IMPORTANCE IS NOT GAUGE-INVARIANT: MEASURING, NOT ESTIMATING, UNIT REDUNDANCY IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Structured pruning ranks attention groups and FFN channels by importance proxies such as weight magnitude or activation traffic. We show these proxies do not measure the network's function: residual transformers admit per-group *gauge freedoms* — function-preserving rescalings of (Q, K), (V, O), and FFN $(\mathrm{up}, \mathrm{down})$ pairs — under which magnitude-based rankings reorder almost arbitrarily (Spearman 0.24–0.30) while no output bit changes (max logit difference 0.0). Interventional ablation rankings are exactly invariant under the bit-exact gauge on Llama-3.1-8B (rank survival 1.000); on QK-normalized architectures the gauge is only ε-approximate, and the observed rank change (Qwen3: 0.936) tracks the induced function drift. Across six backbones from five institutions, estimation-based proxies correlate with interventional importance at $|\rho_s| < 0.26$; *exactly* invariant proxies (Taylor, Fisher) do no better ($|\rho_s| \leq 0.50$; a Taylor top-k keep-set yields $3.7\times$ the ΔNLL of random, 8.73 vs. 2.37) — invariance is necessary, not sufficient. A pre-registered mechanism hypothesis (QK-norm predicts proxy sign) is falsified by our own test, and gauge canonicalization does not rescue the proxies. Measurement has pitfalls of its own: on Mistral-7B's flat redundancy landscape, extreme solo-score selection is catastrophic (ΔNLL $+7.3$ vs. $+0.6$ random) until one mid-budget score refresh restores safety. We therefore develop MAPS (Measured Allocation of Pruning Structure), a measurement-based pruner with a single primary hyperparameter that selects heterogeneous units (KV groups, FFN groups, whole layers) by *measured* candidate comparison. Without recovery training, at a nominal keep ratio of ρ=0.875 it retains 99.5% of zero-shot MC (0.7570 vs. 0.7605) and 99.8% of GSM8K on OLMo-3-32B. Across four backbones and three budgets, measurement-based selection tops every comparison: MAPS takes 10 of 12 backbone–budget cells (up to $+7$ points over the best structured baseline), a fully-measured depth search (SLEB) takes the deepest hub cell — a dose–response in the amount of measurement — and estimation-based criteria (magnitude, Wanda-style, Taylor, FLAP-style) trail throughout, with score families backbone-dependent under a fixed allocator. An allocation-only ablation (global per-parameter greedy over the same scores, even with one mid-budget remeasurement) still trails random: *where* you remove matters as much as *what* you score, and *how much* you measure decides who wins.

1 INTRODUCTION

Structured pruning of large language models must answer one question before any weight is touched: *which units matter?* The field's default currency for this question is the importance proxy — weight magnitude, activation statistics, and their products — inherited from a decade of CNN pruning and applied, largely unchanged, to transformers whose KV groups, FFN channels, and layers are the units of interest. Proxies are attractive because they are cheap. This paper asks whether they are *measurements* — functions of the network's behavior at all — and answers, with a construction rather than an opinion: **no**.

The argument proceeds in four steps, each an experiment. **First** (§2), residual transformers admit per-group *gauge freedoms*: rescalings of (Q, K), (V, O), and FFN $(\mathrm{up}, \mathrm{down})$ pairs that preserve

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.

the function — with power-of-two scales, bit-exactly. Under a random gauge that changes no output bit, magnitude and traffic rankings are destroyed ($\rho_s \leq 0.30$) while causal-ablation rankings are exactly invariant. **Second** (§3), a six-backbone map shows the proxies are near-uncorrelated with interventional importance everywhere ($|\rho_s| < 0.26$ for traffic/magnitude-style proxies), while their *practical* effect swings from strongly helpful to strongly harmful with no architectural predictor we could find: our pre-registered mechanism hypothesis is falsified by its own test, and we publish the falsification. **Third** (§4), the last cheap escape is closed: canonicalizing the gauge changes the correlations by less than 0.02. **Fourth** (§3, 5), measurement itself has a failure mode — on flat landscapes, exact solo scores are noise, and one-shot extreme selection is catastrophic until the *plan itself* is measured.

These findings assemble into a hierarchy, not a winning score: gauge invariance is necessary but not sufficient, and even interventional solo ablations miss interactions on flat landscapes — so the pruning decision itself must be measured. (We use *measured* for scores that evaluate a removal intervention and *proxy* for scores that do not.) This motivates **MAPS** (Measured Allocation of Pruning Structure), which proposes candidates from cheap unit-level interventions and commits only after measuring the candidate plans. It selects heterogeneous units — KV groups, FFN groups, whole layers — using only measured quantities: causal solo scores, and measured comparisons among a fixed set of candidate plans. Its single primary hyperparameter is the keep ratio ρ; it needs no gradients, no canonicalization, and no recovery training. Holding its scores fixed and varying only the allocator reproduces the literature's failure modes on demand: the global per-parameter greedy — even granted one mid-budget score re-measurement — trails *random* at the shallowest budget on three of four backbones, and allocation alone changes zero-shot MC by over 20 points on three of four backbones (Table 3). With measurement at both levels, MAPS takes 10 of 12 backbone–budget cells against every structured baseline — by up to $+7$ zero-shot points over the best of them — and the two ceded cells fall to depth-only removal at the deepest budget, never to a score-estimation method. It retains **99.5% of zero-shot accuracy and 99.8% of GSM8K on OLMo-3-32B at a nominal keep-ratio budget of** ρ=0.875 without recovery, while converting the budget into real deployment resources (-24% VRAM, -27% KV-cache at ρ=0.625) that unstructured sparsity does not deliver (Table 5).

