COMPRESSION HYPERCUBE: A UNIFIED FRAMEWORK FOR PRUNING, QUANTIZATION, AND DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Model compression is critical for deploying large language models (LLMs), yet selecting an effective strategy remains largely empirical: pruning, quantization, and knowledge distillation (KD) are often studied in isolation, with limited guidance on when each method is reliable. We propose the *Compression Hypercube*, a structured framework that characterizes compression as an interaction between method, compression level, recovery strategy, initialization state, and data modality. We identify three key properties governing compression behavior — sequential redundancy, loss-landscape curvature, and compression-induced distributional shift — and derive local bounds linking structural properties to compression error. Across synthetic environments and pre-trained and instruction-tuned LLMs across three model families (2B–14B), we observe substantial variation in outcomes across operators and recovery conditions. Weight-only 8-bit quantization remains near-lossless across models, while depth pruning leads to significant degradation even after recovery. Recovery behavior is regime-dependent: SFT is favored under severe depth pruning, whereas a separate 50%-sparsity comparison favors the tested reverse-KL and on-policy KD objectives. Additional measurements of real-kernel quantization, structured and low-rank pruning, and KV-cache compression broaden the operator coverage while exposing the importance of evaluation and implementation conditions.

1 INTRODUCTION AND PRIOR ART

Large language models (LLMs) are deployed across environments ranging from cloud to edge (Achiam et al., 2023; Lu et al., 2024; Lin et al., 2024), making inference cost a central constraint (Goel et al., 2025; Zhuang et al., 2025; Niu et al., 2026). Model compression addresses this through pruning (Han et al., 2015; Frankle & Carbin, 2019; Frantar & Alistarh, 2023; Ma et al., 2023; Men et al., 2025; Gromov et al., 2025), quantization (Frantar et al., 2022; Lin et al., 2024; Xiao et al., 2023), and knowledge distillation (KD) (Hinton et al., 2015; Sanh et al., 2019; Gu et al., 2024; Ko et al., 2025). Despite extensive progress, selecting a compression strategy remains largely empirical: the same nominal compression level can lead to near-lossless behavior or severe degradation depending on the method, recovery procedure, and model state.

Prior work[1] provides complementary but individually incomplete explanations. Classical pruning theory relies on local second-order loss approximations to estimate parameter importance (LeCun et al., 1989; Hassibi & Stork, 1992; Frantar & Alistarh, 2023; Dong et al., 2019); representation analysis via SVCCA and CKA reveals per-layer redundancy (Raghu et al., 2017; Kornblith et al., 2019); and systematic comparisons across methods show quantization is generally more reliable than pruning at moderate compression (Kuzmin et al., 2023; Zhou et al., 2025). However, these lines of work typically analyze one axis at a time — method choice, recovery strategy, or model initialization — without studying their interactions. In practice, outcomes depend on all of these factors jointly, and these interactions are especially relevant to LLMs that have undergone instruction tuning or RLHF alignment (Ouyang et al., 2022; Rafailov et al., 2023), where initialization state should be measured rather than assumed to have one sensitivity direction.

[1] A detailed discussion of related work is in Appendix A.

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.

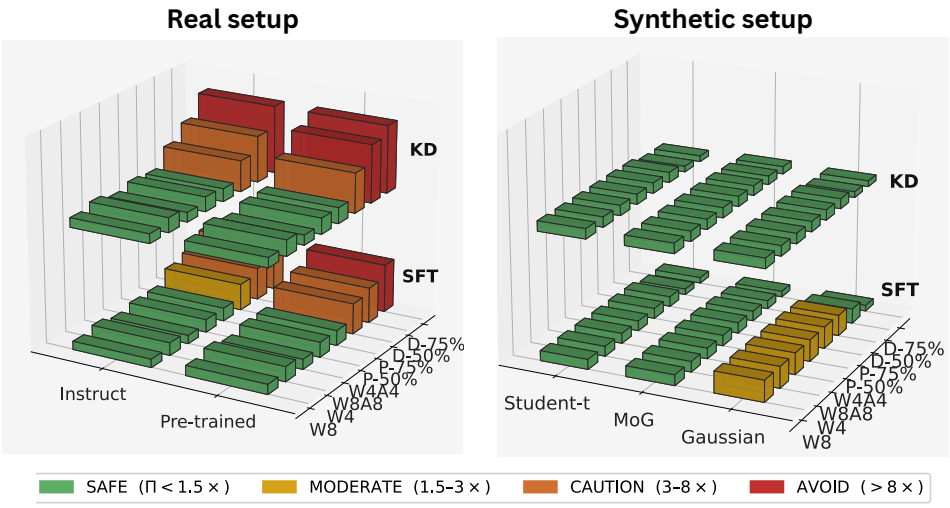

Figure 1: **The Compression Hypercube** $\mathcal{H} = \mathcal{M} \times \mathcal{C} \times \mathcal{R} \times \mathcal{I} \times \mathcal{D}$. Each configuration is defined by five design dimensions: compression method ($\mathcal{M}$), compression level ($\mathcal{C}$), recovery strategy ($\mathcal{R}$), initialization ($\mathcal{I}$), and data distribution ($\mathcal{D}$). Fixing a subset of dimensions corresponds to analyzing a lower-dimensional slice of the full design space.

We formalize this multi-dimensional interaction as the *Compression Hypercube* $\mathcal{H} = \mathcal{M} \times \mathcal{C} \times \mathcal{R} \times \mathcal{I} \times \mathcal{D}$ (method, compression level, recovery, initialization, data; Figure 1). Each point in this space defines a complete compression configuration, and the framework captures both main effects and interactions between dimensions. Within this space, we identify three key properties governing compression behavior: bottleneck sequential redundancy ($\underline{S}$), loss-landscape curvature ($\lambda_{\max}$), and compression-induced distributional shift.

Our study combines local perturbation analysis with systematic empirical evaluation on both synthetic data and real LLMs across three architecture families (2B–14B). Post-recovery PPL ratios range from near-lossless under weight-only INT8 ($\Pi \approx 1.00\times$) to severe degradation under depth pruning ($\Pi \approx 2$–$8\times$ after recovery), consistent with the curvature analysis in Section 3. At Wanda (Sun et al., 2024) 50% sparsity, KD with properly configured objectives (reverse-KL, on-policy) outperforms SFT for recovery by large margins, though this ranking reverses under depth pruning. A 4-feature composite diagnostic provides an initial cross-model sensitivity estimate without trial compression runs, though within-model rankings are driven by compression level.

Positioning. This work does not propose new compression algorithms. It provides a structured way to analyze existing methods jointly, revealing phenomena invisible to single-axis studies — for example, that KD outperforms SFT at Wanda-50% sparsity only under properly configured objectives, and that structural properties ($\underline{S}$, $\lambda_{\max}(F)$) contribute to a composite pre-deployment sensitivity estimate without trial compression runs.

Contributions.

- We introduce the *Compression Hypercube*, a five-dimensional framework that captures interactions between compression method, recovery strategy, initialization state, compression level, and data distribution.
- We provide a unified perturbation analysis relating parameter- and activation-level compression effects via local second-order bounds (Theorems 3.1–3.2).
- We empirically demonstrate, across synthetic setups and real LLMs (2B–14B), that compression outcomes vary with operator, data regime, and recovery protocol.
- We clarify the conditions under which curvature and bottleneck sequential redundancy bound local damage, and show why these diagnostics alone do not predict multi-layer removal.
- We report an exploratory leave-one-model-out selection study that achieves zero top-1 regret for 6/9 models with ≥ 8 configurations, and show why additive base-model features cannot reorder candidates within a single model.

2 METHODOLOGY

2.1 THE COMPRESSION HYPERCUBE

Definition 2.1 (Compression Hypercube)**.** Let $f_\theta : \mathcal{X} \to \mathcal{Y}$ denote an autoregressive language model with parameters $\theta \in \mathbb{R}^P$. We define the compression design space as $\mathcal{H} = \mathcal{M} \times \mathcal{C} \times \mathcal{R} \times \mathcal{I} \times \mathcal{D}$, where $\mathcal{M} = \{U, D, Q\}$ denotes the compression method (unstructured pruning, depth pruning, quantization), $\mathcal{C} \subset [0, 1)$ the compression level, $\mathcal{R} = \{\varnothing, \mathrm{SFT}, \mathrm{KD}\}$ the recovery strategy, $\mathcal{I}$ the initialization state (pre-trained or instruction-tuned), and $\mathcal{D}$ the data distribution.

Each configuration $(m, c, r, i, d) \in \mathcal{H}$ defines a compressed model $\tilde{f}_{\theta_c}$ obtained from f_θ. We evaluate configurations using the objective $\mathcal{J}(m, c) = \mathcal{E}_m(c) - \gamma \mathcal{S}_m(c)$, where $\mathcal{E}_m(c)$ is task error, $\mathcal{S}_m(c)$ is system-level speedup, and γ controls the accuracy–latency tradeoff. We additionally report the perplexity ratio $\Pi = \mathrm{PPL}(\tilde{f})/\mathrm{PPL}(f)$, where $\mathrm{PPL}(\cdot) = \exp(\ell(\cdot))$ and ℓ denotes cross-entropy loss. This formulation separates design variables from evaluation metrics and enables controlled analysis across configurations.

Comparison design. A controlled comparison fixes a base checkpoint, evaluation set, and recovery budget while changing one coordinate. Our empirical tables serve three distinct purposes: the synthetic grid changes innovation distribution under a fixed architecture to probe sensitivity to input statistics; the eight paired checkpoints support initialization and operator comparisons at common labels; the 14B experiments confirm these patterns at larger scale.

For each checkpoint, pre-recovery PPL is divided by that checkpoint's uncompressed PPL. Post-recovery ratios use the same denominator, while the synthetic recovery table divides by the compressed model instead. Key interactions include: operator by severity, operator by recovery (SFT versus KD in severe regimes), recovery objective at Wanda-50%, data shape in the synthetic models, initialization within each paired model family, and real-kernel deployment benchmarks. These slices show why a one-dimensional "compression percentage" can conceal large differences.

2.2 COMPRESSION OPERATORS

Let $\tilde{\theta} = \theta + \delta\theta$ denote the post-compression parameters. Each compression method induces a characteristic perturbation structure:

Unstructured pruning zeros out a subset S^c of individual parameters: $\delta\theta_i = -\theta_i$ for $i \in S^c$, $\delta\theta_i = 0$ otherwise, yielding a sparse perturbation whose squared norm equals the pruned-weight energy $\|\theta_{S^c}\|_2^2$.

Depth pruning removes entire residual blocks. For $h_{\ell+1} = h_\ell + F_\ell(h_\ell)$, removal gives $\tilde{h}_{\ell+1} = h_\ell$, with activation perturbation $\delta a = -F_\ell(h_\ell)$. Unlike parameter-level pruning, this is an activation-level perturbation whose magnitude depends on the residual branch energy.

Quantization maps each parameter to a nearest point in a discrete set, with $|\delta\theta_i| \leq \Delta/2$ where Δ is the quantization step size. This produces a dense but bounded perturbation, typically much smaller in norm than pruning perturbations at comparable compression ratios.

These perturbation structures — sparse (pruning), activation-level (depth), dense-bounded (quantization) — define the basis for the unified analysis in Section 3.

2.3 SYNTHETIC DISTRIBUTIONAL ENVIRONMENTS

To isolate the effect of data statistics from architecture, we design a controlled synthetic environment. We train 8-layer, 64-dimensional transformers on three autoregressive AR(1) processes sharing the same dynamics ($x_t = 0.70\, x_{t-1} + \varepsilon_t$) but differing in noise distribution: **Gaussian** ($\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$) serves as the baseline unimodal case; **MoG** ($\varepsilon_t \sim \frac{1}{K}\sum_k \mathcal{N}(\mu_k, \sigma^2)$) proxies multi-domain corpora with clustered structure; **Student-t** ($\varepsilon_t \sim t_3 \cdot s$) proxies heavy-tailed corpora with occasional extreme values. With centered mixtures and variance-matched innovations, these distributions can share a stationary mean and variance while differing in tails, enabling controlled comparison of compression sensitivity under different distributional regimes. Full details are in Appendix E.

