QUANTIZATION-INDUCED DECISION CHANGES HAVE REGIME-DEPENDENT INTERVENTION SPECTRA

Anonymous authors
Paper under double-blind review

ABSTRACT

Quantization can preserve aggregate task performance while replacing individual predictions. However, neither score changes nor disagreement rates reveal how the associated behavioral effects are organized across representation directions. To address this gap, we introduce intervention-spectrum analysis, which tracks paired-decision transfer along ordered shared directions and normalizes aggregate prefix transfer by the complete task-supported transfer rate. Using directions fitted to other folds' full-precision states, we project each held-out paired-state difference onto successive prefixes. On disagreeing inputs, we add these projections at their native magnitude to the quantized residual state, testing agreement with the original full-precision prediction regardless of correctness. Across five checkpoints and three quantizers, final-site task-supported interventions outperform dimension- and norm-matched ambient controls in fixed-choice answer selection and teacher-forced full-vocabulary prediction. Despite this shared directional advantage, quantizer-level summaries show that the first 16 directions attain 91.8–92.1% of the complete task-supported one-way transfer rate in fixed choice, versus 12.6–15.1% in full vocabulary. This contrast persists under plain pooled-PCA, while candidate restriction within the full-vocabulary computation does not reproduce compact accumulation. These findings distinguish directional advantage from concentration: effective transfer can remain broadly distributed along a shared representation ordering. An anonymous code repository is available at https://anonymous.4open.science/r/quantization-behavioral-drift-CD71/.

1 INTRODUCTION

Quantization is commonly evaluated by how well it preserves aggregate task performance at lower numerical precision (Frantar et al., 2023; Lin et al., 2024; Xiao et al., 2023; Li et al., 2024). However, preserving task performance does not ensure that a quantized model makes the same decisions as its full-precision counterpart. Improvements and degradations can offset one another, while substitutions between incorrect predictions leave accuracy unchanged. Similar scores can therefore conceal substantial decision replacement. For quantization to be understood as a change in model behavior, these replacements must be examined beyond their frequency or net effect on accuracy: *their internal organization also matters*.

Existing measurements address different aspects of this problem. Disagreement identifies which predictions change, while reconstruction errors characterize differences in internal representations. State interventions test *paired-decision transfer*: whether adding or subtracting a state component makes the intervened prediction match the paired model's original prediction (Hanna et al., 2023; Makelov et al., 2024). However, a high transfer rate using all retained directions in a representation space does not reveal whether a few leading directions already attain most of that rate. Effective transfer could concentrate in a short ordered prefix or accumulate gradually over many directions. We therefore study this accumulation in fixed-choice answer selection and teacher-forced full-vocabulary prediction to distinguish the efficacy of a complete intervention from the concentration of transfer along shared representation directions.

To measure this concentration, we introduce *intervention-spectrum analysis* using the natural state difference between paired full-precision and quantized models on the same input (Figure 1). For

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.

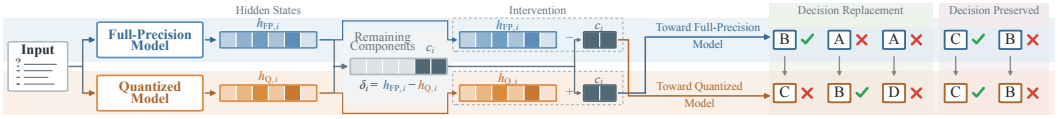


Figure 1: **From replacement to spectra.** Similar aggregate scores can conceal changed decisions. Projected paired differences support interventions whose normalized transfer can concentrate in short prefixes or accumulate broadly within the chosen space, ordering, and operator.

each regime, full-precision states outside the evaluation fold define an ordered basis of directions supported across its tasks. Each held-out input supplies its paired difference and projection coefficients, so directions are shared within each basis while the intervention component varies by input. On inputs where the original predictions disagree, we separately add each prefix projection at its native magnitude to the quantized residual state and test agreement with the original full-precision prediction, regardless of correctness. Coverage divides each prefix's aggregate transfer rate by the complete task-supported transfer rate. The intervention spectrum shows whether a short prefix attains most of that reference rate or many directions are needed to reach comparable Coverage.

Across five checkpoints and GPTQ, AWQ, and RTN, the final-site results reveal a clear separation: both regimes show directional advantages over matched controls, yet fixed-choice transfer concentrates in a short prefix while full-vocabulary transfer accumulates across hundreds to thousands of directions. This contrast persists under plain pooled-PCA, so it does not require the every-task support filter. Furthermore, restricting candidate cardinality within the same full-vocabulary computation retains low early Coverage without reproducing compact fixed-choice accumulation. Geometric comparisons help interpret the contrast by distinguishing row-local output sensitivity from directional alignment across examples. Together, these findings show that directional advantage alone does not establish compact organization. Our contributions are:

- We introduce intervention-spectrum analysis to measure how paired-decision transfer accumulates across ordered shared directions, using cross-fitted interventions normalized by the complete task-supported transfer rate.
- We identify a regime-dependent separation between directional advantage and concentration: across five checkpoints and three quantizers, final-site summaries show advantages in both regimes but rapid fixed-choice and broad full-vocabulary accumulation under the representation-derived ordering.
- We characterize the distinction between local output sensitivity and cross-example directional concentration. With candidate cardinality fixed at two, fixed-choice decision axes are more concentrated across examples than controlled full-vocabulary axes in all five models, providing complementary geometric evidence for the spectrum contrast.

2 Related Work

Post-Training Quantization. Post-training quantization reduces numerical precision while aiming to preserve model quality. GPTQ uses approximate second-order information, while AWQ protects weights associated with salient activations (Frantar et al., 2023; Lin et al., 2024). Related methods use adaptive rounding, activation scaling, or calibration-aware optimization (Nagel et al., 2020; Yao et al., 2022; Dettmers et al., 2022; Xiao et al., 2023; Shao et al., 2024). Other approaches introduce structured transformations, sparse representations, or vector quantization (Chee et al., 2023; Dettmers et al., 2024; Kim et al., 2024; Egiazarian et al., 2024), and LQER approximates weight quantization error with low-rank factors (Zhang et al., 2024).

Behavior Beyond Aggregate Scores. Quantization effects vary across languages, context lengths, and reasoning tasks (Marchisio et al., 2024; Mekala et al., 2025; Kurtic et al., 2025; Liu et al., 2025; Li et al., 2024), and extend to bias and alignment (Hua et al., 2026; Waldis et al., 2026). Rababah et al. (2026) compare correct-prediction overlap with aggregate scores and examine attention-weight distortions. Statistically-lossless quantization distinguishes task performance from distributional

fidelity (Helcig et al., 2026); its Expected Acceptance Rate measures token agreement under optimal coupling, distinct from deterministic top-one agreement and intervention-based transfer.

Causal Analysis of Compressed Models. Chang et al. (2025) connect input-dependent quantization sensitivity to residual-stream magnitude and error amplification, using cross-model activation patching to locate performance-critical components. Zhou et al. (2026) distinguish precision-dependent failure modes through component and activation/error-subspace analyses, residual patching, and targeted interventions; related work examines compression-sensitive reasoning components (Zhang et al., 2026). However, identifying sensitive components or effective interventions does not establish how transfer accumulates along ordered shared directions, or whether a short prefix attains most of the transfer rate obtained with all retained directions.

Representation Intervention and Output Sensitivity. Activation patching, causal tracing, and subspace interventions test how internal representations affect outputs (Vig et al., 2020; Meng et al., 2022; Geiger et al., 2022; Ravfogel et al., 2021; Li et al., 2023; Todd et al., 2024; Hanna et al., 2023; Palumbo et al., 2025). Such tests address functional effects that probe accuracy alone need not establish (Ravichander et al., 2021). However, interventions may introduce off-distribution states or recruit pathways different from those used naturally (Zhang & Nanda, 2024; Makelov et al., 2024). Naturally paired states avoid synthetic input corruption, but projecting their differences can still yield off-distribution states and non-additive responses.

Output-sensitive geometry provides a complementary perspective. The predictive Fisher describes local output-distribution sensitivity (Martens, 2020), and C-PTQ uses a Fisher-weighted objective for channel scaling during multimodal language-model quantization (Li et al., 2026). Other work studies subspaces of learned lens-correction maps (Lombardo et al., 2026) and residual decompositions anchored to predicted-token unembedding directions (Guda, 2026). However, local sensitivity alone does not establish whether output-sensitive directions align across examples.

3 METHOD

In this section, we first define the paired decisions and intervention components, then specify the transfer spectrum and the geometric analyses used to interpret it (Figure 2).

3.1 DECISION REGIMES

We pair decisions and states at identical inputs and contexts. For decision i, FP and Q denote the full-precision and quantized models. The *fixed-choice* regime scores 2–13 single-token option labels (A–M), not the full option texts, at the final prompt position; the *full-vocabulary* regime selects from the complete vocabulary at a fixed teacher-forced next-token position. Hats distinguish original paired predictions $\hat{y}_{\mathrm{FP},i}$ and $\hat{y}_{\mathrm{Q},i}$ from intervened predictions. At residual-stream site ℓ, paired states $h^{(\ell)}_{\mathrm{FP},i}, h^{(\ell)}_{\mathrm{Q},i} \in \mathbb{R}^d$ have width d and naturally occurring difference

$$\delta^{(\ell)}_i = h^{(\ell)}_{\mathrm{FP},i} - h^{(\ell)}_{\mathrm{Q},i}. \tag{1}$$

The difference is defined on all inputs. Transfer is evaluated on changed decisions $\mathcal{F} = \{i : \hat{y}_{\mathrm{FP},i} \neq \hat{y}_{\mathrm{Q},i}\}$, and disruption on initially agreeing decisions $\mathcal{A} = \{i : \hat{y}_{\mathrm{FP},i} = \hat{y}_{\mathrm{Q},i}\}$. In full-vocabulary contexts, teacher forcing supplies both models with gold preceding tokens; this context prefix differs from a basis prefix. Transfer matches the original paired prediction and need not improve accuracy. Task collection and decision interface change together, so the comparison concerns these evaluated regimes. Appendices A and C.1 specify prompts, token scoring and row selection; Appendix A.2 gives fold assignment and basis fitting.

3.2 TASK-SUPPORTED INTERVENTION

Five-fold cross-fitting fits directions to full-precision states from four folds and evaluates interventions on the held-out fold f. Disagreement restricts transfer evaluation, not basis fitting. We center training states within each task and pool them with equal task weights for principal component analysis (PCA). We then retain directions that pass a fixed variance-support threshold in every task, pre-

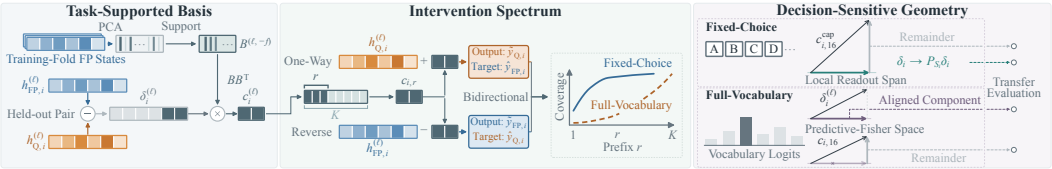


Figure 2: **Overview of intervention-spectrum analysis.** Training-fold full-precision states define the basis; held-out paired differences supply ordered-prefix interventions. Geometry provides complementary tests. Local projection uses the complete paired difference δ_i; local removal starts from capped top 16 (Appendix B.4). Curves are schematic.

serving their pooled-PCA order. The basis, ordering and thresholds use no intervention outcomes; held-out transfer tests their behavioral relevance (Appendix A.2).

The retained orthonormal basis $B^{(\ell,-f)} \in \mathbb{R}^{d\times K_{\ell,f}}$ has task-supported basis dimension $K_{\ell,f}$ for the given model, regime, site and fold; model and regime indices are suppressed, and $-f$ excludes the evaluation fold. For row i in fold $f(i)$, its own paired difference supplies the coefficients of the *task-supported component*:

$$c_i^{(\ell)} = B^{(\ell,-f(i))}\big(B^{(\ell,-f(i))}\big)^\top \delta_i^{(\ell)}. \tag{2}$$

To measure paired-decision transfer, we apply $c_i^{(\ell)}$ in opposite directions:

$$\widetilde{h}_{\mathrm{Q},i}^{(\ell)} = h_{\mathrm{Q},i}^{(\ell)} + c_i^{(\ell)}, \qquad \widetilde{h}_{\mathrm{FP},i}^{(\ell)} = h_{\mathrm{FP},i}^{(\ell)} - c_i^{(\ell)}. \tag{3}$$

The two interventions in Equation (3) are evaluated in separate forward runs. Write $\widetilde{y}_{\mathrm{Q},i}(v_i)$ and $\widetilde{y}_{\mathrm{FP},i}(v_i)$ for predictions after adding v_i to Q or subtracting it from FP. For row-wise interventions $v = \{v_i\}_{i\in\mathcal{F}}$, $R_{\mathrm{one}}(v)$ measures one-way (also called forward) transfer to the original full-precision prediction, and $R_{\mathrm{rev}}(v)$ measures reverse transfer to the original quantized prediction. Bidirectional transfer, $R_{\mathrm{bi}}(v)$, requires both runs to succeed on the same row.

Only the selected token's residual state is patched at the block input; nonlinear computation proceeds normally. Matching predictions need not imply matching computation paths. Replacing this state with its counterpart need not transfer predictions: receiver weights and other token states at that site are retained. Primary spectra use projected paired differences at their native magnitude, without equalizing intervention norms across prefix sizes. Auxiliary norm matching is specified separately.

The primary spectra retain all directions that pass the support criterion. Separate fixed-choice directional controls and local-readout ablations use a basis capped at 48 directions, which need not be the first 48 directions of the uncapped construction (Appendix A.2). Table A.4 distinguishes the input component and norm treatment of each comparison.

3.3 Intervention Spectrum

The intervention spectrum plots transfer against the number of included directions; it is not a PCA eigenvalue spectrum. For a fixed model, regime and site, suppress ℓ and write B and K for each row's fold-specific basis and dimension. Let $B_{1:r}$ contain its first r columns; the prefix intervention is $c_{i,r} = B_{1:r}B_{1:r}^\top\delta_i$. Each prefix is applied to the original receiving state in a separate run, without accumulating patches from earlier prefixes.

For transfer criterion E (one-way, reverse, or same-row bidirectional), write $R_E(r) = R_E(\{c_{i,r}\}_{i\in\mathcal{F}})$ and $R_{E,\mathrm{supp}} = R_E(\{c_i\}_{i\in\mathcal{F}})$. The vertical axis divides the aggregate prefix transfer rate by the complete task-supported transfer rate on the same naturally changed rows with the same weights. Taking this ratio after aggregation, rather than averaging cell-level ratios, defines the *intervention spectrum*:

$$\mathrm{Coverage}_E(r) = R_E(r)/R_{E,\mathrm{supp}}.$$

High Coverage at a short prefix indicates concentrated transfer; a broad profile reaches comparable Coverage only after many directions are included. Coverage normalizes rates, not patch norms; concentration is specific to the space, ordering, and operator.

Each complete task-supported endpoint uses all retained directions in the row's fold; it need not correspond to one common integer K across folds. The supplemental plain-PCA reference instead uses the fold-local space retaining 95% of training-state variance. Neither is the complete paired difference δ_i. Candidate-restricted spectra use each candidate set's own complete task-supported denominator with matched population and aggregation. Appendix A.4.1 specifies zero denominators and unavailable prefixes.

Nested subspaces need not produce nested success sets, so Coverage can be nonmonotonic and exceed one; the complete endpoint need not be a maximum. This rate ratio is not explained variance or a cumulative fraction of transferred rows. An observed-grid crossing is the first tested prefix reaching a threshold, not an intrinsic or minimum sufficient dimension; grid non-crossing does not bound unmeasured prefixes. EventRetention asks which complete task-supported endpoint successes also transfer at the prefix; Appendix C.6 defines its row-overlap summary and aggregation.

Same-rank random prefixes test ordered directions against alternatives in the same task-supported space. In both regimes, each random-prefix patch is rescaled row by row to the corresponding ordered-prefix patch norm at the measured ranks. In the main removal analysis, the injected remainder is $\delta_i - c_{i,r}$: the selected component is excluded from the complete paired-difference intervention. The matched-random control projects δ_i onto a random rank-r subspace of the same task-supported space and rescales the result to match $\|c_{i,r}\|_2$ row by row. Removal specificity is the transfer rate under random removal minus that under ordered removal. The injected remainders are not norm-matched (Appendices A.3 and A.4).

3.4 Decision-Sensitive Geometry

Geometry-aligned interventions measure transfer along output-sensitive directions; cross-example analysis examines whether local decision axes align across inputs. For fixed-choice row i with m_i candidates, $z^{\mathrm{FP}}_{i,a}(h)$ is candidate a's logit after full-precision continuation from state h. With reference candidate a_0, the local readout span is

$$\mathcal{S}_i = \mathrm{span}\{\nabla_h (z^{\mathrm{FP}}_{i,a}(h) - z^{\mathrm{FP}}_{i,a_0}(h))\}_{a \neq a_0}.$$

The gradients are evaluated at $h^{(\ell)}_{\mathrm{FP},i}$, with $\mathrm{rank}(\mathcal{S}_i) \leq m_i - 1$. For binary decisions, one axis per row need not imply concentration across rows: the axes can point in different directions. The cross-example analysis measures their distribution separately from these local ranks.

Full-vocabulary predictive-Fisher geometry uses full-precision token-log-probability gradients with respect to the residual state. The cross-fitted estimate $\widehat{F}^{(\ell,-f)}$ is an uncentered, equal-task gradient second moment. Fisher-95 is its leading eigenspace retaining 95% of the estimated trace. Local-readout rank describes a row-specific candidate interface; predictive-Fisher rank summarizes training-fold output-distribution sensitivity. They are not a common dimensional scale. Both use full-precision gradients independently of intervention outcomes. Reverse and bidirectional transfer criteria reuse the component without symmetrizing its gradient geometry (Appendix C.5).

Let P_U denote orthogonal projection onto subspace U. Local and Fisher projections act on the complete paired difference: $v^{\mathrm{local}}_i = P_{\mathcal{S}_i}\delta_i$, and $P_F\delta_i$ for a Fisher eigenspace F. Removal starts from a prefix-16 component. Local removal uses $c^{\mathrm{cap}}_{i,16} - P_{\mathcal{S}_i}c^{\mathrm{cap}}_{i,16}$; Fisher-95 removal uses $c_{i,16} - P_F c_{i,16}$, where F is the Fisher-95 eigenspace and $c_{i,16}$ is the uncapped ordered prefix-16 component. Norm restoration rescales each remainder to its starting component's norm. Fisher-16 compares its native projection with its row-wise rescaling to the uncapped prefix-16 patch norm (Appendix C.5). These controls do not equalize norms across prefixes in the primary spectrum.

4 Experiments and Results

Experimental Setup. We test the regime contrast across five 7B–14B instruction-tuned checkpoints from Qwen, Llama, and Mistral (Qwen et al., 2024; Grattafiori et al., 2024; Jiang et al., 2023; Mistral AI Team, 2024), with GPTQ, AWQ, and round-to-nearest (RTN) using 4-bit weights and 16-bit activations (W4A16). The five fixed-choice tasks are ARC-Challenge (ARC), HellaSwag (HS), PIQA (PI), SQuAD2 answerability (S2), and TruthfulQA-MC (TQ) (Clark et al., 2018; Zellers et al., 2019; Bisk et al., 2020; Rajpurkar et al., 2018; Lin et al., 2022). The three full-vocabulary

tasks are LAMBADA (LM), TriviaQA (TR), and GSM8K (G8) (Paperno et al., 2016; Joshi et al., 2017; Cobbe et al., 2021). These tasks expose two decision interfaces under a common intervention framework. Appendix A specifies checkpoint revisions, task counts, quantization groups and calibration conditions.

