WHOSE WORD SURVIVES COMPRESSION?

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training quantization can preserve answer accuracy while changing the computation that selects between retrieved context and parametric memory. We test causal transfer of source-arbitration interventions under GPTQ INT4 with JuICE, a dual-run attention-head intervention, on Llama-3.1-8B and Gemma-2-9B base. A question-aware matcher audit reveals that apparent transfer can be a same-entity wrap artifact rather than a transferred causal object. On Llama, the original-matcher (un-audited) context-K10 transfer ($+15$ percentage points (pp), 95% CI $[+8, +23]$) is one such wrap: the audited number is $+1$ $[+0, +3]$ (null), with 17 of 17 original INT4 flips naming the same context entity. The locked Llama parametric pair (clean K2) remains detected at n=100 on both precisions under gold-PAR-only counting (-8 $[-14, -3]$ on FP16; -7 $[-12, -2]$ on INT4), confirmed head-specific by a matched random-K2 control. On Gemma, identified context-suppression H^- remains detected versus a locked random-K control but is attenuated on INT4 (inclusive -19 $[-27, -11]$ vs. -1 $[-5, +2]$ on FP16; -11 $[-18, -4]$ vs. -1 $[-5, +3]$ on INT4); context-K10 enlarge is a small real effect ($+7$ $[+2, +12]$ / $+5$ $[+1, +9]$) but not INT4-specific, and the original-matcher (un-audited) raw peak L27_H6 is dead as a wrap artifact. Conversely, Gemma's clean parametric pair attenuates to an in-cell transfer null on INT4 despite a detected FP16 effect—the mirror image of Llama's result. A same-sign Δ is therefore not evidence of transfer without question-aware matching, entity confirmation, and a matched random-head control.

1 INTRODUCTION

When retrieved context contradicts a model's parametric memory, a small set of attention heads can be used, at test time, to steer whose word the model trusts (Li et al., 2025). This paper asks whether that residual-stream geometry is an FP16 object or a property of the model that transfers under compression: can apparent FP16→INT4 transfer be trusted from output labels alone? Accuracy can fall from forgetting, terse decoding, or a change in source arbitration; a transferred intervention distinguishes those cases. Our investigation proceeds in two distinct stages: Stage 1 establishes *that* arbitration behavior shifts under quantization, providing behavioral motivation on a third-party Instruct GPTQ checkpoint; Stage 2 asks *why* and *what circuit* is responsible by identifying and transferring locked attention-head interventions on self-quantized base checkpoints (Grattafiori et al., 2024; Rivière et al., 2024). Crucially, the two stages are not mechanistically linked—they evaluate different checkpoints and share no heads—serving respectively as behavioral motivation and causal transfer test.

The contribution is a causal, head-level transfer test under this GPTQ INT4 recipe (Frantar et al., 2023), paired with a question-aware matcher audit that re-evaluates the output labels. We demonstrate that behavioral efficacy is not, by itself, evidence that the same causal object transfers under compression: without a question-aware matcher, apparent transfer can be a same-entity wrap artifact rather than a true circuit.

Can apparent FP16→INT4 transfer be trusted? On Llama, the FP16-identified context K10 effect ($+8$ $[+2, +14]$) transfers onto INT4 as a large same-sign output ($+15$ $[+8, +23]$) under the original matcher, but the question-aware audit reveals this to be a same-entity wrap artifact. The audited INT4 effect is a null $+1$ $[+0, +3]$, with 17 of 17 original flips naming the same context entity in a role phrase. Native INT4 K10 leaves only a thin leftover, and reverse transfer is mostly a scorer artifact, showing that apparent behavioral transfer can mask a dead causal object.

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

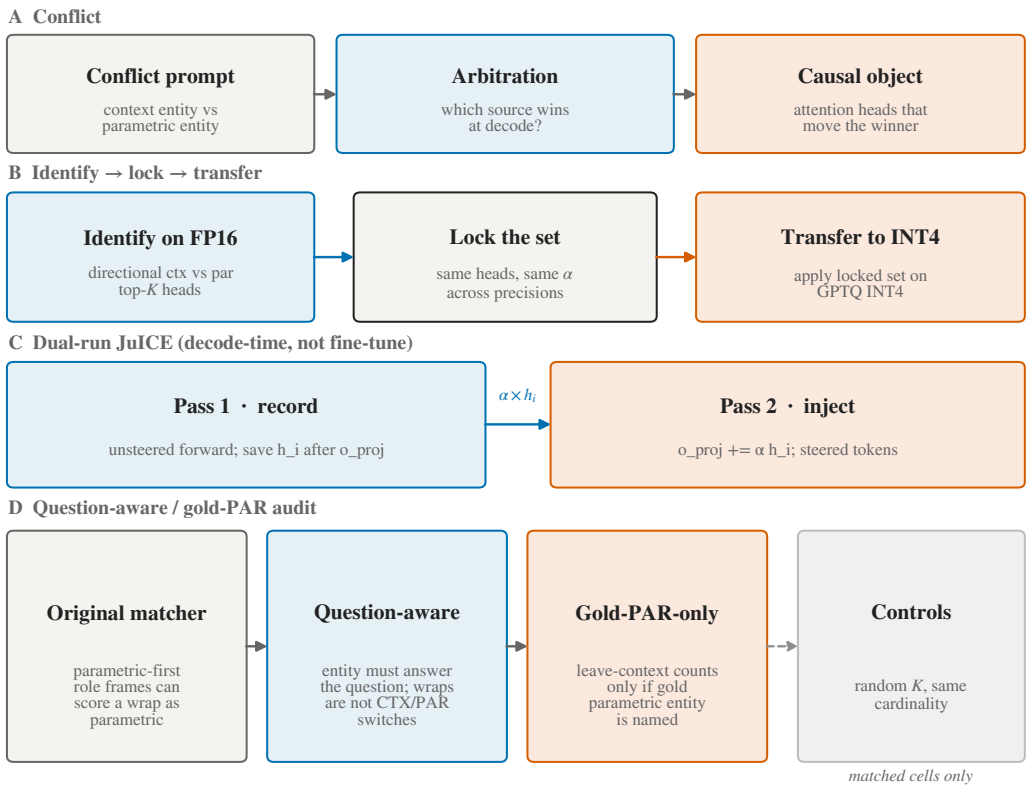


Figure 1: Pipeline. (A) A conflict prompt pits a context entity against a parametric entity. (B) Heads are identified on FP16, locked, and applied unchanged on GPTQ INT4. (C) Steering is dual-run JuICE: Pass 1 records h_i after o_{proj}; Pass 2 injects αh_i. (D) The original matcher can score a same-entity wrap as parametric; the question-aware matcher requires the named entity to answer the question; gold-PAR-only counts a leave-context move only if the gold parametric entity is named. Random-K controls (dashed arrow) are a parallel verification reported only for matched cells.

Does a locked intervention still elicit the intended entity? On Llama, the locked parametric pair (clean K2) remains detected at n=100 under conservative gold-PAR-only counting on both precisions (-8 $[-14, -3]$ on FP16; -7 $[-12, -2]$ on INT4), and a matched random-K2 control (null on both precisions) confirms specificity against generic intervention magnitude.

Is transfer uniform across models and directions? On Gemma, the original-matcher (un-audited) raw parametric peak (L27_H6) is dead as a wrap artifact (hard zero at n=50; $+0$ $[-3, +3]$ at n=100 INT4). However, two objects remain detected at n=100: identified context-suppression H^- versus a locked random-K control (though the effect is attenuated on INT4), and a small context-K10 enlarge ($+7$ $[+2, +12]$ on FP16; $+5$ $[+1, +9]$ on INT4) whose paired INT4 specificity contrast is null ($+2$ $[+0, +5]$). Gemma's own clean parametric pair (L31_H7, L39_H6), despite a comparable detected FP16 effect, fails to transfer (-2 $[-6, +2]$, null)—the reverse of Llama's outcome, showing that which direction of arbitration survives GPTQ is model-dependent rather than fixed.

The full picture of audited findings, missing controls, and evaluated cells is consolidated in Table 3.

2 Related Work

When retrieved context contradicts parametric memory, a language model must decide whose word to trust (Xie et al., 2024). JuICE (Li et al., 2025) is a test-time, training-free dual-run intervention on a small set of attention heads. The original JuICE models are base checkpoints, not Instruct, a distinction developed in Methods (§3.2). We do not revive a memory-head versus context-head dichotomy (Jin et al., 2024); the transfer test is whether a locked intervention remains the same

causal object after GPTQ. Prior compression work measures recall without a conflicting passage (Wang et al., 2026) or RAG accuracy without constructed conflicts (Yazan et al., 2024). Neither tests whether an FP16-identified arbitration intervention still acts on the same object after INT4. A drop in closed-book recall, or a stable RAG score, can occur without that intervention remaining detected. We double-filter on facts that remain answerable at both precisions and then transfer locked JuICE.