2.4 REAL-LLM EXPERIMENTAL SETUP

Our primary comparison uses eight LLMs from three families: LLaMA-3.2-3B and LLaMA-3.1-8B (Dubey et al., 2024), Qwen-2.5-7B (Qwen et al., 2025), and Gemma-2-2B (Team et al., 2024), with pre-trained and instruction-tuned (IT) variants for each of the four base checkpoints. Qwen3-14B (Yang et al., 2025) extends the evaluation to 14B scale (Appendix F).

Compression configurations include: **Unstructured pruning** via Wanda (Sun et al., 2024) at 50% and 75% sparsity; **Depth pruning** via ShortGPT-style layer removal (Men et al., 2025) at 50% and 75%; **Quantization** with RTN (W8A8, W4A4, W8A16, W4A16) and calibrated methods (AWQ (Lin et al., 2024), BNB-NF4 (Dettmers et al., 2023); Table 3). Recovery uses LoRA-SFT (r=8, α=16) (Hu et al., 2022) and three KD variants: forward-KL, reverse-KL/MiniLLM (Gu et al., 2024), and on-policy (student generated output – SGO) distillation (Ko et al., 2025), across $N \in \{50, 250, 1000, 4000\}$ C4 samples (Raffel et al., 2020). Degradation is measured via WikiText-2 (Merity et al., 2016) perplexity and zero-shot accuracy on HellaSwag (Zellers et al., 2019), ARC-Challenge (Clark et al., 2018), WinoGrande (Sakaguchi et al., 2021), and PIQA (Bisk et al., 2020).

3 THEORETICAL BOUNDS

We state local perturbation results (proofs in Appendix B). They characterize an unrecovered model and do not guarantee performance after recovery or for multi-layer deletion.

Theorem 3.1 (Generic compression error)**.** *Let $\mathcal{L}$ be twice differentiable along $\{\theta + t\delta\theta : t \in [0, 1]\}$. Then*

$$\mathcal{L}(\theta + \delta\theta) - \mathcal{L}(\theta) = \nabla\mathcal{L}(\theta)^\top \delta\theta + \int_0^1 (1 - t)\, \delta\theta^\top H(\theta + t\delta\theta)\delta\theta\, dt.$$

Error depends on both perturbation norm and its alignment with curvature: small perturbations along high-curvature directions can be worse than large perturbations in flat regions.

Theorem 3.2 (Single-block removal)**.** *For a fixed input, let $a = h_\ell + F_\ell(h_\ell)$, $\tilde{a} = h_\ell$, and let ϕ_ℓ denote the downstream loss with all subsequent weights fixed. If ϕ_ℓ is twice differentiable on the segment between a and $\tilde{a}$ and $\lambda_{\max}(\nabla^2\phi_\ell(z)) \le \Lambda_\ell$ throughout it, then*

$$\phi_\ell(\tilde{a}) - \phi_\ell(a) \le -\nabla\phi_\ell(a)^\top F_\ell(h_\ell) + \tfrac{1}{2}\Lambda_\ell \|F_\ell(h_\ell)\|_2^2.$$

The bound applies to one block with the incoming state held fixed; sequentially removing many blocks changes later states and requires a separate analysis.

Corollary 3.3 (Normalized adjacent states)**.** *If $\|h_\ell(x)\|_2 = \|h_{\ell+1}(x)\|_2 = 1$ for every calibration input x, then $\mathbb{E}_x\|F_\ell(h_\ell(x))\|_2^2 = 2(1 - S_\ell)$, where $S_\ell = \mathbb{E}_x[h_\ell(x)^\top h_{\ell+1}(x)]$. If additionally $\nabla\phi_\ell(a)^\top F_\ell(h_\ell) = 0$ pointwise and $\Lambda_\ell \ge 0$ is uniform, the expected loss increase is at most $\Lambda_\ell(1 - S_\ell)$. Without these additional conditions, layer cosine similarity alone gives no such loss bound.*

The measured adjacent-state cosine is a diagnostic, not a certificate of safe multi-block removal. A low minimum similarity identifies a possible bottleneck only if the removed set contains the relevant layer.

Proposition 3.4 (Curvature sensitivity)**.** *For a symmetric Hessian H and any radius $r > 0$,*

$$\sup_{\|\delta\theta\|_2 = r} \frac{1}{2}\delta\theta^\top H\delta\theta = \frac{1}{2}\lambda_{\max}(H)r^2.$$

$\lambda_{\max}(H)$ quantifies worst-case sensitivity: models with higher curvature are more vulnerable to perturbations of the same magnitude.

Comparative regimes. Quantization induces bounded perturbations ($\|\delta\theta_Q\| \ll \|\delta\theta_{\text{prune}}\|$), giving lower expected error (Theorem 3.1). Depth pruning concentrates error at the activation level: for a single removed block, Theorem 3.2 bounds the loss increase by downstream curvature Λ_ℓ and residual magnitude; under the conditions of Corollary 3.3, the latter relates to $(1-S_\ell)$, but the bound does not extend to multi-block removal. High $\lambda_{\max}(H)$ amplifies perturbation sensitivity

Table 1: Pre-recovery PPL ratio on the three synthetic AR(1) distributions (lower is better). W8A16 and W4A16 use weight-only quantization.

Modality	Prune-50%	Prune-75%	Depth-50%	Depth-75%	Q-W8A8	Q-W4A4	Q-W8A16	Q-W4A16
Gaussian	1.00	1.01	1.38	2.67	1.00	1.12	1.00	1.00
MoG	1.00	1.04	1.06	1.32	1.00	1.28	1.00	1.00
Student-t	0.99	1.00	1.37	2.01	1.00	0.99	1.00	1.00

Table 2: Recovered-to-compressed PPL ratio for synthetic models at $N = 4000$ (lower is better; values below 1 improve on the compressed model).

	Prune-50%		Depth-50%		Depth-75%		Q-W8A8	
Modality	KD	SFT	KD	SFT	KD	SFT	KD	SFT
Gaussian	1.00	**2.66**	0.78	**2.16**	0.47	1.74	1.00	**2.68**
MoG	1.00	0.96	0.95	0.92	0.80	0.76	1.00	0.97
Student-t	1.02	1.01	0.76	0.76	0.55	0.55	1.00	1.00

(Proposition 3.4). These bounds also inform recovery strategy selection: KD preserves the teacher's full output distribution, providing richer gradient signal than point-wise SFT when the compressed model retains enough structure for the KL objective to be informative (Appendix B.6). Since the logit-space Hessian is identical for CE and forward-KL (Theorem B.5), the advantage of forward-KL KD over SFT lies entirely in the gradient target, not in local curvature. Reverse-KL and on-policy objectives have different Hessian structure and may confer additional benefits beyond what this analysis covers.

4 Results

Notation. $\Pi = \text{PPL}(\tilde{f})/\text{PPL}(f)$: perplexity ratio (compressed/original; Π=1 means no degradation). S_ℓ: layerwise cosine similarity; $\underline{S} = \min_\ell S_\ell$ (bottleneck redundancy). $\lambda_{\max}(H)$: largest Hessian eigenvalue. W8/W4: weight-only INT8/INT4; W8A8/W4A4: weight-and-activation.

4.1 Hypercube Results: Synthetic Setup

We evaluate an 8-layer TinyTransformer across three data modalities and eight compression configurations with two recovery methods (details in Appendix E).

Compression configuration impact. Table 1 reports pre-recovery PPL ratios. Weight-only quantization remains at $1.00\times$ across all modalities. Depth pruning depends on data: MoG ($\underline{S}$=0.91) reaches only $1.06\times$ at depth-50%, while Student-t ($\underline{S}$=0.50) reaches $1.37\times$ — directionally matching the $\underline{S}$ ordering at this severity, though the pattern does not hold at depth-75% (see below). Unstructured pruning at 50% is essentially lossless ($\Pi \leq 1.00$) for all three modalities, but at 75% differences emerge: MoG shows slightly more degradation ($1.04\times$ vs. $1.01\times$ and $1.00\times$), consistent with its clustered structure creating sharper activation boundaries that sparse weight matrices struggle to preserve. Activation quantization (W4A4) produces variable degradation (0.99–$1.28\times$), whereas weight-only W4 remains lossless ($1.00\times$), confirming that activation discretisation is the dominant error source in low-bit regimes — a pattern that recurs at scale in Section 4.2. At depth-75%, Gaussian suffers most ($2.67\times$) despite intermediate $\underline{S}$=0.76, because its unimodal structure concentrates representational energy in a few critical layers; removing them destroys information that the remaining layers cannot redistribute.

Recovery behavior. Table 2 reports recovery ratios (recovered/compressed PPL; < 1 means improvement). For Gaussian depth-50%, KD achieves $0.78\times$ vs. SFT's $2.16\times$, recovering to $\sim 1.08\times$ post-recovery baseline. MoG shows smaller SFT–KD separation since compressed models are already near-baseline. KD shows lower (better) recovery ratios than SFT for Gaussian depth configurations, while MoG slightly favors SFT and Student-t shows ties. These distributional differences demonstrate that the data axis of the hypercube can alter method rankings even when architecture and compression level are held constant.

Table 4: Pre-recovery PPL ratios for eight core LLMs. W8/W4: weight-only INT8/INT4; W+A8: weight and activation INT8. Bold marks ratios above $100\times$.

	Unstructured Pruning		Depth Pruning		Quantization		
Model	50%	75%	50%	75%	W8	W4	W+A8
LLaMA-3.2-3B	1.98	**161**×	**141**×	**588**×	1.00	2.33	7.33
LLaMA-3.2-3B-IT	1.73	**104**×	62.10×	**875**×	1.00	2.20	1.90
LLaMA-3.1-8B	1.86	$>$**10**$^4\times$	**277**×	**1453**×	1.00	1.73	45.09
LLaMA-3.1-8B-IT	1.79	**1208**×	**211**×	**698**×	1.00	1.72	6.98
Qwen-2.5-7B	1.41	$>$**10**$^4\times$	$\gg 10^6\times^\star$	**7756**×	1.01	63.55	2.41
Qwen-2.5-7B-IT	1.38	$>$**10**$^4\times$	$\gg 10^6\times^\star$	$>$**10**$^4\times$	1.00	50.22	2.69
Gemma-2-2B	2.49	23.84	$>$**10**$^5\times$	$>$**10**$^4\times$	1.03	6.45	1.35
Gemma-2-2B-IT	1.32	16.85	**831**×	$>$**10**$^6\times$	1.00	4.74	1.10

*Qwen-2.5-7B/IT depth-50% ratios: $2.88/3.54 \times 10^9$.

4.2 Hypercube Results: Real LLMs

On real LLMs, a *synthetic-to-real transfer gap* emerges for depth pruning: synthetic models recover to near $1.0\times$ post-recovery, but real LLMs remain at 2.3–$8.3\times$ — a gap absent for quantization. Quantization perturbations are dense and bounded ($|\delta\theta_i| \leq \Delta/2$), so recovery landscapes are similar across scales. Depth pruning removes entire residual blocks whose energy depends on the learned representation hierarchy: synthetic 8-layer transformers develop relatively uniform layer representations ($\underline{S} \geq 0.50$), while real LLMs exhibit highly non-uniform profiles — e.g., Qwen-2.5-7B has $\underline{S}$=0.122 — meaning that even removing the most redundant half of layers disrupts the representational pipeline beyond what LoRA fine-tuning can recover.

Compression configuration impact. Table 4 reports pre-recovery PPL ratios for eight real models across all compression configurations. Weight-only INT8 (W8) is consistently stable, achieving $\Pi = 1.00$–$1.03\times$ across all models — validating the theoretical prediction that bounded, small perturbations lead to minimal degradation. Unstructured pruning at 50% gives moderate degradation (1.32–$2.49\times$), while 75% sparsity leads to much larger ratios ($16.9\times$ to above $10^4\times$).

Depth pruning causes the most severe degradation: median $554\times$ at depth-50%, with Qwen-2.5-7B ($\underline{S}$=0.122) reaching $2.88\times10^9\times$. This is consistent with Theorem 3.2, where error scales with both the residual branch magnitude and downstream curvature. The W4 results use naive RTN; calibrated methods (AWQ, BNB-NF4) dramatically reduce degradation (Table 3). W8A8 shows model-dependent behavior: LLaMA-3.1-8B reaches $45.09\times$ and LLaMA-3.2-3B $7.33\times$, while several models remain near $1.0\times$, supporting the weight-only vs. activation quantization distinction from the synthetic experiments.

