98
	W3	0.70	0.52	0.18	0.48	0.13	0.01	0.12	0.98

Table 19: Generation readout on the Refuse and Conflict test items (480 each), split by whether the output also gives an answer (the judge's third field, Appendix D). Noticed = share of the class's items labelled with that class (the recall of Table 9); it is the sum, before rounding, of No ans. (noticed, no answer given) and Answered (noticed but answered; unlike NBA in Table 9, only outputs labelled with the item's own class count). Ans. given = share of the class's items whose output gives an answer, noticed or not. Point estimates.

Model	Var.	Class	ΔB	Split of ΔB		Δ answer given
				No answer	Answered	
Claims						
Llama-3.1-8B	W4	Refuse	−0.06 [−0.09, −0.03]	0.02 [−0.03, 0.07]	−0.09 [−0.14, −0.03]	−0.02 [−0.07, 0.03]
Llama-3.1-8B	W4	Conflict	−0.08 [−0.12, −0.03]	0.00 [−0.02, 0.03]	−0.08 [−0.13, −0.03]	−0.01 [−0.04, 0.01]
Llama-3.1-8B	NF4	Conflict	−0.10 [−0.14, −0.06]	−0.01 [−0.03, 0.02]	−0.09 [−0.14, −0.05]	−0.01 [−0.04, 0.01]
Granite-4.1-8B	W3	Refuse	−0.13 [−0.17, −0.10]	0.07 [0.02, 0.11]	−0.20 [−0.25, −0.15]	−0.07 [−0.11, −0.02]
Partial cases						
Qwen3-4B	W3	Answer	−0.19 [−0.23, −0.15]	–	−0.19 [−0.23, −0.15]	−0.07 [−0.10, −0.04]
Llama-3.1-8B	W3	Answer	−0.26 [−0.31, −0.22]	–	−0.26 [−0.31, −0.22]	−0.15 [−0.19, −0.12]
Llama-3.1-8B	W3	Conflict	−0.11 [−0.16, −0.06]	0.08 [0.05, 0.12]	−0.19 [−0.25, −0.14]	−0.18 [−0.21, −0.14]
Granite-4.1-8B	W3	Conflict	−0.20 [−0.24, −0.16]	0.00 [−0.01, 0.02]	−0.20 [−0.25, −0.16]	−0.01 [−0.03, 0.00]

Table 20: The claims and partial cases of Table 17, with ΔB split by whether the output also gives an answer. No answer / Answered: change vs. BF16 in the share of the class's items that are labelled with the class and give no answer / give an answer; the two sum to ΔB before rounding. Δ answer given: change in the share of the class's items whose output gives an answer, whatever its label. For Refuse and Conflict, No answer is the stricter reading that counts only outputs that notice the problem and do not answer. For Answer, a matching label means that the output raises no flag, so ΔB lies in Answered (no entry under No answer). Brackets: 95% paired-bootstrap CIs, not adjusted for multiplicity.