3 METHODS

3.1 STUDY SCOPE AND CHECKPOINTS

Stage 1 is behavioral (Llama-3.1-8B-Instruct); Stage 2 is identify-and-steer on Llama-3.1-8B base and Gemma-2-9B base (Rivière et al., 2024; Grattafiori et al., 2024). Table 1 lists every file. The Instruct INT4 artifact is a third-party community GPTQ, whereas the base INT4 artifacts are self-quantized; these are distinct files. Gemma has no Stage 1 Instruct GPTQ. Throughout, n denotes the number of evaluation items, while N_{CALIB} denotes the GPTQ calibration-set size.

Table 1: Checkpoints used in this paper. Rows are parallel across the two model families; "not tested" marks the absent Gemma Stage 1 behavioral condition.

Role	Model	Precision	Checkpoint
Stage 1 behavioral	Llama-3.1-8B-**Instruct**	FP16	`meta-llama/Llama-3.1-8B-Instruct`
Stage 1 behavioral	Llama-3.1-8B-**Instruct**	GPTQ-INT4	`hugging-quants/Meta-Llama-3.1-8B-Instruct-GPTQ-INT4` (third-party)
Stage 2 ID + steer	Llama-3.1-8B **base**	FP16	`meta-llama/Llama-3.1-8B`
Stage 2 transfer (headline)	Llama-3.1-8B **base**	GPTQ-INT4	self-quantized `llama31_8b_base_gptq_int4_ncalib128` (N_{CALIB}=128)
Stage 2 robustness / native ID / recovery	Llama-3.1-8B **base**	GPTQ-INT4	self-quantized `llama31_8b_base_gptq_int4` (N_{CALIB}=64; retained for comparison)
Stage 1 behavioral	Gemma-2-9B-**IT**	not tested	—
Stage 2 ID + steer	Gemma-2-9B **base**	FP16	`google/gemma-2-9b`
Stage 2 transfer (headline)	Gemma-2-9B **base**	GPTQ-INT4	self-quantized `gemma2_9b_base_gptq_int4_ncalib128` (N_{CALIB}=128)

3.2 INSTRUCT VERSUS BASE MECHANISTIC TARGET

JuICE was developed on base models (Li et al., 2025). We use $\texttt{dir_par} = \texttt{par_pos} - \texttt{ctx_pos}$ as the directional parametric score. On Instruct with a chat template, parametric ID and steering fail (Table 2): this score is suppressed by $\sim 1130\times$. This is a model-plus-prompt confound, so all INT4 transfer cells use base-identified heads.

FP16 base must steer before transfer. Table 6 reports question-aware context K10 at $+4$ $[+0, +8]$ (null) and clean K2 at gold-PAR-only -8 $[-14, -3]$ (inclusive -15 $[-22, -7]$); those locked heads are then applied to INT4 without re-identification.

3.3 GPTQ RECIPE AND EVALUATION

Stage 2 GPTQ uses gptqmodel 2.2, transformers 4.51.3, 4-bit quantization, group size 128, `sym=True`, and `desc_act=True`. C4 is truncated to 512 tokens, with headline N_{CALIB}=128; Llama N_{CALIB}=64 is a robustness check only. Gemma uses the same recipe with `GetHeadsGemma2Attention` (GQA $16/8$, `head_dim`=256, logit softcap 50), eager attention, gptqmodel `BACKEND.TORCH`, and greedy T=0. Data is ConflictQA popular (Xie et al., 2024). Stage 1 double-filters n=250 to n=164 items correct in closed-book mode at both precisions. Stage 2 uses the seed-42 n=100 IDs for Llama headline and diagnostic cells and for Gemma headlines; Gemma n=50 is a robustness check. Reverse-transfer cells remain on the original n=50 IDs. Head identification uses $|\mathcal{D}|$=4 shared calibration IDs, three-type JuICE completions, and `FAST_POS_ALPHAS`=$(1, 3, 5)$ and `FAST_NEG_ALPHAS`=$(-1, -2)$.

3.4 JUICE INTERVENTION AND SCORING

Steering is applied at decode time by dual-run JuICE (Figure 1), not by fine-tuning.

Table 2: Instruct (chat template) versus base (plain JuICE prompts). Parametric directional ID is suppressed on Instruct. In-precision steering (Cell 3: FP16 ID → FP16 apply) numbers are from the original matcher; QA FP16 context K10 is $+4$ $[+0, +8]$.

Metric (unit)	Instruct (chat)	Base (plain JuICE prompts)
Raw enlarge Jaccard (unitless)	$\mathbf{0.54}$	$\mathbf{0.053}$
Peak raw `scores_pos` (score units; ctx / par)	0.0080 / 0.0077	**0.036 / 0.058**
Best `dir_par` (score units)	$\mathbf{2.68 \times 10^{-4}}$	$\mathbf{0.304}$ ($\sim$1130×)
Directional top-10 Jaccard (unitless)	0	0
Parametric Cell-3 at $\alpha{+}{=}1$ (pp)	0 / -2 ($K{=}2$ / $K{=}3$)	$-\mathbf{6}$ to $-\mathbf{8}$
Context Cell-3 at $\alpha{+}{=}1$ (pp)	$+6$ to $+8$	$+6$ to $+10$

Pass 1 records per-head outputs after `o_proj`; pass 2 adds α times the saved vector. We use context sets with $K{=}10$, parametric sets with clean/noisy $K{=}2$ and $K{=}3$, and $K{=}1$ on Gemma's raw parametric peak. Default transfer is enlarge-only, $\alpha{+}{=}1$, $\alpha{-}{=}0$; suppression uses $\alpha{+}{=}0$, $\alpha{-}{=}1$. Evaluation uses greedy `juice_coherent` decoding with `max_new_tokens`=16. The output labels are `parametric`, `context`, `rejects_context`, and `other`; we define `follows_context` as the event that the label equals `context`. Here, H^- denotes a suppression intervention with $\alpha_+{=}0$ and $\alpha_-{=}1$, and `exam-meta` denotes the multiple-choice-format text sometimes appended to a context-entity answer. A letter heuristic stayed at 100% while $\alpha{=}3/5$ `other` hit 59–66% on the $n{=}100$ question-aware grid (Table 16), so we report label counts. Scorer brevity and parametric-first behavior are documented in Appendix A.

3.5 Question-aware matching and gold-PAR-only counting

Stage 1 already passed `question=` into `label_conflict_response`. Stage 2 originally did not. Without the question, the matcher checks parametric gold first, including role-phrase aliases. Completions that assert the *context* entity in a role frame ("Edina is the capital of Namibia") were therefore labeled parametric. Passing the gold question strips echoed question tokens and relabels those wraps as context.

We report two Δ conventions after that audit. Inclusive Δ counts every baseline-context item that leaves the context label under steering. Gold-PAR-only Δ counts a leave-context move as success only when the steered completion mentions the gold parametric entity; leave-other and wrap items remain suppressed. Llama clean K2 uses gold-PAR-only as the conservative headline (-8 / -7) because the goal of a parametric-enlarge intervention is eliciting the gold parametric entity rather than mere disorientation; inclusive Δ is an upper bound (-15 / -12 in Table 8). Conversely, Gemma identified H^- is a context-suppression intervention whose goal is displacing retrieved context, making inclusive Δ the natural headline (-19 / -11); under gold-PAR-only counting it is -15 / -10 (Table 13). Crucially, the alternative convention also excludes zero on both precisions for both models, meaning the choice of convention does not change any detection call, only which number is foregrounded. Context-enlarge cells use inclusive Δ plus a same-entity wrap count. All Stage 2 Δ and CIs in Results are question-aware unless an original-matcher number is marked as such.

Paired bootstrap, 1000 draws, seed 42, percentile 95% CI, throughout. Stage 1 resamples 164 items. Llama Stage 2 Δ is steered minus baseline `follows_context` on 100 eval IDs; Gemma headlines use 100, with $n{=}50$ as robustness. Reverse transfer uses 50. "Excludes 0" is the detection rule. CIs are whole percentage points; gold-PAR-only n=100 FP16 K2 reports $[-14, -3]$ (bootstrap hi -2.975). Random-K excludes the FP16 identified set. Gemma suppress specificity is $\Delta_{\text{id}} - \Delta_{\text{random}}$; DiD is $\text{spec}_{INT4} - \text{spec}_{FP16}$. We do not infer specificity from overlapping cell CIs. Because we report paired-bootstrap 95% CIs across conditions, models, precisions, and matchers without a formal multiple-comparisons correction, the pre-registered gate and subsequent entity-level audits serve as the primary safeguards against false positives.