Table 3: Pre-recovery PPL ratios for RTN, AWQ (Lin et al., 2024), and BNB-NF4 at 4-bit weights (lower is better).

Model	RTN 4-bit	AWQ 4-bit	BNB-NF4
LLaMA-3.2-3B	1.63×	1.07×	1.01×
LLaMA-3.2-3B-IT	1.75×	1.07×	1.06×
LLaMA-3.1-8B	1.64×	1.07×	1.08×
LLaMA-3.1-8B-IT	1.65×	1.05×	1.05×
Gemma-2-2B	8.88×	1.25×	1.03×
Gemma-2-2B-IT	8.94×	1.05×	1.03×
Qwen-2.5-7B-IT	12.4×	1.06×	1.06×

Table 3 compares calibrated quantization methods against naive RTN at 4-bit precision. AWQ and BNB-NF4 dramatically reduce degradation for RTN-sensitive architectures: Gemma-2-2B drops from $8.88\times$ to $1.25\times$ with AWQ, and Qwen-2.5-7B-IT from $12.4\times$ to $1.06\times$. The large RTN gap reflects high weight kurtosis (κ_W=8.97 and 21.9 respectively), which makes naive rounding particularly destructive.

Recovery: KD vs. SFT. Table 5 compares LoRA-SFT against three KD variants (forward-KL, reverse-KL/MiniLLM, on-policy/SGO) on Wanda-pruned models at 50% sparsity. All three KD variants outperform SFT by large margins: LLaMA-3.2-3B recovers to 2.58–$2.65\times$ with KD vs. $13.9\times$ with SFT; Gemma-2-2B to 0.62–$0.67\times$ vs. $52.0\times$. For 7B-scale models, KD achieves sub-baseline recovery (0.85–$0.86\times$), suggesting that the uncompressed teacher's distributional signal can push the pruned model below its own pre-compression perplexity under our evaluation protocol.

Table 6: **Base-model predictors of compression outcome (no compression required).** Individual Spearman rank correlations (ρ) between each base-model feature and $\log_{10}$(PPL ratio) across 9 models $\times$ available configurations (n=87). No individual feature is significant ($p > 0.28$). However, a 4-feature ridge regression (bottom row) achieves zero top-1 regret for models with ≥ 8 configurations (Table 7), though within-model rankings are determined by compression level (Appendix F.5).

Predictor (base-model only)	Spearman ρ	p-value	n	Notes
Individual features (n=87, all configs):				
$\lambda_{\max}(F)$ (Fisher curvature)	0.117	0.28	87	Not significant
λ_{ratio} ($\lambda_{\max}/\lambda_{\text{mean}}$)	0.086	0.43	87	Not significant
$\|g\|_{\max}$ (gradient norm)	0.079	0.47	87	Not significant
$\underline{S}$ (bottleneck S-score)	-0.024	0.82	87	Not significant
$\overline{S}$ (mean S-score)	0.020	0.85	87	Not significant
Weight kurtosis (κ_W)	0.057	0.60	87	Not significant
Composite predictor (leave-one-model-out ridge, λ=1):				
4-feature ridge†	—	—	87	**Top-1 regret = 0.0***

† Features: $\underline{S}$, λ_{ratio}, κ_W, nominal compression level.
* For models with ≥ 8 configs (6/9); see Table 7.

Table 5: Separate Wanda-50% recovery experiment at $N = 1000$: PPL relative to the uncompressed model. Compare objectives within this batch only.

		Knowledge Distillation		
Model	SFT	F-KL	R-KL	SGO
LLaMA-3.2-3B	13.9×	2.65×	2.58×	3.92×
Gemma-2-2B	52.0×	**0.62**×	0.67×	0.97×
LLaMA-3.1-8B	1.71×	0.86×	0.85×	0.86×
Qwen-2.5-7B	1.63×	**0.85**×	0.85×	0.85×

Under forward-KL with default hyperparameters (Table 15, Appendix), SFT appears competitive; the gap emerges once reverse-KL, on-policy objectives, and tuned hyperparameters are included. The mechanism is consistent with the analysis in Appendix B.6: KD transfers the teacher's soft-label distribution with richer gradient information than SFT's point-wise cross-entropy objective. The data-size sweep (Table 16, Appendix) is noisy and single-seed; KD is lower in most but not all cells. Even with the best recovery, depth-50% remains at 2.31–8.32×, indicating that these models lack the residual redundancy for safe depth pruning regardless of recovery strategy. Notably, for depth pruning SFT outperforms KD on all eight models (Table 15, Appendix; sign test p=0.004), reversing the ranking at Wanda-50% — catastrophic structural collapse from layer removal makes the teacher's distributional signal less useful than direct next-token supervision.

Structural predictors. Table 6 shows that no individual base-model feature correlates significantly with $\log_{10}(\Pi)$ across all 87 model–configuration pairs ($p > 0.28$). However, a 4-feature ridge regression using $\underline{S}$, curvature ratio, kurtosis, and nominal compression level achieves **zero top-1 regret** in leave-one-model-out configuration selection for 6/9 models (aggregate regret 0.161; Table 7). Within a held-out model, the base-model features are constants, so the predicted ranking among that model's candidates is driven entirely by compression level; the structural features shift predicted severity *across* models, identifying which are more compression-sensitive (Appendix F.5). The diagnostic requires only a single forward/backward pass ($\sim$5 min on one GPU, no trial compression needed).

Initialization state effects. The empirical Fisher curvature (Table 18, Appendix) shows that instruction tuning *reduces* $\lambda_{\max}(F)$ for LLaMA-3.2-3B (1.05 $\rightarrow$ 0.94) and Gemma-2-2B (275 $\rightarrow$ 140), but *increases* it for LLaMA-3.1-8B (3.92 $\rightarrow$ 4.12) and Qwen-2.5-7B (4.28 $\rightarrow$ 9.48). The sign is therefore not universal (2/4 families show smoothing). Where $\lambda_{\max}$ decreases, the result is consistent with alignment acting as implicit regularisation; the exceptions suggest that alignment effects on curvature are model- and scale-dependent.

Additional operator families. The supplementary measurements cover three other operator families (Tables 19, 20, and 21, Appendix F). At 50% retained rank, SVD-LLM (Wang et al., 2025) shows pre-recovery PPL ratios of 128× on LLaMA-3.2-3B and 10.7× on Gemma-2-2B, with post-KD ratios moving in opposite directions. FLAP structured pruning (An et al., 2024) at 25% removal is less catastrophic than depth pruning at the same fraction. StreamingLLM KV-cache eviction (Xiao et al., 2024) remains within 1.133× at 75% eviction, whereas KIVI (Liu et al., 2024) KV-INT2 reaches 5.65–8.77×. Qwen3-14B (Table 23, Appendix) confirms the INT8 near-baseline and depth-50% severe degradation patterns at larger scale.

Table 7: **Predict-then-verify: configuration selection from base-model features.** Leave-one-model-out ridge regression (λ=1) using four base-model features ($\underline{S}$, curvature ratio, weight kurtosis κ, nominal compression level) to predict $\log_{10}(\Pi)$ across all available configurations per model. Top-1 regret = $\log_{10}(\Pi_{\text{picked}}) - \log_{10}(\Pi_{\text{oracle}})$; 0.0 means the predictor selects the oracle-best configuration. **Models with ≥ 8 configurations achieve zero regret; models with only 3 configurations fail**, indicating that sufficient configuration coverage is required for the composite predictor to generalise.

Held-out Model	n configs	Spearman ρ	Top-1 Regret	Rank
Well-covered models (≥ 8 configs):				
Gemma-2-2B	14	0.248	**0.0**	1
Gemma-2-2B-IT	17	0.124	**0.0**	1
Gemma-3-4B-IT	8	0.655	**0.0**	1
LLaMA-3.2-3B	14	0.412	**0.0**	1
LLaMA-3.2-3B-IT	17	0.284	**0.0**	1
Qwen3.5-4B	8	0.546	**0.0**	1
Sparse-coverage models (3 configs):				
LLaMA-3.1-8B	3	—	0.18	2
LLaMA-3.1-8B-IT	3	—	0.20	2
Qwen-2.5-7B-IT	3	—	1.07	2
Aggregate	87	0.378	0.161	67%

Note: The three sparse-coverage models (LLaMA-8B variants, Qwen-7B-IT) have only pruning configurations in the training pool, so the predictor lacks exposure to quantisation and depth-pruning outcomes for similarly-sized models. With ≥ 8 configurations per model, the predictor achieves perfect top-1 selection. This suggests the diagnostic is reliable when the training pool spans the full Hypercube, and that expanding configuration coverage for 7B–8B models would likely restore zero regret.

Table 8: A100-80GB measured peak GPU memory, latency, and throughput for two LLaMA-Instruct models under FP16 and BNB kernels.

		LLaMA-3.2-3B-IT			LLaMA-3.1-8B-IT		
Backend	BS	Mem (GB)	ms/tok	tok/s	Mem (GB)	ms/tok	tok/s
FP16	1	6.53	33.5	29.9	17.7	39.9	25.1
	8	7.21	4.8	207.9	18.6	5.6	179.5
	16	7.98	2.5	399.0	19.6	3.1	321.0
BNB-INT8	1	3.73	152.7	6.5	9.26	188.8	5.3
	8	4.54	22.6	44.3	10.4	27.0	37.0
	16	5.46	11.7	85.5	11.7	14.0	71.6
BNB-NF4	1	2.40	55.9	17.9	5.96	65.5	15.3
	8	3.05	9.0	111.7	6.78	10.4	96.0
	16	3.85	4.6	217.6	7.83	5.6	178.5

5 Discussion

The results support a practical pre-deployment protocol: (1) compute four base-model features — $\underline{S}$, curvature ratio $\lambda_{\max}(F)/\bar{\lambda}(F)$, weight kurtosis κ_W, and nominal compression level — from a single forward/backward pass ($\sim$5 min on one GPU); (2) compare across models via 4-feature ridge regression (Table 7), noting that within a model the candidate ranking is set by compression level; (3) select the configuration matching the deployment objective: W8 for quality, calibrated W4 (AWQ/NF4) for memory, or depth pruning only when large speedup justifies the accuracy loss and $\underline{S}$ is well above the range observed here (all evaluated LLMs have $\underline{S} \leq 0.55$, where depth-50% recovery remains poor).

W8 as the safest default. Weight-only INT8 is the most stable configuration, achieving $\Pi = 1.00$–$1.03\times$ across all models with no recovery needed. Real-kernel benchmarks (Table 8) show BNB-NF4 achieves $\sim 37\%$ memory footprint relative to FP16, making it preferred when the goal is fitting models into GPU VRAM rather than improving throughput. INT8 is $\sim 4.6\times$ slower than FP16 at batch size 1 due to dequantization overhead, and the ratio persists at larger batch sizes ($\sim 4.7\times$ at batch 16). Figure 2 confirms that no configuration achieves both $> 1.5\times$ speedup and > 0.60 mean zero-shot accuracy, highlighting the accuracy–latency tradeoff the hypercube's objective $\mathcal{J}(m, c) = \mathcal{E}_m(c) - \gamma \mathcal{S}_m(c)$ is designed to navigate.