Both regimes use held-out cross-fitting at three block-input residual-stream sites, near normalized depths 0.52, 0.77, and 1.00 (Appendix Table A.1). Early, middle, and final label these sampled sites relatively; the final site is a block input, not the output logits. Teacher forcing fixes full-vocabulary contexts, with all GSM8K positions from one problem assigned to one fold. Rates average calibrations within checkpoint/task, tasks equally, then checkpoints. Checkpoints are cross-model replication units; task/calibration cells are repeated measurements and folds are analysis partitions. Source-grouped intervals condition on the evaluated checkpoints and fitted representations (Appendix Tables B.4, C.4, and C.13); they exclude refitting, Fisher sampling-seed and checkpoint-population uncertainty. Plotted checkpoint min–max ranges are not confidence bands. Appendix A.4.1 distinguishes these intervals from cell-level bootstrap estimates; Appendices A.3 and C.1 specify patch and target positions.

Decision Replacement. Small net score changes can conceal substantial replacement of individual decisions (Table 1; Figure 3). For full-vocabulary AWQ, exact-token performance changes by only -0.16 percentage points, while 11.89% of predictions change. The transition breakdown explains this discrepancy: improvements and degradations partly cancel, while substitutions between incorrect predictions change decisions without affecting accuracy. Replacement varies across checkpoints and tasks (Appendix Figure B.1 and Table C.1); the following analyses examine how paired-state interventions transfer the changed decisions.

Directional Advantage. Matched controls test whether transfer depends on direction selection when subspace dimension and patch norm are held fixed. At the final sampled site, task-supported components outperform matched ambient-random components in both regimes across GPTQ, AWQ, and RTN, including under bidirectional transfer (Table 2; Appendix Tables B.1 and C.2). This directional advantage does not establish whether most transfer concentrates in a short prefix.

These control contrasts are interpreted within each regime: fixed-choice uses the capped companion, whereas full-vocabulary uses the uncapped construction. Full-precision-fitted and quantized-state-fitted bases yield similar aggregate transfer in the tested constructions. Complementary-space controls, depth trends and calibration sensitivity are reported in Appendix Tables B.1, C.3 and B.3, and Figure B.5; GPTQ source-grouped intervals appear in Appendix Table C.4. We next compare the primary uncapped spectra in both regimes to distinguish directional advantage from concentration.

Transfer Concentration. Directional advantages coexist with sharply different accumulation profiles (Figure 4). At the final sampled site, quantizer-level summaries over five checkpoints show that the first 16 directions attain 91.8–92.1% of the complete task-supported one-way transfer rate for fixed-choice decisions, compared with 12.6–15.1% for full-vocabulary decisions. Thus, a short shared prefix attains nearly all of the fixed-choice reference rate, whereas full-vocabulary transfer accumulates more gradually along the ordering.

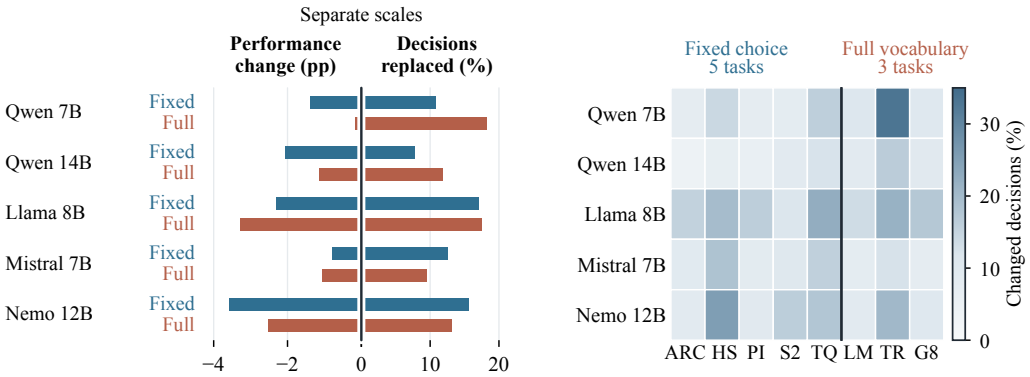
In the final-site GPTQ uncapped analysis, the complete task-supported interventions transfer 86.3% and 70.0% of changed fixed-choice and full-vocabulary decisions, respectively (Appendix B.3.1). Both regimes therefore support substantial transfer, but differ in how early their ordered prefixes attain it. The same qualitative separation holds bidirectionally (Table 3; within-regime intervals in Appendix Tables B.4 and C.4). Full prefix curves and their complete task-supported reference rates appear in Figure 5 and Appendix Figure C.1.

The success-set overlap provides a complementary view of this contrast. At the final site under GPTQ, prefix-16 one-way EventRetention is 88.5% for fixed-choice and 12.8% for full-vocabulary decisions (Appendix Table C.16). The fixed-choice prefix therefore captures most successes of its complete endpoint, whereas the full-vocabulary prefix captures a much smaller share. Prefix-only and endpoint-only successes also occur, showing why aggregate Coverage alone cannot identify which decisions are transferred.

Removal tests how complete paired-difference transfer depends on the leading component. Removing the ordered top-16 component reduces transfer more than matched-random removal, with

Table 1: **Small net score changes coexist with substantial decision replacement.** Five-checkpoint task-balanced summaries. FP/Q: accuracy or next-token exact match (%); Δ: Q minus FP (pp); Changed: disagreement (%); Range: checkpoint range. CW/WC/WW: correct-to-wrong/wrong-to-correct/wrong-to-different-wrong shares among changed decisions (%).

	Fixed-choice			*Full-vocabulary*		
	GPTQ	AWQ	RTN	GPTQ	AWQ	RTN
FP	71.97	71.97	71.97	62.75	62.75	62.75
Q	69.95	70.94	70.57	61.12	62.59	61.70
Δ	−2.03	−1.03	−1.40	−1.63	−0.16	−1.05
Changed	**12.78**	**9.88**	**11.30**	**14.07**	**11.89**	**14.81**
Range	[7.73–17.12]	[5.59–13.80]	[6.87–14.74]	[9.53–18.22]	[8.31–18.50]	[10.59–20.42]
CW/WC/WW	50.7/33.1/16.2	47.3/37.7/15.0	49.5/35.6/15.0	32.8/20.4/46.9	28.9/25.1/46.0	30.2/22.4/47.4



(a) Checkpoint summaries. **(b)** Task-level decision changes.

Figure 3: **Similar scores, different decisions.** (a) GPTQ checkpoint summaries with Fixed/Full regime subrows. Zero alignment is visual only: performance change (pp) and decision replacement (%) use independent scales; bar lengths are not comparable across sides. (b) Task-level replacement; task codes are defined in the experimental setup.

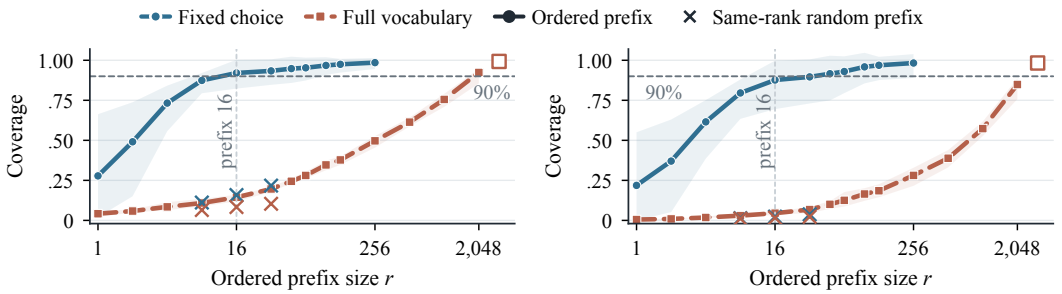
a much larger contrast in fixed-choice decisions (Table 3). The control matches the removed subspace dimension and component norm, but not the injected remainder norm, so it does not isolate direction from remainder magnitude. Condition-level exceptions and low-count cells are reported in Appendix B.5.

Basis Size and Tasks. The contrast persists beyond basis truncation and task averaging. Fixed-choice accumulation remains rapid without a basis cap, despite retaining hundreds of directions (Figure 5). Under final-site GPTQ, the separation persists at approximately matched basis fractions using measured prefixes (Appendix Table C.5). Prefix-16 Coverage is also high for each of the five fixed-choice tasks and low for each of the three full-vocabulary tasks (Appendix Table C.6). These comparisons show that neither cap saturation nor a single task accounts for the aggregate contrast, although condition-level crossings remain heterogeneous (Appendix B.3.1).

Basis Construction. Plain pooled-PCA preserves the contrast without the every-task support filter: prefix-32 one-way Coverage reaches 90.5–93.1% for fixed-choice and 17.6–19.6% for full-vocabulary decisions across quantizers (Appendix Figure B.6). Each curve uses its own complete transfer rate within the fold-local PCA space retaining 95% of training-state variance. At rank 32, ordered projections outperform same-rank random subspaces with matched row-wise patch norms in every quantizer-level mean comparison across both regimes and all three transfer criteria (Appendix Table B.10); some task–condition–seed contrasts are non-positive. Thus, the alternative construction retains both contrasting accumulation profiles and directional advantages at the tested rank, without establishing ordering invariance. The AWQ contrast also persists on the same four-checkpoint sub-

Table 2: **Directional advantage over matched controls.** Final-site GPTQ transfer on changed decisions (%), averaged across five checkpoints. Advantage: task-supported minus dimension- and row-norm-matched ambient random (pp). Fixed-choice uses the capped-48 companion; full vocabulary uses the uncapped basis.

Regime	Criterion	Task-supported	Ambient random	Advantage
Fixed-choice	One-way	**81.4**	10.3	**71.1**
	Bidirectional	**67.2**	0.7	**66.5**
Full-vocabulary	One-way	**70.0**	57.3	**12.7**
	Bidirectional	**49.7**	29.9	**19.7**



(a) One-way transfer.

(b) Bidirectional transfer.

Figure 4: **Regime-dependent intervention spectra.** Final-site GPTQ Coverage relative to the complete task-supported transfer rate. Filled curves/bands: five-checkpoint aggregate/ranges. Open squares use checkpoints available at that prefix, with the same subset and weights for prefix and complete transfer. Crosses: same-rank random prefixes of the primary uncapped bases, with row-wise patch norms matched to the ordered prefix (Appendices A.4 and C.3.1).

set in both regimes, excluding the checkpoint affected by baseline-replay filtering (Appendix B.5.3; Table B.9).

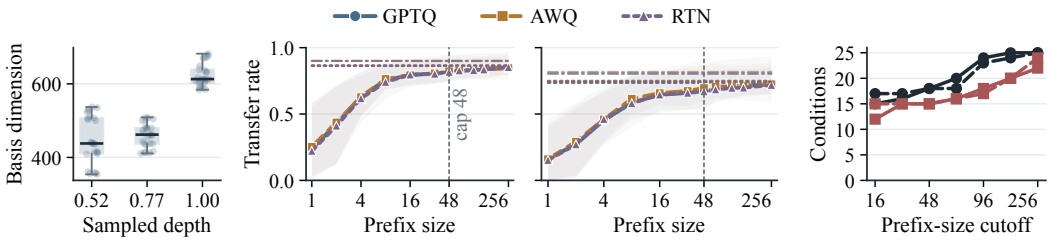
Candidate Restriction. Candidate restriction tests whether fewer readout choices reproduce compact accumulation within the full-vocabulary computation. At $\kappa = 2$, the original full-precision and quantized winners on disagreeing inputs are fixed before intervention; only the candidate set changes, with the basis, patches and patched logits held fixed. Relative to each set's own complete task-supported transfer rate, prefix-16 one-way Coverage is 14.42% for $\kappa = 2$ and 14.36% for the full vocabulary (ratios of equal-block means; Appendix Table C.11). Restriction changes absolute transfer, yet early Coverage remains low across all five checkpoints and under reverse and bidirectional criteria. Thus, reducing downstream cardinality alone does not reproduce fixed-choice concentration within the evaluated computation and prefix range. This comparison does not establish statistical equivalence or control differences in tasks, prompts and answer encoding (Appendix C.4).

Decision Geometry. Output-sensitive transfer and cross-example axis alignment help interpret concentration along the primary representation ordering; they measure different properties. For fixed-choice decisions, projecting the complete paired difference onto the row-local candidate-logit span attains 94.5–95.2% of this analysis's own complete paired-difference one-way transfer rate. Removing that span's component from capped top 16 leaves little bidirectional transfer, even after norm restoration (Appendix B.4). The span varies by row, so this high projection transfer does not imply one shared low-dimensional space.

For full-vocabulary decisions, matched Fisher-16 projections exceed uncapped ordered prefix-16 one-way transfer at the same rank and row-wise patch norm. Across quantizer-level point estimates, the paired advantage ranges from 14.38 to 16.40 percentage points (Appendix Table C.13; absolute transfer rates in Table C.12, Part B). Conditional paired source-bootstrap intervals are positive for each quantizer and all three transfer criteria. Fisher-16 is not restricted to the original task-supported span: it changes subspace selection and directional priority, not only ordering within

Table 3: **Concentration and removal across regimes.** Coverage: aggregate prefix-16 transfer rate divided by the complete task-supported transfer rate (%). Removal contrast: matched-random minus ordered-remainder transfer (pp) after uncapped top-16 removal from the paired difference. Removed-component rank/norm are matched; remainder norms are unconstrained.

		Top-16 Coverage (%)		Removal Contrast (pp)	
Quantizer	Criterion	Fixed-choice	Full-vocabulary	Fixed-choice	Full-vocabulary
GPTQ	One-way	92.1	14.4	67.1	4.3
	Bidirectional	87.7	4.5	61.7	6.1
AWQ	One-way	91.8	15.1	68.0	4.0
	Bidirectional	87.7	5.0	62.7	5.4
RTN	One-way	92.1	12.6	67.0	2.9
	Bidirectional	87.0	2.8	61.9	3.9



(a) Basis size. **(b)** One-way. **(c)** Bidirectional. **(d)** 90% crossings.

Figure 5: **Fixed-choice concentration persists without a basis cap.** (a) Fold basis dimensions. (b,c) Final-site transfer: quantizer means/ranges; dotted/dash-dot lines denote complete task-supported transfer rates/complete paired-difference transfer rates. The cap-48 guide marks prefix size 48, not equivalence to the separately selected capped basis. (d) 90% crossings: black circles/red squares denote one-way/bidirectional; solid/dashed denote equal-task/pooled-event aggregation.

that space. Broad representation-ordered accumulation can therefore coexist with stronger output-sensitive transfer. Fisher-95 projection also supports substantial transfer; its removal leaves less transfer than same-rank random removal after restoring final patch norms, for all three criteria (Appendix C.5). Fisher-16 tests one matched rank, not a complete Fisher-ordered spectrum.

A one-dimensional local axis can point in different directions across examples. Holding candidate cardinality at two gives each row a single contrast axis, yet fixed-choice axes are more concentrated across examples in all five models (Appendix Table C.14). The participation-ratio dimension D_{PR} measures how broadly unit decision axes are dispersed across examples; the median of the five model-level controlled/fixed ratios is $90.85\times$. Task collections and answer encodings remain distinct: fixed-choice labels are reused, whereas full-vocabulary candidate tokens can vary by row. A separate stable-row analysis associates shared output coordinates with greater axis alignment (Appendix Table C.15). The primary cross-example geometry analysis does not require paired prediction disagreement; models, not example pairs, are replication units (Appendix C.5.1).

5 Conclusion

We introduced intervention-spectrum analysis to characterize how paired-decision transfer accumulates across shared representation directions. At the final sampled site, both regimes exhibit directional advantages, yet fixed-choice transfer concentrates in short prefixes while full-vocabulary transfer accumulates broadly. This contrast shows that complete-reference efficacy alone cannot characterize transfer concentration. Geometric comparisons reinforce this distinction by separating local output sensitivity from directional alignment across examples. Together, these results reveal behavioral organization under the chosen paired difference, representation space, ordering, and native-magnitude operator. The evidence covers W4A16 group-128 quantization of 7–14B instruction-tuned checkpoints, with jointly varying tasks and interfaces and teacher-forced full-vocabulary contexts.

AI USE STATEMENT

Generative AI tools were used for language editing and grammatical corrections. The authors take full responsibility for the final content of this paper.

REPRODUCIBILITY STATEMENT

Experimental protocols and intervention definitions are provided in Section 3 and Appendices A–C. The supplementary materials provide frozen evaluation and calibration inputs, configurations, saved outcomes, sufficient counts, and analysis code for checking the reported summaries. Appendix D specifies input reconstruction and the distinction between outcome verification and end-to-end replay. The code is available at: https://anonymous.4open.science/r/quantization-behavioral-drift-CD71/

REFERENCES

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7432–7439, 2020. doi: 10.1609/aaai.v34i05.6239. URL https://ojs.aaai.org/index.php/AAAI/article/view/6239.

Ting-Yun Chang, Muru Zhang, Jesse Thomason, and Robin Jia. Why do some inputs break low-bit LLM quantization? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3410–3429, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.168. URL https://aclanthology.org/2025.emnlp-main.168/.

Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher M. De Sa. QuIP: 2-bit quantization of large language models with guarantees. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0196. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/0df38cd13520747e1e64e5b123a78ef8-Abstract-Conference.html.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. URL https://arxiv.org/abs/1803.05457.

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

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. GPT3.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c3ba4962c05c49636d4c6206a97e9c8a-Abstract-Conference.html.

Tim Dettmers, Ruslan A. Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, and Dan Alistarh. SpQR: A sparse-quantized representation for near-lossless llm weight compression. In *International Conference on Learning Representations*, 2024.

Vage Egiazarian, Andrei Panferov, Denis Kuznedelev, Elias Frantar, Artem Babenko, and Dan Alistarh. Extreme compression of large language models via additive quantization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12284–12303, 2024. URL https://proceedings.mlr.press/v235/egiazarian24a.html.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.17323.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7324–7338, 2022. URL https://proceedings.mlr.press/v162/geiger22a.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Nelson Guda. Geometric and behavioral stratification in transformer residual streams, 2026. URL https://arxiv.org/abs/2608.12447v3.

Michael Hanna, Roberto Zamparelli, and David Mareček. The functional relevance of probed information: A case study. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 835–848. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.eacl-main.58. URL https://aclanthology.org/2023.eacl-main.58/.

Michael Helcig, Eldar Kurtić, and Dan Alistarh. Statistically-lossless quantization of large language models, 2026. URL https://arxiv.org/abs/2605.02404v2.

Stanley Z. Hua, Sanae Lotfi, and Irene Y. Chen. Uncertainty drives social bias changes in quantized large language models. *arXiv preprint arXiv:2602.06181*, 2026. URL https://arxiv.org/abs/2602.06181.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

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, 2017. doi: 10.18653/v1/P17-1147.

Sehoon Kim, Coleman Richard Charles Hooper, Amir Gholami, Zhen Dong, Xiuyu Li, Sheng Shen, Michael W. Mahoney, and Kurt Keutzer. SqueezeLLM: Dense-and-sparse quantization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 23901–23923, 2024. URL https://proceedings.mlr.press/v235/kim24f.html.

Eldar Kurtic, Alexandre Noll Marques, Shubhra Pandit, Mark Kurtz, and Dan Alistarh. "Give Me BF16 or Give Me Death"? Accuracy-Performance Trade-Offs in LLM Quantization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26872–26886. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1304. URL https://aclanthology.org/2025.acl-long.1304/.