The pre-registered Llama gate required same-sign transfer of context K10 and parametric clean K2 at coherence ≥ 0.90 under the original matcher; that gate passed. The question-aware audit later nulls context K10 transfer and leaves clean K2 as the Llama headline. Gemma headlines are identified H^- versus random K and context K10 enlarge at $n{=}100$; L27_H6 is a wrap artifact. The scope of unrun cells is consolidated in Limitations.

Table 3: Synthesis of locked interventions. Original/inclusive is the original matcher or inclusive question-aware Δ as marked. Audited effect is the headline estimand after wrap/gold-PAR audit. Unresolved cells are separated into Table 10. All CIs are percentile 95%, 1000 paired bootstrap draws, seed 42.

Model	Intervention	Original / inclusive	Audited effect	Control / check	Verdict
Llama	Ctx K10 enlarge	orig. $+8$ $[+2, +14]$ / $+15$ $[+8, +23]$	QA $+4$ $[+0, +8]$ / $+1$ $[+0, +3]$ (both null)	random K10 $+0$ $[-4, +4]$ / -1 $[-4, +2]$; 17/17 INT4 wraps	wrap artifact
Llama	Clean K2 (L20_H14+L26_H15)	incl. QA -15 $[-22, -7]$ / -12 $[-19, -6]$; orig. -8 $[-14, -3]$ / -4 $[-8, -1]$	gold-PAR -8 $[-14, -3]$ / -7 $[-12, -2]$	matched random K2 (L20_H24+L3_H19), n=100 N_{CALIB}=128: -1 $[-3, +0]$ FP16 / -1 $[-3, +0]$ INT4 (both null)	detected; head-specific
Gemma	Identified H^-	incl. QA -19 $[-27, -11]$ / -11 $[-18, -4]$	gold-PAR -15 $[-23, -8]$ / -10 $[-17, -4]$	random -1 $[-5, +2]$ / -1 $[-5, +3]$; spec. -18 $[-27, -10]$ / -10 $[-18, -2]$; DiD $+8$ $[+1, +16]$	detected; attenuated
Gemma	Ctx K10 enlarge	n=50 QA $+4$ $[+0, +10]$ (null); INT4 n=50 **not run**	n=100 QA $+7$ $[+2, +12]$ / $+5$ $[+1, +9]$	random $+0$ $[+0, +0]$ / $+3$ $[+0, +7]$; spec. $+7$ $[+2, +12]$ / $+2$ $[+0, +5]$ (null); DiD -5 $[-9, -1]$	detected; not INT4-specific
Gemma	Clean K2 (L31_H7+L39_H6)	incl. QA -6 $[-12, -1]$ / -2 $[-6, +2]$	gold-PAR -4 $[-9, +1]$ / -2 $[-6, +2]$ (both null)	matched random K2 (L41_H11+L7_H3): $+0$ $[+0, +0]$ / $+2$ $[+0, +5]$; spec. -6 / -4 $[-8, -1]$	transfer null; not detected
Gemma	L27_H6	orig. n=50 -12 $[-22, -4]$ / -8 $[-16, -2]$	QA hard 0 (n=50 both; n=100 FP16); INT4 n=100 $+0$ $[-3, +3]$	INT4 random K1 (L40_H15) $+0$ $[-6, +6]$ (n=50) / $+1$ $[-2, +5]$ (n=100); FP16 random K1 **not run**	wrap artifact

Table 4: Stage 1 memorization rate. MR* is the primary behavioral estimand (shared 164).

Precision	Closed-book	MR (unfiltered)	MR* (shared 164)
FP16	182/250 (72.8%)	92/250 (36.8%)	**76/164 (46.3%)**
INT4	175/250 (70.0%)	48/250 (19.2%)	**37/164 (22.6%)**

4 RESULTS

4.1 STAGE 1: BEHAVIORAL ARBITRATION ON INSTRUCT

Stage 1 asks, before any head is identified: when a fact is preserved under INT4, does conflict arbitration still look like FP16? We compare FP16 Instruct to a third-party GPTQ-INT4 Instruct checkpoint on 250 ConflictQA popular items, then restrict to the 164 items closed-book correct at both precisions.

MR* on the shared 164 is the primary estimand (Table 4). It falls by 23.8 pp $[-31.7, -17.1]$ after quantization, excluding zero. Closed-book performance is nearly stable (-7 items), whereas unfiltered MR falls by -17.6 pp, so the arbitration shift does not disappear under the double filter. Cluster-robust logit AME on `follows_context` is $+10.3$ pp $[+3.1, +17.5]$ (Table 5); corrected label counts in Appendix Tables 19–20 show a parametric collapse and a rise in context. Over half of the parametric-recall drop (39 items) shifts into `rejects_context` ($+9$) and `other` ($+13$) rather than context ($+17$; Tables 19–20), so this collapse is only partially an arbitration shift toward context. A first-pass substring scorer inflated the AME to $+20.6$ pp by counting denials as context, so it is not used here. Llama-3.2-1B-Instruct was a null screen and is not mixed into the 8B claim.

4.2 CAN APPARENT FP16→INT4 TRANSFER BE TRUSTED?

On Llama-3.1-8B base, apparent forward transfer of context K10 is a wrap artifact. Under the original matcher, FP16 context K10 was $+8$ $[+2, +14]$ and its INT4 N_{CALIB}=128 transfer was $+15$ $[+8, +23]$. After the question-aware audit, both cells are null: $+4$ $[+0, +8]$ on FP16 and $+1$ $[+0, +3]$ on INT4 (Table 7, Figure 2A). A random-K10 control confirms the null ($+0$ $[-4, +4]$ on FP16; -1 $[-4, +2]$ on INT4). Entity audit of the 17 original INT4 transfer flips finds 17/17 same-entity wraps: the unsteered baseline already generated a full-sentence role-frame naming the context

Table 5: Stage 1 `follows_context` AME on the shared 164 items. The cluster-robust logit row is the primary estimand.

Estimand	Point	95% CI	excl. 0?
Cluster-robust logit AME	$+\mathbf{10.3}$ **pp**	$[+\mathbf{3.1}, +\mathbf{17.5}]$	yes
Raw ΔP(`context`)	$+10.4$ pp	$(52 \to 69)/164$	—
Item-averaged AME	$+9.8$ pp	—	—
Mixed VB (Variational Bayes) AME	$+16.1$ pp	$[+5.9, +26.2]$	yes

Table 6: FP16 base steering on the 100 Llama eval items, question-aware matcher. Gold-PAR-only Δ for clean K2 counts only CTX→gold-PAR-name flips; inclusive Δ is an upper bound. Letter heuristic coherence is 100% in every row (not shown).

Condition	Heads	fc	Δ [95% CI]	P/C/O
baseline	—	67%	—	23/67/10
context K10	directional ctx top-10	71%	$+4$ $[+0, +8]$ (null)	17/71/12
parametric clean K2 (incl.)	**L20_H14, L26_H15**	52%	-15 $[-22, -7]$	31/52/17
parametric clean K2 (gold-PAR)	same pair	—	$\mathbf{-8}$ $[\mathbf{-14}, \mathbf{-3}]$	9 gold-PAR / 7 leave*
random context K10	seed 42	67%	$+0$ $[-4, +4]$	19/67/14

*Row 4 reports gold-PAR / leave-context transition counts rather than the standard P/C/O format.

entity (mislabeled parametric by the original matcher), while steering shortened the completion to a bare entity name that the matcher then correctly labeled context—creating a spurious baseline-to-steered positive gain. Native INT4 K10 applied in-precision remains a thin leftover ($+4$ $[+1, +8]$); reverse transfer of those heads onto FP16 is mostly a scorer artifact (see Appendix A for details on both). Behavioral transfer alone is therefore not sufficient evidence of causal-object transfer when a scorer can be satisfied by role-wrapped outputs.

4.3 Does a locked intervention still elicit the intended entity?

Llama clean K2 (L20_H14+L26_H15) is detected at n=100 when success is restricted to gold parametric names (Figure 2B). Under gold-PAR-only counting, the locked FP16 pair is -8 $[-14, -3]$, and its INT4 transfer is -7 $[-12, -2]$. Counting every leave-context move yields inclusive upper bounds of -15 $[-22, -7]$ on FP16 and -12 $[-19, -6]$ on INT4. We treat n=100 as the headline because the n=50 FP16 gold-PAR-only cell includes 0. A matched random-K2 control (L20_H24, L3_H19; n=100, N_{CALIB}=128) confirms this specificity directly: the random pair is null on both precisions (-1 $[-3, +0]$ gold-PAR-only and inclusive), yielding a specificity contrast of -6 to -14 pp that rules out generic intervention magnitude. Other parametric proxies and variations—including noisy K2/K3, native INT4 clean K2, the raw parametric peak L1_H4, suppression H^-, and an α-grid rescale—fail to recover a stronger verified parametric object without high-alpha scorer collapse, as detailed in Appendix B. A locked intervention can therefore continue to elicit the intended entity after quantization, provided it is verified at the entity level. Mechanistically, both L20_H14 and L26_H15 have negative or near-zero raw parametric scores (`par_pos` ≤ 0 on both precisions; Table 9) and rank highly under directional scoring primarily through strong context suppression (`ctx_pos` ≈ -0.2 to -0.3); the surviving effect is better characterized as disinhibition of baseline parametric memory via context suppression than direct parametric amplification.