Depth pruning remains high-risk. All evaluated LLMs have $\underline{S} \leq 0.55$; depth-50% post-recovery ratios remain at 2.31–$8.32\times$, indicating insufficient residual redundancy for safe depth pruning. Depth pruning may be viable for architectures designed with layer redundancy or where throughput

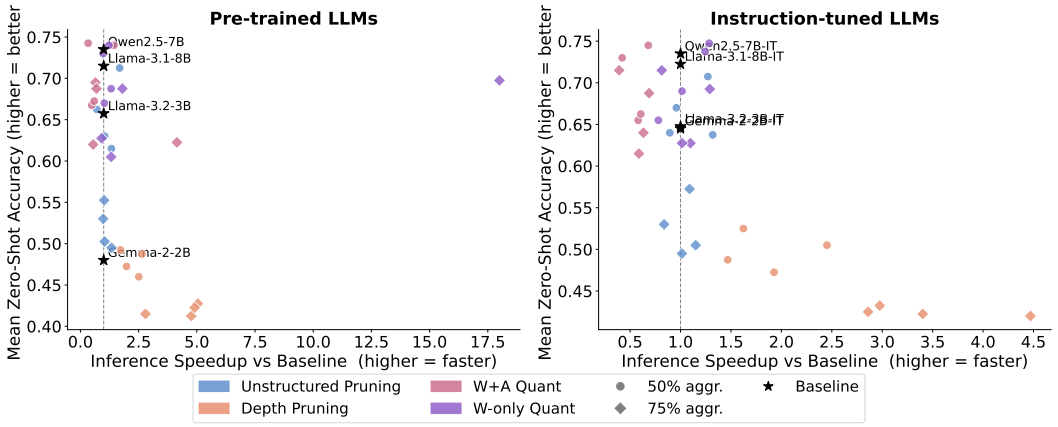

Figure 2: Pareto frontier: speedup vs. accuracy. No configuration achieves both $> 1.5\times$ speedup and > 0.60 mean zero-shot accuracy. W-only INT8 best preserves quality, while depth-75% yields the highest speedup at substantial accuracy cost.

is the primary objective (Tables 13, 14, Appendix); the reversal of the KD advantage in this regime (Section 4.2) further constrains viable recovery pipelines. Structured pruning (FLAP; An et al. 2024) is less catastrophic at 25% removal (Table 20, Appendix), though SFT recovery remains essential.

Composing compression methods. Compression methods are often applied sequentially. Our composition analysis (Table 24, Appendix) shows that the interaction is *not additive*: Wanda-50% followed by W4 yields degradation worse than the sum of individual degradations, because pruning concentrates weight energy into fewer parameters, amplifying sensitivity to subsequent quantization noise. Weight-only INT8 after moderate pruning should add limited further overhead, since W8 perturbations are bounded and small ($\Pi \approx 1.00\times$ alone; Theorem 3.1).

Limitations. Our evaluation covers 2B–14B parameters; extending beyond 30B remains an important direction. The local second-order bounds serve as qualitative worst-case characterizations rather than precise predictors under severe structural changes such as depth pruning at 75%. Recovery uses a single seed and C4 training with WikiText-2 evaluation; the domain mismatch may inflate SFT recovery ratios more than KD ratios, since the teacher's distributional signal is less domain-sensitive than point-wise labels. Within a held-out model, the composite diagnostic ranking is largely driven by compression level; the base-model features shift predictions across models but do not reorder configurations within one. Extended results including recovery details (Table 15) and zero-shot accuracy (Table 12) are in Appendix F.

6 Conclusion

We introduced the Compression Hypercube, a five-dimensional framework for systematically analyzing LLM compression, and evaluated it across synthetic environments and real LLMs (2B–14B, three architecture families). Post-recovery PPL ratios span from near-lossless under weight-only INT8 ($\Pi \approx 1.00\times$) to severe degradation under depth pruning ($\Pi \approx 2$–$8\times$ after recovery), with theoretical bounds correctly predicting the qualitative ordering. Depth pruning causes the most severe degradation across all evaluated models ($\underline{S} \leq 0.55$); high weight kurtosis (κ_W) is associated with large RTN-quantization gaps. KD with proper objectives (reverse-KL, on-policy) outperforms SFT at Wanda-50% sparsity in one experiment batch; this advantage disappears under forward-KL with default hyperparameters and reverses entirely for depth pruning, where SFT dominates across all eight models. A 4-feature composite diagnostic achieves zero top-1 regret in leave-one-model-out configuration selection for 6/9 models, providing an initial cross-model sensitivity estimate ($\sim$5 min, one GPU, no trial compression). The resulting workflow — measure structural redundancy and curvature, compare models via the composite score, select by accuracy–latency objectives — provides an initial screening step, though within-model candidate rankings are driven by compression level rather than the structural features.

AI USE STATEMENT

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

In preparing this paper, we made limited and clearly scoped use of generative AI tools — specifically Claude, ChatGPT, and Grammarly. Their use included language editing, presentation, auxiliary scripting, and consistency checks.

The authors conceived and conducted the experiments, reviewed any AI-assisted ancillary code, and take full responsibility for the final content of this work — including all text, claims, and artifacts produced with the aid of generative AI — and for its correctness.

ETHICS STATEMENT

We are not aware of any ethical concerns specific to the methods or experiments presented here.

REPRODUCIBILITY STATEMENT

Section 2 reports the synthetic data generation process (details in Appendix E). Real experimental setup is in Section 2.4, with full specifications (LoRA hyperparameters, evaluation context length, hardware, BOS handling) in Appendix C. Recovery uses C4 samples while evaluation uses WikiText-2; the domain mismatch is intentional but means sub-baseline ratios ($\Pi < 1.0$) may partly reflect domain adaptation. Specific results were independently verified: (i) Gemma-2-2B BOS sensitivity confirmed (PPL drops $15.8\times$ with BOS); (ii) Qwen3-14B baseline PPL independently measured at 10.58 (bf16), consistent with fp16 baseline of 11.17. A matched-budget KD–SFT reproduction and independent latency audit are identified as priorities in Appendix C. Source code for all real and synthetic experiments is supplied under the supplementary material.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Yongqi An, Xu Zhao, Tao Yu, Ming Tang, and Jinqiao Wang. Fluctuation-based adaptive structured pruning for large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 10865–10873, 2024.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.

Sungmin Cha and Kyunghyun Cho. Why knowledge distillation works in generative models: A minimal working explanation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=hg5UGeAr1Q.

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.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

Zhen Dong, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Hawq: Hessian aware quantization of neural networks with mixed-precision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 293–302, 2019.

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

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJl-b3RcF7.

Elias Frantar and Dan Alistarh. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International conference on machine learning*, pp. 10323–10337. PMLR, 2023.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022.

Yash Goel, Ayan Sengupta, and Tanmoy Chakraborty. Position: Enough of scaling LLMs! lets focus on downscaling. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=CYJlJgEzZs.

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

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28, 2015.

Babak Hassibi and David Stork. Second order derivatives for network pruning: Optimal brain surgeon. *Advances in neural information processing systems*, 5, 1992.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

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

Jongwoo Ko, Tianyi Chen, Sungnyun Kim, Tianyu Ding, Luming Liang, Ilya Zharkov, and Se-Young Yun. Distillm-2: A contrastive approach boosts the distillation of llms. In *International Conference on Machine Learning*, pp. 31044–31062. PMLR, 2025.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Andrey Kuzmin, Markus Van Baalen, Yuwei Ren, Markus Nagel, Jorn Peters, and Tijmen Blankevoort. Pruning vs quantization: Which is better? *Advances in Neural Information Processing Systems*, 36, 2023. URL https://arxiv.org/abs/2307.02973. arXiv:2307.02973.

Yann LeCun, John Denker, and Sara Solla. Optimal brain damage. *Advances in neural information processing systems*, 2, 1989.

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. *Proceedings of machine learning and systems*, 6: 87–100, 2024.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. Kivi: A tuning-free asymmetric 2bit quantization for kv cache. *arXiv preprint arXiv:2402.02750*, 2024.

Zhenyan Lu, Xiang Li, Dongqi Cai, Rongjie Yi, Fangming Liu, Xiwen Zhang, Nicholas D. Lane, and Mengwei Xu. Small language models: survey, measurements, and insights, September 2024. URL http://arxiv.org/abs/2409.15790. arXiv:2409.15790.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems*, 36:21702–21720, 2023.

Xin Men, Mingyu Xu, Qingyu Zhang, Qianhao Yuan, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. Shortgpt: Layers in large language models are more redundant than you expect. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20192–20204, 2025.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Chenxu Niu, Wei Zhang, Jie Li, Yongjian Zhao, Tongyang Wang, Xi Wang, and Yong Chen. Tokenpowerbench: Benchmarking the power consumption of llm inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 32582–32590, 2026.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

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, January 2025. URL http://arxiv.org/abs/2412.15115. arXiv:2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

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.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. A simple and effective pruning approach for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=PxoFut3dWW.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Xin Wang, Yu Zheng, Zhongwei Wan, and Mi Zhang. Svd-llm: Truncation-aware singular value decomposition for large language model compression. In *International Conference on Learning Representations*, volume 2025, pp. 19299–19319, 2025.

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 *International conference on machine learning*, pp. 38087–38099. PMLR, 2023.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, volume 2024, pp. 21875–21895, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

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, 2019.

Zihan Zhou, Simon Kurz, and Zhixue Zhao. Revisiting pruning vs quantization for small language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 12055–12070, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.645. URL https://aclanthology.org/2025.findings-emnlp.645/.

Boqin Zhuang, Jiacheng Qiao, Mingqian Liu, Mingxing Yu, Ping Hong, Rui Li, Xiaoxia Song, Xiangjun Xu, Xu Chen, Yaoyao Ma, et al. Beyond benchmarks: The economics of ai inference. *arXiv preprint arXiv:2510.26136*, 2025.

APPENDIX CONTENTS

A Related Work **14**

B Proofs of Theoretical Results **16**

B.1 Proof of Theorem 3.1 (Generic Compression Error) 16

B.2 Supporting Results for Unstructured Pruning 16

B.3 Proof of Theorem 3.2 and Corollary 3.3 16

B.4 Supporting Results for Quantization 17

B.5 Proof of Proposition 3.4 (Curvature-Amplified Compression Sensitivity) 17

B.6 CE and KD Logit-Space Analysis 18

C Detailed Experimental Setup **18**

D Model Profiles **19**

E Synthetic Environment Details **20**

F Additional Measurements **21**

F.1 Model degradation and task accuracy 21

F.2 Measured inference behavior 21

F.3 Recovery objectives and data size 23

F.4 Low-rank, structured, and KV-cache compression 23

F.5 Exploratory leave-one-model-out selection 26

F.6 Additional checkpoints and operator combinations 26

G Interpretation of the Synthetic Comparison **26**

A RELATED WORK

Model compression methods. Model compression is primarily achieved through pruning, quantization, and knowledge distillation (KD). Pruning removes parameters either at the unstructured level (Han et al., 2015; Frankle & Carbin, 2019; Frantar & Alistarh, 2023) or by removing higher-level architectural components such as layers or attention blocks (Ma et al., 2023; Men et al., 2025; Gromov et al., 2025). Early approaches rely on magnitude-based heuristics, while more recent one-shot methods, including SparseGPT and Wanda, estimate parameter importance using local curvature approximations or activation statistics (Frantar & Alistarh, 2023; Sun et al., 2024). These approaches enable high sparsity without full retraining, but their effectiveness depends on how well the importance metric aligns with the loss landscape. In particular, structured pruning methods such as layer removal can introduce large changes in intermediate representations, leading to performance degradation that is not always proportional to the number of removed parameters.

Quantization reduces model precision, typically to 8-bit or 4-bit representations, using post-training techniques such as GPTQ, AWQ, and SmoothQuant (Frantar et al., 2022; Lin et al., 2024; Xiao et al., 2023). These methods aim to minimize quantization error under calibration data and often preserve model quality in practice, especially when weight distributions are well-behaved and activation scaling is properly handled. Compared to pruning, quantization induces bounded perturbations

across parameters rather than removing them entirely, which has been associated with more stable performance across a range of settings.

Knowledge distillation transfers information from a larger teacher model to a smaller or compressed student (Hinton et al., 2015; Sanh et al., 2019; Gu et al., 2024; Ko et al., 2025). It is commonly used as a post-compression recovery mechanism, where the student is trained to match the output distribution of the teacher. Variants differ in the choice of divergence, temperature scaling, and training objectives, but their effectiveness depends on the alignment between the teacher distribution and the student model capacity.

Theory of compression and local approximations. Theoretical analysis of compression has traditionally focused on local approximations of the loss function. Classical methods such as Optimal Brain Damage and Optimal Brain Surgeon use second-order Taylor expansions to estimate the impact of removing parameters (LeCun et al., 1989; Hassibi & Stork, 1992). These approaches rely on curvature information, typically through the Hessian or its approximations, to identify parameters that can be removed with minimal loss increase. More recent work extends this perspective to large models through Hessian-aware pruning and quantization methods (Dong et al., 2019; Frantar & Alistarh, 2023), although exact curvature computation remains intractable in practice and is often approximated.

This line of work provides a useful characterization of parameter-level perturbations, but it does not directly address structural modifications such as layer removal, where the effect is better understood at the level of intermediate representations rather than individual parameters.