Var.	Class	ΔB	Holm p	ΔP	KBF / Lost	Excess loss	Status
H *probe (residual stream; main-text claims)*							
W8	Answer	−0.01 [−0.03, 0.01]	1.000	−0.02 [−0.03, 0.00]	0.03 / 0.00	0.01 [−0.01, 0.03]	–
W8	Refuse	−0.01 [−0.02, 0.01]	1.000	0.00 [−0.01, 0.01]	0.00 / 0.01	0.00 [−0.02, 0.01]	–
W8	Conflict	0.00 [−0.02, 0.02]	1.000	0.01 [0.00, 0.02]	0.04 / 0.00	0.00 [−0.02, 0.02]	–
W4	Answer	−0.03 [−0.06, 0.00]	0.617	−0.05 [−0.09, −0.02]	0.08 / 0.01	0.02 [−0.02, 0.07]	–
W4	Refuse	0.01 [−0.01, 0.04]	1.000	−0.02 [−0.05, 0.01]	0.02 / 0.00	0.03 [0.00, 0.07]	–
W4	Conflict	0.05 [0.01, 0.09]	0.143	0.05 [0.03, 0.07]	0.10 / 0.01	0.05 [0.01, 0.09]	–
NF4	Answer	−0.04 [−0.06, −0.01]	0.401	−0.03 [−0.06, 0.00]	0.09 / 0.01	0.00 [−0.04, 0.04]	–
NF4	Refuse	−0.02 [−0.05, 0.00]	1.000	−0.01 [−0.03, 0.02]	0.05 / 0.01	−0.02 [−0.04, 0.02]	–
NF4	Conflict	0.00 [−0.03, 0.04]	1.000	0.00 [−0.02, 0.03]	0.14 / 0.02	0.00 [−0.03, 0.05]	–
W3	Answer	−0.19 [−0.23, −0.15]	<0.001	−0.06 [−0.10, −0.03]	0.25 / 0.05	−0.13 [−0.18, −0.08]	partial
W3	Refuse	0.04 [0.02, 0.07]	0.031	−0.04 [−0.07, −0.02]	0.01 / 0.01	0.09 [0.05, 0.12]	–
W3	Conflict	−0.06 [−0.11, −0.02]	0.119	0.01 [−0.01, 0.04]	0.26 / 0.02	−0.06 [−0.11, −0.02]	–
M *probe (MLP output)*							
W8	Answer	−0.01 [−0.03, 0.01]	1.000	0.00 [−0.01, 0.02]	0.02 / 0.01	−0.01 [−0.03, 0.01]	–
W8	Refuse	−0.01 [−0.02, 0.01]	1.000	0.00 [−0.01, 0.01]	0.01 / 0.01	0.00 [−0.02, 0.01]	–
W8	Conflict	0.00 [−0.02, 0.02]	1.000	0.00 [−0.01, 0.01]	0.04 / 0.00	0.00 [−0.02, 0.02]	–
W4	Answer	−0.03 [−0.06, 0.00]	0.617	−0.04 [−0.08, −0.01]	0.08 / 0.01	0.01 [−0.03, 0.06]	–
W4	Refuse	0.01 [−0.01, 0.04]	1.000	−0.03 [−0.05, 0.00]	0.02 / 0.00	0.04 [0.00, 0.07]	–
W4	Conflict	0.05 [0.01, 0.09]	0.143	0.01 [−0.01, 0.04]	0.11 / 0.00	0.05 [0.02, 0.09]	–
NF4	Answer	−0.04 [−0.06, −0.01]	0.401	−0.02 [−0.05, 0.01]	0.09 / 0.01	−0.01 [−0.05, 0.03]	–
NF4	Refuse	−0.02 [−0.05, 0.00]	1.000	−0.01 [−0.03, 0.01]	0.05 / 0.01	−0.01 [−0.04, 0.02]	–
NF4	Conflict	0.00 [−0.03, 0.04]	1.000	−0.01 [−0.03, 0.02]	0.14 / 0.02	0.01 [−0.03, 0.05]	–
W3	Answer	−0.19 [−0.23, −0.15]	<0.001	−0.06 [−0.10, −0.03]	0.25 / 0.05	−0.13 [−0.18, −0.08]	partial
W3	Refuse	0.04 [0.02, 0.07]	0.031	−0.05 [−0.08, −0.02]	0.01 / 0.01	0.09 [0.05, 0.13]	–
W3	Conflict	−0.06 [−0.11, −0.02]	0.119	0.00 [−0.02, 0.03]	0.25 / 0.03	−0.06 [−0.10, −0.01]	–

Table 21: All RQ3 rows for Qwen3-4B (generation readout; W8, W4, NF4 and W3), for the **H** probe (top) and the **M** probe (bottom). ΔB and its Holm p (the model's family of 30 McNemar tests) depend only on behaviour, so they are identical in both blocks. Other columns as in Table 17; KBF / Lost are rates. Claims in the main text are decided on **H**; **M** is shown for completeness. Brackets: 95% paired-bootstrap CIs.