Jiameng Li, Han Zhou, and Matthew B. Blaschko. C-PTQ: Fisher-weighted channel-wise sensitivity for post-training quantization of MLLMs, 2026. URL https://arxiv.org/abs/2607.21076v1.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/81b8390039b7302c909cb769f8b6cd93-Abstract-Conference.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, 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, 2024. URL https://proceedings.mlsys.org/paper_files/paper/2024/hash/42a452cbafa9dd64e9ba4aa95cc1ef21-Abstract-Conference.html.

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

Ruikang Liu, Yuxuan Sun, Manyi Zhang, Haoli Bai, Xianzhi Yu, Tiezheng Yu, Chun Yuan, and Lu Hou. Quantization hurts reasoning? an empirical study on quantized reasoning models. *arXiv preprint arXiv:2504.04823*, 2025. URL https://arxiv.org/abs/2504.04823.

Gianfranco Lombardo, Giuseppe Trimigno, and Stefano Cagnoni. A geometric perspective on next-token prediction in large language models: Three emerging phases, 2026. URL https://arxiv.org/abs/2605.09011v1.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Ebt7JgMHv1.

Kelly Marchisio, Saurabh Dash, Hongyu Chen, Dennis Aumiller, Ahmet Üstün, Sara Hooker, and Sebastian Ruder. How does quantization affect multilingual llms? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15928–15947. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.935. URL https://aclanthology.org/2024.findings-emnlp.935/.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020. URL https://jmlr.org/papers/v21/17-678.html.

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. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.479. URL https://aclanthology.org/2025.emnlp-main.479/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1262. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016. URL https://arxiv.org/abs/1609.07843. Introduces the WikiText corpus.

Mistral AI Team. Mistral nemo. Mistral AI Research Release, July 2024. URL https://mistral.ai/news/mistral-nemo/. Released July 18, 2024.

Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. Up or down? adaptive rounding for post-training quantization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7197–7206, 2020. URL https://proceedings.mlr.press/v119/nagel20a.html.

Nils Palumbo, Ravi Mangal, Zifan Wang, Saranya Vijayakumar, Corina S. Pasareanu, and Somesh Jha. Validating mechanistic interpretations: An axiomatic approach. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 47509–47544, 2025. URL https://proceedings.mlr.press/v267/palumbo25a.html.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1525–1534, 2016. doi: 10.18653/v1/P16-1144.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Baha Rababah, Shahzeb Qamar, Lorenz Sparrenberg, Rafet Sifa, Murat Kantarcioglu, Cuneyt Gurcan Akcora, and Carson K. Leung. The illusion of equivalency: Statistical characterization of quantization effects in LLMs, 2026. URL https://arxiv.org/abs/2607.08734v2.

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.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2124. URL https://aclanthology.org/P18-2124/.

Shauli Ravfogel, Grusha Prasad, Tal Linzen, and Yoav Goldberg. Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. In *Proceedings of the 25th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.conll-1.15/.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.295. URL https://aclanthology.org/2021.eacl-main.295/.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqian Li, Kaipeng Zhang, Peng Gao, Yu Qiao, and Ping Luo. Omniquant: Omnidirectionally calibrated quantization for large language models. In *International Conference on Learning Representations*, 2024.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *International Conference on Learning Representations*, 2024.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Andreas Waldis, Vagrant Gautam, Anne Lauscher, Dietrich Klakow, and Iryna Gurevych. Aligned probing: Relating toxic behavior and model internals. *Transactions of the Association for*

Computational Linguistics, 14:271–291, 2026. doi: 10.1162/tacl.a.613. URL https://aclanthology.org/2026.tacl-1.14/.

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, 2023. URL https://proceedings.mlr.press/v202/xiao23c.html.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/adf7fa39d65e2983d724ff7da57f00ac-Abstract-Conference.html.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

Cheng Zhang, Jianyi Cheng, George Anthony Constantinides, and Yiren Zhao. LQER: Low-rank quantization error reconstruction for llms. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 58763–58779, 2024. URL https://proceedings.mlr.press/v235/zhang24j.html.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/06a52a54c8ee03cd86771136bc91eb1f-Abstract-Conference.html.

Nan Zhang, Eugene Kwek, Yusen Zhang, Ngoc-Hieu Nguyen, Prasenjit Mitra, and Rui Zhang. When reasoning meets compression: Understanding the effects of LLMs compression on large reasoning models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=2za3iNkwXn.

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/.

A Experimental and Analysis Protocol

A.1 Models, Quantization, and Decision Regimes

Each observation pairs a full-precision model with its quantized counterpart under an identical context and regime-specific decision interface. The fixed-choice regime contains 24,869 decisions from five tasks with 2–13 candidates per row. The full-vocabulary regime contains 15,417 teacher-forced token decisions from three tasks, each scored over the complete vocabulary. Both regimes evaluate the same five instruction-tuned checkpoints. Residual states are collected at the block inputs nearest normalized depths 0.52, 0.77, and 1.00; ties are resolved toward the later block. Table A.1 lists checkpoint revisions and block indices. Target-token mappings are tokenizer-specific and remain fixed across the quantization conditions of each checkpoint. Table A.2 lists fixed-choice tasks, row counts and candidate cardinalities; Appendix C.1 specifies the full-vocabulary protocol. Table A.3 summarizes quantization and calibration settings.

Fixed-Choice Decision Interface. The default prompt is `Question: {stem}\nChoices:\n{choices}\nThe correct answer is`; inputs with context prepend `Context: {context}`, and HellaSwag uses its stem as that context. Each choice line is `{label}. {option}`, with labels A–M assigned in candidate order. A candidate is scored by the next-token logit of the string consisting of one leading space followed by its label, evaluated at the final non-padding prompt position; ties select the lowest candidate index. For every evaluated tokenizer, each such label string maps to one tokenizer token. The tokenizer is loaded from the full-precision snapshot in Table A.1 with `use_fast=True`, `local_files_only=True`, and `trust_remote_code=True`; label encoding uses `add_special_tokens=False`. Numeric IDs and tokenizer resources are not packaged; the fixed revisions and this encoding procedure specify their reconstruction.

The primary intervention evaluations use no chat template, automatically inserted special tokens, or KV caching. Fixed-choice batching uses masked right-padding, with the end-of-sequence token only as a padding fallback when no pad token is defined; predictions are read at the final non-padding position. Full-precision and quantized models use identical row identifiers and candidate sets. Candidate cardinality ranges from 2 to 13, and the scorer rejects larger rows.

ARC-C uses all 299 rows of the frozen validation split of `allenai/ai2_arc`, configuration `ARC-Challenge`, revision `210d026faf99`; its source IDs agree across the 25 quantization conditions. SQuAD2 uses all 11,873 validation rows of `rajpurkar/squad_v2`, revision `3ffb306f725f`. Its options are `unanswerable`, then `answerable`, so A and B encode those alternatives respectively; the gold label is B when the source answer list is nonempty. The released normalized rows retain source order, question, context, options and gold labels. Fold assignment partitions this fixed evaluation set rather than selecting it.

Quantization Conditions. All conditions use 4-bit weights, 16-bit activations, and input-axis weight groups of 128. GPTQ and AWQ each use separate C4 and WikiText calibration sets (Frantar et al., 2023; Lin et al., 2024; Raffel et al., 2020; Merity et al., 2016), giving ten conditions per method. Their checkpoint configurations and tensor-header inventories identify quantized query, key, value and output attention projections and the gate, up and down MLP projections in every transformer block; embedding and output-head weights are float16. RTN contributes five calibration-free conditions and stores dequantized floating-point weights. Its runtime contract quantizes every `torch.nn.Linear` module inside `model.model.layers`, excluding embeddings and the output head. Calibration variants share a parent checkpoint and are treated as paired conditions. The study evaluates model behavior, not packed-kernel performance.

Each frozen calibration source contains 4,194,304 normalized characters in 256 chunks. C4 uses a seeded streaming shuffle of English training documents; WikiText uses reconstructed training articles and a seeded article permutation. For each checkpoint tokenizer, deterministic concatenation is encoded without special tokens; the first 262,144 IDs form 128 sequences of 2,048 tokens, with no further sampling. The same token pack is used for GPTQ and AWQ. Released packs and configurations specify the exact inputs and are bound to the evaluated checkpoints; full model and intervention-tensor replay has the separate scope described in Appendix D.

A.2 Cross-Fitted Task-Supported Construction

Both regimes use deterministic five-fold assignments with seed 4242 to separate basis construction from intervention evaluation.

Let T denote the number of tasks, with $T = 5$ for the fixed-choice decision regime and $T = 3$ for the full-vocabulary token decision regime. Let $h^{(\ell)}_{s,t,i} \in \mathbb{R}^d$ be the residual-stream state at sampled site ℓ and the evaluated prompt/token position for model $s \in \{\mathrm{FP}, Q\}$, task t, and row i. Each regime uses one five-fold assignment shared across checkpoints and quantizers, and the basis used to intervene on fold f is fitted only on the other four folds. For full-vocabulary token decisions, a deterministic hash of task and source ID assigns the 15,417 rows to folds while keeping the four GSM8K positions from each problem together. For fixed-choice decisions, the 24,869 rows are stratified by gold-label index, sorted within each stratum by a deterministic hash of dataset, row ID, and label, and assigned round-robin modulo five. Let I_t^{-f} be the training-fold row set for task t and $n_t = |I_t^{-f}|$; the suppressed f in n_t always refers to this training subset. Define $X^{(\ell,-f)}_{s,t} \in \mathbb{R}^{n_t \times d}$

Table A.1: **Evaluated checkpoints and sampled block inputs.** Revisions are shown as 12-character prefixes.

Checkpoint	Repository	Revision	Blocks	Total
Qwen-7B	`Qwen/Qwen2.5-7B-Instruct`	`a09a35458c70`	14/21/27	28
Qwen-14B	`Qwen/Qwen2.5-14B-Instruct`	`cf98f3b3bbb4`	24/36/47	48
Llama-8B	`meta-llama/Llama-3.1-8B-Instruct`	`0e9e39f249a1`	16/24/31	32
Mistral-7B	`mistralai/Mistral-7B-Instruct-v0.3`	`c170c708c41d`	16/24/31	32
Nemo-12B	`mistralai/Mistral-Nemo-Instruct-2407`	`04d8a90549d2`	20/30/39	40

Table A.2: **Fixed-choice evaluation set.** Variable-cardinality tasks show the maximum observed candidate cardinality; the complete interface spans 2–13 candidates. The full-precision and quantized models are evaluated on identical rows.

Task	Rows	Candidates / row
ARC-C	299	≤ 5
HellaSwag	10,042	4
PIQA	1,838	2
SQuAD2-Ans.	11,873	2
TruthfulQA-MC	817	≤ 13
Total	24,869	2–13

by stacking the corresponding residual-stream rows, define $\mu_{s,t}^{(\ell,-f)} = n_t^{-1} \sum_{i \in I_t^{-f}} h_{s,t,i}^{(\ell)}$, and let $\mathbf{1} \in \mathbb{R}^{n_t}$ be the all-ones vector.

For side s, center the training-fold matrix separately within each task,

$$C_{s,t}^{(\ell,-f)} = X_{s,t}^{(\ell,-f)} - \mathbf{1}\mu_{s,t}^{(\ell,-f)\top}, \tag{4}$$

and form the equal-task pooled matrix

$$\widetilde{C}_s^{(\ell,-f)} = \mathrm{concat}_{t=1}^{T} \left(C_{s,t}^{(\ell,-f)} / \sqrt{n_t} \right). \tag{5}$$

The $1/\sqrt{n_t}$ factor removes direct weighting by task sample count in the pooled second moment; it does not equalize task covariance traces or eliminate the influence of higher-variance tasks. Here concat stacks task matrices by rows and Var denotes variance across rows. A singular value decomposition (SVD) computed in float32 yields PCA candidate directions u_j, indexed by j. If k is the temporary PCA retention dimension selected to attain the pooled variance target $\eta = 0.95$, u_j is retained only when every task satisfies

$$\gamma_{t,j} = \frac{\mathrm{Var}(C_{s,t}^{(\ell,-f)} u_j)}{\sum_{j'=1}^{k} \mathrm{Var}(C_{s,t}^{(\ell,-f)} u_{j'})} \geq 0.1/k. \tag{6}$$

The quantity $\gamma_{t,j}$ is the fraction of task t variance along candidate direction j within the temporary PCA space. The 95% pooled-variance target defines the candidate dimension k before filtering; the retained task-supported dimension K is determined afterward, so the filtered space need not retain 95% of pooled variance. With $\tau = 0.10$, each retained direction must contribute at least 10% of that task's average per-direction variance within the k-dimensional candidate space, not 10% of its total variance. The capped fixed-choice control construction uses at most $K_{\max} = 48$ retained directions. When more pass, the 48 strongest by $\min_t \gamma_{t,j}$ are kept and returned to pooled-PCA order; when fewer pass, the actual dimension is used without padding. The uncapped cross-regime construction retains every direction that passes the same support rule, $\gamma_{t,j} \geq \tau/k$. The capped basis therefore need not coincide with the first equally many uncapped directions. The primary pooled-variance target $\eta = 0.95$ and cross-task support fraction $\tau = 0.10$ are fixed a priori, independently of intervention outcomes. Training-fold activation variance defines this task-balanced reference space and order, not PCA of the paired difference or optimization of held-out transfer. Variance support need not imply shared semantics. The fixed-basis task-omission check changes only endpoint aggregation, not basis fitting (Appendix B.5, Fixed-Basis Task Omission).

Table A.3: **Quantization conditions.** Calibration is a paired design axis for GPTQ/AWQ and is not applicable to RTN.

Method	W/A	Group	Calibration	Conditions
GPTQ	4/16	128	C4, WikiText	10
AWQ	4/16	128	C4, WikiText	10
RTN	4/16	128	none	5

A.3 Transfer Criteria and Controls

Paired-decision transfer and stable-row disruption distinguish transfer of changed decisions from disturbance of agreeing ones. For row i, site ℓ, and held-out fold $f(i)$, define

$$\delta_i^{(\ell)} = h_{\mathrm{FP},i}^{(\ell)} - h_{\mathrm{Q},i}^{(\ell)}, \qquad c_i^{(\ell)} = B^{(\ell,-f(i))}\big(B^{(\ell,-f(i))}\big)^\top \delta_i^{(\ell)}. \tag{7}$$

Only the input residual stream of the selected transformer block at the final non-padding prediction position is changed. In separate forward runs, the same component is added to the quantized state and subtracted from the full-precision state; bidirectional transfer requires both runs to match their respective paired predictions on the same row. No scalar coefficient is fitted. The complete paired-difference intervention uses δ; its success rate is the complete paired-difference transfer rate. For the following shorthand, fix ℓ and $f(i)$ and suppress their indices.

Let $\mathcal{F} = \{i : \hat{y}_{\mathrm{FP},i} \neq \hat{y}_{\mathrm{Q},i}\}$ be naturally changed rows within a task–condition–site cell. Write $v = \{v_i\}_{i\in\mathcal{F}}$, $c = \{c_i\}_{i\in\mathcal{F}}$ and $\delta = \{\delta_i\}_{i\in\mathcal{F}}$ for row-indexed intervention collections. The cell-level rates for the three transfer criteria are

$$R_{\mathrm{one}}(v) = \Pr[\widetilde{y}_{\mathrm{Q},i}(v_i) = \hat{y}_{\mathrm{FP},i} \mid i \in \mathcal{F}], \tag{8}$$

$$R_{\mathrm{rev}}(v) = \Pr[\widetilde{y}_{\mathrm{FP},i}(v_i) = \hat{y}_{\mathrm{Q},i} \mid i \in \mathcal{F}], \tag{9}$$

$$R_{\mathrm{bi}}(v) = \Pr[\text{both transfer events} \mid i \in \mathcal{F}]. \tag{10}$$

Here $\widetilde{y}_{\mathrm{Q},i}(v_i)$ follows addition to the quantized state and $\widetilde{y}_{\mathrm{FP},i}(v_i)$ follows subtraction from the full-precision state. Bidirectional transfer requires both events on the same row. Appendix A.4.1 specifies how these cell-level one-way, reverse one-way and bidirectional rates are aggregated.

Stable-row disruption is the corresponding decision-change rate on $\mathcal{A} = \{i : \hat{y}_{\mathrm{FP},i} = \hat{y}_{\mathrm{Q},i}\}$. Transfer measures alignment with the paired model's decision; its accuracy effect depends on which model was correct before intervention.

Let $B_{\mathrm{Q}}^{(\ell,-f)}$ denote the basis fitted from quantized training-fold states by the same out-of-fold task-supported construction; after fixing ℓ and f, write it as B_{Q}. Write $v_{\mathrm{amb}} = \{v_{\mathrm{amb},i}\}_{i\in\mathcal{F}}$ for the row-wise ambient random interventions. It, the complement $\delta_i - c_i$, and the component $B_{\mathrm{Q}}B_{\mathrm{Q}}^\top\delta_i$ are rescaled row by row to match $\|c_i\|_2$. Ambient random bases have the same fold-specific dimension and use seeds 11, 23, and 47. The complete paired difference is not norm matched. For transfer criterion E, the relative task-supported advantage is computed after aggregating absolute rates:

$$A_{\mathrm{rel},E} = \frac{R_E(c) - R_E(v_{\mathrm{amb}})}{R_E(\delta) - R_E(v_{\mathrm{amb}})}. \tag{11}$$

The ratio is reported as NA when the denominator is below 0.05. It is not clipped and may exceed one.

A.4 Ordered Prefixes, Removal, and Statistical Aggregation

Ordered prefixes measure accumulation within the task-supported space; removal tests dependence on leading directions.

For retained basis $B = [b_1, \ldots, b_K]$, where K is the task-supported basis dimension, define $c_{i,r} = B_{1:r}B_{1:r}^\top\delta_i$. For $E \in \{\mathrm{one}, \mathrm{rev}, \mathrm{bi}\}$, define $R_E(r) := R_E(\{c_{i,r}\}_{i\in\mathcal{F}})$, $R_{E,\mathrm{supp}} := R_E(\{c_i\}_{i\in\mathcal{F}})$, and $\mathrm{Coverage}_E(r) = R_E(r)/R_{E,\mathrm{supp}}$; the intervention spectrum is this Coverage curve over ordered prefix size r. The capped companion one-way and bidirectional analyses use $r \in \{1, 2, 4, 8, 16, 32, 48\}$. Their random prefixes rotate the capped basis, retain r directions, and

match projection norms row by row using seeds 101, 211, and 307. At ordered prefix size 48, the rotation spans the complete capped basis, so that companion's random control is recorded as NA.

The uncapped fixed-choice grid is $r \in \{1, 2, 4, 8, 16, 32, 48, 64, 96, 128, 256\}$. A prefix is evaluated only when every fold has at least r supported directions; the fold-local complete task-supported endpoint is evaluated separately from this grid. The full-vocabulary uncapped grid is $r \in \{1, 2, 4, 8, 16, 32, 48, 64, 96, 128, 256, 512, 1024, 2048, 3072, 4096, 5120\}$.

The fixed-choice random points in Figure 4 belong to the primary uncapped construction, not the capped companion. At prefixes 8, 16 and 32, leading columns of a Gaussian-QR rotation of the complete uncapped basis supply a rank-matched subspace; its projected paired difference is rescaled to the same row's ordered-prefix patch norm. Each rotation is shared within model, site, fold and seed across tasks and quantization conditions. Seeds 101, 211 and 307 are averaged within task–condition cells before calibration, equal-task and five-checkpoint aggregation. Random and ordered rates use the same naturally changed population and the same uncapped complete task-supported denominator. Full-vocabulary controls are specified in Appendix C.3.1.