Representation redundancy and layer-wise structure. A complementary line of work studies redundancy in neural representations. Methods such as SVCCA and CKA measure similarity between layer representations and show that deep networks often learn correlated intermediate features (Raghu et al., 2017; Kornblith et al., 2019). This observation has motivated layer-dropping and depth-reduction approaches, which assume that redundant layers can be removed with limited impact. However, these methods typically rely on empirical similarity measures and do not provide explicit bounds on how representation changes translate into loss differences. As a result, the connection between representation redundancy and compression-induced error remains largely heuristic.

Empirical evaluations and practical observations. Several studies have compared compression methods empirically across tasks and model scales. These works often report that quantization preserves performance more reliably than pruning, particularly at moderate compression levels (Kuzmin et al., 2023; Zhou et al., 2025). However, the observed behavior varies across models and tasks, and comparisons are typically performed under fixed settings without systematically varying recovery strategies or model initialization.

For large language models, instruction tuning and alignment procedures change the model's initialization state (Ouyang et al., 2022; Rafailov et al., 2023). The direction of any resulting change in compression sensitivity remains an empirical question. KD can also change generation quality and distributional coverage in different ways (Cha & Cho, 2025); our PPL comparisons do not directly measure that trade-off.

Gap. Existing work provides insights into individual aspects of compression, including parameter importance, curvature sensitivity, and representation redundancy. Prior unified benchmarks (Kuzmin et al., 2023; Zhou et al., 2025) compare methods under fixed settings but do not systematically vary recovery strategy, initialization state, or data distribution alongside method choice; as a result, they identify *which* method is best under their conditions but not *why* the ranking changes under different conditions. Standard second-order sensitivity analyses (LeCun et al., 1989; Hassibi & Stork, 1992; Dong et al., 2019) characterize parameter-level perturbations but do not extend to activation-level structural changes (depth pruning) or to the interaction between compression-induced damage and recovery-strategy choice. Our study combines a conditional single-block activation bound with controlled comparisons of operator, recovery, initialization, and data conditions. The empirical design has incomplete coverage of the full Cartesian product.

B PROOFS OF THEORETICAL RESULTS

All Taylor statements below are local around the evaluated parameter or activation point. We do not assume global Lipschitz continuity of the transformer. When a second-order remainder is used, the statement explicitly assumes twice continuous differentiability on the relevant line segment.

B.1 PROOF OF THEOREM 3.1 (GENERIC COMPRESSION ERROR)

Define $g(t) = \mathcal{L}(\theta + t\,\delta\theta)$. By the fundamental theorem of calculus, $g(1) - g(0) = \int_0^1 g'(t)\,dt$. Since $g'(t) = \nabla\mathcal{L}(\theta + t\,\delta\theta)^\top \delta\theta$, expand the gradient along the same segment:

$$\nabla\mathcal{L}(\theta + t\,\delta\theta) = \nabla\mathcal{L}(\theta) + \int_0^t H(\theta + s\,\delta\theta)\,\delta\theta\,ds.$$

Substituting and swapping the order of integration ($\int_0^1 \int_0^t a(s)\,ds\,dt = \int_0^1 (1-s)a(s)\,ds$) gives the integral remainder form, completing the proof.

Corollary (used in subsequent results). If $\lambda_{\max}(H(\theta + t\,\delta\theta)) \le \Lambda$ for all $t \in [0,1]$ and θ is a stationary point, then $\mathcal{L}(\theta + \delta\theta) - \mathcal{L}(\theta) \le \frac{1}{2}\Lambda\|\delta\theta\|_2^2$, obtained by bounding the integrand and integrating $(1-t)$ over $[0,1]$. ■

B.2 SUPPORTING RESULTS FOR UNSTRUCTURED PRUNING

Lemma B.1 (Exact parameter error of unstructured pruning)**.** *For any deterministic or random pruning mask $M \in \{0,1\}^{m\times n}$,*

$$\|\tilde{W} - W\|_F^2 = \sum_{i,j}(1 - M_{ij})W_{ij}^2.$$

If M is random, $\mathbb{E}_M[\|\tilde{W} - W\|_F^2] = \sum_{i,j}\mathbb{P}(M_{ij} = 0)W_{ij}^2$.

Proof. Since $\tilde{W} = M \odot W$, we have $\tilde{W} - W = -(1 - M) \odot W$. Therefore $\|\tilde{W} - W\|_F^2 = \sum_{i,j}(1 - M_{ij})^2 W_{ij}^2$. Because $M_{ij} \in \{0,1\}$, $(1 - M_{ij})^2 = 1 - M_{ij}$. Linearity of expectation gives the random-mask case. □

Theorem B.2 (Loss bound for unstructured pruning)**.** *Let θ be a stationary point of $\mathcal{L}$. Let unstructured pruning zero out the parameter set S^c with $|S^c| = cP$, inducing perturbation $\delta\theta_i = -\theta_i$ for $i \in S^c$ and $\delta\theta_i = 0$ otherwise. Assume $\mathcal{L}$ is twice continuously differentiable on $\{\theta + t\,\delta\theta : t \in [0,1]\}$ and $\lambda_{\max}(H(\theta + t\,\delta\theta)) \le \Lambda_U$. Then*

$$\mathcal{L}(\theta + \delta\theta) - \mathcal{L}(\theta) = \int_0^1 (1-t)\,\theta_{S^c}^\top H_{S^c,S^c}(\theta + t\,\delta\theta)\,\theta_{S^c}\,dt \;\le\; \frac{\Lambda_U}{2}\|\theta_{S^c}\|_2^2.$$

Proof. Apply Theorem 3.1: since θ is stationary, the first-order term vanishes. The perturbation is supported on S^c, so $\delta\theta^\top H\,\delta\theta = \theta_{S^c}^\top H_{S^c,S^c}\,\theta_{S^c}$. Bounding the Hessian by Λ_U and integrating $(1-t)$ gives the bound. □

Practical implication. Unstructured pruning is safe when the removed coordinates have small curvature-weighted energy. Magnitude pruning is a crude proxy; Hessian-weighted methods (OBS/SparseGPT) approximate the criterion more directly.

B.3 PROOF OF THEOREM 3.2 AND COROLLARY 3.3

Proof of Theorem 3.2. Set $v = -F_\ell(h_\ell)$ and $g(t) = \phi_\ell(a + tv)$. Applying the fundamental theorem of calculus twice yields

$$\phi_\ell(a + v) - \phi_\ell(a) = \nabla\phi_\ell(a)^\top v + \int_0^1 (1-t)v^\top\nabla^2\phi_\ell(a + tv)v\,dt.$$

The eigenvalue assumption bounds the integral by $\Lambda_\ell \|v\|^2/2$. Substituting $v = -F_\ell(h_\ell)$ proves the claim. No stationarity of an intermediate activation is assumed.

Proof of Corollary 3.3. The actual residual branch satisfies $F_\ell(h_\ell) = h_{\ell+1} - h_\ell$. Under the stated unit-norm assumption, its squared norm is $2 - 2h_{\ell+1}^\top h_\ell$ pointwise; expectation gives $2(1 - S_\ell)$. Under the separate orthogonality and uniform-curvature assumptions, expectation of the theorem's right-hand side is $\Lambda_\ell(1 - S_\ell)$. For cosine similarities measured after separate normalization of unnormalized states, this exact residual identity does not hold. ■

B.4 Supporting Results for Quantization

Lemma B.3 (Deterministic quantization error bound)**.** *Let Q_b be a b-bit symmetric uniform quantizer over range $[-R, R]$ with step $\Delta_b = 2R/(2^b - 1)$ and W_{clip} the entrywise clipping of W to $[-R, R]$. If the clipping endpoints are grid points and clipping maps to those endpoints,*

$$\|Q_b(W) - W\|_F^2 \;\le\; \frac{N\Delta_b^2}{4} + \|W - W_{\text{clip}}\|_F^2 \;=\; \frac{NR^2}{(2^b-1)^2} + \|W - W_{\text{clip}}\|_F^2,$$

where $N = mn$ is the total number of entries.

Proof. Decompose: $Q_b(W) - W = (Q_b(W_{\text{clip}}) - W_{\text{clip}}) + (W_{\text{clip}} - W)$. When clipping endpoints are grid points, the rounding residual r and the clipping residual s have disjoint support, so $\|Q_b(W) - W\|_F^2 = \|r\|_F^2 + \|s\|_F^2$. For unclipped entries, nearest-neighbor rounding ensures $|r_{ij}| \le \Delta_b/2$, giving $\|r\|_F^2 \le N\Delta_b^2/4$. Substituting $\Delta_b = 2R/(2^b - 1)$ completes the proof. □

Theorem B.4 (Loss bound for quantization)**.** *Let quantization induce $\delta\theta_Q = Q_b(\theta) - \theta$. Suppose θ is stationary and $\lambda_{\max}(H(\theta + t\,\delta\theta_Q)) \le \Lambda_Q$ along the segment. Then*

$$\mathcal{L}(\theta + \delta\theta_Q) - \mathcal{L}(\theta) \;\le\; \frac{\Lambda_Q}{2} \sum_i \left[\frac{N_i R_i^2}{(2^b-1)^2} + \|W_i - W_{i,\text{clip}}\|_F^2 \right].$$

Proof. Apply Theorem 3.1: since θ is stationary, $\mathcal{L}(\theta + \delta\theta_Q) - \mathcal{L}(\theta) \le (\Lambda_Q/2)\|\delta\theta_Q\|_2^2$. The squared perturbation decomposes across matrices: $\|\delta\theta_Q\|_2^2 = \sum_i \|Q_b(W_i) - W_i\|_F^2$. Apply Lemma B.3 to each matrix and sum. □

Practical implication. Quantization is safe when clipping error is small. 8-bit quantization is often nearly lossless because the rounding term is suppressed by $(2^8 - 1)^2 = 65025$. 4-bit quantization may fail when weight ranges are dominated by outliers.

B.5 Proof of Proposition 3.4 (Curvature-Amplified Compression Sensitivity)

For a symmetric Hessian H, the Rayleigh quotient satisfies $\lambda_{\max}(H) = \sup_{\|v\|_2=1} v^\top H v$. For $\|\delta\theta\|_2 = r$, write $\delta\theta = rv$ with $\|v\|_2 = 1$:

$$\sup_{\|\delta\theta\|_2 = r} \frac{1}{2}\delta\theta^\top H\,\delta\theta = \frac{r^2}{2} \sup_{\|v\|_2=1} v^\top H v = \frac{1}{2}\lambda_{\max}(H)\, r^2.$$

If $\lambda_{\max}(H_{\text{instruct}}) > \lambda_{\max}(H_{\text{pre}})$, the worst-case value is strictly larger for the instruction-tuned model. ■

Remark. This is a worst-case sensitivity statement. A larger $\lambda_{\max}$ does not guarantee that every compression perturbation direction suffers more damage, only that the worst-case ceiling is higher. This bound does not predict instruction-following behavior or the effect of a particular compression operator.

B.6 CE AND KD LOGIT-SPACE ANALYSIS

Theorem B.5 (Gradient and Hessian of CE and forward-KL)**.** *For $p = \mathrm{softmax}(z)$, one-hot label y, and teacher distribution q:*

$$\nabla_z \mathcal{L}_{\mathrm{CE}}(z; y) = p - y, \qquad \nabla_z \mathcal{L}_{\mathrm{FKL}}(z; q) = p - q.$$

Both losses share the same logit-space Hessian:

$$\nabla_z^2 \mathcal{L}_{\mathrm{CE}} = \nabla_z^2 \mathcal{L}_{\mathrm{FKL}} = \mathrm{diag}(p) - pp^\top,$$

which is positive semidefinite.