Var.	Class	ΔB	Holm p	ΔP	KBF / Lost	Excess loss	Status
H *probe (residual stream; main-text claims)*							
W8	Answer	0.00 [−0.01, 0.01]	1.000	0.01 [−0.01, 0.04]	0.00 / 0.00	0.00 [−0.01, 0.02]	–
W8	Refuse	0.00 [−0.02, 0.01]	1.000	−0.01 [−0.03, 0.01]	0.03 / 0.00	0.00 [−0.02, 0.03]	–
W8	Conflict	0.00 [−0.02, 0.02]	1.000	0.02 [0.00, 0.04]	0.13 / 0.00	0.00 [−0.02, 0.02]	–
W4	Answer	−0.01 [−0.03, 0.01]	1.000	−0.06 [−0.10, −0.03]	0.03 / 0.00	0.05 [0.01, 0.09]	–
W4	Refuse	−0.08 [−0.11, −0.04]	0.001	−0.04 [−0.06, −0.01]	0.16 / 0.02	−0.04 [−0.09, 0.01]	–
W4	Conflict	−0.04 [−0.07, 0.00]	0.506	0.02 [0.00, 0.04]	0.49 / 0.02	−0.04 [−0.07, 0.00]	–
NF4	Answer	0.01 [−0.01, 0.03]	1.000	−0.01 [−0.05, 0.02]	0.01 / 0.01	0.02 [0.00, 0.06]	–
NF4	Refuse	−0.04 [−0.07, 0.00]	0.557	−0.04 [−0.07, −0.01]	0.11 / 0.02	0.00 [−0.04, 0.05]	–
NF4	Conflict	−0.01 [−0.04, 0.03]	1.000	0.01 [−0.01, 0.04]	0.40 / 0.00	−0.01 [−0.04, 0.03]	–
W3	Answer	−0.06 [−0.09, −0.03]	0.002	−0.12 [−0.16, −0.08]	0.08 / 0.03	0.06 [0.01, 0.10]	–
W3	Refuse	0.04 [−0.01, 0.08]	1.000	−0.05 [−0.08, −0.02]	0.12 / 0.02	0.09 [0.03, 0.14]	–
W3	Conflict	0.26 [0.21, 0.31]	<0.001	−0.03 [−0.06, 0.00]	0.19 / 0.01	0.29 [0.24, 0.34]	–
M *probe (MLP output)*							
W8	Answer	0.00 [−0.01, 0.01]	1.000	0.01 [0.00, 0.03]	0.00 / 0.00	0.00 [−0.01, 0.01]	–
W8	Refuse	0.00 [−0.02, 0.01]	1.000	0.00 [−0.01, 0.01]	0.03 / 0.01	0.00 [−0.02, 0.02]	–
W8	Conflict	0.00 [−0.02, 0.02]	1.000	0.01 [0.00, 0.02]	0.13 / 0.00	0.00 [−0.02, 0.02]	–
W4	Answer	−0.01 [−0.03, 0.01]	1.000	−0.03 [−0.06, 0.00]	0.02 / 0.01	0.02 [−0.01, 0.05]	–
W4	Refuse	−0.08 [−0.11, −0.04]	0.001	0.01 [−0.01, 0.04]	0.16 / 0.01	−0.08 [−0.11, −0.04]	claim
W4	Conflict	−0.04 [−0.07, 0.00]	0.506	0.02 [0.00, 0.05]	0.50 / 0.01	−0.04 [−0.07, 0.00]	–
NF4	Answer	0.01 [−0.01, 0.03]	1.000	0.01 [−0.02, 0.03]	0.01 / 0.00	0.01 [−0.01, 0.03]	–
NF4	Refuse	−0.04 [−0.07, 0.00]	0.557	0.00 [−0.03, 0.02]	0.11 / 0.02	−0.03 [−0.06, 0.01]	–
NF4	Conflict	−0.01 [−0.04, 0.03]	1.000	0.01 [−0.01, 0.04]	0.40 / 0.01	−0.01 [−0.04, 0.03]	–
W3	Answer	−0.06 [−0.09, −0.03]	0.002	0.00 [−0.03, 0.04]	0.07 / 0.02	−0.06 [−0.09, −0.02]	claim
W3	Refuse	0.04 [−0.01, 0.08]	1.000	−0.03 [−0.06, 0.00]	0.12 / 0.01	0.06 [0.01, 0.11]	–
W3	Conflict	0.26 [0.21, 0.31]	<0.001	0.00 [−0.03, 0.03]	0.19 / 0.01	0.26 [0.22, 0.32]	–

Table 22: All RQ3 rows for Qwen3-8B; layout and columns as in Table 21.