Fixed-choice removal evaluates ordered prefix sizes 16 and 32; the full-vocabulary analysis additionally includes 64 (Appendix C.3.2). The true remainder $\delta_i - c_{i,r}$ is evaluated without rescaling. A matched control removes a random-within-task-supported component with the same per-row removed norm. Writing $\delta = \{\delta_i\}_{i\in\mathcal{F}}$, $c_r = \{c_{i,r}\}_{i\in\mathcal{F}}$, and $\tilde{c}_r^{\text{matched}} = \{\tilde{c}_{i,r}^{\text{matched}}\}_{i\in\mathcal{F}}$, for $E \in \{\text{one}, \text{bi}\}$,

$$L_E^{(r)} = R_E(\delta) - R_E(\delta - c_r), \qquad \text{Spec}_E^{(r)} = R_E(\delta - \tilde{c}_r^{\text{matched}}) - R_E(\delta - c_r). \tag{12}$$

$L_E^{(r)}$ is the complete paired-difference transfer-rate loss, while $\text{Spec}_E^{(r)}$ is removal specificity relative to the matched-random removal control. This removal comparison is distinct from the relative task-supported advantage $A_{\text{rel},E}$.

The matched control fixes the rank and row-wise norm of the removed component. It does not match or restore the norm of the remaining intervention, so removal specificity is interpreted as an operational contrast under this removal operator.

Table A.4 summarizes the intervention objects. When a projection subscript is a basis matrix, it denotes projection onto that matrix's column span. Each vector is added to the quantized state or subtracted from the full-precision state in separate runs. A shared name such as "prefix 16" does not make capped and uncapped constructions interchangeable.

A.4.1 Aggregation, Uncertainty, and Missingness

Task-balanced transfer rates are computed within task from concatenated held-out predictions; fold-specific rates are not averaged. GPTQ/AWQ calibration variants are then averaged within checkpoint and task, and the T tasks are equally macro-averaged. All five checkpoint values remain visible, and each quantizer is summarized separately over this evaluated checkpoint set. Condition-level and row-pooled results are labeled separately from this primary aggregation.

The earlier fixed-choice ordered-prefix and removal intervals are cell-level estimates: they use 5,000 paired resamples of naturally changed rows within each task–condition cell, percentile 95% confidence intervals (CI), and seed 4242 with deterministic cell offsets.

For the GPTQ full-vocabulary anchor, the source-grouped protocol uses 5,000 paired bootstrap replicates (seed 20260821) with checkpoints fixed; all four GSM8K positions from a problem share a fold and bootstrap cluster. The same construction is applied to saved AWQ and RTN outcomes for top-16 one-way and bidirectional Coverage and removal specificity (Table C.4). Resamples use shared multiplicities across conditions and quantizers; GPTQ and AWQ average two calibration variants per checkpoint, while RTN has one calibration-free condition. Coverage divides the aggregate prefix transfer rate by the aggregate complete task-supported transfer rate within each resample.

Absolute transfer is zero when a nonempty changed-decision population has no successful transfers. Conditional rates are NA for empty populations; Coverage is NA when its reference rate is zero. The relative-advantage denominator guard is specified in Appendix A.3. Unavailable prefix sizes are also NA. Depth diagnostics use the mean within-block Spearman correlation across three sites

Table A.4: **Intervention inputs and norm treatment.** Indices for site and fold are suppressed. The primary spectrum and its reference retain native projection magnitudes; separately matched geometry comparisons do not change that operator.

Comparison	Input component	Norm / reference
Primary prefix	$P_{B_{1:r}}\delta_i$	Native; reference $P_B\delta_i$
Complete task-supported intervention	$P_B\delta_i$	Native complete supported space
Plain-PCA endpoint	$P_{\mathrm{PCA}_{95}}\delta_i$	Native; fold-local 95%-variance space
Complete paired-difference intervention	δ_i	Native complete difference
Local projection	$P_{\mathcal{S}_i}\delta_i$	Native; own complete paired-difference transfer rate
Local removal	$c_{i,16}^{\mathrm{cap}} - P_{\mathcal{S}_i}c_{i,16}^{\mathrm{cap}}$	Unrestored or restored to $\|c_{i,16}^{\mathrm{cap}}\|$
Fisher-16 projection	$P_{F_{16}}\delta_i$	Original or matched to uncapped $\|c_{i,16}\|$
Fisher removal	$c_{i,16} - P_{F_{95}}c_{i,16}$	Unrestored or restored to $\|c_{i,16}\|$
Primary ordered removal	$\delta_i - c_{i,r}$	Remaining norm unconstrained
Candidate restriction	Fixed patch, forward run and logits	Only candidate readout changes; own endpoint

and 100,000 within-block permutations with plus-one correction. Folds are used for cross-fitting and calibration variants for sensitivity analysis; neither is treated as independent replication.

For the primary fixed-choice prefix-16 Coverage and the Fisher-16 Matched minus uncapped prefix-16 contrast, an additional paired source bootstrap uses 5,000 attempted draws (seed 20260923). Within each task, one source-ID multiplicity is shared across checkpoints, calibration conditions, quantizers and both interventions; the four GSM8K positions share their problem's multiplicity. Original predictions fix each cell's changed-decision population. Each draw recomputes its conditional transfer rates, averages calibrations, tasks and the five fixed checkpoints, and then forms Coverage or the paired rate difference. Percentile 95% intervals condition on these checkpoints, fitted bases or geometry and source-ID universes; they do not include representation refitting, Fisher sampling-seed variation or checkpoint-population uncertainty. An empty changed-decision cell makes the affected quantizer's aggregate transfer rates and their derived Coverage or paired difference undefined in that draw; the other quantizers remain separately evaluable. This leaves 5,000/4,972/4,999 valid fixed-choice draws for GPTQ/AWQ/RTN and all 5,000 Fisher draws. No undefined rate is replaced by zero. Tables B.4 and C.13 report these intervals separately from the earlier cell-level intervals.

The complementary geometry protocols are specified with their results: Appendix B.4 defines the row-local readout span, projection of the complete paired difference, and removal from the capped prefix-16 component. Appendix C.5 gives the one-sample-per-training-row Fisher estimator and intervention comparisons.

B Fixed-Choice Additional Results

B.1 Decision Changes and Task-Level Heterogeneity

Decision replacement remains frequent in checkpoint–task cells with small or positive accuracy change (Figure B.1): 31 of 125 task–condition cells have $\Delta\mathrm{Acc} > 0$. Changed decisions include correct-to-wrong, wrong-to-correct, and wrong-to-different-wrong transitions. The task-level breakdown shows why replacement must be measured separately from net accuracy, including conditions in which quantization improves accuracy.

B.2 Directional Organization and Control Analyses

Task-supported-minus-ambient contrasts replicate across quantizers and transfer criteria, with larger aggregate advantages at later sites (Table B.1). Non-positive task–condition cells occur mainly at the earliest site; final-site comparisons and depth-wise exception counts retain this heterogeneity



(a) Accuracy change after quantization.



(b) Fraction of examples whose decision changes.

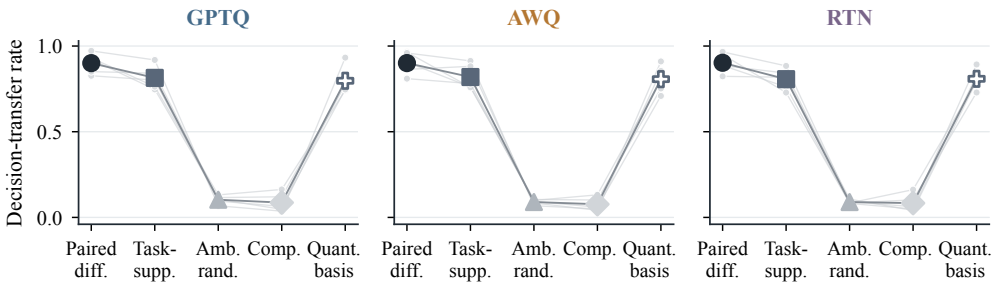
Figure B.1: **Descriptive decision changes by task and checkpoint.** Cells average calibration variants within checkpoint–task pairs; checkpoint summaries weight the five tasks equally. Task codes ARC/HS/PI/S2/TQ denote ARC-C/HellaSwag/PIQA/SQuAD2/TruthfulQA.

(Tables B.2 and B.3). For GPTQ, task-supported minus ambient-random transfer is 22.4, 59.1 and 71.1 percentage points one way, and 14.6, 48.7 and 66.5 bidirectionally, from the earliest to final sampled site. These contrasts use unrounded rates. The depth breakdown distinguishes an aggregate final-site advantage from an advantage in every task–condition cell.

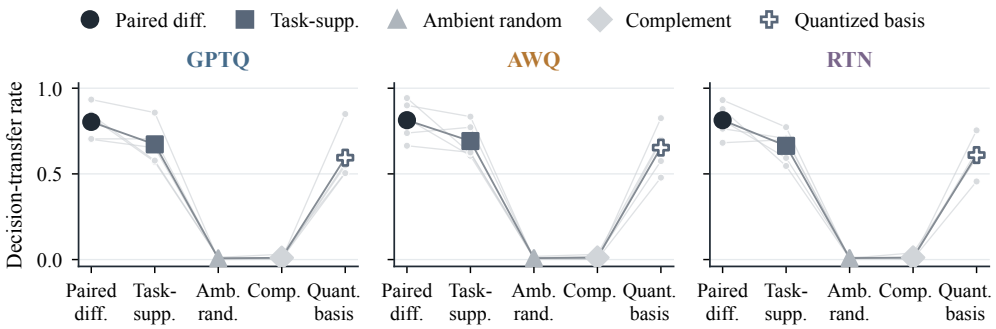
The fixed-choice control construction is capped at 48 directions; means follow calibration-within-checkpoint/task, equal-task and equal-checkpoint aggregation. Table B.1 uses this control run's own complete paired-difference transfer rate at every site. Local-readout and uncapped analyses retain their separate run-specific references and ratio denominators (Appendix D).

At the final site, task-supported components carry much of the complete paired-difference transfer, whereas fold-dimension-matched, rowwise norm-matched ambient random components carry little (Figure B.2). Complementary-component results and results from bases fitted on quantized states show that the pattern persists across matched intervention norms and both basis constructions. Similar aggregate final-site rates show that the observed directional advantage extends to both tested precision-state constructions.

The early weak or non-positive task-supported-minus-ambient contrasts coexist with stable-row disruption and unavailable relative comparisons (Figure B.3). The three sampled sites characterize relative efficacy across depth; finer localization lies outside this sampling resolution.



(a) One-way transfer.



(b) Bidirectional transfer.

Figure B.2: **Task-supported directions carry more transfer.** Ambient-random controls match fold-specific basis dimension and row-wise intervention norm. Pale lines and points show five checkpoints; larger connected markers show quantizer means, with component shapes defined in the legend. Directional organization holds for bases fitted from either precision state.

B.3 Intervention Spectra and Removal

The uncapped analysis tests fixed-choice concentration in the complete task-supported basis. The capped control construction shows the same within-regime ordering of transfer advantages; Table B.11 summarizes its exceptions.

B.3.1 Uncapped Task-Supported Directions

Removing the cap tests whether rapid fixed-choice accumulation is a direct consequence of limiting the basis. The uncapped prefix curves separate concentration from cap saturation. Prefix 16 attains 91.8–92.1% one-way and 87.0–87.7% bidirectional Coverage across quantizers under the representation-derived ordering. Transfer continues to increase beyond prefix 48, and the complete task-supported transfer rate remains below the complete paired-difference transfer rate. High short-prefix Coverage is therefore relative to the task-supported endpoint, not the complete paired difference.

The comparison uses the same rows, folds and sampled sites as the capped controls; Appendix A.4 defines the evaluated grid. The uncapped construction retains every full-precision task-supported direction that passes the fold-specific criterion, without padding. Task-supported basis dimensions range from 354 to 682 across sampled sites. Across the five checkpoints and their five held-out folds, the final-site uncapped full-precision bases contain 584 to 682 supported directions (mean 624.080). Figure 5a shows these fold-level dimensions. Dimensions of bases fitted from quantized states are provided in the derived data, while intervention prefixes are defined by the full-precision basis.

Table B.1: **Fixed-choice directional controls across depth.** Transfer proportions on changed decisions average five checkpoint task macros. The basis is capped at 48; ambient random matches its dimension and row norm. Each complete paired-difference transfer rate comes from the same run and population as its control entries.

Quantizer	Normalized depth	Paired difference	Task-supported	Ambient random	Complement	Quantized-state basis
One-way transfer						
GPTQ	0.52	0.412	0.310	0.086	0.106	0.302
	0.77	0.798	0.694	0.102	0.130	0.680
	1.00	0.9003	0.814	0.103	0.086	0.797
AWQ	0.52	0.417	0.322	0.071	0.093	0.328
	0.77	0.763	0.669	0.095	0.098	0.667
	1.00	0.901	0.819	0.090	0.077	0.807
RTN	0.52	0.412	0.315	0.078	0.089	0.310
	0.77	0.800	0.700	0.094	0.101	0.679
	1.00	0.902	0.807	0.090	0.083	0.809
Reverse one-way transfer						
GPTQ	0.52	0.366	0.279	0.061	0.079	0.268
	0.77	0.763	0.663	0.076	0.092	0.624
	1.00	0.8803	0.790	0.071	0.068	0.726
AWQ	0.52	0.384	0.295	0.059	0.071	0.295
	0.77	0.761	0.656	0.069	0.086	0.641
	1.00	0.897	0.815	0.077	0.082	0.778
RTN	0.52	0.376	0.297	0.063	0.080	0.284
	0.77	0.760	0.693	0.077	0.087	0.626
	1.00	0.884	0.785	0.071	0.070	0.731
Bidirectional transfer						
GPTQ	0.52	0.209	0.152	0.006	0.011	0.139
	0.77	0.625	0.496	0.009	0.016	0.451
	1.00	0.8035	0.672	0.007	0.010	0.595
AWQ	0.52	0.217	0.169	0.007	0.006	0.173
	0.77	0.605	0.475	0.008	0.006	0.452
	1.00	0.813	0.692	0.008	0.011	0.653
RTN	0.52	0.229	0.166	0.007	0.008	0.159
	0.77	0.636	0.522	0.006	0.006	0.451
	1.00	0.813	0.664	0.009	0.012	0.610

The final-site GPTQ complete task-supported transfer rates are 86.26%/73.57% one-way/bidirectional for fixed-choice and 69.95%/49.66% for full-vocabulary token decisions, using calibration-within-checkpoint and equal-task aggregation over five checkpoints.

Table B.4 reports source-paired uncertainty for primary uncapped prefix-16 Coverage, conditional on the five checkpoints and fitted bases. The capped alternate-fold comparison uses a separate construction (Appendix Table B.7).

For final-site GPTQ, the fixed-choice macro curve first reaches 90% Coverage at tested prefix size 16 one way and 48 bidirectionally. These macro-curve crossings are distinct from the condition-level counts below. Two condition-level summaries count observed 90%-coverage crossings (Table B.5). The pooled-event version weights tasks by their changed-decision counts, whereas the equal-task version gives all five tasks equal weight within each condition. Both keep C4 and WikiText as separate calibration conditions and are not checkpoint-population summaries. Table B.5 lists every final-site condition that remains below threshold, including AWQ Qwen-14B with WikiText calibration under both summaries.

A crossing is the first evaluated grid prefix reaching the threshold; censored means no crossing was observed on the tested grid through prefix size 256, without bounding unmeasured prefixes. All conditions not listed in Table B.5, Block B, reach the bidirectional 90% target by that prefix under both summaries.

Table B.2: **Final-site task–condition control summary.** Counts retain calibration-specific checkpoint–task cells and use strict inequalities; equalities, if any, are not counted. Ambient denotes the ambient-random control. FP-basis and Q-basis are task-supported bases fitted to full-precision and quantized training-fold states, respectively.

Quantizer	Criterion	N	Task-supported > ambient	Task-supported > complement	Q-basis > FP-basis
GPTQ	One-way	50	50	50	19
	Reverse	50	50	50	10
	Bidirectional	50	50	50	11
AWQ	One-way	50	50	50	17
	Reverse	50	50	50	13
	Bidirectional	50	50	50	16
RTN	One-way	25	25	25	11
	Reverse	25	25	25	1
	Bidirectional	25	25	25	7

Table B.3: **Task–condition control exceptions across normalized depth.** Counts retain calibration-specific checkpoint–task cells; ambient denotes the ambient-random control. FP-basis and Q-basis are task-supported bases fitted to full-precision and quantized training-fold states, respectively.

Quantizer	Depth	N	Task-supported ≤ ambient		Q-basis > FP-basis	
			One-way	Bidirectional	One-way	Bidirectional
GPTQ	0.52	50	7	9	18	12
	0.77	50	0	0	18	15
	1.00	50	0	0	19	11
AWQ	0.52	50	10	14	21	17
	0.77	50	0	1	19	13
	1.00	50	0	0	17	16
RTN	0.52	25	4	6	7	6
	0.77	25	0	1	4	2
	1.00	25	0	0	11	7

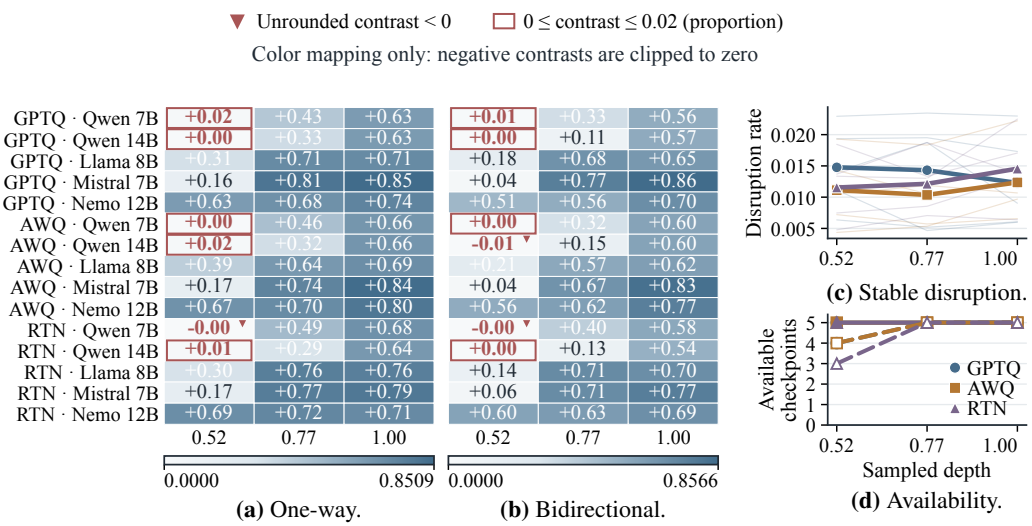
Table B.4: **Paired uncertainty for fixed-choice prefix-16 Coverage.** Final-site aggregate rate ratios (%), with conditional source-bootstrap 95% intervals. Prefix and complete task-supported transfer rates share the same resampled source weights; checkpoints and fitted bases remain fixed (Appendix A.4.1).

Quantizer	One-way	Reverse	Bidirectional
GPTQ	92.06 [90.53, 93.61]	92.81 [91.54, 94.06]	87.66 [85.70, 89.62]
AWQ	91.81 [90.19, 93.47]	92.64 [91.41, 93.80]	87.66 [86.01, 89.28]
RTN	92.10 [90.27, 94.12]	92.66 [91.06, 94.34]	87.05 [84.91, 89.25]

In Figure 5, fold-dimension boxes show medians, IQRs and 1.5-IQR whiskers. Final-site curves average calibrations within checkpoint and weight tasks equally; bands span checkpoint minima and maxima, and the cap-48 guide marks prefix size 48; it does not identify the separately selected capped basis. Crossing counts keep the 25 calibration conditions separate and distinguish pooled-event from equal-task aggregation.