Proof. For any distribution r, let $\mathcal{L}_r(z) = -\sum_k r_k \log p_k = -r^\top z + \log \sum_j e^{z_j}$. Then $\nabla_z \mathcal{L}_r = p - r$. For CE, $r = y$; for forward-KL, $\mathrm{KL}(q\|p) = -H(q) + \mathcal{L}_q(z)$ (where $H(q)$ is independent of z), so $\nabla_z \mathrm{KL}(q\|p) = p - q$. The Hessian follows from $\partial p_i / \partial z_j = p_i(\mathbf{1}[i = j] - p_j)$, giving $J_{\mathrm{softmax}} = \mathrm{diag}(p) - pp^\top$. Positive semidefiniteness holds since $v^\top(\mathrm{diag}(p) - pp^\top)v = \mathrm{Var}_{k\sim p}(v_k) \geq 0$. □

Theorem B.6 (KD as cross-entropy with a softened target)**.** *Let $\mathcal{L}_\alpha = \alpha\,\mathrm{CE}(y, p) + (1-\alpha)\,\mathrm{KL}(q\|p)$ and $r_\alpha = \alpha y + (1-\alpha)q$. Then $\mathcal{L}_\alpha = \mathrm{CE}(r_\alpha, p) - (1-\alpha)H(q)$, so $\nabla_z \mathcal{L}_\alpha = p - r_\alpha$ and $\nabla_z^2 \mathcal{L}_\alpha = \mathrm{diag}(p) - pp^\top$.*

Proof. $\mathrm{KL}(q\|p) = -H(q) + \mathrm{CE}(q, p)$. Substituting: $\mathcal{L}_\alpha = \alpha\,\mathrm{CE}(y, p) + (1-\alpha)\,\mathrm{CE}(q, p) - (1-\alpha)H(q) = \mathrm{CE}(\alpha y + (1-\alpha)q, p) - (1-\alpha)H(q)$. Apply Theorem B.5 with $r = r_\alpha$. □

Implication. CE and forward-KL KD have *identical* local curvature in logit space; their difference is entirely in the gradient target. Forward-KL KD preserves teacher behavior absent from hard labels, which is useful when recovery data are scarce. This equivalence does not extend to reverse-KL or on-policy KD, whose logit Hessians differ.

C DETAILED EXPERIMENTAL SETUP

Compression configurations.

- **Unstructured pruning**: Wanda (Sun et al., 2024) at 50% and 75% sparsity
- **Depth pruning**: ShortGPT-style layer removal (Men et al., 2025) at 50% and 75%
- **Quantization**: RTN (W8A8, W4A4, W8A16, W4A16); calibrated methods (AWQ (Lin et al., 2024), BNB-NF4, BNB-INT8)
- **Structured pruning**: FLAP (An et al., 2024) at 25%
- **Other**: SVD-LLM (Wang et al., 2025) at 50% retained rank; StreamingLLM (Xiao et al., 2024) cache eviction and KIVI (Liu et al., 2024) KV quantization on a 2048-token suffix-PPL protocol

Recovery fine-tuning. LoRA-SFT (r=8, α=16, learning rate 2×10^{-4}, AdamW) on C4 recovery corpus; forward-KL KD from frozen uncompressed parent; reverse-KL KD (MiniLLM (Gu et al., 2024)); on-policy SGO distillation (Ko et al., 2025). Data sizes: $N \in \{50, 250, 1000, 4000\}$ C4 samples. All recovery runs use a single random seed; recovery sensitivity across seeds and hyperparameters is discussed in the Limitations.

Reproducibility priorities. The most consequential follow-up is a matched recovery rerun for the nominally identical Wanda-50% rows in the two batches. It should freeze the checkpoint hash, tokenizer and BOS policy, exact C4 examples and their order, LoRA placement and optimization schedule, random seeds, and WikiText-2 evaluation code before comparing SFT, forward-KL, reverse-KL, and on-policy KD at equal token budgets. Reporting confidence intervals across seeds would distinguish objective effects from training variance. The same audit should re-evaluate Gemma-2-2B with a consistent BOS convention and reproduce latency under one documented backend per quantizer. These steps would turn the present conditional observations into a stronger predictive study without changing the useful five-axis organization of the experiments.

Table 9: Model profiles. $\overline{S}$ and $\underline{S}$: mean and minimum S-score of adjacent hidden-state cosine similarities (measured at baseline). $\lambda_{\max}$: largest Hessian eigenvalue (power iteration). *IT* = Instruct (RLHF-aligned). PPL on WikiText-2.

Model	State	Params	Layers	Base PPL	$\lambda_{\max}$	$\overline{S}$	$\underline{S}$
LLAMA-3.2-3B	Pretrained	3.2B	28	10.20	24.30	0.90	0.36
LLAMA-3.2-3B-IT	Instruct	3.2B	28	15.00	70.00	0.91	0.39
LLAMA-3.1-8B	Pretrained	8.0B	32	8.10	45.70	0.91	0.34
LLAMA-3.1-8B-IT	Instruct	8.0B	32	9.40	40.90	0.91	0.41
QWEN2.5-7B	Pretrained	7.6B	28	8.70	22.10	0.94	0.12
QWEN2.5-7B-IT	Instruct	7.6B	28	9.90	33.30	0.94	0.16
GEMMA-2-2B	Pretrained	2.6B	26	194.00	576.50	0.80	0.51
GEMMA-2-2B-IT	Instruct	2.6B	26	27.80	223.70	0.87	0.53

Evaluation. Perplexity on WikiText-2 (Merity et al., 2016) with context length 2048 and stride 2048 (non-overlapping). Zero-shot accuracy on HellaSwag, ARC-Challenge, WinoGrande, and PIQA via lm-evaluation-harness. Latency and memory on NVIDIA A100-80GB at batch sizes 1, 8, and 16. The Gemma-2-2B evaluation does not prepend a BOS token (standard HuggingFace pipeline), which produces anomalously high baseline PPL for the pretrained variant (Section D). Pre-deployment diagnostics from a 50-sample C4 calibration batch: $\underline{S}$ (adjacent hidden-state cosine similarities), empirical Fisher $\lambda_{\max}(F)$ and curvature ratio $\lambda_{\text{ratio}} = \lambda_{\max}(F)/\bar{\lambda}(F)$, weight kurtosis κ_W.

Omitted paradigms. Our evaluation does not cover mixed-precision quantization, progressive or iterative pruning, channel-wise structured pruning beyond FLAP, activation-aware quantization (SmoothQuant), or distillation from a partially compressed intermediate teacher. These represent important production configurations and extending the framework to cover them is a direction for future work.

Implications for comparing configurations. The evidence suggests a simple reporting discipline for future compression studies. First, state the operator and its implementation, since RTN W4, AWQ, NF4, and weight-plus-activation quantization are materially different treatments despite similar bit labels. Second, report both the pre-recovery and final model under the same tokenizer and evaluation set. Selecting the best run among different data budgets is a useful upper envelope, but it should be accompanied by fixed-budget results when comparing recovery algorithms. Third, separate memory capacity, measured decoding speed, and task quality. A configuration can preserve PPL and reduce VRAM while running more slowly on the available kernel, as the BNB-INT8 measurements demonstrate. The hypercube serves as a record of these conditions and a guide to the comparisons still missing; it does not by itself identify a globally optimal compressor.

D MODEL PROFILES

Table 9 lists the eight paired LLMs with their structural diagnostics. $\overline{S}$ and $\underline{S}$ are the mean and minimum layer-wise S-scores measured on a 50-sample C4 calibration batch before any compression. $\lambda_{\max}$ values in this table are from a power-iteration Hessian estimator; empirical Fisher values for the paired models are reported in Table 18. A separate BOS-enabled baseline check for pretrained Gemma-2-2B gave PPL 11.0, versus 172.8 in the main no-BOS experiment; the BOS-enabled instruct baseline was 15.9. Compressed checkpoints were not all reevaluated under the BOS-enabled protocol, so the corrected baseline alone cannot yield corrected compression ratios. The reported Gemma ratios retain the original denominator and should not support cross-family or instruction-tuning comparisons.

Table 10: Sequential redundancy (S-scores) for the three synthetic AR(1) environments. Values are means over multiple seeds. Layer pairs i–$(i{+}1)$ measure cosine similarity of adjacent hidden states.

Modality	$\overline{S}$	$\underline{S}$	$\lambda_{\max}(H)$	L0–L1	L1–L2	L2–L3	L3–L4	L4–L5	L5–L6	L6–L7
Gaussian	0.95	0.76	1.31	0.76	0.93	0.97	0.99	1.00	1.00	1.00
MoG	0.98	0.91	1.19	0.91	0.98	0.99	1.00	1.00	1.00	1.00
Student-t	0.91	0.50	2.12	0.50	0.97	0.96	0.96	0.99	1.00	1.00

E Synthetic Environment Details

Symbol	Definition
$f_\theta : \mathcal{X} \to \mathcal{Y}$	Neural language model, parameters $\theta \in \mathbb{R}^P$
$h_\ell(x) \in \mathbb{S}^{d-1}$	Unit-normalized hidden state after layer ℓ
$F_\ell(h_\ell)$	Residual branch output; $h_{\ell+1} = h_\ell + F_\ell(h_\ell)$
$S(l, l{+}1)$	Layer-l sequential redundancy: $\mathbb{E}_x[\cos(h_l, h_{l+1})]$
$\underline{S}$	Bottleneck redundancy: $\min_l S(l, l{+}1)$
$\delta\theta = \tilde{\theta} - \theta$	Compression-induced parameter perturbation
$H = \nabla^2_\theta \mathcal{L}$	Loss Hessian; $\lambda_{\max}(H)$ is its largest eigenvalue
λ_{ratio}	Curvature ratio: $\lambda_{\max}(F)/\bar{\lambda}(F)$ where $\bar{\lambda} = \text{tr}(F)/d$
S^c	Pruned parameter index set (Section 2.2; distinct from S_ℓ)
$\mathcal{S}_m(c)$	System-level speedup (Section 2.1; distinct from S_ℓ)
$z_S, z_T;\ p_S, p_T$	Student/teacher logits and softmax probabilities

Table 10 reports the layer-wise S-scores for the three synthetic modalities. The MoG environment has the highest minimum S-score ($\underline{S} = 0.91$), while Student-t has the lowest ($\underline{S} = 0.50$); these values alone do not order the observed depth-75% losses. The layer-pair scores show that S-scores are low only at the first layer pair (L0–L1) and rise to near 1.0 at deeper layers for all modalities. Figure 3 shows the marginal and joint distributions for all three environments, and Figure 4 illustrates the distribution geometry of the recovered models under SFT versus KD.

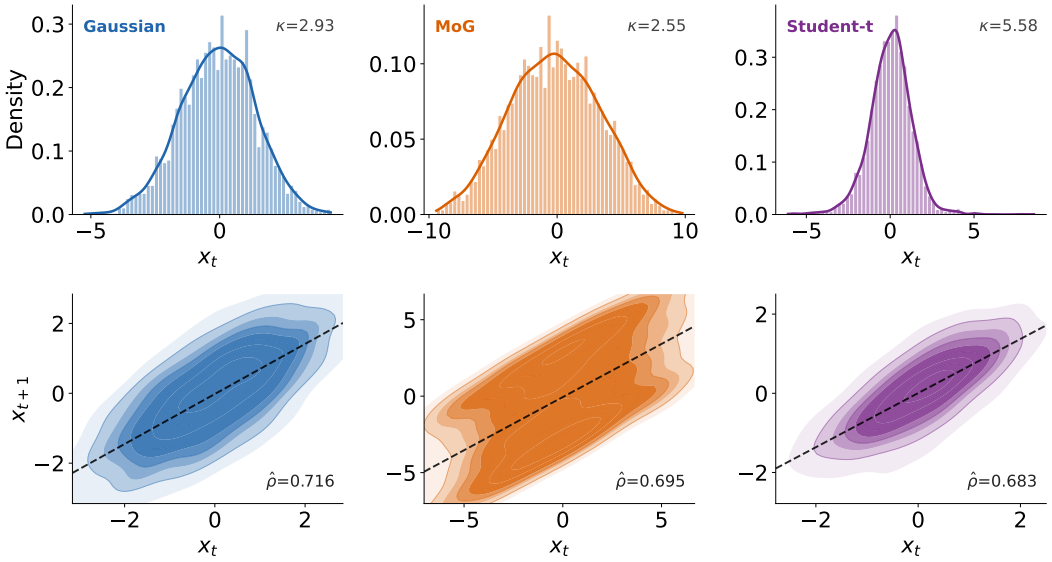

Figure 3: **Distribution Geometry of Synthetic Environments.** Three AR(1) processes ($\rho = 0.70$), differing only in noise family. Row 0: marginal density (histogram + KDE, kurtosis annotated). Row 1: joint density $p(x_t, x_{t+1})$ with KDE contours and regression line. Gaussian (left): tight diagonal joint, high S-scores. MoG (middle): multi-cluster structure with high bottleneck redundancy $\underline{S} = 0.91$. Student-t (right): heavy-tail scatter, with $\underline{S} = 0.50$; the figure depicts data geometry rather than recovery performance.