4.4 Is transfer uniform across models and directions?

On Gemma-2-9B base, identified context-suppression H^- is detected but attenuated on INT4. At n=100, inclusive Δ is -19 $[-27, -11]$ on FP16 versus random -1 $[-5, +2]$, and -11 $[-18, -4]$ on INT4 versus random -1 $[-5, +3]$ (Figure 4A). Specificity excludes 0 at both precisions (-18 $[-27, -10]$ and -10 $[-18, -2]$), but the DiD is $+8$ $[+1, +16]$, confirming FP16 specificity is larger. Context K10 enlarge is a small real effect (FP16 $+7$ $[+2, +12]$; INT4 $+5$ $[+1, +9]$) but is not INT4-specific: while FP16 specificity excludes zero ($+7$ $[+2, +12]$), the paired INT4 specificity contrast is $+2$ $[+0, +5]$ (null, includes 0) and the DiD is -5 $[-9, -1]$, confirming specificity is lost on INT4 (Figure 4B). Gemma context K10 enlarge was run on INT4 at n=100, and on FP16 at both n=100 and n=50; the INT4 transfer of this condition at n=50 was not run. The original-matcher (un-

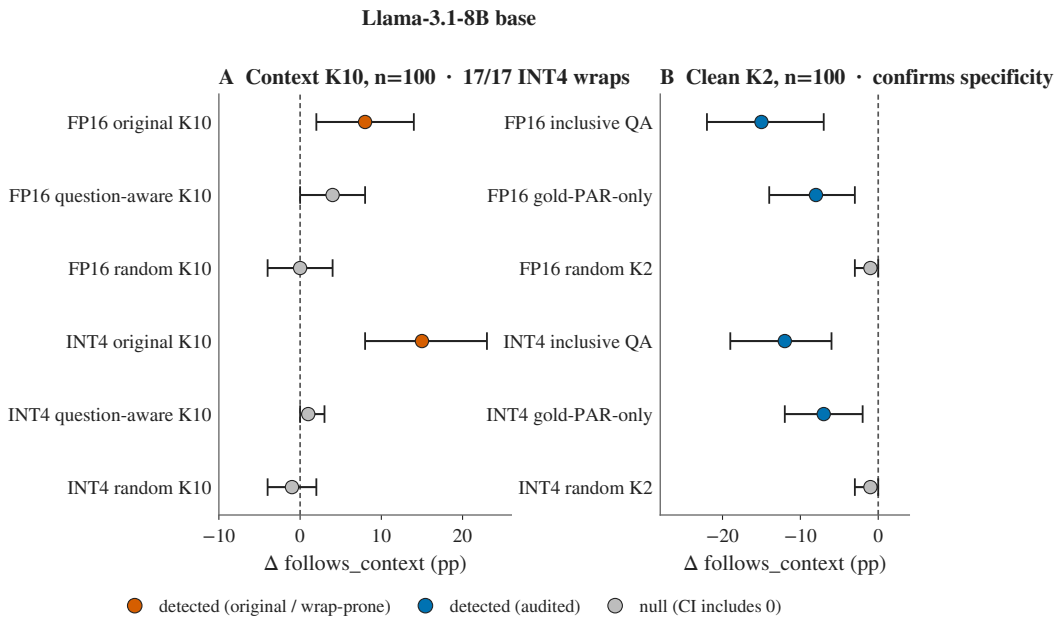


Figure 2: Llama-3.1-8B, n=100. (A) Original-matcher context K10 is $+8$ $[+2, +14]$ on FP16 and $+15$ $[+8, +23]$ on INT4 N_{CALIB}=128; question-aware cells are $+4$ $[+0, +8]$ and $+1$ $[+0, +3]$ (both null). Random K10 is $+0$ $[-4, +4]$ / -1 $[-4, +2]$. All 17 original INT4 matcher flips are same-entity wraps. (B) Locked clean K2 (L20_H14+L26_H15): gold-PAR-only -8 $[-14, -3]$ / -7 $[-12, -2]$; inclusive QA -15 $[-22, -7]$ / -12 $[-19, -6]$. Matched random-K2 (L20_H24+L3_H19, n=100, N_{CALIB}=128): -1 $[-3, +0]$ both precisions, null — confirms specificity.

Table 7: Forward transfer of FP16 context K10 onto canonical INT4 (N_{CALIB}=128), question-aware matcher. This cell is a wrap artifact (see §4.2); original-matcher INT4 K10 was $+15$ $[+8, +23]$.

Condition	INT4 fc	INT4 Δ [95% CI]	excl. 0?	FP16 Δ [CI]	P/C/O
baseline	74%	—	—	—	11/74/15
context K10	75%	$+1$ $[+0, +3]$	no	$+4$ $[+0, +8]$	9/75/16
random context K10	73%	-1 $[-4, +2]$	no	$+0$ $[-4, +4]$	10/73/17
native INT4 K10	78%	$+4$ $[+1, +8]$	yes	—	8/78/14

FP16 baseline fc is 67% (P/C/O 23/67/10; Table 6).

audited) raw parametric peak, L27_H6, is dead as a wrap artifact: question-aware Δ is a hard zero at n=50 on both precisions, and $+0$ $[-3, +3]$ at n=100 INT4 (Figure 4C). The original L27_H6 at n=50 was compared against an INT4 random K1 control, but the corresponding FP16 random K1 control was not run at n=50. Finally, Gemma clean-K2 enlarge (L31_H7, L39_H6) transfers as an in-cell null: at n=100, INT4 inclusive Δ is -2 $[-6, +2]$ (gold-PAR-only -2 $[-6, +2]$, both spanning zero), attenuating from FP16 (inclusive -6 $[-12, -1]$; gold-PAR -4 $[-9, +1]$). The matched random-K2 control (L41_H11, L7_H3) is a hard zero on FP16 ($+0$ $[+0, +0]$) and $+2$ $[+0, +5]$ on INT4. Although the paired INT4 specificity contrast excludes zero (-4 $[-8, -1]$), this is an artifact of the random control drifting toward context ($+2$), not evidence that the identified pair's -2 effect is distinguishable from noise; we do not claim specificity is confirmed for an in-cell null.

In both families, directional re-ranking exhibits a blind spot. That score (`par_pos − ctx_pos`) promotes heads that strongly suppress context tokens over heads that actually boost parametric tokens (Figure 3, Table 9). On Llama, L1_H4 is the raw parametric peak (`par_pos`$= +0.0577$) but ranks only #29 under directional ranking. Both Llama's clean parametric pair (L20_H14, L26_H15) and Gemma's (L31_H7, L39_H6) consist of context suppressors, while L27_H6 is Gemma's raw positive peak.

Table 8: Llama clean-K2 expansion (question-aware matcher). Gold-PAR-only Δ is the conservative headline; inclusive Δ is an upper bound (1000 draws, seed 42).

Condition	n=50 gold-PAR	n=100 gold-PAR	n=100 incl.	excl. 0 at n=100?
FP16 clean K2	-6 $[-14, +0]$	$\mathbf{-8}$ $[\mathbf{-14}, \mathbf{-3}]$	-15 $[-22, -7]$	yes
INT4 clean K2 transfer	-8 $[-16, -2]$	$\mathbf{-7}$ $[\mathbf{-12}, \mathbf{-2}]$	-12 $[-19, -6]$	yes
FP16 random K2 control	—	-1 $[-3, +0]$	-1 $[-3, +0]$	no
INT4 random K2 control	—	-1 $[-3, +0]$	-1 $[-3, +0]$	no

Random control heads: L20_H24, L3_H19 (seed 42, excluding all identified context-K10 and parametric-K2 heads); n=100, N_{CALIB}=128, identical steering protocol (α_+=1, α_-=0) and eval set as the headline pair.