Var.	Class	ΔB	Holm p	ΔP	KBF / Lost	Excess loss	Status
H *probe (residual stream; main-text claims)*							
W8	Answer	0.01 [−0.01, 0.03]	1.000	0.00 [−0.01, 0.01]	0.02 / 0.00	0.01 [−0.01, 0.03]	–
W8	Refuse	−0.01 [−0.02, 0.01]	1.000	0.00 [−0.01, 0.01]	0.02 / 0.00	0.00 [−0.02, 0.01]	–
W8	Conflict	0.00 [−0.02, 0.02]	1.000	−0.01 [−0.02, 0.00]	0.06 / 0.00	0.00 [−0.02, 0.03]	–
W4	Answer	0.03 [0.01, 0.06]	0.270	−0.03 [−0.06, −0.01]	0.03 / 0.00	0.07 [0.03, 0.10]	–
W4	Refuse	−0.06 [−0.09, −0.03]	0.002	0.01 [−0.01, 0.04]	0.10 / 0.01	−0.06 [−0.09, −0.03]	claim
W4	Conflict	−0.08 [−0.12, −0.03]	0.012	−0.01 [−0.03, 0.00]	0.30 / 0.00	−0.06 [−0.10, −0.01]	claim
NF4	Answer	−0.04 [−0.07, −0.01]	0.081	−0.01 [−0.03, 0.01]	0.07 / 0.01	−0.03 [−0.06, 0.00]	–
NF4	Refuse	−0.03 [−0.06, 0.00]	0.937	−0.01 [−0.04, 0.01]	0.08 / 0.02	−0.02 [−0.05, 0.02]	–
NF4	Conflict	−0.10 [−0.14, −0.06]	<0.001	−0.01 [−0.03, 0.01]	0.30 / 0.00	−0.09 [−0.13, −0.05]	claim
W3	Answer	−0.26 [−0.31, −0.22]	<0.001	−0.04 [−0.08, −0.01]	0.34 / 0.05	−0.22 [−0.28, −0.17]	partial
W3	Refuse	0.03 [−0.01, 0.06]	1.000	−0.03 [−0.05, 0.00]	0.06 / 0.02	0.05 [0.01, 0.10]	–
W3	Conflict	−0.11 [−0.16, −0.06]	<0.001	−0.03 [−0.05, 0.00]	0.42 / 0.03	−0.08 [−0.14, −0.03]	partial
M *probe (MLP output)*							
W8	Answer	0.01 [−0.01, 0.03]	1.000	0.00 [−0.02, 0.01]	0.02 / 0.00	0.01 [−0.01, 0.03]	–
W8	Refuse	−0.01 [−0.02, 0.01]	1.000	0.00 [−0.01, 0.01]	0.02 / 0.00	0.00 [−0.02, 0.01]	–
W8	Conflict	0.00 [−0.02, 0.02]	1.000	0.00 [−0.01, 0.01]	0.06 / 0.00	0.00 [−0.02, 0.02]	–
W4	Answer	0.03 [0.01, 0.06]	0.270	−0.04 [−0.06, −0.01]	0.04 / 0.00	0.07 [0.03, 0.11]	–
W4	Refuse	−0.06 [−0.09, −0.03]	0.002	0.00 [−0.02, 0.03]	0.09 / 0.02	−0.06 [−0.09, −0.03]	claim
W4	Conflict	−0.08 [−0.12, −0.03]	0.012	0.00 [−0.02, 0.02]	0.30 / 0.00	−0.08 [−0.11, −0.03]	claim
NF4	Answer	−0.04 [−0.07, −0.01]	0.081	0.00 [−0.03, 0.03]	0.07 / 0.01	−0.04 [−0.06, 0.00]	–
NF4	Refuse	−0.03 [−0.06, 0.00]	0.937	−0.02 [−0.04, 0.01]	0.07 / 0.03	−0.01 [−0.05, 0.03]	–
NF4	Conflict	−0.10 [−0.14, −0.06]	<0.001	0.02 [0.00, 0.04]	0.30 / 0.00	−0.10 [−0.14, −0.06]	claim
W3	Answer	−0.26 [−0.31, −0.22]	<0.001	−0.04 [−0.08, −0.01]	0.33 / 0.05	−0.22 [−0.28, −0.17]	partial
W3	Refuse	0.03 [−0.01, 0.06]	1.000	−0.05 [−0.08, −0.01]	0.06 / 0.02	0.08 [0.03, 0.12]	–
W3	Conflict	−0.11 [−0.16, −0.06]	<0.001	−0.01 [−0.04, 0.01]	0.42 / 0.03	−0.10 [−0.15, −0.05]	claim