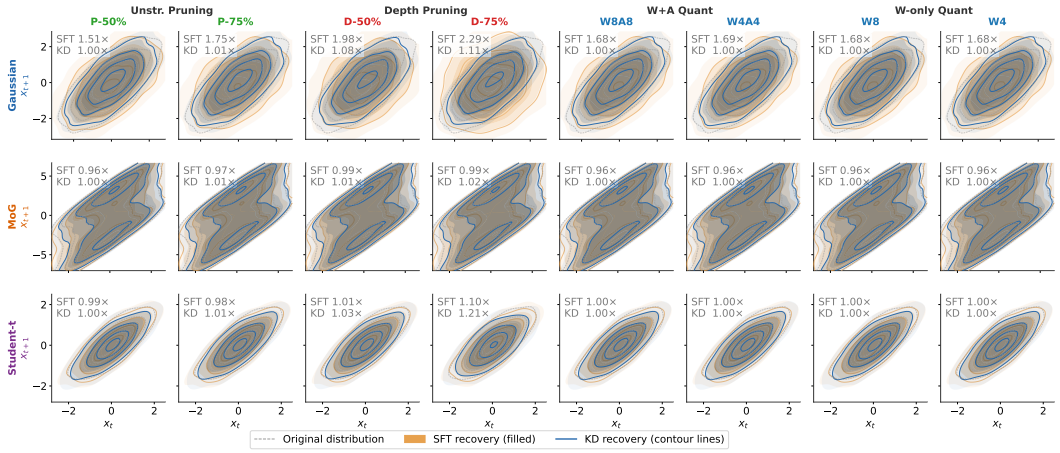

Figure 4: Distribution geometry of recovery fine-tuning strategies.

Table 11: **Pre- vs. Post-Recovery PPL Ratios and Recovery Necessity across all configurations.** *Pre-RFT*: PPL ratio immediately after compression, before any recovery fine-tuning. *Post-RFT (best)*: lowest PPL ratio achieved across all recovery methods (SFT, KD, KD+CE) and data sizes ($N \in \{50, 250, 1\text{k}\}$ C4 samples). $\Delta\Pi$: median pre-RFT − median post-RFT (benefit of recovery). Median computed over n=8 models (LLaMA-3.2-3B, LLaMA-3.1-8B, Qwen-2.5-7B, Gemma-2-2B, and their Instruct counterparts). $\Pi = 1.0\times$ means no degradation relative to uncompressed. **Bold** = consistently safe across all models. Statistical tests: KW = Kruskal-Wallis (global); MW = Mann-Whitney U (one-sided). $^{***}p < 0.001$, n.s. $p \geq 0.05$.

Config	Method	Setting	Pre-RFT Π	Post-RFT Π	$\Delta\Pi$ (benefit)	Safety
Weight-only quantization — W8: 62% recovery-optional; W4: 0% optional						
W-only INT8	Weight-only Quant.	8-bit	$1.00\times$	$\mathbf{1.00}\times$	≈ 0	**Safe**
W-only INT4	Weight-only Quant.	4-bit	$3.54\times$	$1.28\times$	$2.25\times$	Caution
Weight+Activation quantization — high pre-RFT variability; 0% recovery-optional						
W+A INT8	W+A Quant.	8-bit	$2.55\times$	$\mathbf{1.00}\times$	$1.55\times$	**Safe**
W+A INT4	W+A Quant.	4-bit	$> 10^5\times^{\dagger}$	$1.28\times$	$\gg 10^5\times$	Caution
Unstructured pruning — median $\Delta\Pi = 0.52\times$ (50%) and $681\times$ (75%); 0% optional						
Unstr. 50%	Unstr. Pruning	50%	$1.76\times$	$1.23\times$	$0.52\times$	Caution
Unstr. 75%	Unstr. Pruning	75%	$684\times^{\ddagger}$	$2.55\times$	$681\times$	Caution
Depth pruning — 0% optional; recovery reduces damage but does not restore safety						
Depth 50%	Depth Pruning	50%	$554\times$	$3.38\times$	$550\times$	Unsafe
Depth 75%	Depth Pruning	75%	$4{,}605\times$	$7.62\times$	$4{,}596\times$	Unsafe
Statistical tests on recovery benefit ($\Delta\Pi$) across method groups						
Depth vs. Quant (MW, one-sided)	$p < 0.001$ ***	effect $r = 0.81$	(depth recovery benefit $\gg$ quant)			
Prune vs. Quant (MW, one-sided)	$p = 0.35$ n.s.	(no significant difference at this sample size)				
Global across all groups (K-W)	$p < 0.001$ ***					

F Additional Measurements

F.1 Model degradation and task accuracy

Table 11 provides the per-configuration pre- and post-recovery PPL ratios; Table 12 gives the per-task accuracy rather than another visualization of the same aggregate. The post-recovery column selects among reported runs and is not a fixed-budget comparison.

F.2 Measured inference behavior

Table 13 reports accuracy and latency for one checkpoint; Table 8 in the main text instead benchmarks BNB kernels at multiple batch sizes. Their quantization backends and baselines differ, so the speedups should not be pooled. Table 14 summarizes additional model sizes.

Table 12: **Task-specific zero-shot accuracy by model and compression configuration.** Each model block shows the uncompressed baseline (shaded) followed by the best post-recovery (best-RFT) result for each compression configuration. *Best RFT*: highest mean zero-shot accuracy over all recovery methods (∅, SFT, KD) and data sizes $N \in \{50, 250, 1\text{k}, 4\text{k}\}$. W8/W4=weight-only INT8/INT4; W+A8=weight+activation INT8.

Model	Config	HellaSwag	ARC-C	WinoGrande	PIQA	Mean
LLaMA-3.2-3B	Baseline	0.67	0.44	0.71	0.81	0.66
	Unstr. 50%	0.59	0.39	0.75	0.79	0.63
	Unstr. 75%	0.48	0.31	0.60	0.62	0.50
	Depth 50%	0.46	0.21	0.60	0.62	0.47
	Depth 75%	0.36	0.23	0.57	0.50	0.41
	W8	0.68	0.44	0.74	0.82	0.67
	W4	0.68	0.39	0.67	0.77	0.63
	W+A8	0.68	0.43	0.75	0.81	0.67
LLaMA-3.2-3B-IT	Baseline	0.64	0.46	0.75	0.74	0.65
	Unstr. 50%	0.62	0.45	0.75	0.73	0.64
	Unstr. 75%	0.49	0.29	0.63	0.61	0.51
	Depth 50%	0.48	0.30	0.62	0.55	0.49
	Depth 75%	0.38	0.27	0.52	0.52	0.42
	W8	0.65	0.46	0.77	0.74	0.66
	W4	0.68	0.47	0.70	0.66	0.63
	W+A8	0.64	0.46	0.77	0.75	0.66
LLaMA-3.1-8B	Baseline	0.70	0.51	0.80	0.85	0.71
	Unstr. 50%	0.69	0.44	0.73	0.79	0.66
	Unstr. 75%	0.56	0.29	0.61	0.66	0.53
	Depth 50%	0.49	0.29	0.61	0.58	0.49
	Depth 75%	0.30	0.28	0.58	0.49	0.41
	W8	0.71	0.56	0.82	0.83	0.73
	W4	0.70	0.48	0.79	0.82	0.70
	W+A8	0.78	0.53	0.84	0.82	0.74
LLaMA-3.1-8B-IT	Baseline	0.72	0.54	0.81	0.82	0.72
	Unstr. 50%	0.65	0.47	0.79	0.77	0.67
	Unstr. 75%	0.50	0.39	0.55	0.68	0.53
	Depth 50%	0.48	0.32	0.67	0.63	0.53
	Depth 75%	0.42	0.22	0.54	0.55	0.43
	W8	0.73	0.58	0.83	0.81	0.74
	W4	0.70	0.53	0.82	0.81	0.71
	W+A8	0.72	0.56	0.83	0.81	0.73
Qwen-2.5-7B	Baseline	0.71	0.57	0.81	0.85	0.73
	Unstr. 50%	0.68	0.56	0.79	0.82	0.71
	Unstr. 75%	0.54	0.40	0.56	0.71	0.55
	Depth 50%	0.47	0.31	0.55	0.62	0.49
	Depth 75%	0.40	0.21	0.54	0.56	0.43
	W8	0.72	0.57	0.82	0.85	0.74
	W4	0.64	0.54	0.75	0.82	0.69
	W+A8	0.72	0.58	0.81	0.85	0.74
Qwen-2.5-7B-IT	Baseline	0.72	0.60	0.78	0.84	0.73
	Unstr. 50%	0.67	0.58	0.77	0.81	0.71
	Unstr. 75%	0.54	0.37	0.61	0.77	0.57
	Depth 50%	0.46	0.29	0.57	0.70	0.50
	Depth 75%	0.38	0.23	0.53	0.54	0.42
	W8	0.72	0.62	0.82	0.83	0.75
	W4	0.62	0.57	0.77	0.81	0.69
	W+A8	0.72	0.62	0.80	0.84	0.74
Gemma-2-2B	Baseline	0.45	0.31	0.52	0.64	0.48
	Unstr. 50%	0.65	0.47	0.62	0.72	0.61
	Unstr. 75%	0.42	0.34	0.63	0.59	0.49
	Depth 50%	0.45	0.22	0.55	0.62	0.46
	Depth 75%	0.38	0.22	0.55	0.54	0.42
	W8	0.66	0.50	0.83	0.76	0.69
	W4	0.62	0.37	0.69	0.74	0.60
	W+A8	0.64	0.47	0.78	0.80	0.67
Gemma-2-2B-IT	Baseline	0.68	0.44	0.70	0.76	0.65
	Unstr. 50%	0.64	0.47	0.67	0.78	0.64
	Unstr. 75%	0.43	0.26	0.63	0.66	0.49
	Depth 50%	0.49	0.33	0.47	0.60	0.47
	Depth 75%	0.41	0.21	0.51	0.57	0.42
	W8	0.69	0.50	0.77	0.80	0.69
	W4	0.61	0.43	0.67	0.80	0.63
	W+A8	0.67	0.50	0.69	0.79	0.66

Table 13: Inference throughput and zero-shot accuracy for LLaMA-3.2-3B-Instruct on an NVIDIA A100-80GB GPU (batch size 1, sequence length 512). Speedup relative to uncompressed BF16 baseline. Zero-shot accuracy: mean over HellaSwag, ARC-Challenge, WinoGrande, and PIQA. No recovery fine-tuning applied. W4 = weight-only INT4; W+A8 = weight+activation INT8. Memory is peak GPU VRAM during inference. *Note*: W+A quantization on this hardware configuration uses slower INT8 CUDA kernels; depth pruning achieves the highest speedup by eliminating transformer layers entirely.

Config	Speedup↑	ms/tok↓	Mem (GB)↓	ZS-Acc↑
Baseline (BF16, uncompressed)	1.00×	58.90	6.00	0.65
Unstr. Prune 50%	1.32×	44.70	12.00	0.41
Unstr. Prune 75%	1.15×	51.20	12.00	0.38
Depth Prune 50%	1.47×	40.20	9.40	0.37
Depth Prune 75%	3.40×	17.30	8.10	0.37
W-only INT8 (W8)	0.78×	75.60	12.00	0.65
W-only INT4 (W4)	1.10×	53.70	12.00	0.59
W+A INT8	0.58×	101.50	12.00	0.58
W+A INT4	0.59×	100.30	12.00	0.37

Table 14: **Inference speedup by model size group.** Mean speedup vs. baseline.

Compression Configuration	∼3B models	∼7–8B models
Unstructured Prune 50%	1.07	0.93
Unstructured Prune 75%	1.06	0.88
Depth Prune 50%	**1.93**	**1.69**
Depth Prune 75%	**3.17**	**2.71**
W8 (weight-only, 8-bit)	1.04	1.18
W4 (weight-only, 4-bit)	1.09	0.87

F.3 Recovery objectives and data size

Table 15 reports the per-model recovery results under SFT and default forward-KL KD across four data budgets. Table 5 in the main text reports the separate KD-objective comparison at 50% Wanda sparsity with stronger KD variants. SFT values for some nominally matching model and sparsity labels differ across these experiment batches; comparisons must be read within each batch. Table 16 reports the data-size sweep under its own fixed setup, and Table 17 changes calibration distribution.