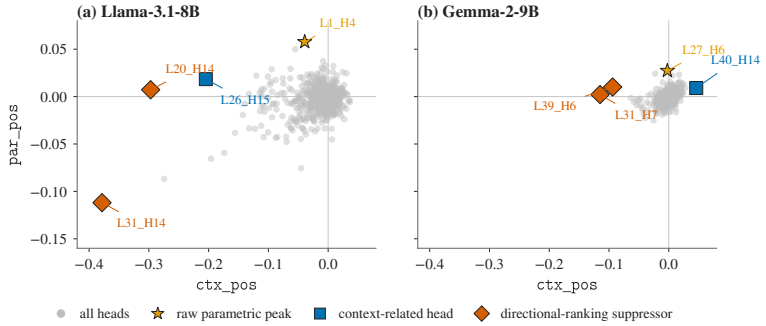


Figure 3: FP16 identification scatter for Llama-3.1-8B and Gemma-2-9B on a shared axis scale. Light-gray points are all heads. Gold stars mark raw parametric peaks, blue-teal squares context-related heads, and coral diamonds directional suppressors.

Under this GPTQ INT4 recipe, transfer is therefore intervention-, direction-, model-, and evaluation-dependent; the audited findings and controls are synthesized in Table 3.

5 Discussion

This evaluation distinguishes between a behavioral effect on the scorer, an audited causal effect, and structural circuit preservation. Our evidence supports the presence of audited causal effects for the surviving interventions, not structural circuit preservation. The two families reject a category-level account of transfer: the evidence supports neither "context survives compression" nor "parametric memory breaks." Rather, whether a particular locked intervention remains detected after GPTQ is model- and direction-dependent.

The cross-family regularity is methodological: a same-sign Δ is not evidence of transfer without a question-aware matcher, entity-level confirmation, and a matched random-head control. Llama context K10 and Gemma L27_H6 both produced large same-sign outputs that an entity-level audit revealed to be role-phrase wrap artifacts. Llama clean K2 remains detected after the entity audit and is confirmed head-specific by a matched random-K2 control (null on both precisions), ruling out generic intervention magnitude. Directional ranking alone does not establish a source-specific circuit without these checks. Furthermore, calibration size ($|\mathcal{D}|$=4) is a two-edged result: while observed head overlap sitting inside the INT4 noise floor (context K10 J=0.33 vs. noise mean 0.27 $[0.05, 0.82]$; noisy parametric K2 J=1.0 vs. 0.95) supports that head identity is fragile across compression, it also means locked heads identified on FP16 may partly reflect noisy sample draws rather than stable circuits—though this stability holds only for the noisy parametric filter; native re-identification of the clean pair itself yields disjoint heads (J=0.00), so the headline clean-K2 finding should not be read as structurally stable under re-identification.

Crucially, closing the Gemma clean-K2 transfer cell reveals a striking cross-family asymmetry in parametric transfer. On Llama-3.1-8B, the locked clean parametric pair robustly survives INT4 transfer under conservative gold-PAR counting (-8 $[-14, -3]$ FP16 vs. -7 $[-12, -2]$ INT4; matched random control null at -1 $[-3, +0]$), while context K10 transfer collapses as a wrap artifact. On Gemma-2-9B, the opposite pattern holds: raw parametric enlarge (L27_H6) is an audited wrap artifact, and clean parametric transfer attenuates to an in-cell null (-2 $[-6, +2]$), whereas context suppression (H^-: -19 / -11) and context enlarge (K10: $+7$ / $+5$) survive. Quantization transfer is thus doubly selective rather than uniformly preserving across model architectures.

Table 9: Directional re-ranking vs. raw parametric enlarge. L1_H4 is the raw peak on both precisions and is ranked #29 / #14 by `dir_par`.

Head	Role	FP16 `par_pos` / `ctx_pos` / `dir_par` (rank)	INT4 `par_pos` / `ctx_pos` / `dir_par` (rank)
L1_H4	raw parametric enlarge **#1** on both precisions	$+\mathbf{0.0577}$ / -0.0392 / $+0.0969$ (**#29**)	$+\mathbf{0.0578}$ / -0.0581 / $+0.1159$ (**#14**)
L20_H14	directional #1; in clean + noisy K2	$+0.0072$ / -0.297 / $+\mathbf{0.304}$ (**#1**)	$-\mathbf{0.0436}$ / -0.301 / $+0.257$ (**#1**)
L31_H14	noisy K2 (ctx-suppression)	-0.112 / -0.378 / $+0.267$ (#2)	-0.104 / -0.360 / $+0.256$ (#2)
L26_H15	clean K2	$+0.0184$ / -0.205 / $+0.223$ (#3)	-0.0019 / -0.185 / $+0.183$ (#4)

6 Conclusion

If a source-arbitration intervention is computed in FP16 residual-stream geometry, GPTQ INT4 does not apply a uniform tax to that object. However, these findings evaluate this specific GPTQ INT4 recipe (4-bit, group size 128), not quantization in general. Under this recipe, Llama clean K2 is detected at n=100 under gold-PAR-only counting, while Gemma clean K2 attenuates to a transfer null (-2 $[-6, +2]$). Gemma identified H^- versus random K is detected but attenuated on INT4. Gemma context-K10 enlarge is a small real effect, but its paired INT4 specificity contrast is null ($+2$ $[+0, +5]$). Both Llama context K10 and Gemma L27_H6 fail to transfer as wrap artifacts. This establishes a cross-family asymmetry: Llama preserves parametric steering while corrupting context arbitration, whereas Gemma preserves context suppression while attenuating parametric transfer to an in-cell null. Which intervention remains detected is therefore model-, direction-, and matcher-specific, not determined by source category.

7 Limitations

The following limitations define the scope of the claims.

1. Calibration size. Head identification uses only four calibration items. At this sample size, head-set drift is not identifiable: observed FP16-versus-INT4 overlap sits entirely within the INT4-versus-INT4 Jaccard noise floor for every specification we computed. We also do not evaluate sensitivity to the calibration seed.

2. Audit conventions and missing controls. The Llama clean-K2 headline uses conservative gold-PAR-only counting, while the Gemma H^- headline uses inclusive Δ. The matcher audit is a post-hoc rescore of saved completions, not a rewrite of JSON labels; counts include both automatic matches and human-checked entity audits. Several cells and controls were not run, explicitly limiting our claims: (A) Llama context K5 was not expanded to n=100; (B) Gemma context K10 enlarge INT4 transfer at n=50 was not run; (C) the corresponding FP16 random K1 control for Gemma L27_H6 was not run at n=50 (consolidated in Table 10).

3. GPTQ-only scope. All INT4 results use GPTQ at 4-bit, group size 128. We did not test AWQ, bitsandbytes, or other recipes; claims are specific to this GPTQ recipe.

4. Checkpoint and stage scope. Stage 1 and Stage 2 evaluate distinct artifacts: third-party Instruct GPTQ versus self-quantized base checkpoints. Stage 1 includes a 1B null screen, and the AME has a post-selection caveat. Stage 2 is base-only; Gemma Instruct MR*, native INT4 re-identification, and reverse transfer were not run and are not inferred from Llama. Steered deltas are paired against their session baseline, as unsteered behavior shifts across calibration checkpoints.

Table 10: Unresolved cells: missing-data conditions omitted from the headline synthesis.

Model	Intervention	Status	Reason / Details
Llama	Ctx K5	**not expanded**	n=50 QA $+8$ $[+2, +16]$ / $+2$ $[+0, +8]$ (N64); random $+2$ $[+0, +6]$; **not expanded** to n=100
Gemma	Ctx K10 enlarge (n=50)	**not run**	INT4 transfer not run at n=50; n=100 INT4 exists and is the headline ($+5$ $[+1, +9]$)
Gemma	FP16 random K1 (n=50)	**not run**	Control for L27_H6 at n=50 not run; only INT4 random K1 (L40_H15) exists

REPRODUCIBILITY STATEMENT

(This section is recommended and does not count toward the page limit.)

Paired bootstrap is the detection protocol throughout. Stage 1 resamples the 164 double-filtered items; Llama Stage 2 resamples 100 eval items, while Gemma headlines resample 100 and Gemma robustness resamples 50. Both use 1000 draws and a percentile 95% CI, with seed 42 throughout (random-K additionally uses seeds 42–45). Decode is greedy at $T{=}0$. Head-identification calibration IDs, the eval split, locked head sets, and the GPTQ recipe ($\mathrm{N}_{\mathrm{CALIB}}$, group size, `sym`, `desc_act`) are specified in §3. Question-aware rescoring and gold-PAR-only counting are specified in §3.5. Code and data will be released in anonymized form with the submission as anonymous supplementary material. This statement does not itself contain the reproduction recipe; the recipe is §3–§4.