Table 23: All RQ3 rows for Llama-3.1-8B; layout and columns as in Table 21.

Var.	Class	ΔB	Holm p	ΔP	KBF / Lost	Excess loss	Status
H *probe (residual stream; main-text claims)*							
W8	Answer	0.01 [−0.01, 0.02]	1.000	0.03 [0.00, 0.05]	0.01 / 0.00	0.01 [−0.01, 0.02]	–
W8	Refuse	−0.01 [−0.03, 0.01]	1.000	0.01 [−0.01, 0.03]	0.02 / 0.01	−0.01 [−0.02, 0.01]	–
W8	Conflict	0.01 [−0.01, 0.04]	1.000	0.00 [−0.01, 0.01]	0.11 / 0.01	0.02 [−0.01, 0.05]	–
W4	Answer	−0.01 [−0.03, 0.01]	1.000	0.03 [0.01, 0.06]	0.05 / 0.00	−0.01 [−0.03, 0.01]	–
W4	Refuse	0.02 [0.00, 0.05]	1.000	−0.01 [−0.03, 0.01]	0.04 / 0.00	0.03 [0.00, 0.06]	–
W4	Conflict	0.01 [−0.04, 0.05]	1.000	0.00 [−0.02, 0.01]	0.33 / 0.01	0.01 [−0.03, 0.06]	–
NF4	Answer	−0.03 [−0.05, 0.00]	0.558	0.00 [−0.02, 0.03]	0.06 / 0.01	−0.03 [−0.05, 0.00]	–
NF4	Refuse	0.00 [−0.04, 0.03]	1.000	−0.01 [−0.03, 0.01]	0.08 / 0.01	0.01 [−0.03, 0.05]	–
NF4	Conflict	−0.03 [−0.07, 0.01]	1.000	−0.01 [−0.03, 0.00]	0.38 / 0.01	−0.02 [−0.06, 0.03]	–
W3	Answer	−0.02 [−0.05, 0.00]	1.000	0.00 [−0.04, 0.03]	0.04 / 0.02	−0.02 [−0.04, 0.03]	–
W3	Refuse	−0.13 [−0.17, −0.10]	<0.001	−0.01 [−0.03, 0.02]	0.19 / 0.03	−0.13 [−0.17, −0.08]	claim
W3	Conflict	−0.20 [−0.24, −0.16]	<0.001	−0.04 [−0.06, −0.01]	0.70 / 0.01	−0.16 [−0.21, −0.11]	partial
M *probe (MLP output)*							
W8	Answer	0.01 [−0.01, 0.02]	1.000	−0.01 [−0.02, 0.01]	0.01 / 0.00	0.01 [0.00, 0.04]	–
W8	Refuse	−0.01 [−0.03, 0.01]	1.000	0.00 [−0.02, 0.01]	0.02 / 0.01	0.00 [−0.02, 0.02]	–
W8	Conflict	0.01 [−0.01, 0.04]	1.000	0.00 [−0.01, 0.00]	0.10 / 0.01	0.02 [−0.01, 0.04]	–
W4	Answer	−0.01 [−0.03, 0.01]	1.000	0.04 [0.00, 0.07]	0.04 / 0.01	−0.01 [−0.03, 0.01]	–
W4	Refuse	0.02 [0.00, 0.05]	1.000	0.01 [−0.02, 0.03]	0.04 / 0.01	0.02 [0.00, 0.05]	–
W4	Conflict	0.01 [−0.04, 0.05]	1.000	−0.02 [−0.04, 0.00]	0.32 / 0.02	0.03 [−0.02, 0.07]	–
NF4	Answer	−0.03 [−0.05, 0.00]	0.558	0.00 [−0.03, 0.03]	0.05 / 0.02	−0.03 [−0.05, 0.01]	–
NF4	Refuse	0.00 [−0.04, 0.03]	1.000	−0.01 [−0.03, 0.01]	0.08 / 0.01	0.01 [−0.03, 0.05]	–
NF4	Conflict	−0.03 [−0.07, 0.01]	1.000	−0.01 [−0.03, 0.00]	0.37 / 0.02	−0.02 [−0.06, 0.03]	–
W3	Answer	−0.02 [−0.05, 0.00]	1.000	0.00 [−0.04, 0.03]	0.05 / 0.01	−0.02 [−0.04, 0.03]	–
W3	Refuse	−0.13 [−0.17, −0.10]	<0.001	−0.01 [−0.04, 0.01]	0.19 / 0.03	−0.12 [−0.16, −0.07]	claim
W3	Conflict	−0.20 [−0.24, −0.16]	<0.001	−0.03 [−0.06, −0.01]	0.70 / 0.03	−0.16 [−0.21, −0.12]	partial