The supplemental final-site uncapped one-way prefix-16 and bidirectional prefix-32 removal checks identify one task–condition cell with a non-positive interval lower bound for either complete paired-difference transfer-rate loss or removal specificity. Table B.6 reports both intervals; low event count is not used to exclude cells.

The fixed-choice removal contrasts in Table 3 use the final-site uncapped top-16 removal analysis. In the exception table, n_{chg} counts naturally changed rows.



(a) One-way. **(b)** Bidirectional. **(c)** Stable disruption. **(d)** Availability.

Figure B.3: **Directional controls.** (a,b) Task-supported minus ambient-random transfer-rate differences in proportions ($0.02 = 2$ percentage points). Values are rounded to two decimals; -0.00 denotes a negative value that rounds to zero. Triangles indicate unrounded negative point estimates, not statistical significance; red outlines mark $0 \leq$ contrast ≤ 0.02. Only the color mapping clips negatives to zero. (c) Five pale checkpoint traces per quantizer and dark quantizer means for disruption on initially agreeing rows. (d) Number of checkpoints, out of five, with valid relative-advantage estimates $A_{\mathrm{rel},E}$: the denominator $R_E(\delta) - R_E(v_{\mathrm{amb}})$ must be at least 0.05. Solid/filled denotes one-way; dashed/open denotes bidirectional. A zero count means no valid estimates, not zero transfer.

Table B.5: **Fixed-choice crossings on the evaluated grid.** A counts first observed crossings; B lists all final-site bidirectional 90% non-crossings under either aggregation. NR means not reached on the tested grid through prefix size 256, not a bound on unmeasured prefixes.

A. Crossing-count summary

	Pooled events				Equal task / condition			
Target	≤ 16	≤ 32	≤ 256	NR	≤ 16	≤ 32	≤ 256	NR
One-way 80%	25/25	25/25	25/25	0/25	25/25	25/25	25/25	0/25
One-way 90%	17/25	17/25	25/25	0/25	15/25	16/25	25/25	0/25
Bidirectional 80%	17/25	17/25	25/25	0/25	17/25	17/25	25/25	0/25
Bidirectional 90%	15/25	15/25	24/25	1/25	12/25	15/25	22/25	3/25

B. Conditions with tested-grid non-crossings

Quantizer	Checkpoint	Calibration	Task-supported basis dimension	Pooled bidirectional 90%	Equal-task bidirectional 90%
AWQ	Qwen-14B	WikiText	670	NR	NR
	Qwen-7B		603	256	NR
GPTQ	Qwen-7B	C4	603	256	NR

B.4 Local Readout Geometry

The fixed-choice interface defines a compact candidate-logit span. For each example i, let m_i denote its fixed-choice candidate count. Exact differentiation through the full-precision continuation gives the local span $\mathcal{S}_i$, with $\mathrm{rank}(\mathcal{S}_i) \leq m_i - 1$, evaluated at the paired full-precision state. Its final-site mean rank is 2.406. Mean subspace overlap with the capped top-16 basis rises from 0.054 to 0.148 to 0.216 across the sampled sites. For each row, overlap sums squared inner products between the two orthonormal bases and divides by the smaller rank; a zero-rank span contributes zero. These full-precision geometry summaries use no behavioral-flip filter: duplicate quantization-condition

Table B.6: **Supplemental uncapped removal exception.** Final-site bidirectional prefix-32 estimates are proportions with paired-bootstrap 95% CI. The transfer-rate loss is measured relative to the complete paired-difference intervention. This cell is the sole interval exception in the one-way prefix-16 and bidirectional prefix-32 checks defined above.

Quantizer	Checkpoint	Calibration	Task	Metric	n_{chg}	Estimate [CI]
GPTQ	Qwen-14B	WikiText	ARC-C	Complete transfer-rate loss	10	0.200 [-0.300, 0.700]
				Removal specificity		0.200 [-0.233, 0.600]

copies of the same geometry are removed, folds are averaged within model, and models are equally weighted. Transfer rates retain their separate naturally changed population and task/calibration weights. This diagnostic differs from projected energy and cross-example squared-axis alignment. Figure B.4a–b keeps tasks separate.

Local projection uses $v_i^{\mathrm{local}} = P_{\mathcal{S}_i}\delta_i$, while local-span removal uses $v_i^{\mathrm{remove}} = c_{i,16}^{\mathrm{cap}} - P_{\mathcal{S}_i}(c_{i,16}^{\mathrm{cap}})$, with norm-restored variants rescaled to $\|c_{i,16}^{\mathrm{cap}}\|$. The saved local worker returns zero when the remainder norm is at most 10^{-12} or the target norm is nonpositive; otherwise it multiplies by target norm divided by remainder norm. The starting component comes from the capped fixed-choice basis without initial scaling, and is distinct from the uncapped spectrum's prefix. Figure B.4 reports absolute transfer on naturally changed rows. On initially agreeing rows, local projection of the complete paired difference produces quantized-model disruption rates of 0.009, 0.008 and 0.008 for GPTQ, AWQ and RTN, respectively (fractions, with the same task/calibration/checkpoint aggregation). Across quantizers, norm-restored removal leaves 0.09–0.85% of the capped top-16 bidirectional reference, using means of within-checkpoint remainder/reference ratios (Figure C.4). The local-projection ratio uses this analysis's complete paired-difference transfer rate, not task-supported Coverage or top-16 retention. Projection and removal test the same row-local span using different input vectors. Together with its compact rank, these results support row-local output-sensitive geometry; cross-example concentration is tested separately.

The uncapped spectrum and capped local-readout ablations use different task-supported constructions. Appendix B.5.5 reports the complementary scalar activation diagnostics and plain-PCA prefix-16 energy fraction.

B.5 Robustness, Calibration Sensitivity, and Exceptions

Calibration, task weighting, fold assignment and basis construction test different sources of sensitivity; exceptions and activation diagnostics delimit the interpretation.

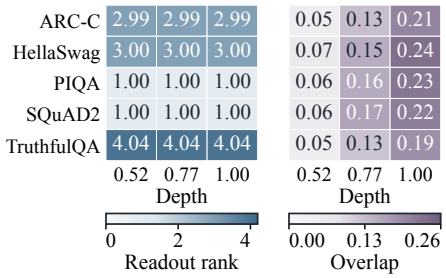
Fixed-Basis Task Omission. With the fitted basis held fixed, omitting any one task from aggregation leaves a positive mean within-block Spearman correlation between sampled depth and task-supported-minus-ambient transfer, for both one-way and bidirectional criteria and every quantizer. This tests task-weighting sensitivity of the directional-advantage trend, not refitting or generalization to an unseen task.

B.5.1 Calibration Sensitivity

Many paired calibration outcomes are close, but some differ materially, including GPTQ Qwen-14B and Nemo behavior results. GPTQ and AWQ are each evaluated with separate C4 and WikiText calibration sets; Figure B.5 compares them within the same checkpoint–task cell. This paired comparison measures sensitivity to calibration data within a checkpoint. Points near the equality diagonal indicate similar outcomes for the two calibration sets; larger departures identify condition-specific sensitivity. The figure reports behavioral discrepancies and absolute transfer, not the sensitivity of complete normalized spectra.

B.5.2 Cross-Fit Sensitivity

Prefix-16 absolute transfer varies little under the evaluated deterministic fold assignments. This check uses the capped task-supported companion for fixed-choice and the uncapped construction for



(a) Readout rank. **(b)** Overlap.

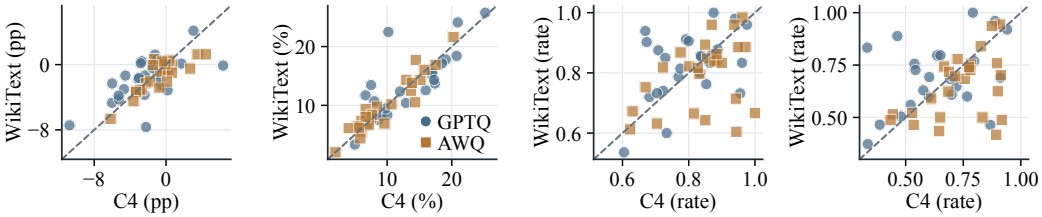
Intervention	GPTQ	AWQ	RTN
Full Δ	90.05	90.16	90.32
Capped 16	79.10	78.61	79.53
Projection	85.09	85.48	85.98
Removal	17.30	13.84	15.41
Restored	17.71	14.18	15.97

(c) One-way transfer (%).

Intervention	GPTQ	AWQ	RTN
Full Δ	80.45	81.50	81.30
Capped 16	64.02	65.11	64.15
Projection	62.22	67.92	65.45
Removal	0.06	0.53	0.05
Restored	0.08	0.55	0.05

(d) Bidirectional transfer (%).

Figure B.4: **Local readout geometry and intervention transfer.** (a,b) Row-local rank and overlap with the capped prefix-16 basis. (c,d) Five-checkpoint mean absolute transfer (%). Full Δ: complete paired difference; Capped 16: capped companion prefix 16; Projection: $P_{\mathcal{S}_i}\delta_i$; Removal: local-span component removed from capped prefix 16; Restored: that remainder rescaled to its starting prefix norm. Checkpoint min–max ranges are retained in the released `appendix_h_readout_ablation.csv`; only means are displayed.



(a) Accuracy shift. **(b)** Disagreement. **(c)** One-way. **(d)** Bidirectional.

Figure B.5: **Calibration sensitivity of behavior and transfer.** Points pair GPTQ/AWQ C4/WikiText checkpoint–task results: (a)–(b) descriptive accuracy changes (points) and changed decisions (%); (c)–(d) absolute transfer rates. Dashed diagonals mark equal calibration results.

full vocabulary. The rank-16 construction rule and patch procedure are held fixed; each assignment uses its corresponding cross-fitted bases. Seeds 4243 and 4244 reassign 79.7–80.4% of rows relative to 4242 across the four regime–seed comparisons. At the final site, the largest absolute change across one-way, reverse and same-row bidirectional macro rates is 1.06 percentage points, with mean absolute change 0.24 points (Table B.7). These are absolute prefix-transfer sensitivity results, not complete spectra or independent model replications; verification of the alternate-fold fitting procedure has the scope stated in Appendix D.

Final-site values are unweighted means over condition–task blocks: 125 fixed-choice and 75 full-vocabulary blocks per assignment. This capped fixed-choice baseline differs from the uncapped EventRetention construction in Table C.16.

B.5.3 Plain Pooled-PCA Construction

The plain pooled-PCA comparison tests basis construction at the final sampled block input of each checkpoint. It uses out-of-fold, task-centered, equal-task full-precision PCA without the every-task support filter, retaining the fold-local 95%-variance space. The measured prefix grid is $\{1, 2, 4, 8, 16, 32, 48, 64, 96, 128, 256, 512, 1024, 2048\}$, with a prefix available only when its fold's dimension is at least that large. Complete endpoints project onto each fold's full retained PCA space, whose dimensions range from 999–1,191 in fixed choice and 2,447–3,140 in full vocabulary. These endpoints differ from filtered task-supported and complete paired-difference interventions, and are not defined by prefix 2,048.

Figure B.6 shows rapid fixed-choice and broad full-vocabulary accumulation under this alternative construction. At prefix 32, one-way Coverage is 90.5–93.1% in fixed choice and 17.6–19.6% in full vocabulary across quantizer summaries (Table B.8). At prefix 32, fixed-choice Coverage is higher in every checkpoint–quantizer comparison for all three transfer criteria. Projection-operator

Table B.7: **Prefix-16 absolute transfer under alternate fold assignments.** Values are percentages; fixed-choice uses the capped companion and full vocabulary the uncapped construction.

	Fixed-choice decision			Full-vocabulary token decision		
Fold seed	One-way	Reverse one-way	Bidirectional	One-way	Reverse one-way	Bidirectional
4242	79.05	77.33	64.52	9.91	10.11	2.19
4243	78.36	76.68	63.45	9.94	10.17	2.20
4244	78.97	77.26	64.34	9.92	10.14	2.19

comparison confirms that the separately stored plain-PCA prefix-16 basis and the first 16 complete-basis directions span numerically equivalent subspaces in all 50 regime/model/fold cells. Its plotted outcomes use the same complete-spectrum evaluation as the other prefixes.

For fixed-choice AWQ Mistral-7B, retaining rows whose full-precision and quantized baseline predictions both reproduce excludes 133/24,869 (0.53%) C4 rows and 138/24,869 (0.55%) WikiText rows; 24,736 and 24,731 remain, respectively. These exclusions affect evaluation only, not basis fitting, the primary uncapped spectrum or the controlled candidate-restriction analysis in Appendix C.4. The two affected entries are calibration conditions of one model. Prefix and complete-endpoint rates use the same retained population; exact row masks and per-task counts accompany the results.

Table B.9 separates excluded changed and agreeing rows using the frozen original predictions. The exclusion fraction among changed rows varies by task, so its fraction among all rows does not bound its effect on conditional transfer. Removing Mistral-7B from both AWQ regimes leaves the same four checkpoints: prefix-32 one-way Coverage is 92.13% for fixed-choice and 19.78% for full-vocabulary decisions, compared with 93.10% and 19.59% in the five-checkpoint summaries. Reverse and bidirectional four-checkpoint Coverage is 91.51%/86.56% for fixed choice and 20.24%/7.85% for full vocabulary. These are ratios of aggregate prefix and own-endpoint rates, with calibration-within-checkpoint/task, equal-task and equal-checkpoint weights. The contrast persists in this checkpoint-exclusion analysis; it does not estimate the omitted rows' intervention outcomes.

At prefix 1,024, fixed-choice rates use the four available checkpoints and the same subset for the denominator; prefix 2,048 is unavailable in fixed choice and remains below every full-vocabulary endpoint dimension. Missing results are not zero or clamped prefixes. The spectra measure endpoint-normalized accumulation without the support filter. Fold-specific prefix fractions are available in the accompanying summaries; no matched-fraction intervention is inferred from them.

Within-PCA Rank-32 Control. To test direction choice separately from patch magnitude at one prefix, Table B.10 compares ordered rank-32 projections with random rank-32 subspaces of the same fold-local retained PCA space. It covers both regimes, all five checkpoints and GPTQ/AWQ/RTN at the final sampled site. For each condition, fold and seed (1701, 1702, 1703), a Gaussian coefficient matrix is orthonormalized and mapped through the retained PCA basis; a second orthonormalization gives the random basis. Each random projection of δ_i is rescaled to that row's ordered-rank-32 patch norm. A target norm at most 10^{-12} returns zero; a nonzero target with a random projection at most this threshold is rejected. The supplied row-level audit reports neither branch on the evaluated rows and maximum relative norm error below 10^{-6}.

The control uses the same rows, folds, original predictions and rank eligibility as the ordered PCA evaluation, including its fixed replay-exclusion mask. We average calibration conditions within checkpoint/task, tasks equally, and then the five checkpoints within quantizer; random-seed means follow this hierarchy. Task-pooled summaries in the supplement use a different estimand. Seeds describe random-subspace variation, not additional checkpoint replications, and no confidence intervals are inferred from them. Appendix D describes the source and count checks.

Ordered transfer exceeds the seed-mean random transfer in every quantizer-level comparison, including reverse and same-row bidirectional transfer. This aggregate advantage is not uniform at task level: among 600 task–condition–seed comparisons, 11 one-way, 12 reverse and 15 bidirectional contrasts are nonpositive and remain included. This is a within-PCA rank-32 comparison, with

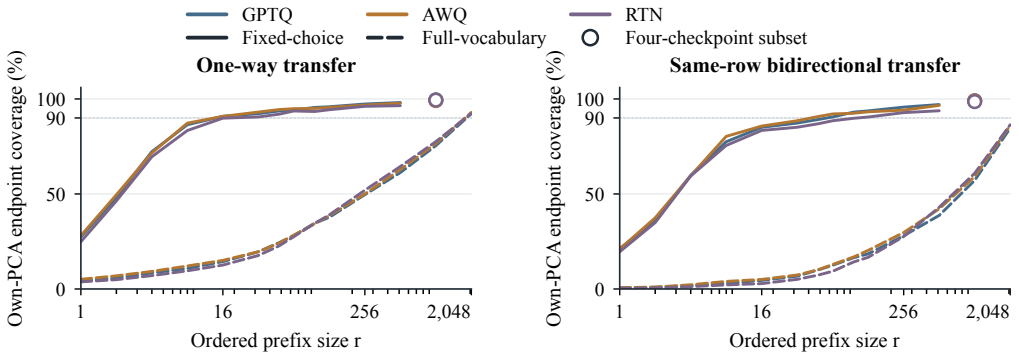


Figure B.6: **Final-site concentration under plain pooled-PCA.** Coverage divides the aggregate prefix transfer rate by the complete transfer rate in its fold-local 95%-variance PCA space on identical eligible rows. Colors denote quantizers; line styles denote regimes. Isolated hollow circles are the fixed-choice prefix-1,024 estimates from four available checkpoints; numerator and endpoint denominator use the same subset and weights. They are separate from the five-checkpoint curves, not additional endpoints.

Table B.8: **Final-site plain-PCA endpoints and prefix-32 Coverage.** Percentages use equal calibration, task and checkpoint weights within quantizer. F/R/BI denote one-way, reverse and same-row bidirectional criteria. Each complete endpoint uses its fold-local 95%-variance PCA space.

		Absolute complete endpoint (%)			Prefix-32 Coverage (%)		
Regime	Quantizer	F	R	BI	F	R	BI
Fixed-choice	GPTQ	87.35	84.46	75.66	92.27	92.24	87.31
	AWQ	86.89	86.63	76.51	93.10	92.77	88.53
	RTN	88.79	85.01	77.57	90.53	91.38	85.11
Full-vocabulary	GPTQ	70.06	68.88	49.82	19.60	19.13	6.74
	AWQ	70.83	70.06	51.74	19.59	19.38	7.36
	RTN	63.82	65.20	45.01	17.62	17.82	5.07

no ambient-endpoint or full alternative-order comparison; the ordered spectrum retains its native, prefix-dependent magnitudes.

B.5.4 Cell-Level Heterogeneity and Availability Limits

Table B.11 consolidates behavioral, control and threshold exceptions across their stated diagnostic populations. Negative-specificity counts use the strict < 0 rule, including near-zero floating-point negatives; they do not count statistically significant negative effects. These counts pool all sampled sites, whereas interval and availability limits use the final site. Depth is normalized; ambient denotes the ambient-random control and CI the confidence interval. Zero-count categories remain in the accompanying data; Tables B.2 and B.3 provide the final-cell and depth-wise breakdowns.

B.5.5 Conventional Activation Diagnostics

The diagnostics address two different questions: whether a scalar separates changed from unchanged decisions, and whether its magnitude co-varies with intervention transfer. We examine the paired state difference's Euclidean norm, relative norm (divided by the mean of the full-precision and quantized state norms), cosine drift, and projected energy fraction in the first 16 directions of a fold-local plain pooled-PCA basis. The energy fraction divides squared projected norm by squared paired-difference norm, with a numerical stabilizer; this basis omits the every-task support filter. These quantities are descriptive diagnostics, distinct from intervention-spectrum Coverage and EventRetention.

Table B.9: **Replay exclusions in the two affected PCA conditions.** AWQ Mistral-7B only. Changed/agreeing refers to frozen original predictions; the last column divides excluded changed rows by original changed rows. Both intervention numerator and own-endpoint denominator use the retained population.