NOTATION

n	Number of evaluation items; Llama Stage 2 uses $n{=}100$; Gemma headlines use $n{=}100$ with $n{=}50$ robustness; reverse transfer uses $n{=}50$.
N_{CALIB}	GPTQ calibration-set size; headline 128, robustness check 64.
K	Number of intervened heads; $K{=}10$ for context, $K{=}2$ for parametric pairs, and $K{=}1$ for Gemma's raw peak.
H^-	Context-suppression intervention with $\alpha_+{=}0$ and $\alpha_-{=}1$.
fc	`follows_context`, the fraction labeled `context`.
`ctx_pos`, `par_pos`, `dir_par`	Context score, parametric score, and `par_pos` − `ctx_pos` directional score.
`exam-meta`	Multiple-choice-format text appended to some context-entity completions; treated as a scorer artifact when diagnosed.
gold-PAR-only	Conservative Δ: CTX→leave counts as success only if the steered text mentions the gold parametric entity.

AI USE DISCLOSURE

(This section is **required** and does not count toward the page limit.)

Cursor-based generative AI tools were used during implementation to draft and refactor analysis code, debug execution and LaTeX issues, generate and revise figure layouts, inspect rendered outputs, and edit manuscript language. The authors selected the experiments and evaluation protocol, ran and inspected the analyses, checked every reported result against saved outputs, and take responsibility for the methods, claims, interpretations, and conclusions.

REFERENCES

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations*, 2023.

Aaron Grattafiori et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Tug-of-war between knowledge: Exploring and resolving knowledge conflicts in retrieval-augmented language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 16867–16878, 2024.

Gaotang Li, Yuzhong Chen, and Hanghang Tong. Taming knowledge conflicts in language models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 34074–34104. PMLR, 13–19 Jul 2025.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Qianli Wang, Mingyang Wang, Nils Feldhus, Simon Ostermann, Yuan Cao, Hinrich Schuetze, Sebastian Moeller, and Vera Schmitt. Through a compressed lens: Investigating the impact of quantization on factual knowledge recall. In *Proceedings of the 6th Workshop on Trustworthy Natural Language Processing*, 2026.

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*, 2024. URL https://openreview.net/forum?id=auKAUJZMO6.

Mert Yazan, Suzan Verberne, and Frederik Situmeang. The impact of quantization on retrieval-augmented generation: An analysis of small LLMs. In *Proceedings of the Workshop on Information Retrieval's Role in RAG Systems*, pp. 77–81, 2024.

A REVERSE TRANSFER AND NATIVE IDENTIFICATION

Forward transfer under the original matcher showed that FP16's context K10 still moved the scorer on INT4. After the question-aware audit that cell is a wrap. Native INT4 K10 applied in-precision remains a thin leftover ($+4$ $[+1, +8]$ on n=100). Reverse transfer of those heads onto FP16 (Cell A) was run on the original n=50 IDs, $\alpha+$=1, $\alpha-$=0 unless noted, with the same paired bootstrap (1000 draws, seed 42). The native INT4 set is L7_H26, L14_H10, L7_H21, L4_H5, L3_H3, L31_H20, L4_H11, L14_H23, L8_H29, L9_H0. Jaccard versus FP16 K10 is 0.33 (overlap 5/15: L7_H26, L4_H5, L31_H20, L14_H23, L9_H0). Baselines are reused from prior sessions.

Table 11: Identify×apply triangulation for context K10. Original-matcher numbers are in parentheses. Cell B is a thin leftover; Cell A is not a stable reverse circuit.

Cell	Identify → apply	fc (QA)	QA Δ [95% CI]	Comparator
A reverse (n=50)	INT4 K10 → FP16	76%	$+8$ $[-2, +20]$ null	orig. $+48$ $[+34, +62]$; FP16-ID K10 on FP16 (QA): $+4$ $[+0, +8]$
B native (n=100)	INT4 K10 → INT4 N_{CALIB}=128	78%	$+4$ $[+1, +8]$ yes	orig. $+16$ $[+8, +24]$; FP16-ID K10 on INT4 (QA): $+1$ $[+0, +3]$

Letter-heuristic coherence remained at 100% in both cells. Cell B is not a reproduction of the original $+16$ object: after wrap correction it is $+4$ $[+1, +8]$ with about three real CTX switches. Inspecting original-matcher `other` labels still shows terse names or short fragments, not the fluent rambling of high-alpha parametric collapse.

Cell A is a rejected same-circuit hypothesis. Reverse transfer is mostly a scorer artifact: entity-confirmed detection persists at α=0.5 ($+12$ $[+2, +24]$, 6 real CTX) but not at α=1 ($+8$ $[-2, +20]$, null) or α=0.25 ($+10$ $[+0, +20]$, null despite 6 real CTX) — we do not treat this as a stable bidirectional effect. The original-matcher α-sweep (Table 14) recorded a large positive at every gain with exam-meta 36–58%. Question-aware P/C/R/O is Table 15. On INT4's own ID, 9/10 of the native K10 heads are context-specific (`ctx_pos`>0 and `par_pos`≤ 0); L8_H29 is the exception. Two INT4-only heads are context-negative when scored on FP16 ID. Steering that set on FP16 is not the same K10.

The scorer has a pipeline property that the question-aware matcher corrects. Some parametric-to-context flips occur when the unsteered baseline already named the context entity in a full sentence, labeled `parametric` because the matcher checks parametric gold first, and steering shortens the completion to a bare name that then labels `context`. We therefore report label counts, wrap counts, and gold-PAR-only Δ where the claim is parametric.

Gold-PAR-only counting recodes leave-other and wrap as still-suppressed. There are minor item-level caveats: item 1353 (steered "General Aung San") is human gold-PAR but missed by automatic mention after question-stripping and does not change detection. One native-K10 leftover involves a Tunis⊂Tunisia substring. Entity dumps are `entity_audit_k2_k10.txt` and `entity_audit_hminus.txt`.

B PARAMETRIC RECOVERY ATTEMPTS: FULL DETAIL

On FP16, the clean K2 pair remains detected at n=100 under gold-PAR-only counting (-8 pp, $[-14, -3]$; 9 gold-PAR / 7 leave). Inclusive Δ is -15 $[-22, -7]$. On INT4, the identical locked pair is -7 $[-12, -2]$ gold-PAR-only (7 / 5) against inclusive -12 $[-19, -6]$. All diagnostic arms below use the n=100, N_{CALIB}=128 reruns unless noted, question-aware.

0. Transfer of the FP16 clean pair. The starting test applies L20_H14 and L26_H15 at $\alpha+$=1, $\alpha-$=0. Inclusive INT4 Δ is -12 $[-19, -6]$ (P/C/O 18/62/20). Gold-PAR-only is the headline. The original-matcher n=50 result was the borderline -4 $[-10, +0]$ null that motivated expansion.

1. Alpha grid on the same clean pair. If INT4 merely rescaled the same axis, a different $\alpha+$ should restore a cleaner object. Table 16 reports question-aware $\alpha+ \in \{0.5, 1, 2, 3, 5\}$ with $\alpha-$=0

and n=100. The letter heuristic remains at 100% at every gain. At α=3 and 5, `other` is 59–66%: fluent-looking rambling, blank-slot, and off-entity completions, not parametric answers. The -70 pp detections at α=3/5 are scorer collapse. Rescale does not restore a stronger parametric object.

2. L1_H4, the raw parametric peak. L1_H4 is the only INT4 (and FP16) head with `par_pos` above 5× the identification noise floor. On the n=100 INT4 rerun, steering K=1 at $\alpha+$=1, $\alpha-$=0 yields -3 $[-7, +0]$, which includes 0; P/C/O is 14/71/15. The raw peak remains a null control rather than reproducing the clean-pair effect.

3. Native INT4 re-identification. The native clean filter returns L16_H30 and L17_H27. Applied on the N_{CALIB}=128 checkpoint at K=2, $\alpha+$=1, $\alpha-$=0 on 100 items, this pair yields -6 $[-11, -2]$ (P/C/O 17/68/15). This is detected under inclusive counting and is not the locked FP16 pair.

4. Context-suppression as a parametric proxy. H− is the FP16 directional context K10, enlarge empty, $\alpha+$=0, $\alpha-$=1. On 100 items the question-aware delta is -4 $[-8, -1]$, with P/C/O 16/70/14. This is a detected suppression proxy, not evidence that the parametric clean pair is stronger.

Jaccard vs. $|\mathcal{D}|$=4 noise. Head-set overlap is not identifiable at this calibration size. Twenty-five calibration-item resamples place every observed FP16-versus-INT4 Jaccard we computed within the INT4-versus-INT4 noise floor (context K10 J=0.33 versus noise mean 0.27 $[0.05, 0.82]$; noisy parametric K2 J=1.0 versus noise already about 0.95, though this stability holds only for the noisy filter; native re-identification of the clean pair yields disjoint heads with J=0.00). The Jaccard noise uses 25 draws.