Table 24: All RQ3 rows for Granite-4.1-8B; layout and columns as in Table 21.

H Forced-Choice Readout

We also recorded a forced-choice readout (Appendix C). We report it descriptively only and do not use it as evidence for any claim, because at BF16 it almost never selects Conflict (Figure 10). Its BF16 Conflict recall is 0.01, 0.04, 0.02 and 0.14 on the four models (no CIs stored). Its BF16 accuracy is 0.54 [0.51, 0.56] (Qwen3-4B), 0.61 [0.58, 0.63] (Qwen3-8B), 0.42 [0.40, 0.45] (Llama-3.1-8B) and 0.63 [0.60, 0.65] (Granite-4.1-8B). On Llama-3.1-8B the readout mostly picks "A" (Answer recall 0.93). At W3, accuracy is 0.38 [0.36, 0.41], 0.47 [0.44, 0.49], 0.33 [0.31, 0.36] and 0.58 [0.55, 0.61].

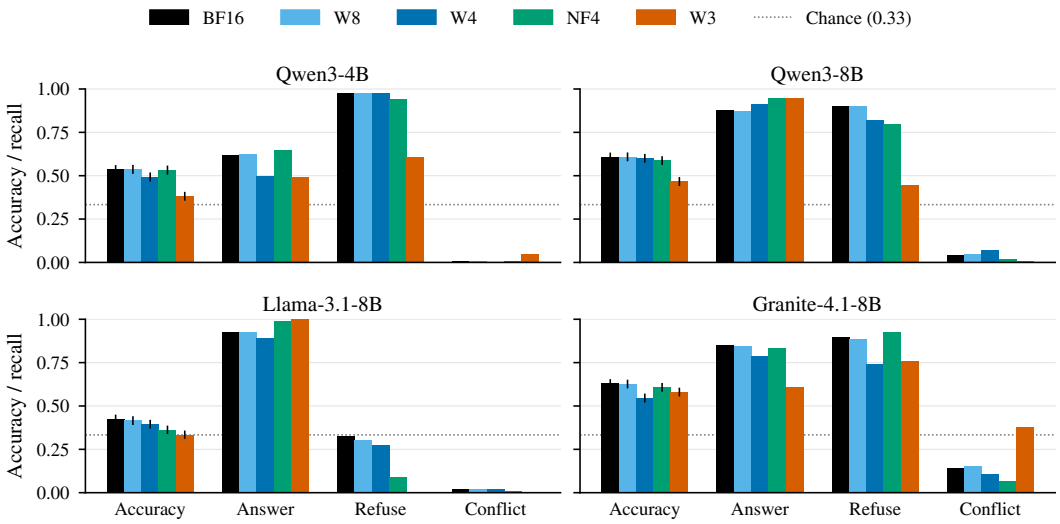

Figure 10: Forced-choice (A/B/C) readout per variant: accuracy (95% bootstrap CI) and per-class recall (no CIs stored). Shown for completeness only: this readout is descriptive and is not used as evidence for any claim, because at BF16 it almost never selects Conflict. Dotted line: chance (0.33).

I Reproducibility

Hardware and software. Each run uses one NVIDIA A10 GPU (24 GB; 23.68 GB reported) with driver 580.105.08, inside a Docker image based on CUDA 13.0. Two physical GPUs of the same type were used across runs. All 24 runs (four models × six variants) record identical software versions (Table 25).

Component	Version
Python	3.12.3
PyTorch	2.14.1+cu130 (CUDA 13.0)
Transformers	5.19.0
Accelerate	1.15.0
GPTQModel	7.6.0
bitsandbytes	0.50.2
scikit-learn	1.9.1
NumPy	2.2.6
SciPy	1.16.2
pandas	2.3.3
huggingface_hub	1.33.0
GPU	NVIDIA A10, 24 GB
NVIDIA driver	580.105.08
OS	Linux 6.8.0 x86_64, glibc 2.39

Table 25: Software and hardware, identical in all 24 runs (recorded by each run).