Calibration	Task	Original rows	Original changed	Excluded changed	Excluded agreeing	Changed excluded (%)
C4	ARC-C	299	23	0	0	0.00
	HellaSwag	10,042	1,522	75	50	4.93
	PIQA	1,838	132	5	3	3.79
	SQuAD2-Ans.	11,873	483	0	0	0.00
	TruthfulQA-MC	817	113	0	0	0.00
WikiText	ARC-C	299	25	0	0	0.00
	HellaSwag	10,042	1,438	63	51	4.38
	PIQA	1,838	124	4	9	3.23
	SQuAD2-Ans.	11,873	727	0	0	0.00
	TruthfulQA-MC	817	113	5	6	4.42

Table B.10: **Ordered PCA directions exceed norm-matched random subspaces at rank 32.** Absolute transfer (%) on the canonical naturally changed population. Random entries average three seeds after the same calibration/task/checkpoint aggregation as Table B.8. The subspaces share rank and row-wise patch norm; F/R/BI denote one-way, reverse and same-row bidirectional transfer.

		F		R		BI	
Regime	Quantizer	Ordered	Random	Ordered	Random	Ordered	Random
Fixed-choice	GPTQ	80.60	14.67	77.90	10.61	66.06	1.61
	AWQ	80.89	14.14	80.37	11.42	67.73	2.02
	RTN	80.38	13.35	77.69	11.28	66.01	1.73
Full-vocabulary	GPTQ	13.73	7.21	13.18	6.94	3.36	1.11
	AWQ	13.87	7.56	13.58	7.45	3.81	1.15
	RTN	11.24	5.60	11.62	5.84	2.28	0.64

Table B.12 reports the final site. Discrimination is the mean within-block area under the receiver operating characteristic curve (AUROC) for naturally changed versus unchanged decisions. Transfer associations use each diagnostic's mean on naturally changed rows and the corresponding absolute one-way plain-PCA prefix-16 transfer rate. Blocks are condition–task–site combinations: 125 per site in the fixed-choice regime and 75 in the full-vocabulary regime. Equal block weighting gives GPTQ/AWQ/RTN weights of 40/40/20, distinct from separately reported quantizer summaries elsewhere. Partial Spearman correlations residualize both ranked variables on model, task, and quantizer indicators within regime and site. They describe adjusted associations, not effects of manipulating energy or norm.

Activation distances have above-chance mean discrimination at the final site, yet their associations with prefix transfer differ across regimes. The adjusted energy association is near zero for fixed-choice one-way transfer and positive for full-vocabulary transfer. For the latter, it is also positive for reverse and bidirectional transfer, at 0.504 and 0.468. The full-vocabulary task-supported prefix-16 baseline gives a similar adjusted one-way energy association of 0.422. This baseline is the uncapped top-16 task-supported intervention, not a predictive-Fisher projection. Energy therefore co-varies with measured transfer in this comparison. These associations do not separate the contributions of projected energy, intervention norm, and decision margin to transfer.

The diagnostic summaries cover all three sampled sites and all three transfer criteria. Raw-correlation intervals use 1,000 block-bootstrap resamples; they do not resample independent checkpoints, and no adjusted-correlation intervals are supplied. These correlations retain the original prefix-16 transfer targets; they are not recomputed against the separate complete-PCA run in Figure B.6. They do not test alternative-ordering spectra. Weight-reconstruction error is outside this comparison because no directly comparable frozen record is available.

Table B.11: **Low-count, negative, censored, and unavailable results.** Intervention, threshold and interval counts use the capped $K_{\max} = 48$ construction. Each row retains its stated diagnostic population; the supplemental uncapped exception is separate (Table B.6).

Category	Event / criterion	Setting	Count/N
Observed reversals	Positive ΔAcc	–	31/ 125
	Task-supported $\leq$ ambient	One-way, depth .52	21/ 125
	Task-supported $\leq$ ambient	Bidirectional, depth .52	29/ 125
	Task-supported $\leq$ ambient	Bidirectional, depth .77	2 / 125
	Negative removal specificity	One-way, $r = 16$	43/ 375
	Negative removal specificity	Bidirectional, $r = 32$	34/ 375
Capped-spectrum limits	80% not reached	One-way	4 / 375
	90% not reached	One-way	4 / 375
	80% not reached	Bidirectional	21/ 375
	90% not reached	Bidirectional	23/ 375
Interval / availability limits	CI lower ≤ 0 or NA	One-way, $r = 8$	4 / 125
	CI lower ≤ 0 or NA	Bidirectional, $r = 16$	2 / 125
	Specificity CI lower ≤ 0	One-way, $r = 16$	3 / 125
	Specificity CI lower ≤ 0	Bidirectional, $r = 32$	21/ 125

Table B.12: **Conventional diagnostics and final-site transfer.** Fixed/Full denote the two regimes; adjusted coefficients are descriptive point estimates without supplied intervals.

	Mean AUROC		Raw Spearman		Adjusted Spearman	
Diagnostic	Fixed	Full	Fixed	Full	Fixed	Full
Paired-difference norm	0.574	0.654	−0.439	0.107	−0.231	−0.194
Relative paired-difference norm	0.582	0.660	−0.437	−0.034	−0.175	−0.069
Cosine drift	0.583	0.661	−0.434	−0.030	−0.168	−0.076
Plain-PCA prefix-16 energy	0.614	0.492	0.213	0.531	−0.042	0.414

C Full-Vocabulary Additional Results

C.1 Decision Protocol and Decision Discrepancies

The full-vocabulary protocol measures fixed-context next-token decisions, not sequence-level answer accuracy. Each intervention row contains a fixed token prefix and target-token position; both models receive the same ground-truth prefix and select the top-scoring token over the complete vocabulary.

The evaluation contains 5,153 LAMBADA token decisions, 5,000 TriviaQA token decisions, and 5,264 GSM8K token decisions. The datasets contribute final-word, canonical-answer, and multi-position reasoning/final-answer targets, respectively.

Target positions follow a frozen character-level specification that is converted identically for every tokenizer: after tokenizing the complete frozen text without automatically inserted special tokens, the target is the first tokenizer token whose character span ends after the selected raw-word start, and the evaluated prefix ends immediately before that token. LAMBADA uses all 5,153 examples from the English test split and selects the final non-whitespace word. TriviaQA uses the `rc.nocontext` validation split and the canonical answer; a deterministic question-ID hash selects one answer word, repeated question IDs are resolved by content hash, and the first 5,000 tokenizer-compatible unique questions under the fixed hash ranking are retained. For GSM8K, n counts non-whitespace words before the last `####` marker after removing `<<...>>` calculator annotations. Zero-based reasoning-word indices use Python nearest-integer rounding (ties to even) in $\mathrm{round}((n-1)q)$ for $q \in \{0.25, 0.50, 0.75\}$, and the fourth position selects the first non-whitespace final-answer word after `####`. Exactly 1,316 problems satisfy the all-tokenizer criterion and contribute all four positions.

Target selection was specified before full-precision or quantized collection, disagreement measurement, intervention transfer, or controlled-candidate analysis. The resulting 15,417 rows are evalu-

ated under every checkpoint–calibration condition. A prediction is correct when its full-vocabulary top-one token equals the selected target token, so "token performance" is exact next-token accuracy at the selected positions, not sequence-level answer accuracy.

The analysis uses the source-grouped five-fold map in Appendix A.2; all four positions from one GSM8K problem remain in the same fold. The held-out fold supplies intervention rows, and the other four folds determine the representation basis.

The checkpoints, quantization conditions and sampled sites follow Appendix A; rates follow Appendix A.4.1. Final-site intervals use the source-grouped bootstrap defined there, with all four GSM8K positions from one problem resampled together; they quantify source-example variation conditional on the five evaluated checkpoints. AWQ and RTN top-16 Coverage and removal contrasts use the same construction (Table C.4). Paired Fisher-16 advantages also have conditional intervals for each quantizer (Table C.13); their resampling protocol is specified separately in Appendix A.4.1.

C.1.1 Changed-Decision Rates and Exact-Token Performance

Small net exact-token accuracy changes coexist with substantial full-vocabulary decision replacement across GPTQ, AWQ, and RTN (Table C.1; Figure 3). The GPTQ C4 and WikiText calibration variants preserve this qualitative discrepancy pattern. Table C.1 retains five-checkpoint task means and equal-task macros; Table 1 separates correctness reversals from wrong-to-different-wrong replacements.

C.2 Cross-Quantizer Directional Organization

At the final sampled site, directional organization replicates across GPTQ, AWQ, and RTN: task–condition rates give positive task-supported minus ambient-random advantages in 30/30 GPTQ, 30/30 AWQ and 15/15 RTN cells on both one-way and bidirectional criteria, after averaging the three ambient-random seeds within each cell. These counts describe final-site cell-level consistency; Table C.2 gives the aggregate control chain and separate complete paired-difference transfer rates.

Bases fitted from either precision state yield similar aggregate transfer across quantizers and transfer criteria (Table C.2). The full control chain includes reverse one-way transfer and stable-row disruption, alongside the three-site GPTQ trend (Table C.3). These comparisons extend Table 2. They describe how directional advantage varies with depth and how interventions affect initially agreeing decisions.

The uncapped construction and intervention criteria follow Appendices A.2 and A.3. Transfer is evaluated on naturally changed decisions at all three sites, while disruption uses initially agreeing rows. Point estimates average calibrations within checkpoint and tasks equally over five checkpoints; transfer is in percentages and control contrasts in percentage points. For Table C.3, the final-site task-supported advantage has 95% source-grouped intervals: one-way [12.08, 13.33]; bidirectional [19.02, 20.43]. These intervals quantify source-example variation conditional on the five evaluated checkpoints.

C.3 Intervention Spectra and Removal

To distinguish broad accumulation from low transfer overall, we report absolute prefix rates alongside their Coverage of the complete task-supported endpoint. Full-vocabulary transfer accumulates across large prefixes, and ordered-removal contrasts grow with removed prefix size. Table C.7 combines Coverage relative to the complete task-supported transfer rate with condition-level observed crossings and censoring; Table C.8, Block A, gives the removal contrasts. The macro curve describes aggregate accumulation; condition-level crossings preserve the variation around that summary. At partially available prefixes in Figure 4, numerator and denominator use the same available checkpoint subset and weights.

For Table 3, task-supported basis dimension K has mean [checkpoint range]: fixed-choice 624 [588–675]; full-vocabulary 2,681 [2,449–3,139]. These summaries average each checkpoint's five folds before summarizing across checkpoints, unlike the fold-level range in Appendix B.3.1. Table C.4 reports full-vocabulary source-grouped intervals for all three quantizers. Fixed-choice Coverage

Table C.1: **Full-vocabulary task and macro summaries.** Task rows are descriptive means over the five checkpoints after calibration averaging; macro rows weight the three tasks equally.

Quantizer	Task	FP exact	Q exact	Δ (pp)	Changed (%)
GPTQ	GSM8K	69.41	68.95	-0.46	11.10
	LAMBADA	73.77	71.49	-2.28	10.22
	TriviaQA	45.08	42.92	-2.16	20.89
	Macro	62.75	61.12	-1.63	14.07
AWQ	GSM8K	69.41	68.81	-0.60	9.37
	LAMBADA	73.77	73.48	-0.29	8.36
	TriviaQA	45.08	45.47	+0.39	17.96
	Macro	62.75	62.59	-0.16	11.89
RTN	GSM8K	69.41	68.21	-1.20	13.78
	LAMBADA	73.77	72.68	-1.09	10.50
	TriviaQA	45.08	44.22	-0.87	20.14
	Macro	62.75	61.70	-1.05	14.81

Table C.2: **Final-site cross-quantizer intervention control chain.** Transfer columns are percentages of naturally changed decisions; stable-disruption columns are percentages of initially agreeing decisions. Values follow calibration-within-checkpoint and equal-task aggregation.

Quantizer	Component	One-way	Reverse	Bidirectional	Stable quantized	Stable full-precision
GPTQ	Paired difference	81.33	80.89	67.67	1.48	1.40
	Task-supported	69.95	68.78	49.66	1.78	1.64
	Ambient random	57.25	54.51	29.93	1.58	1.50
	Complement	25.40	23.64	8.17	2.57	2.54
	Quantized-state basis	70.06	68.89	49.81	1.69	1.56
AWQ	Paired difference	80.61	80.81	67.95	1.43	1.50
	Task-supported	70.77	70.07	51.73	1.49	1.58
	Ambient random	57.59	56.23	31.66	1.40	1.41
	Complement	25.69	24.86	9.36	2.29	2.24
	Quantized-state basis	70.55	70.21	51.77	1.51	1.58
RTN	Paired difference	72.44	73.88	58.15	1.68	1.60
	Task-supported	63.66	65.13	44.92	1.80	1.69
	Ambient random	50.86	51.60	25.90	1.53	1.44
	Complement	20.36	20.77	6.55	2.44	2.39
	Quantized-state basis	63.40	64.55	44.23	1.71	1.66

intervals appear in Table B.4; fixed-choice removal contrasts remain descriptive. Coverage divides aggregate rates, whereas removal contrasts average rate differences.

The concentration gap persists at every approximately matched basis fraction in Table C.5, using observed prefix-transfer rates without interpolation. The normalized observed 90% crossing ratios $r_{90}/\overline{K}$ below supplement the matched-fraction and task-wise comparisons; they retain each regime's evaluated prefix grid.

Mean task-supported basis dimensions across checkpoint fold means are 624.08 and 2,680.84. The following GPTQ final-site condition-level $r_{90}/\overline{K}$ summaries instead use each checkpoint's own fold-mean dimension. Both regimes use equal-task rates within each condition and the separately evaluated fold-local complete endpoint. Fixed-choice one-way transfer has 10/10 observed-grid crossings, median 1.99% [1.25–15.79]; bidirectional transfer has 9/10, median 5.02% [1.36–42.11], with no observed crossing for Qwen-7B/C4 through its largest tested fraction, 42.11% (Table B.5). Full-vocabulary one-way transfer has 10/10 crossings, median 80.87% [75.97–97.87], and bidirectional transfer has 2/10 at 97.87%; the other 8/10 have no crossing through their largest tested fractions, 75.97–83.61%. Medians and ranges include observed crossings only; censoring follows the evaluated-grid convention in Appendix C.3.1. Task weights, threshold and endpoint definitions are aligned; task collections and grids differ.

Table C.3: **Full-vocabulary GPTQ controls across depth.** Transfer rates are percentages; task-supported-minus-ambient contrasts are percentage points.

Site	Paired difference	Task-supported	Ambient random	Complement	Task-supported − ambient
One-way transfer					
Early	32.3	**27.5**	22.6	15.3	**4.9**
Middle	59.0	**51.5**	42.1	21.6	**9.4**
Final	81.3	**70.0**	57.3	25.4	**12.70**
Bidirectional transfer					
Early	12.1	**8.9**	5.0	3.1	**3.9**
Middle	36.7	**27.4**	15.4	5.2	**12.0**
Final	67.7	**49.7**	29.9	8.2	**19.73**

Table C.4: **Full-vocabulary source-grouped uncertainty.** Directional advantage: task-supported minus matched ambient-random transfer (pp). Prefix-16 Coverage: prefix-16 divided by aggregate complete task-supported transfer rate (%). Removal specificity: matched-random minus ordered remainder transfer after prefix-16 removal (pp). Intervals quantify source-example variation conditional on the five checkpoints.

	One-way		Bidirectional	
Metric	Estimate	95% source interval	Estimate	95% source interval
GPTQ				
Directional advantage (pp)	12.70	[12.08, 13.33]	19.73	[19.02, 20.43]
Prefix-16 Coverage (%)	14.42	[13.77, 15.08]	4.51	[4.07, 4.97]
Removal specificity (pp)	4.27	[3.79, 4.76]	6.12	[5.53, 6.71]
AWQ				
Prefix-16 Coverage (%)	15.08	[14.39, 15.76]	5.02	[4.53, 5.52]
Removal specificity (pp)	4.02	[3.47, 4.57]	5.36	[4.73, 6.01]
RTN				
Prefix-16 Coverage (%)	12.61	[11.79, 13.45]	2.84	[2.39, 3.31]
Removal specificity (pp)	2.91	[2.34, 3.49]	3.90	[3.22, 4.58]

Task-wise summaries preserve high early Coverage in the fixed-choice collection and low early Coverage in the full-vocabulary collection under final-site GPTQ (Table C.6). Each fixed-choice task attains at least 81.9/71.3% one-way/bidirectional Coverage at prefix 16; the three full-vocabulary tasks attain 12.6–15.9% one-way and 3.5–5.5% bidirectionally. The contrast also persists at prefix 64.

C.3.1 Uncapped Intervention-Spectrum Curves

The uncapped curves measure transfer after adding the projection of δ_i onto the first r task-supported directions in pooled-variance order. Grid availability and the separately evaluated fold-local complete task-supported endpoint follow Appendix A.4; this endpoint, not the complete paired difference, normalizes Coverage.

Same-rank random prefixes within the task-supported space use sizes 8, 16, and 32 and seeds 101, 211, and 307. Each random projection is rescaled to the corresponding ordered-prefix patch norm on the same row. These comparisons test direction choice at matched prefix rank and norm; they neither equalize norms across prefix sizes nor match the main removal remainder norms. The separate plain-PCA rank-32 control is reported in Appendix B.5.3.

In the final-site condition-level task averages, the ordered prefixes outperform the averaged same-rank random prefixes for both one-way and bidirectional transfer at ordered prefix sizes 8, 16, and 32. These aggregates include task cells with non-positive contrasts; the released task-level records preserve those exceptions. The condition-level directional advantage coexists with a slowly accumulating ordered curve: early prefixes yield higher transfer than their recorded same-rank random alternatives, and larger tested prefixes attain higher transfer rates. Figure C.1 displays the sampled

Table C.5: **GPTQ final-site Coverage at approximately matched basis fractions.** Here $\overline{K}$ is the mean across checkpoint fold means (624.08 fixed-choice; 2,680.84 full-vocabulary), not a fold-specific dimension. Coverage entries are percentages.

Pair	Regime and prefix size	$r/\overline{K}$ (%)	One-way (%)	Bidirectional (%)
1	Fixed-choice, 16	2.56	92.1	87.7
	Full-vocabulary, 64	2.39	28.1	12.5
2	Fixed-choice, 32	5.13	93.4	89.6
	Full-vocabulary, 128	4.77	37.8	18.5
3	Fixed-choice, 64	10.26	95.3	93.0
	Full-vocabulary, 256	9.55	49.7	28.1
4	Fixed-choice, 128	20.51	97.5	96.9
	Full-vocabulary, 512	19.10	61.4	38.8
5	Fixed-choice, 256	41.02	98.5	98.3
	Full-vocabulary, 1,024	38.20	75.5	57.3

Table C.6: **GPTQ final-site task-wise intervention Coverage (%).** Each task first averages calibration variants within checkpoint, then reports the ratio of the five-checkpoint mean prefix transfer rate to the mean complete task-supported transfer rate.

Regime	Task	Prefix 16		Prefix 64	
		One-way	Bidirectional	One-way	Bidirectional
Fixed-choice	ARC-C	92.4	89.3	94.5	97.0
	HellaSwag	93.8	90.3	95.1	92.6
	PIQA	95.2	93.1	96.2	93.1
	SQuAD2-Ans.	97.1	94.2	98.2	95.7
	TruthfulQA-MC	81.9	71.3	92.4	86.6
Full-vocabulary	GSM8K	12.6	3.5	31.4	16.5
	LAMBADA	14.6	4.5	25.8	9.5
	TriviaQA	15.9	5.5	27.1	11.4

prefix curves, same-rank random-prefix controls, and separately evaluated complete task-supported transfer rates across sampled sites.

Observed one-way 90% crossings occur at large tested prefixes; most bidirectional conditions have no crossing on their available grid (Table C.7). NR denotes no observed crossing through the largest available tested prefix, $r_{\max}$. Nonmonotonicity permits an unmeasured smaller prefix to cross and then fall, so these entries do not bound the first crossing over all integer prefixes. Each fold-local complete endpoint has Coverage one by definition, not an observed crossing at a common scalar K. Unavailable prefixes are NA and never interpolated.