5. Matched random-K2 control and historical check. A matched random-K2 control (L20_H24, L3_H19; seed 42, excluding all identified context-K10 and parametric-K2 heads) at n=100, N_{CALIB}=128 under the identical steering protocol (α_+=1, α_-=0) produces a null effect on both precisions (-1 $[-3, +0]$ gold-PAR-only and inclusive, identical on FP16 and INT4), confirming specificity directly. The historical random parametric-K2 check (L13_H2, L16_H20; seed 44; $+4$ $[+0, +10]$, null) was run only at n=50 on the N_{CALIB}=64 checkpoint and was not a matched control for the headline; it is superseded by this n=100 matched control.

Taken together, this is gold-PAR-only transfer of the locked pair at n=100, not a blanket parametric recovery. Higher-gain detections are accompanied by 36–66% `other` outputs and are scorer collapse.

C GEMMA-2-9B IDENTIFICATION, TRANSFER, AND SUPPRESS-SPECIFICITY

Figure 4 and Tables 12–13 are the n=100 Gemma enlarge and H^- displays cited from §4.4.

Table 12: Gemma-2-9B enlarge cells after the question-aware audit. L27_H6 is a wrap; context K10 enlarge is a small detected enlarge. Clean-K2 INT4 transfer is an in-cell null (-2 $[-6, +2]$). Letter heuristic 100%.

Condition	Heads / notes	n=50 Δ [CI]	n=100 Δ [CI]	excl. 0 at n=100?
FP16 ctx K10 enlarge	directional top-10	$+4$ $[+0, +10]$	$\mathbf{+7}$ $[\mathbf{+2}, \mathbf{+12}]$	yes
INT4 ctx K10 enlarge	same locked set	—	$\mathbf{+5}$ $[\mathbf{+1}, \mathbf{+9}]$	yes
FP16 par K2 clean	L31_H7, L39_H6	-2 $[-10, +4]$	-6 $[-12, -1]$	yes
INT4 par K2 clean	same locked pair	—	-2 $[-6, +2]$	no
FP16 L27_H6	raw peak	$+0$ $[+0, +0]$	$+0$ $[+0, +0]$	no
INT4 L27_H6	same locked head	$+0$ $[+0, +0]$	$+0$ $[-3, +3]$	no
INT4 random K=1	L40_H15	$+0$ $[-6, +6]$	$+1$ $[-2, +5]$	no
FP16 random-K10 enlarge	seed 42	$+0$ $[+0, +0]$	$+0$ $[+0, +0]$	no
INT4 random-K10 enlarge	same seed	—	$+3$ $[+0, +7]$	no
FP16 random K=2	L41_H11, L7_H3	—	$+0$ $[+0, +0]$	no
INT4 random K=2	same pair	—	$+2$ $[+0, +5]$	no

Table 13: Gemma identified H^- versus locked random K. Inclusive Δ counts every leave-context move; gold-PAR-only Δ uses the Llama-K2 convention. n=50 is robustness.

Condition	n=50 incl. [CI]	n=100 incl. [CI]	n=100 gold-PAR [CI]	excl. 0 at n=100?
FP16 identified H−	-14 $[-26, -4]$	$\mathbf{-19}$ $[\mathbf{-27}, \mathbf{-11}]$	$\mathbf{-15}$ $[\mathbf{-23}, \mathbf{-8}]$	yes
INT4 identified H−	-8 $[-16, +0]$	$\mathbf{-11}$ $[\mathbf{-18}, \mathbf{-4}]$	$\mathbf{-10}$ $[\mathbf{-17}, \mathbf{-4}]$	yes
FP16 random H−	-4 $[-10, +0]$	-1 $[-5, +2]$	—	no
INT4 random H−	-2 $[-10, +4]$	-1 $[-5, +3]$	—	no
specificity FP16	-10 $[-22, +2]$	-18 $[-27, -10]$	—	yes
specificity INT4	-6 $[-18, +6]$	-10 $[-18, -2]$	—	yes
DiD	$+4$ $[-4, +14]$	$+8$ $[+1, +16]$	—	yes

We repeat the base-model identify×apply protocol on Gemma-2-9B using the same four calibration IDs, the same GPTQ recipe ($\mathrm{N_{CALIB}}$=128), and the same paired bootstrap. Headlines use n=100; n=50 is robustness. Native INT4 re-identification and reverse transfer were not run; those cells remain empty. Clean-K2 INT4 transfer is an in-cell null (-2 $[-6, +2]$, Table 12).

Identification. FP16 directional context and parametric top-10 sets are fully separated (Jaccard 0.0; noise-proxy median 0.0063). The directional context K10 is L40_H14, L17_H7, L24_H15, L22_H0, L32_H11, L25_H0, L3_H9, L8_H9, L7_H6, L29_H5. The directional / clean parametric K=2 pair is L31_H7, L39_H6—identical, and both are context suppressors (`ctx_pos` -0.115 / -0.094; `par_pos` 0.002 / 0.010, at or near the noise floor). They rank as parametric only because `directional_par = par_pos − ctx_pos` is large when `ctx_pos` is large-negative. The raw positive parametric peak is L27_H6 (`par_pos`= + 0.027), the Llama L1_H4 analog. Locked random-K10 heads (reused for suppress controls): L41_H11, L7_H3, L1_H9, L17_H13, L15_H13, L14_H7, L9_H1, L6_H9, L35_H9, L5_H10. Identification ranking is unchanged by the matcher audit; only the steered labels change.

L27_H6 is a wrap. Original-matcher L27_H6 was -12 on FP16 and -8 on INT4 at n=50. Question-aware Δ is a hard zero at n=50 on both precisions (1000/1000 draws exactly zero) and $+0$ $[-3, +3]$ at n=100 INT4. The original-matcher (un-audited) transfer was role-phrase wrapping of the same entity.

Context K10 enlarge is a small real effect at n=100. Question-aware context K10 enlarge is $+7$ $[+2, +12]$ on FP16 and $+5$ $[+1, +9]$ on INT4 at n=100, against random-K10 $+0$ $[+0, +0]$ / $+3$ $[+0, +7]$. At n=50 the FP16 cell is $+4$ $[+0, +10]$ (null). This is a small enlarge, not a Llama-style wrap-inflated $+15$.

Identified H^- versus random K. Identified H− is -19 $[-27, -11]$ on FP16 and -11 $[-18, -4]$ on INT4 at n=100 (inclusive). Locked random H− is -1 on both precisions. Specificity and DiD exclude 0 at n=100 and include 0 at n=50 (Table 13). Gold-PAR-only Δ remains detected at n=100 (-15 / -10). We report this as the Gemma headline, not L27_H6.

D Detailed diagnostic tables

The historical N_{CALIB}=64 versus 128 comparison was run on the original 50-item set and is not repeated here; the expanded n=100 gold-PAR-only comparison is Table 8. Context K5 was run only at n=50 (FP16 and INT4 N_{CALIB}=64) and was not expanded to the n=100 audit.

Table 14: Cell A α-sweep, original matcher: INT4 K10 heads applied on FP16 (n=50). Exam-meta is the fraction of context labels with exam-meta dump. Kept for comparison with Table 15.

$\alpha+$	fc	Δ [95% CI]	excl. 0?	Coh	exam-meta	P/C/O
0.25	50%	$\mathbf{+34}$ $[\mathbf{+20}, \mathbf{+50}]$	yes	100%	**58**%	21/25/4
0.5	52%	$\mathbf{+36}$ $[\mathbf{+22}, \mathbf{+50}]$	yes	100%	**52**%	18/26/6
1.0	64%	$\mathbf{+48}$ $[\mathbf{+34}, \mathbf{+62}]$	yes	100%	36%	12/32/6

Table 15: Cell A α-sweep, question-aware matcher: INT4 K10 heads applied on FP16 (n=50). Real CTX is the entity-confirmed context-entity switch count among QA flips; R = `rejects_context`, where the completion explicitly denies the premise (e.g., "X is not the capital of Y"). Table 14 uses the original (un-audited) matcher, which has no `rejects_context` category and folds denial completions into parametric, so counts are not directly comparable cell-by-cell.

$\alpha+$	fc	Δ [95% CI]	excl. 0?	real CTX	P/C/R/O
baseline	68%	—	—	—	12/34/0/4
0.25	78%	$+10$ $[+0, +20]$	no	6	5/39/1/5
0.5	80%	$\mathbf{+12}$ $[\mathbf{+2}, \mathbf{+24}]$	yes	6	4/40/1/5
1.0	76%	$+8$ $[-2, +20]$	no	5	4/38/3/5

Table 16: Rescale repair: α-grid on FP16 clean K2 applied to INT4 (N_{CALIB}=128, n=100), question-aware matcher.