Hyperparameters and seeds. Table 26 lists the probe, analysis and judge settings. Seed 42 is used for sampling the GPTQ calibration documents and for all bootstrap resampling, including the direction null; data handling and probe fitting are deterministic. The random-label control uses shuffle seeds 0–4. All CIs use 10,000 bootstrap resamples.

Runtime. Extracting both features at every layer for all 7,188 instances took 45–103 minutes per model and variant (2,706–6,190 s). GPTQ quantization took 12–23 minutes per variant (750–1,384 s); NF4 quantizes at load time in 12–93 s. Probe training and all analyses run on the CPU.

Setting	Value
Probe	standard scaler; multinomial logistic regression, C=1, L-BFGS, at most 2,000 iterations; the fit is deterministic
Calibration	sigmoid, fitted on validation with the regression frozen; all test predictions use the calibrated probabilities
Layer selection	uncalibrated validation accuracy, per feature
Random-label control	best **H** and **M** layers, shuffled training labels, seeds 0–4, uncalibrated
RQ2 plateau	BF16 layers within 0.03 of the best validation accuracy
Re-scaled probe	scaler refit on unlabelled quantized training activations
Direction null	50 bootstrap refits of the BF16 probe
Ridge map	penalty $m \cdot \mathrm{tr}(Q^\top Q)/d$ with $m \in \{10^{-3}, \ldots, 10^{2}\}$, where Q holds the centred quantized training activations and d is their dimension; m is chosen by reconstruction error on validation activations
No-change tolerance	$\lvert\Delta\rvert \leq 0.02$ and CI covers 0
Bootstrap	10,000 resamples, percentile 95% CI, paired, seed 42
McNemar	exact below 25 discordant pairs, else χ^2 with continuity correction
Holm correction	one family per RQ and model; RQ1: 30 tests (5 variants $\times$ 3 classes $\times$ 2 features); RQ3: 30 tests (5 variants $\times$ 3 classes $\times$ 2 readouts)
LLM judge	temperature 0, seed 0, 100 output tokens
Perplexity	WikiText-2 test, non-overlapping 2,048-token windows

Table 26: Probe, analysis and judge settings.