The GPTQ full-vocabulary anchor includes 282 unavailable task–condition prefix-grid entries across the three sampled sites. Table C.7 reports the final-site crossings and censored conditions. Negative contrasts remain in the aggregates; Appendix A.4.1 specifies NA handling, calibration aggregation and the source-grouped intervals in Table C.4.

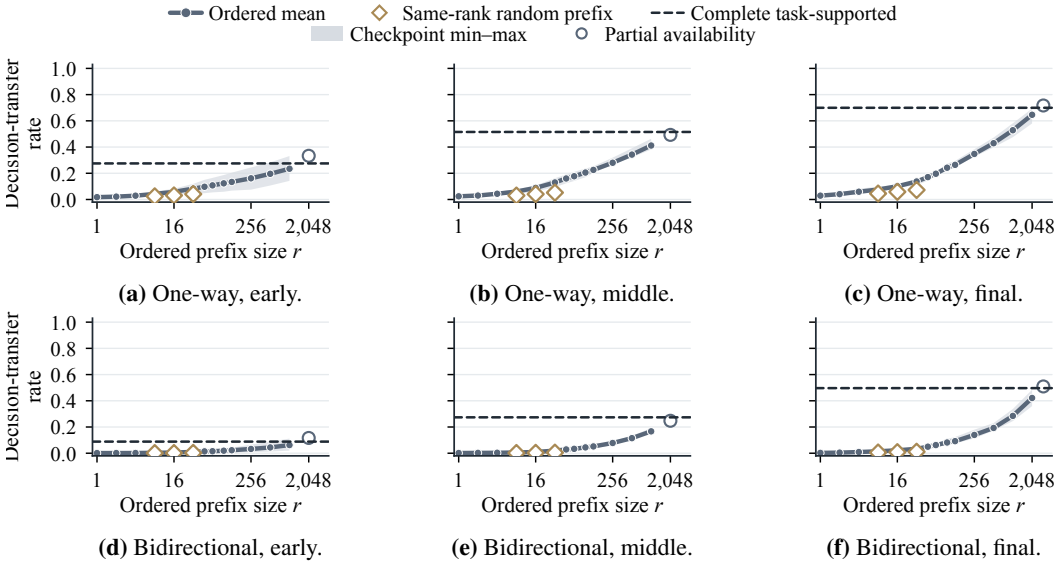
Task-level rates, checkpoint curves, fold-specific task-supported basis dimensions, condition thresholds, and prefix availability preserve the large-prefix availability difference: ordered prefix size 3,072 is available only for the Qwen 2.5–14B GPTQ/AWQ conditions and one RTN condition, and the bidirectional 90% threshold remains censored in most final-site conditions.

C.3.2 Removal and Ordered-Prefix Transfer-Rate Loss

In the descriptive aggregate, ordered removal lowers paired-decision transfer more than matched-random removal, and the contrast grows with removed prefix size. For $r \in \{16, 32, 64\}$, removal intervenes with the remainder $\delta_i - c_{i,r}$. Its transfer-rate loss relative to the complete paired difference measures dependence on the leading ordered prefix. The random comparator projects δ_i into a rank-

Table C.7: **Full-vocabulary spectra and observed-grid crossings.** Coverage uses the complete task-supported transfer rate for the same quantizer and transfer criterion. NR counts conditions not reaching 90% on their tested grid, through the largest available prefix (2,048 or 3,072); it does not bound unmeasured prefixes.

		Coverage (%)					90% threshold	
Quantizer	Criterion	16	64	256	1,024	2,048	Observed-grid crossing	NR
GPTQ	One-way	14.4	28.1	49.7	75.5	92.4	8 at 2,048; 2 at 3,072	0/10
	Bidirectional	4.5	12.5	28.1	57.3	84.9	2 at 3,072	8/10
AWQ	One-way	15.1	28.4	50.0	76.6	93.0	8 at 2,048; 2 at 3,072	0/10
	Bidirectional	5.0	12.9	29.7	59.1	86.0	2 at 3,072	8/10
RTN	One-way	12.6	27.6	52.1	77.7	92.6	4 at 2,048; 1 at 3,072	0/5
	Bidirectional	2.8	9.3	27.7	61.0	86.6	1 at 2,048; 1 at 3,072	3/5



(a) One-way, early. **(b)** One-way, middle. **(c)** One-way, final.

(d) Bidirectional, early. **(e)** Bidirectional, middle. **(f)** Bidirectional, final.

Figure C.1: **Full-vocabulary transfer accumulates across large prefixes.** Absolute decision-transfer rates on naturally changed rows, under GPTQ. Gray lines/filled circles and min–max bands use all five checkpoints; open circles mark means over available subsets. Brown open diamonds show same-rank, row-norm-matched random prefixes within task-supported space at sizes 8, 16 and 32; dashed references show complete task-supported transfer.

r subspace drawn by Gaussian QR within the same task-supported span, rather than selecting basis columns. The subspace is fixed within fold, site, rank and seed; seeds are 101, 211 and 307. Its projected component is rescaled to $\|c_{i,r}\|$ before subtraction from δ_i. A projection or target norm at most 10^{-12} returns a zero removed component; otherwise scaling is the target norm divided by the projection norm, without clipping. Removal specificity compares the resulting matched-random and ordered-prefix remainder rates. Remaining-intervention norms are neither matched nor restored, so this contrast does not isolate direction from all magnitude effects.

The prefix-16 contrast agrees with the gradual prefix curves: the leading prefix attains a small fraction of complete task-supported transfer, and larger tested prefixes attain higher fractions. Removal contrasts include task-cell reversals, which remain in the released task-level records. Table C.8, Block B, reports one-way and bidirectional outcomes as proportions. Multiplication by 100 yields transfer rates in percentages, but transfer-rate loss and removal specificity in percentage points. Figure C.2 plots the latter sampled-site contrasts.

The removal comparisons span task–condition–site cells; Table C.4 gives source-grouped uncertainty for the principal task-supported control, prefix-16 Coverage and prefix-16 removal contrasts.

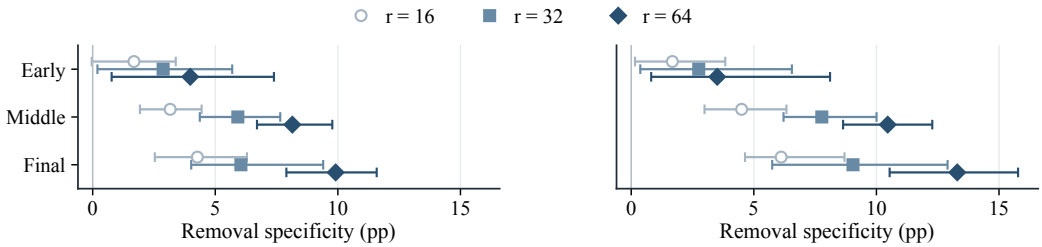
Table C.8: **Full-vocabulary ordered-removal contrasts across quantizers.** Specificity is matched-random minus ordered-prefix remainder transfer. Block A uses percentage points; Block B gives GPTQ transfer rates and contrasts as proportions. Removed-component rank and norm are matched; remainder norm is unconstrained.

A. Cross-quantizer removal specificity (percentage points)

Quantizer	Criterion	Prefix 16	Prefix 32	Prefix 64
GPTQ	One-way	4.27	6.05	9.91
	Bidirectional	6.12	9.04	13.30
AWQ	One-way	4.02	5.38	9.01
	Bidirectional	5.36	7.60	12.13
RTN	One-way	2.91	5.02	9.10
	Bidirectional	3.90	6.26	12.56

B. GPTQ decomposition (proportions)

Prefix size	Criterion	Paired difference	Ordered remainder	Matched-random remainder	Transfer-rate loss	Specificity
16	Bidirectional	0.677	0.582	0.643	0.095	0.061
	One-way	0.813	0.751	0.794	0.062	0.043
32	Bidirectional	0.677	0.536	0.627	0.141	0.090
	One-way	0.813	0.723	0.783	0.091	0.060
64	Bidirectional	0.677	0.465	0.598	0.211	0.133
	One-way	0.813	0.666	0.765	0.148	0.099



(a) One-way removal. **(b)** Bidirectional removal.

Figure C.2: **Ordered-removal specificity (pp).** GPTQ matched-random minus ordered-remainder transfer by depth and removed prefix. Removed-component rank/norm are matched; remainder norms are unconstrained. Markers/whiskers show checkpoint means/min–max; shapes and shades distinguish prefixes 16, 32 and 64.

C.4 Controlled Nested Candidate-Set Analysis

Candidate-Specific Accumulation. Restriction changes the absolute complete task-supported transfer rate, but early endpoint-normalized accumulation remains broad under all three transfer criteria over the evaluated prefix range (Table C.11). Figure C.3b uses the same candidate set, naturally changed rows and calibration/task/checkpoint weights in numerator and denominator. Low early Coverage therefore persists with candidate-specific endpoints. Prefix 2,048 is a measured prefix, not the complete endpoint. Tasks, prompt encoding and representation ordering remain fixed, so this comparison probes cardinality within the full-vocabulary computation rather than an independent task/interface effect.

Candidate restriction also changes output-sensitive geometry. Across restricted candidate sets $\kappa \in \{2, 4, 16, 128\}$, estimated Fisher-95 rank decreases at each tested reduction in cardinality in all 125 fold–condition trajectories (Figure C.3a). These are repeated measurements, not independent checkpoint replications; the separately estimated full-vocabulary anchor is outside this monotonic sequence. Each binary decision has at most rank-one local sensitivity, but differently oriented axes across training examples can produce a high-rank Fisher matrix. The candidate-restricted Fisher-95 rank in Table C.9 therefore describes aggregate sensitivity, not a row-local contrast span.

Table C.9: **Candidate restriction: geometry and absolute transfer.** Fisher ranks average 125 condition–fold trajectories. Transfer gives equal weight to each of the 75 natural-flip condition–task blocks. Dashes indicate unreported full-vocabulary ranks; these are distinct aggregation populations.

	Predictive-Fisher geometry		Complete task-supported transfer (%)		
κ	Mean Fisher-95 rank	Rank fraction	One-way	Reverse one-way	Bidirectional
2	1,550.616	0.3563	73.46	72.87	55.80
4	1,678.840	0.3862	69.16	69.41	50.17
16	1,734.480	0.3992	69.02	68.65	49.60
128	1,782.856	0.4104	69.02	68.57	49.55
$\lvert\mathcal{V}\rvert$	–	–	69.02	68.57	49.54

Table C.10 reports absolute prefix transfer: Part A averages calibrations within checkpoint, tasks equally and then the five checkpoints separately by quantizer; Part B gives equal weight to the 75 blocks, matching the complete-endpoint columns in Table C.9. At $\kappa = 4$, absolute prefix-2,048 rates are within 0.13–0.26 percentage points of the full-vocabulary readout. Bidirectional success is the conjunction of forward and reverse success on the same row.

Controlled Readout Design. The analysis reuses teacher-forced rows at the final sampled site and changes only the candidate set used to read out an already computed patched transformer forward (Figure C.3). Let $\mathcal{V}$ be the model's complete token vocabulary and $\kappa \in \{2, 4, 16, 128, |\mathcal{V}|\}$ denote candidate cardinality. On naturally changed rows, the binary set contains the unpatched full-precision and quantized full-vocabulary winners. On initially agreeing rows, it contains the full-precision winner and the next-highest distinct token under the unpatched full-precision logits. Larger sets append tokens in descending unpatched full-precision-logit order while skipping duplicates. Candidate sets are fixed before intervention. Prompts, tokenization, rows, paired states, δ_i, site, five-fold task-supported basis, prefix vector, patched transformer forward and full-vocabulary logits are identical across κ; only the restricted prediction may vary.

The restricted readout masks non-candidates to negative infinity in the original vocabulary coordinates and takes the first index returned by `torch.topk` with `k=2` on float32 logits. If the full-vocabulary winner from those same logits is a candidate and its logit exactly equals the restricted maximum, that winner is retained; the tolerance is zero. Original paired predictions remain the targets. This tie rule changes neither logits nor patch vectors.

Aggregation. Predictive-Fisher rank averages 125 model–quantizer–calibration–fold trajectories; the rank fraction first divides by each model's residual-stream width. The separately estimated full-vocabulary anchor is not part of the restricted-softmax monotonic sequence. Geometry trajectories and transfer blocks are distinct aggregation units.

Transfer uses 51,135 naturally changed rows in 75 condition–task blocks, excluding 375 unchanged-decision sentinels. Each candidate set has its own fold-local full uncapped task-supported endpoint. The normalized grid $r \in \{1, 2, 4, 8, 16, 32, 48, 64, 96, 128, 256, 512, 1024, 2048\}$ uses the same eligible rows, folds, targets and candidate masks at every prefix and endpoint. Table C.11 divides equal-block mean prefix rates by equal-block mean endpoint rates, giving GPTQ/AWQ/RTN weights of 40/40/20. Linked prefix-to-endpoint counts additionally support EventRetention: mean both-successful rates divided by mean endpoint-successful rates. Neither metric averages cell-level conditional ratios. Quantizer-specific curves average calibrations, tasks and checkpoints equally at each level. All five checkpoints show low prefix-16 Coverage.

The stated tie rule applies to the forward and reverse implementations. The released workers and sufficient counts support candidate-specific Coverage and prefix/endpoint EventRetention; raw endpoint/reverse prediction tensors are not included. Appendix D distinguishes this outcome-level verification from independent prediction replay. Candidate-specific adjacent-prefix transitions are outside the reported analyses.

Table C.10: **Absolute transfer under controlled candidate restriction.** Part A uses quantizer-specific calibration/task/checkpoint macros; Part B uses the equal-block population of the complete endpoints. Candidate-specific Coverage is reported separately in Table C.11.

A. Quantizer-specific forward prefix summary

Quantizer	κ	Prefix 16 (%)	Prefix 2,048 (%)
GPTQ	2	10.86	69.78
	4	10.24	64.88
	$\lvert\mathcal{V}\rvert$	10.08	64.62
AWQ	2	11.36	69.26
	4	10.79	65.93
	$\lvert\mathcal{V}\rvert$	10.67	65.80
RTN	2	8.54	63.69
	4	8.14	59.13
	$\lvert\mathcal{V}\rvert$	8.03	58.96

B. Natural-flip row-aligned paired summary

κ	Prefix size	One-way (%)	Reverse one-way (%)	Bidirectional (%)
2	16	10.60	10.78	2.49
	2,048	68.35	67.62	48.11
$\lvert\mathcal{V}\rvert$	16	9.91	10.11	2.19
	2,048	63.96	63.26	42.44

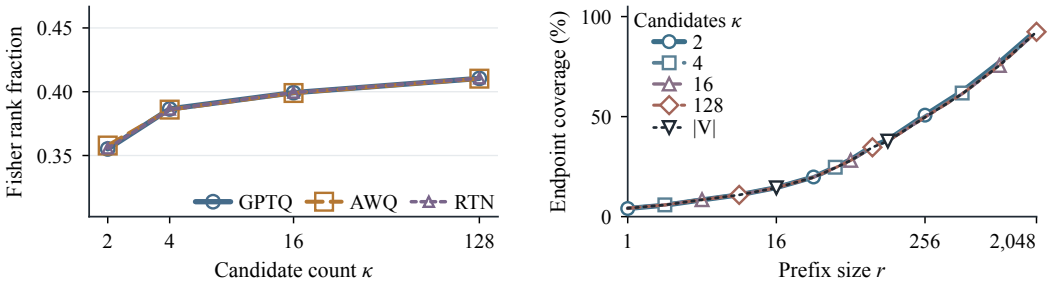
Table C.11: **Candidate-specific endpoint normalization preserves broad accumulation.** Percentages use equal-weight means over the same 75 changed-decision blocks. Coverage divides mean prefix transfer by mean candidate-specific complete-endpoint transfer; EventRetention divides mean both-successful rate by that mean endpoint rate. Ratios follow aggregation, rather than averaging cell-level conditional ratios.

		One-way		Reverse		Bidirectional	
Metric	κ	$r = 16$	$r = 2{,}048$	$r = 16$	$r = 2{,}048$	$r = 16$	$r = 2{,}048$
Coverage (%)	2	14.42	93.04	14.79	92.80	4.46	86.22
	4	14.52	92.76	14.64	92.13	4.41	85.63
	16	14.38	92.67	14.73	92.23	4.42	85.64
	128	14.36	92.67	14.74	92.25	4.42	85.66
	$\lvert\mathcal{V}\rvert$	14.36	92.67	14.75	92.26	4.42	85.67
EventRetention (%)	2	13.14	90.25	13.39	89.95	3.23	81.14
	4	12.85	89.00	12.92	88.68	3.01	79.15
	16	12.71	88.90	12.94	88.56	3.00	78.99
	128	12.70	88.90	12.95	88.57	3.00	79.00
	$\lvert\mathcal{V}\rvert$	12.70	88.90	12.95	88.57	3.00	79.01

C.5 Predictive-Fisher and Cross-Example Axis Geometry

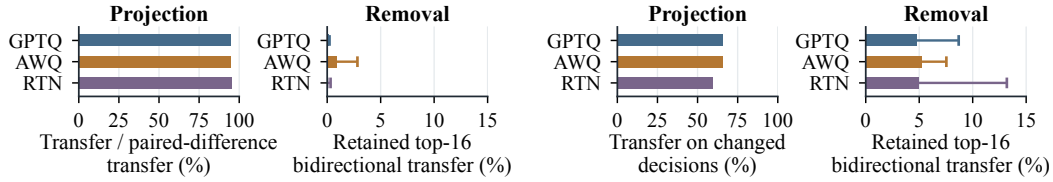
Predictive-Fisher geometry aggregates output sensitivity across training rows and carries substantial full-vocabulary transfer. Matched-binary analysis separately compares local axes across examples. Figure C.4 shows absolute Fisher-95 transfer from projecting the complete paired difference, together with removal comparisons. Removal uses capped prefix 16 for fixed-choice decisions and uncapped prefix 16 for full-vocabulary decisions. For each checkpoint, the bidirectional transfer rate of the norm-restored remainder is divided by the corresponding starting top-16 component's bidirectional transfer rate; these checkpoint-level ratios are then averaged. Zero baselines are NA, and the 100% reference lies outside the plotted range. The local panel reports $P_{\mathcal{S}_i}\delta_i$ transfer relative to its own complete paired-difference transfer rate, separate from Fisher absolute transfer and the removal ratios.

Table C.12 reports Fisher-95 geometry and transfer; Table C.12, Part B, compares uncapped ordered prefix-16, Fisher-16 projections, and rank-matched removal with and without final norm restoration.



(a) Predictive-Fisher rank. **(b)** GPTQ one-way Coverage.

Figure C.3: **Broad accumulation under candidate restriction.** (a) Quantizer-mean predictive-Fisher rank/residual-width fractions. (b) GPTQ one-way candidate-specific Coverage: aggregate prefix transfer rate divided by the same candidate set's complete task-supported transfer rate; all five checkpoints contribute at every plotted prefix.



(a) Fixed-choice geometry. **(b)** Full-vocabulary geometry.

Figure C.4: **Geometry-aligned projections and removal.** (a) Local projection $P_{\mathcal{S}_i}\delta_i$ one-way transfer rate divided by this analysis's own complete paired-difference one-way transfer rate. (b) Absolute Fisher-95 one-way transfer on changed rows (%). Removal shows the norm-restored remainder's bidirectional transfer rate divided by the corresponding starting top-16 component's bidirectional transfer rate. These ratios are computed within checkpoints and then averaged; whiskers span the checkpoint-level ratios.