F.4 Low-rank, structured, and KV-cache compression

SVD-LLM at 50% retained rank had opposite responses to the reported KD recovery on the two tested checkpoints (Table 19). FLAP and depth removal were compared at matched removal fractions within a separate three-model experiment (Table 20); the Gemma recovery ratios rely on its BOS-sensitive baseline and should be read within that experiment. StreamingLLM eviction and KIVI quantization were evaluated using suffix PPL after a 1024-token prefix (Table 21); these results are not directly comparable to full WikiText-2 model PPL. All cells in these three tables have measured outcomes.

Table 15: Best reported post-recovery PPL ratio to each uncompressed baseline across C4 sample counts. The KD column uses default forward-KL; lower is better.

		Depth Pruning 50%		Unstr. Pruning 75%	
Model	ℓ_{orig}	Π_0(SFT)	Π_0(KD)	Π_0(SFT)	Π_0(KD)
LLaMA-3.2-3B	10.92	**6.84**×	25.15×	**2.49**×	3.69×
LLaMA-3.2-3B-IT	16.15	**5.22**×	16.83×	**2.75**×	4.15×
LLaMA-3.1-8B	9.04	**8.32**×	49.08×	**1.85**×	2.50×
LLaMA-3.1-8B-IT	10.91	**7.45**×	42.14×	**2.33**×	2.72×
Qwen-2.5-7B	10.73	**4.43**×	7.80×	**1.63**×	1.85×
Qwen-2.5-7B-IT	11.30	**4.27**×	8.72×	**1.47**×	1.79×
Gemma-2-2B	16.07	**3.62**×	19.36×	**1.52**×	4.60×
Gemma-2-2B-IT	25.09	**2.31**×	6.23×	**0.94**×	1.46×

Table 16: **KD vs. SFT recovery across data sizes.** Recovery ratio ($\ell_{\text{final}}/\ell_{\text{base}}$) for SFT, forward-KL KD, and reverse-KL KD at data sizes $N \in \{50, 250, 1000, 4000\}$ C4 samples, on Wanda-pruned models at 50% and 75% sparsity. **Bold**: lowest ratio within each model, sparsity, and data-size group. These are single-run outcomes; the effect of increasing N is not monotonic in every row.

		N=50			N=250			N=1000			N=4000		
Model	Config	SFT	KD-f	KD-r	SFT	KD-f	KD-r	SFT	KD-f	KD-r	SFT	KD-f	KD-r
Llama-3B	P50	8.3	4.4	**4.3**	26.7	930	**9.0**	43.7	**9.9**	680	235	**8.5**	9.7
	P75	118	**63**	106	154	**59**	621	590	460	**398**	90.8	133	**32.9**
Gemma-2B	P50	4.1	**1.2**	4.0	40.2	**4.0**	5.8	418	**2.9**	3.7	606	**1.9**	2.2
	P75	58.9	**8.9**	14.4	430	**9.5**	11.7	767	**6.2**	7.9	339	**3.0**	4.2

Table 17: **Calibration–evaluation distribution mismatch.** PPL ratio (Π) when Wanda pruning (50%) is calibrated on one distribution and evaluated on another. Diagonal entries (shaded) are matched. Matched calibration has the lowest ratio for two of three evaluation corpora for each model. The Alpaca column favors mismatched calibration for both; Gemma ratios additionally inherit the BOS-sensitive baseline.

Model	Cal. \ Eval.	C4	Alpaca	WikiText-2
LLaMA-3.2-3B	C4	1.89×	0.43×	1.99×
	Alpaca	3.47×	1.55×	3.77×
	WikiText-2	1.94×	0.86×	1.84×
Gemma-2-2B	C4	2.62×	6.81×	3.29×
	Alpaca	3.32×	16.8×	4.19×
	WikiText-2	2.65×	2.23×	3.12×

Table 18: **Empirical Fisher curvature before and after instruction tuning.** $\lambda_{\max}(F)$ is the top eigenvalue of the empirical Fisher, measured on the *uncompressed* base model (positive semidefinite by construction; no compression required). Tax ratio $= \lambda_{\max}(\text{IT})/\lambda_{\max}(\text{pretrained})$; > 1 means instruction tuning increased curvature. The top eigenvalue decreases for two families and increases for two; these measurements do not establish a uniform direction or a causal alignment effect.

Family	**PT** $\lambda_{\max}(F)$	**IT** $\lambda_{\max}(F)$	$\Delta\lambda_{\max}$	**Tax ratio**	**Direction**
LLaMA-3.2-3B	1.050	0.942	−0.108	0.90×	↓
LLaMA-3.1-8B	3.923	4.118	+0.195	1.05×	↑
Gemma-2-2B	275.2	140.0	−135.2	0.51×	↓
Qwen-2.5-7B	4.28	9.48	+5.20	2.22×	↑

Note: $F = \frac{1}{B}\sum_b g_b g_b^\top$ is estimated from 2 calibration layers and 20 gradient batches; $n = 4$ paired families.

Table 19: SVD-LLM at 50% retained rank: PPL ratios before and after the reported KD recovery ($N = 1000$). The two models responded differently; these observations do not establish a curvature mechanism.

Model	$\underline{S}$	Pre (Π)	Post-KD (Π)	Δ
LLaMA-3.2-3B	0.433	128×	218×	+70%
Gemma-2-2B	0.504	10.7×	6.46×	−39%

Table 20: FLAP structured pruning and depth removal at 25% and 50% in a separate three-model experiment. Entries are PPL ratios before and after LoRA-SFT with $N = 1000$ C4 samples. Gemma base-model ratios use the BOS-sensitive baseline.

		25% Removal		50% Removal	
Model	Method	Pre	Post-SFT	Pre	Post-SFT
LLaMA-3.2-3B	FLAP	3.51×	2.54×	848×	5.29×
	Depth	16.9×	2.89×	198×	6.80×
Qwen-2.5-7B	FLAP	2.66×	2.69×	30.0×	4.45×
	Depth	1.93×	2.43×	37,372×	5.67×
Gemma-2-2B	FLAP	2.11×	**0.14**×	17.9×	**0.28**×
	Depth	21.0×	**0.23**×	1.6M×	**0.50**×

Table 21: Suffix PPL ratios on 2048-token sequences after a 1024-token prefix. StreamingLLM evicts cache entries; KIVI quantizes keys and values. This metric is distinct from the full-model WikiText-2 PPL ratio.

Model	Method	Config	Suffix PPL Ratio (Π)
LLaMA-3.2-3B	StreamingLLM	25% evict	1.004×
	StreamingLLM	50% evict	1.028×
	StreamingLLM	75% evict	1.046×
	KIVI	KV-INT4	1.004×
	KIVI	KV-INT2	5.65×
Gemma-2-2B	StreamingLLM	25% evict	1.026×
	StreamingLLM	50% evict	1.129×
	StreamingLLM	75% evict	1.133×
	KIVI	KV-INT4	1.133×
	KIVI	KV-INT2	8.77×

F.5 Exploratory leave-one-model-out selection

The nine-model ridge evaluation used $S_{\min}$, the empirical-Fisher eigenvalue ratio, weight kurtosis, and nominal compression level to predict $\log_{10}(\Pi)$ under leave-one-model-out fitting. Table 22 gives the held-out selection outcomes. Top-1 regret is $\log_{10}(\Pi_{\text{selected}}) - \log_{10}(\Pi_{\text{best}})$ over each model's evaluated candidates. Gemma-3-4B-IT and Qwen3.5-4B are additional checkpoints used in this evaluation, each with eight candidate settings; they are not among the eight paired models used in the primary compression study. Because the candidate-level results and tie-breaking rule for these two checkpoints are not included here, their selection outcomes cannot be independently audited from this paper.

For a fixed held-out model j, an additive linear score $\hat{y}_{j,c} = \beta_0 + \beta_c c + \beta_S S_j + \beta_\lambda \lambda_j + \beta_\kappa \kappa_j$ has $\hat{y}_{j,c_1} - \hat{y}_{j,c_2} = \beta_c(c_1 - c_2)$. Thus the three structural features shift predicted outcome levels across models but do not distinguish candidates within one model; equal nominal levels remain tied unless another rule is specified. The six zero-regret selections, which occur only among well-covered candidate sets, are a descriptive result rather than evidence of a validated structure-aware configuration selector. A meaningful follow-up should compare against the same candidate set ranked by level alone and include method indicators or tested feature-by-method interactions.

Table 22: Reported leave-one-model-out selection outcomes for an additive four-feature ridge model. Regret is the selected minus best observed $\log_{10}$ PPL ratio within the candidate set; zero means the observed best was selected. The aggregate includes all nine held-out models.

Held-out model	Candidates	Top-1 regret	Selected rank
Gemma-2-2B	14	0.00	1
Gemma-2-2B-IT	17	0.00	1
Gemma-3-4B-IT	8	0.00	1
LLaMA-3.2-3B	14	0.00	1
LLaMA-3.2-3B-IT	17	0.00	1
Qwen3.5-4B	8	0.00	1
LLaMA-3.1-8B	3	0.18	2
LLaMA-3.1-8B-IT	3	0.20	2
Qwen-2.5-7B-IT	3	1.07	2

F.6 Additional checkpoints and operator combinations

Qwen3-14B was evaluated on five unrecovered configurations (Table 23). Table 24 reports cross-entropy differences for sequential pruning and quantization; three of four measured interaction terms are positive, but the local empirical-Fisher cross term has the wrong sign in those cases. These measurements do not validate the local approximation for large changes.

G Interpretation of the Synthetic Comparison

The AR(1) process isolates innovation-distribution effects under a small transformer. It does not reproduce the training data, architecture, or learned computations of the real LLMs. The observed post-recovery depth-50% PPL ratios for the evaluated models (2.31–8.32×) therefore should not be extrapolated from the synthetic recovery ratios. Layer similarity is measured on different representation spaces, and neither its absolute scale nor the empirical Fisher is an interchangeable estimate of the downstream curvature in Theorem 3.2. The four-model objective comparison likewise cannot explain differences between experiment batches without matched recovery runs.

Table 23: **Scale extension: Qwen3-14B compression.** PPL ratio (Π) for Qwen3-14B under five compression configurations. Selected 14B configurations: INT8/NF4 preserve WikiText-2 PPL, whereas 50% depth removal is severe.

Config	PPL Ratio (Π)	Notes
Wanda 50%	$1.20\times$	Moderate
Wanda 75%	$11.2\times$	Severe
Depth 50%	$143{,}000\times$	Catastrophic
BNB-INT8	$1.008\times$	Near-lossless
BNB-NF4	$1.034\times$	Near-lossless

Table 24: **Composition interactions: pruning + quantization.** Measured interaction $I = \Delta L_{\text{joint}} - (\Delta L_1 + \Delta L_2)$ (cross-entropy, not PPL) for composing pruning with 4-bit quantization. $I > 0$: super-additive (worse than sum); $I < 0$: sub-additive. Sign match indicates whether the cheap estimator ($I_{\text{pred}} = \delta_1^\top F\, \delta_2$) correctly predicts the interaction sign. The estimator fails at large perturbations ($3/4$ cases show super-additive interactions, but the estimator predicts the correct sign in only $1/4$).

Model	Pair	ΔL_1	ΔL_2	ΔL_{joint}	I_{meas}	I_{pred}	Sign
LLaMA-3.2-3B	Prune50 + Q4	0.689	0.488	1.530	$+0.353$	-2.8×10^{-5}	✗
	Prune75 + Q4	5.513	0.488	5.908	-0.092	-6.6×10^{-4}	✓
Gemma-2-2B	Prune50 + Q4	1.190	2.187	3.662	$+0.286$	-0.060	✗
	Prune75 + Q4	3.602	2.187	7.870	$+2.082$	-0.171	✗

Note: ΔL values are in cross-entropy (nats), not PPL. The cheap bilinear estimator ($\delta^\top F\delta$) uses the empirical Fisher and underestimates the true interaction by orders of magnitude at large perturbations, confirming that local second-order analysis is qualitatively informative but quantitatively unreliable in the severe-compression regime.