Data. We use the benchmark file `instances.zip` of Huq et al. (2026) from their repository (`https://github.com/SyedHuq28/Knowing-Before-Answering-Decoding-Language-Models-for-Reliable-RAG`) at commit `2b338ce`, unchanged and with their split. Its SHA-256 is `7dc9d0d23d8518d23c189decd548e5207130fb4ec1df54fc421889ddf0b55d41`. Every run also records the SHA-256 of the train, validation and test files. GPTQ calibration uses C4 (Raffel et al., 2020); perplexity uses WikiText-2 (Merity et al., 2017).

Code and data availability. Our code is available at `https://github.com/kamil-hassan201/quantized-evidence-probes`. The repository also holds the analysis outputs behind our tables and figures and the small records of each run: environment, quantization settings and checks, perplexity, and probe and behaviour metrics. Activations, quantized checkpoints and generations are not included; they can be regenerated with the code. The repository includes the benchmark file of Huq et al. (2026), redistributed unchanged with the authors' permission; please cite their work when using it.

J Extended Limitations

This section extends the Limitations section of the main text.

Scope. We study four open-weight 4B–8B instruction-tuned models, the most one 24 GB GPU fits, on one English benchmark (Huq et al., 2026) with controlled, synthetic conflicts (entity substitution). Larger models, natural conflicts and other languages may differ.

Quantization. We quantize weights only; activation (e.g., Xiao et al., 2023) and KV-cache quantization, which alter the probed activations directly, are untested, as are AWQ (Lin et al., 2024) and extreme 2-bit methods (e.g., Chee et al., 2023). Our naive GPTQ W2 collapses (WikiText-2 perplexity 11–48$\times$ BF16) and is excluded, so we say nothing about usable 2-bit models. GPTQ uses a PyTorch backend that dequantizes weights to BF16, not deployment's fused kernels, whose numerics may differ slightly, and each variant comes from one quantization run with one calibration sample.

Behaviour measurement. Each model uses one prompt (KRE, selected on BF16 validation with a 128-token cap) and greedy decoding. The test generations use a 512-token cap, which

1.6% of the outputs still hit; we chose it after a first analysis with a 128-token cap, and the Qwen3-4B W3 Conflict claim of that analysis does not hold with the longer outputs (Section 4.3, Appendix G). The generation readout scores whether the output signals the evidence state, not whether the model abstains; if Refuse required abstaining, the two Refuse claims would not hold (Appendix G). We checked the LLM judge against human labels on only 24 outputs (Appendix D), and our behaviour numbers are not comparable with the prompting scores of Huq et al. (2026).

Analysis choices. The one-sided RQ3 probe rule (condition (ii), Section 3.6), the excess-loss measure, the partial category, the plateau average (tolerance 0.03) and the use of the LLM judge as primary parser were fixed after we had seen first results on one or two models; all models and variants were then analysed with these final choices. For **H** with the generation readout, the earlier two-sided rule gives the same claims. The 512-token cap was also chosen after the results with the 128-token cap were known (Appendix G). We added the rotation tests after the main RQ2 results, so they are exploratory; the ridge map, with many parameters relative to the 5,463 training items, is a flexible upper bound, not evidence for a mechanism. We also added the readings of the frozen probe without its calibrator and by AUC after the main RQ2 results, so they are exploratory too (Appendix F). RQ1 and RQ3 read each variant at its own best layer, while RQ2 averages over the BF16 plateau layers (diagnostics at the BF16-selected layer), so their native accuracies differ (at W3: drops of 0.02–0.07 and 0.03–0.04). NF4 versus W4 has no paired test, and each probe is fitted once on the fixed training split (the fit is deterministic), so CIs cover test-set sampling but not training-sample variance.

Comparability and probing. Qwen3-4B and Qwen3-8B differ in post-training as well as size, so their differences are not a pure size effect. Our BF16 Qwen3-4B probe reaches 0.92 $[0.91, 0.93]$, above the 0.89 of Huq et al. (2026), for a reason we could not resolve; numerical precision is unlikely (Appendix A). This offset does not affect our within-setup comparisons. Our "knows but fails to act" results are correlational (Section 5): we do not show that the model computes the evidence state and then ignores it.