Fix a sampled site ℓ and held-out fold f before suppressing those indices. For each training-fold row i in task t, we sample one token $Y_i \sim p_{\mathrm{FP}}(\cdot \mid h_{\mathrm{FP},i})$ and form $g_i = \nabla_h \log p_{\mathrm{FP}}(Y_i \mid h_{\mathrm{FP},i})$. Let $T = 3$ be the number of tasks, let I_t^{-f} be the training rows for task t, and define $n_t = |I_t^{-f}|$. The uncentered equal-task second moment is

$$\widehat{F}^{(\ell,-f)} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{n_t}\sum_{i\in I_t^{-f}} g_i g_i^{\top}. \tag{13}$$

The Monte Carlo estimate uses the same source-grouped five-fold exclusion as the intervention basis. Fisher-95 is the leading eigenspace retaining 95% of $\mathrm{tr}(\widehat{F})$. It is fitted on full-precision training-fold rows and evaluated by intervention on held-out rows. Table C.12 reports its rank, complete paired-difference projection, and removal from the uncapped ordered prefix-16 component on naturally changed rows. Matched removal restores each remainder to that component's norm (Part B). Fisher-16 uses the first 16 eigendirections of the same cross-fitted estimate: Original applies $u_i = P_{F_{16}}\delta_i$, and Matched applies $u_i\|c_{i,16}\|/\|u_i\|$, targeting the uncapped ordered prefix-16 norm. The float32 norm helper returns zero when either projection or target norm is at most 10^{-12}; otherwise it applies this ratio without clipping. Raw Fisher-16 projection norms are absent from the outcome metadata, so branch frequencies and the scale distribution cannot be recovered from those records.

The Fisher-95 random-removal comparator uses a Gaussian-QR subspace of the ambient residual-stream space, matched to the Fisher-95 rank for that model, site and fold. A deterministic seed keyed to model, task, final site and fold fixes one draw across quantizers and calibrations. For target ranks above half the residual width, the projector is computed through the lower-dimensional orthogonal complement, without changing the target subspace rank. Removal starts from the uncapped

task-supported prefix-16 component $c_{i,16}$, with the remainder either injected directly or restored to $\|c_{i,16}\|$. Task and condition rates follow the aggregation below, without an additional seed average. The supplement specifies the random-subspace construction and row-wise norm-restoration operator and includes their implementations.

Table C.13 quantifies uncertainty in the Matched projection contrast using paired saved predictions. The Fisher and ordered-prefix records have identical row IDs, folds, original predictions and changed-decision masks; the comparator predictions also match the primary uncapped prefix-16 records. Positive intervals support the measured transfer advantage at matched rank and final patch norm, without characterizing a complete Fisher-ordered spectrum.

For orthonormal bases A, B, $\mathrm{Overlap}(A, B) = \|A^\top B\|_F^2 / \min(r_A, r_B)$. Ordered task-supported prefixes $B_{1:r}$ are compared with the complete Fisher-95 subspace; the denominators are 16 and 32. This geometric overlap is distinct from paired-difference energy, transfer, Coverage and EventRetention. Full-precision rank and overlap summaries use no behavioral-flip filter: duplicate quantization-condition copies are removed by geometry identity, folds are averaged within model, then models are equally weighted. The repeated rank values in Table C.12 describe the same full-precision geometry, not three independent replications. These rules apply to these geometry summaries only; transfer rates and cross-example binary axes retain their own populations and aggregation.

Fisher-95 projection of the complete paired difference captures substantial full-vocabulary transfer, while its removal from the ordered top-16 component leaves little bidirectional transfer. Without norm restoration, same-rank random removal leaves more transfer than Fisher-95 removal under GPTQ, AWQ and RTN on all three transfer criteria (Table C.12, Part B, unrestored rows); bidirectional outcomes are aligned on the same rows. Together, projection and removal show that broad representation-ordered accumulation can coexist with output-sensitive geometry.

The separation persists when both remainders use the primary ablation's norm-restoration operator (Table C.12, Part B, norm-restored rows). The removed subspaces are rank-matched, and both remainders are restored row-wise to the uncapped top-16 task-supported component norm. In the equal-block pooled comparison, the Fisher-depleted remainder reaches 26.4% of uncapped top-16 task-supported component one-way transfer and 22.8% in reverse, versus 84.9% and 84.0% after same-rank random removal. These ratios divide the corresponding block-mean absolute rates; they are not means of within-checkpoint ratios. Equal block weighting gives GPTQ and AWQ twice RTN's weight because they each have two calibration conditions. Other transfer rows use checkpoint macros; within each quantizer, the balanced block design gives the same norm-restored macro means. The table also reports each quantizer separately, where forward, reverse and bidirectional separations persist. This comparison controls subspace rank and final patch norm for the Fisher-aligned removal contrast.

C.5.1 Cross-Example Binary-Axis Construction

A rank-one interface at each row does not imply a common direction across examples. With candidate cardinality held at two, controlled full-vocabulary axes have larger D_{PR} and D_{90} than fixed-choice axes in all five models (Table C.14). The medians across model-level controlled/fixed ratios are $90.85\times$ and $710.42\times$. Fixed-choice axis concentration also extends across tasks: the median squared cosine between aggregate PIQA and SQuAD2 top axes is 0.7899. Task collections, candidate-token identities and answer encoding are not matched.

Let ξ_i be the unit local decision axis for row i: the rank-one orthonormal candidate-logit-contrast gradient direction for a fixed-choice decision, and the unit-normalized $\kappa = 2$ restricted-softmax categorical Fisher score direction for the controlled full-vocabulary token decision.

$$\nabla_h \log p(a \mid h) = (1 - p_a)\nabla_h(z_a - z_b), \qquad \nabla_h \log p(b \mid h) = -p_a \nabla_h(z_a - z_b). \tag{14}$$

Here $p_a = p(a \mid h)$ is the binary softmax probability of candidate a, with b the other candidate and z_a, z_b their logits. For nonzero scores, Equation 14 shows that unit normalization makes either restricted categorical score direction equal to the normalized candidate-logit contrast axis up to sign. Because every cross-example comparison uses squared-axis alignment or the outer product $\xi_i \xi_i^\top$, this sign is immaterial. For a fixed population of N axes, define

$$M = \frac{1}{N} \sum_{i=1}^{N} \xi_i \xi_i^\top, \qquad D_{\mathrm{PR}} = \frac{1}{\sum_j \lambda_j^2}, \qquad D_{90} = \min \left\{ q : \sum_{j=1}^{q} \lambda_j \geq 0.9 \right\}, \tag{15}$$

Table C.12: **Predictive-Fisher geometry and removal contrasts.** Transfer on changed rows (%). A: rank and Fisher-95 projection of the complete paired difference. B: uncapped task-supported prefix-16 reference and rank-matched removal, with and without restoration to its norm. Fisher-16 Matched scales its complete paired-difference projection to that prefix norm. Shared full-precision geometry: prefix-16/32 overlap .502/.519; quantizer repetition is not independent replication.

A. Predictive-Fisher geometry and complete projection

		Fisher-95 transfer (%)	
Quantizer	Fisher rank [range]	One-way	Bidirectional
GPTQ	1,854 [1,698–1,996]	66.02	44.39
AWQ	1,854 [1,698–1,996]	65.56	45.58
RTN	1,854 [1,698–1,996]	59.61	39.84

B. Projection and rank-matched removal comparisons

Quantizer	Analysis	Variant	Norm treatment	One-way	Reverse one-way	Bidirectional
GPTQ	Projection	Uncapped prefix 16	Original	10.08	10.31	2.24
		Fisher 16	Original	16.90	15.70	6.12
		Fisher 16	Matched	24.46	24.13	12.66
	Removal	Fisher-95	Unrestored	2.20	1.73	0.08
		Random	Unrestored	6.84	6.60	0.95
		Fisher-95	Norm-restored	2.78	2.28	0.12
		Random	Norm-restored	8.60	8.42	1.57
AWQ	Projection	Uncapped prefix 16	Original	10.67	10.65	2.60
		Fisher 16	Original	17.77	17.67	7.40
		Fisher 16	Matched	26.51	27.20	15.56
	Removal	Fisher-95	Unrestored	2.26	1.94	0.05
		Random	Unrestored	7.19	7.13	0.98
		Fisher-95	Norm-restored	2.79	2.51	0.14
		Random	Norm-restored	9.12	9.28	1.76
RTN	Projection	Uncapped prefix 16	Original	8.03	8.63	1.28
		Fisher 16	Original	17.24	17.67	7.75
		Fisher 16	Matched	24.43	25.66	13.91
	Removal	Fisher-95	Unrestored	1.55	1.44	0.02
		Random	Unrestored	5.02	5.54	0.47
		Fisher-95	Norm-restored	1.93	1.93	0.06
		Random	Norm-restored	6.62	7.08	0.86
Overall	Removal	Fisher-95	Norm-restored	2.61	2.30	0.11
		Random	Norm-restored	8.41	8.50	1.50

Table C.13: **Paired Fisher-16 Matched transfer advantage.** Final-site Fisher-16 Matched minus uncapped ordered-prefix-16 transfer (pp), with conditional source-bootstrap 95% intervals. Source multiplicities are shared across both interventions and all calibration/checkpoint copies; the fitted geometries remain fixed (Appendix A.4.1).

Quantizer	One-way	Reverse	Bidirectional
GPTQ	14.38 [13.69, 15.05]	13.82 [13.16, 14.51]	10.42 [9.90, 10.95]
AWQ	15.83 [15.05, 16.61]	16.55 [15.76, 17.36]	12.96 [12.33, 13.61]
RTN	16.40 [15.60, 17.21]	17.02 [16.22, 17.81]	12.63 [12.00, 13.27]

where M is the uncentered second moment of the unit axes, not a centered covariance, λ_j are its trace-normalized eigenvalues in decreasing order, and q counts leading eigenvectors. D_{PR} is its participation-ratio dimension; D_{90} counts directions reaching 90% of its trace. Lower values indicate stronger cross-example axis concentration. These quantities differ from row-local readout rank, predictive-Fisher rank and observed Coverage crossings. Output-coordinate overlap is tested separately through α_o below.

The primary comparison uses equal row counts from two tasks per regime (no additional two-fold partition): PIQA and SQuAD2 answerability for fixed-choice decisions, and LAMBADA and TriviaQA for the controlled full-vocabulary token analysis. Eligibility does not require paired prediction disagreement, unlike the transfer population. Controlled axes are normalized from the saved one-sample restricted-Fisher scores, which must be finite with strictly positive norm in all 25 conditions. This common screen retains 5,140 of 5,153 LAMBADA rows and 4,966 of 5,000 TriviaQA rows, from which 1,800 rows per task are selected in total, not per fold. Fixed-choice rows require two candidates and a rank-one local span; their common eligible pools contain 1,838 PIQA and 11,873 SQuAD2 rows, with 1,800 selected per task.

The 13 and 34 excluded full-vocabulary rows fail the joint finite/nonzero criterion; the saved metadata do not separate exact-zero from nonfinite exclusions. This screen checks the numerical validity of the saved one-sample scores; it does not imply that the binary interface lacks a candidate-logit contrast direction. The two tasks' axes are concatenated for one descriptive second moment per model or condition. Each model is the scientific unit.

Table C.14 reports the exact dimensions and ratios. Fixed columns describe one pooled fixed-choice matrix per model. Controlled columns average condition-level metrics by taking the calibration mean within GPTQ and AWQ and then weighting GPTQ, AWQ, and RTN equally; D_{PR} and D_{90} use the same hierarchy. Every underlying D_{90} is an integer, while its displayed controlled model mean can be fractional. Ratios compare the controlled model mean with the fixed matrix; the five models are the scientific replication units.

The separate stable/matched robustness comparison additionally requires agreement of the full-precision and quantized full-vocabulary predictions across conditions and uses full-precision-defined binary candidates. Its controlled eligible pools contain 2,278 LAMBADA and 881 TriviaQA rows; both regimes use 800 rows per task in total. GSM8K is excluded from the primary geometry because four positions share a source problem and Qwen stable rows have substantial exact-zero one-sample scores; their exact counts are absent from the saved metadata. TruthfulQA's small binary subset is secondary. This additional screen preserves the direction of both concentration differences in five of five models. Its rows supply the coordinate-overlap analysis in Table C.15, separately from the primary concentration population in Table C.14. Each interface has one local contrast axis per row, but fixed-choice option labels are reused and controlled full-vocabulary candidate tokens can vary across examples.

To test a possible geometric correlate inside the controlled regime, let $o \in \{0, 1, 2\}$ be the number of token coordinates shared by two unordered binary candidate pairs, and let α_o be the task-balanced mean of $(\xi_i^\top \xi_j)^2$ conditional on overlap o. Stable full-precision-defined candidates give $\alpha_1 > \alpha_0$ in all five models (Table C.15); the table reports arithmetic means. Including initially disagreeing rows preserves the association: $\alpha_1 > \alpha_0$ in all five model-level summaries. Overlap two also exceeds overlap one in all five models, but it is secondary evidence: across model–condition summaries, overlap-two support can be as low as 44 pairs, compared with minima of 3,931 and 306,534 for overlap one and zero. Example pairs, folds, and the 25 quantization conditions are repeated measurements. Grouping by observed coordinate overlap is an association analysis, not a causal manipulation of coordinate reuse.

C.6 Row-Level EventRetention

Similar aggregate rates can arise from different successful rows. Among rows successfully transferred by the complete endpoint, short prefixes transfer most in fixed-choice decisions but few in full-vocabulary token decisions (Table C.16).

Within each task–condition–site cell and for $E \in \{\mathrm{one}, \mathrm{rev}, \mathrm{bi}\}$, let $S_{E,r}$ and $S_{E,K}$ contain the successful naturally changed rows under prefix r and the fold-specific complete task-supported endpoint, respectively. Both use the same aligned population; K denotes each row's complete basis. The cell-level diagnostic is

$$\mathrm{EventRetention}_E(r) = \frac{|S_{E,r} \cap S_{E,K}|}{|S_{E,K}|}. \tag{16}$$

Macro EventRetention divides the aggregated both-successful rate by the aggregated complete-endpoint-successful rate, using calibration-within-checkpoint/task, equal-task and equal-checkpoint

Table C.14: **Fixed-choice binary axes concentrate more strongly across examples.**

	D_{PR}			D_{90}		
Model	Fixed	Mean controlled	Ratio	Fixed	Mean controlled	Ratio
Llama-8B	1.43	348.98	243.25	2	1,420.83	710.42
Mistral-7B	1.87	169.46	90.85	2	1,316.67	658.33
Nemo-12B	1.47	353.61	241.00	2	1,526.50	763.25
Qwen-14B	1.96	177.38	90.36	2	1,421.50	710.75
Qwen-7B	2.28	189.19	83.12	4	1,293.83	323.46

Table C.15: **Shared candidate tokens accompany stronger stable-row axis alignment.** Entries are arithmetic means α_o of squared-axis alignment on the separate stable-row population, rounded to three significant figures. Calibration means receive equal quantizer weights within task, followed by equal weighting of the two tasks. Overlap $o = 2$ has lower support; example pairs do not supply independent checkpoint replications.

Model	$o = 0$	$o = 1$	$o = 2$
Llama-8B	2.58×10^{-3}	1.01×10^{-1}	8.38×10^{-1}
Mistral-7B	3.69×10^{-3}	1.69×10^{-1}	6.93×10^{-1}
Nemo-12B	1.68×10^{-3}	9.12×10^{-2}	7.09×10^{-1}
Qwen-14B	1.45×10^{-3}	6.07×10^{-2}	7.04×10^{-1}
Qwen-7B	7.85×10^{-3}	9.07×10^{-2}	5.67×10^{-1}

weights (Appendix A.4.1). It is neither a pooled count ratio over all rows nor a mean of cell-level conditional ratios. In Table C.16, each pair of adjacent rows compares the two regimes for the same quantizer and transfer criterion. The four partitions use all naturally changed rows within each cell as their denominator and sum to 100% before rounding and after aggregation.

For the primary spectra, we additionally report the both-successful, prefix-only, endpoint-only, and neither-successful partitions, together with success-to-failure and failure-to-success transitions between adjacent evaluated prefixes. Both regimes use the uncapped task-supported construction. These primary-spectrum transitions are separate from the candidate-specific analysis in Appendix C.4. Cell-level EventRetention is NA when the complete endpoint has no successful rows. EventRetention describes row-level overlap and does not replace aggregate intervention-spectrum Coverage.

At ordered prefix size 16, EventRetention is 80.40–88.98% across fixed-choice transfer criteria and quantizers, compared with 1.86–13.48% for full-vocabulary token decisions (Table C.16). Every reported quantizer–criterion partition contains both prefix-only and endpoint-only successes. At the final site, success-to-failure and failure-to-success transitions both occur in every adjacent-prefix comparison after quantizer-level aggregation; the machine-readable records preserve the complete prefix grid and the single all-site exception outside this final-site statement. Accumulation therefore describes changes in aggregate transfer rates, not nested sets of successfully transferred decisions.

D Reproducibility Scope

Appendix A specifies the checkpoint and tokenizer revisions, task construction, calibration and quantization settings. The supplementary materials provide frozen evaluation rows, calibration token packs, checkpoint configurations and module inventories, together with intervention implementations, outcomes, sufficient counts and analysis code. These materials support reconstruction of the documented inputs and verification of reported transfer rates, paired-outcome intersections and aggregation.

Each analysis uses its own complete paired-difference transfer rate. The primary spectra instead use the complete task-supported transfer rate on matched rows and weights; plain-PCA spectra use their own complete transfer rate in the fold-local PCA space retaining 95% of training-state variance. These three references serve different interventions and are not interchangeable.

Table C.16: **Final-site ordered-prefix-size-16 EventRetention and aligned-row partitions (%).**

Quantizer	Criterion	Regime	EventRetention	Both	Prefix only	Endpoint only	Neither
GPTQ	One-way	Fixed-choice	88.52	76.36	3.06	9.90	10.68
		Full-vocabulary	12.82	8.97	1.11	60.98	28.93
	Reverse	Fixed-choice	88.27	73.26	3.76	9.73	13.25
		Full-vocabulary	13.06	8.99	1.32	59.80	29.89
	Bidirectional	Fixed-choice	80.40	59.15	5.34	14.42	21.09
		Full-vocabulary	3.05	1.51	0.72	48.15	49.61
AWQ	One-way	Fixed-choice	88.11	76.04	3.19	10.26	10.51
		Full-vocabulary	13.45	9.52	1.15	61.25	28.07
	Reverse	Fixed-choice	88.53	75.21	3.49	9.74	11.56
		Full-vocabulary	13.48	9.44	1.21	60.63	28.72
	Bidirectional	Fixed-choice	80.78	60.58	5.16	14.42	19.84
		Full-vocabulary	3.45	1.78	0.81	49.95	47.46
RTN	One-way	Fixed-choice	88.81	76.86	2.85	9.68	10.61
		Full-vocabulary	10.74	6.84	1.19	56.82	35.15
	Reverse	Fixed-choice	88.98	73.70	3.05	9.13	14.12
		Full-vocabulary	11.57	7.53	1.10	57.59	33.78
	Bidirectional	Fixed-choice	81.38	60.45	4.21	13.83	21.50
		Full-vocabulary	1.86	0.83	0.44	44.09	54.64

Verification of saved outcomes and summaries, including candidate-specific Coverage and EventRetention, is distinct from independent tensor-level refitting and intervention replay, which are not included in the supplied materials. Full checkpoint, training-state, fitted-basis and predictive-Fisher tensors are not provided. Independent reconstruction of alternate-fold fitting and raw-record re-estimation of conventional-diagnostic associations are also outside this material scope.