$\alpha+$	Δ [95% CI]	excl. 0?	P/C/O	Coh heuristic
0.5	-4 $[-8, -1]$	yes	14/70/16	100%
1	$\mathbf{-12}$ $[\mathbf{-19}, \mathbf{-6}]$	yes	18/62/20	100%
2	$\mathbf{-44}$ $[\mathbf{-53}, \mathbf{-34}]$	yes	34/30/36	100%
3	$\mathbf{-70}$ $[\mathbf{-79}, \mathbf{-60}]$	yes	37/4/59	100%*
5	$\mathbf{-70}$ $[\mathbf{-79}, \mathbf{-61}]$	yes	30/4/66	100%*

*Letter-heuristic coherence passes (100%) while completions collapse into fluent rambling and blank slots (59–66% `other`).

Table 17: Native INT4 parametric re-identification versus FP16, and question-aware apply results (n=100 where rerun).

Spec	INT4 set	versus FP16	Apply result (QA)
noisy par K2	L20_H14, L31_H14	Jaccard $\mathbf{1.0}$	-4 $[-11, +3]$ null
noisy par K3	L20_H14, L31_H14, L8_H9	J=0.50	-16 $[-25, -6]$ (28% other)
clean K2	**L16_H30, L17_H27**	Jaccard $\mathbf{0.00}$	-6 $[-11, -2]$; P/C/O 17/68/15
L1_H4 K=1	raw peak	—	-3 $[-7, +0]$ null

Table 18: Matcher-audit summary: original-matcher headline versus question-aware call. Wrap = same-entity role-phrase flip. For Llama K2, the original-matcher column is n=100 inclusive Δ (FP16 -8 $[-14, -3]$, INT4 -4 $[-8, -1]$), not gold-PAR-only and not the n=50 INT4 N_{CALIB}=128 original-matcher cell -4 $[-10, +0]$. Gold-PAR-only is the conservative parametric Δ.

Claim	Original matcher	Question-aware	Call after audit
Llama INT4 ctx K10	$+15$ $[+8, +23]$	$+1$ $[+0, +3]$; 17/17 wrap	wrap artifact
Llama FP16 ctx K10	$+8$ $[+2, +14]$	$+4$ $[+0, +8]$; 4 real CTX	null
Llama native INT4 K10	$+16$ $[+8, +24]$	$+4$ $[+1, +8]$	thin leftover
Llama K2 n=100 (FP16/INT4)	orig. incl. -8 $[-14, -3]$ / -4 $[-8, -1]$	gold-PAR -8 $[-14, -3]$ / -7 $[-12, -2]$	detected
Gemma L27_H6	-12 / -8	hard 0 / $+0$ $[-3, +3]$	wrap artifact
Gemma H^- n=100	-13 / -8 vs random detected	-19 vs -1 / -11 vs -1; gold-PAR -15 / -10	detected
Gemma ctx K10 n=100	$+4$ / $+4$ null	$+7$ $[+2, +12]$ / $+5$ $[+1, +9]$	small real

Table 19: Stage 1 label shift on the shared 164 items, corrected scorer.

Label	FP16	INT4	Δ
parametric	76	37	$\mathbf{-39}$
context	52	69	$\mathbf{+17}$
`rejects_context`	13	22	$+9$
other	23	36	$+13$

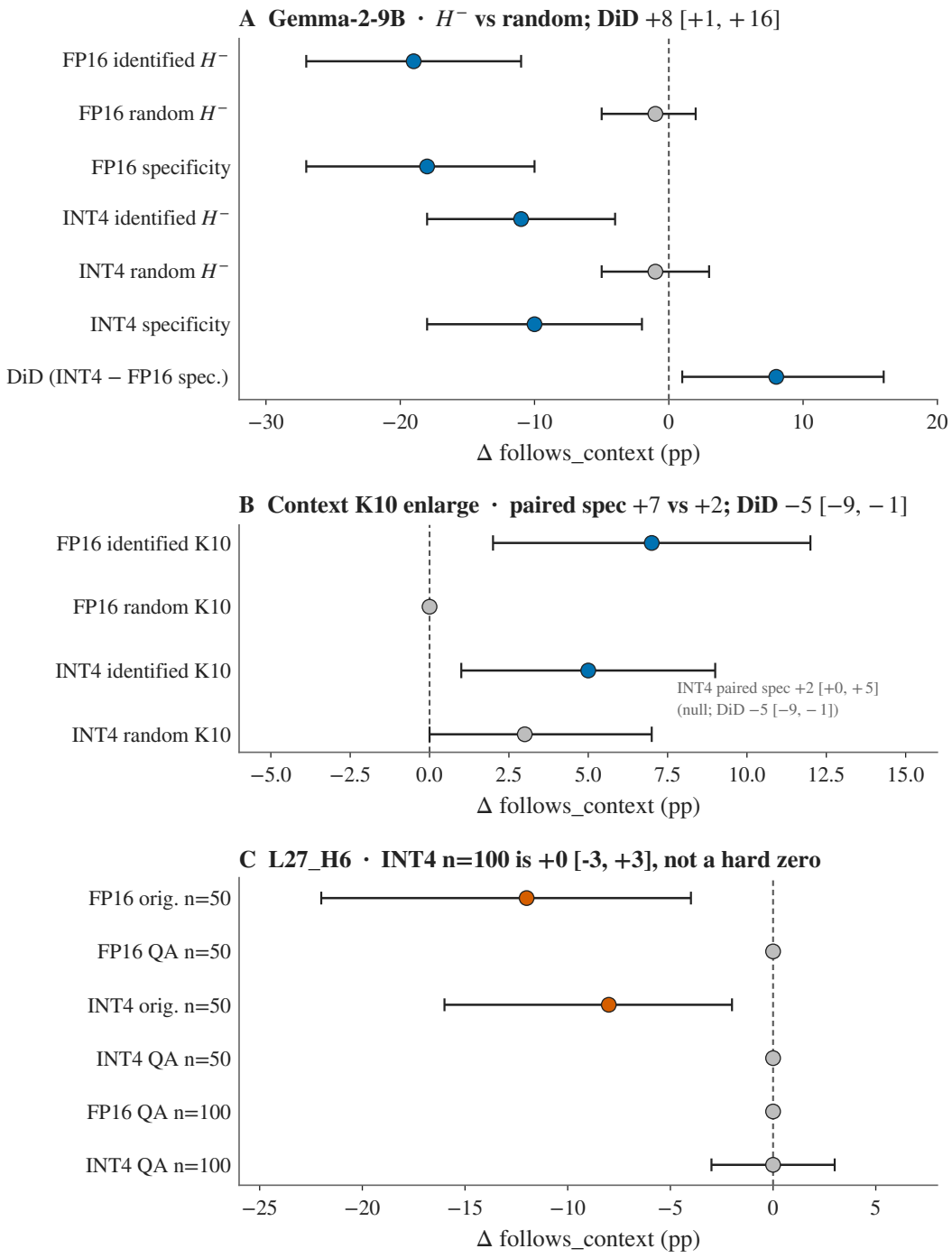


Figure 4: Gemma-2-9B. (A) Identified H^- versus random H^- at n=100: inclusive -19 $[-27, -11]$ vs -1 $[-5, +2]$ on FP16 and -11 $[-18, -4]$ vs -1 $[-5, +3]$ on INT4. Specificity is -18 $[-27, -10]$ / -10 $[-18, -2]$; DiD $+8$ $[+1, +16]$ (attenuation, not hidden). (B) Context K10 enlarge versus random: $+7$ $[+2, +12]$ vs $+0$ $[+0, +0]$ on FP16; $+5$ $[+1, +9]$ vs $+3$ $[+0, +7]$ on INT4; paired specificity $+7$ $[+2, +12]$ vs $+2$ $[+0, +5]$ (null); DiD -5 $[-9, -1]$. (C) L27_H6 original n=50 is -12 $[-22, -4]$ / -8 $[-16, -2]$; question-aware is a hard zero at n=50 on both precisions and at n=100 FP16. INT4 n=100 is $+0$ $[-3, +3]$, not a hard zero.

Table 20: Stage 1 FP16-to-INT4 label transitions on the shared 164 items, corrected scorer.

	→ parametric	→ context	→ `rejects_context`	→ other
parametric →	33	**20**	10	13
context →	3	41	5	3
`rejects_context` →	0	3	5	5
other →	1	5	2	15

The transition matrix reveals asymmetric stability between knowledge sources under quantization: context → context is the most preserved diagonal state ($41/52$ retained), whereas parametric → parametric retention is substantially lower ($33/76$), reflecting parametric recall as the precision-sensitive direction. Crucially, when parametric recall fails, transitions into `rejects_context` (10 items) and `other` (13 items) together outnumber direct shifts into `context` (20 items). As noted in §4.1, this provides transition-level evidence that the memorization collapse is only partially an arbitration shift toward context, with the remainder dispersing into premise denial and off-target completions. Furthermore, the stickiness of `other` → `other` ($15/23$) suggests that disorientation acts as a stable INT4 attractor state rather than transient decoding noise.