Contributions. (1) a gauge theorem with a bit-exact falsifier and a per-family covariance table (§2); (2) a six-backbone empirical map with a pre-registered, falsified mechanism (§3); (3) a gauge-fixing control showing the failure is deeper than the gauge (§4); (4) measurement pitfalls and the allocation lesson (§3, 5); (5) MAPS and a no-recovery application matrix with deployment, transfer, sensitivity, and determinism all measured (§5–6). All conclusions are drawn in the no-recovery regime; Appendix B shows they persist under recovery fine-tuning, which we also show can mask method differences.

2 Importance Is Not Gauge-Invariant

Setup. A decoder block contains grouped-query attention (Ainslie et al., 2023) with n_{kv} KV groups; group g owns query rows $W_Q^{(g)}$, key/value rows $W_K^{(g)}, W_V^{(g)}$, and output columns $W_O^{(:,g)}$ of its r query heads. The SwiGLU (Shazeer, 2020) MLP computes $W_{\text{down}}[\sigma(W_{\text{gate}}x) \odot W_{\text{up}}x]$. We study unit scores in three proxy families — *activation traffic* ($\mathbb{E}|Q_g| + \frac{1}{2}(\mathbb{E}|K_g| + \mathbb{E}|V_g|)$), *weight magnitude*, and *product forms* (Wanda-style $\sum |W| \cdot \mathbb{E}\|x\|$) — against *causal* scores (ΔNLL under a specified unit-removal intervention). "Causal" is used only in this operational, network-intervention sense — no identification of causal effects in the data-generating process is claimed — and is synonymous with "interventional" below.

Proposition 1 (Gauge freedoms)**.** *For any per-group scalars $c_g > 0$, the following preserve the network function exactly (norm-free architectures) or up to the RMSNorm ϵ term (QK-normalized architectures): (i)* QK: $W_Q^{(g)} \to c_g W_Q^{(g)}$, $W_K^{(g)} \to c_g^{-1} W_K^{(g)}$ *(scalars commute with RoPE rotations (Su et al., 2024); under per-head RMSNorm,* $\text{norm}(cq) = cq/\sqrt{c^2\overline{q^2} + \epsilon}$ *equals* $\text{norm}(q)$ *iff* $\epsilon = 0$ *— the* ϵ-clause*); (ii)* VO: $W_V^{(g)} \to c_g W_V^{(g)}$, $W_O^{(:,g)} \to c_g^{-1} W_O^{(:,g)}$*; (iii)* FFN: $W_{\text{up}}^{(G)} \to cW_{\text{up}}^{(G)}$, $W_{\text{down}}^{(:,G)} \to c^{-1} W_{\text{down}}^{(:,G)}$ *(the gate path passes through σ and is* not *a gauge). With $c = 2^s$, every*

	traffic	magnitude	Wanda	causal
Llama-3.1-8B (exact gauge)	0.235	0.298	1.000	**1.000**
Qwen3-8B (ϵ-approx.)	0.964	0.268	1.000	**0.936**

Table 1: **Gauge covariance by score family** (Spearman, before vs. after; attention units). FFN units show the same pattern; magnitude survives partially there on Llama (0.525) because the untransformable gate path contributes to the score.

multiplication is an exponent shift in `bf16`*: on norm-free architectures the transformed network is* bit-identical.

Any score that varies along a gauge orbit is not a function of behavior. Pure weight- and activation-magnitude scores vary under (i)–(iii); product forms pairing a weight with its compensating activation are invariant to these gauges; causal scores are invariant to any function-preserving transform by construction.

E1. We sample $s \sim \mathrm{Unif}\{\pm 1, \pm 2\}$ per group and gauge type (seed 7), apply (i)–(iii) to Llama-3.1-8B and Qwen3-8B, and re-measure all four score families (identical calibration: 6 GSM8K training prompts). Function preservation: Llama $\Delta\mathrm{NLL} = 0.000000$, max logit difference 0.0 (bit-exact); Qwen3 $\Delta\mathrm{NLL} = +3\times10^{-4}$, max logit difference 2.06 (the ϵ-clause in action). Table 1 and Figure 1 report rank survival: on the exact gauge, traffic and magnitude are destroyed by a transformation that changes no output bit, while Wanda's product form and causal scores are exactly invariant. On Qwen3, causal invariance is bounded by the gauge itself (the function drifted), not the estimator; traffic survives because inter-unit spread exceeds the $\pm 4\times$ scramble — yet the quantity is *functionally inert* under the per-head norm. A proxy can be stable and still measure something the function ignores. Gauge dependence is *sufficient* reason to distrust a proxy across architectures, but not the only failure mode: the gauge-invariant Wanda score still fails to track causal importance in practice (§3).

3 A Six-Backbone Map and a Falsified Pre-Registration

We measure, on six backbones from five institutions — Llama-3.1-8B (Grattafiori et al., 2024), Qwen3-8B (Yang et al., 2025), Qwen2.5-7B (Qwen Team, 2024), Mistral-7B-v0.3 (Jiang et al., 2023), OLMo-2-7B (OLMo et al., 2024), Gemma-3-1B (Team et al., 2025) (Table 2) — (M1) Spearman correlation between each proxy and causal ΔNLL over all group units, and (M2) no-recovery ΔNLL at ρ=0.75 under keep-sets (A) proxy top-k, (B) random, (C) causal top-k; identical calibration throughout. The lineup contains a within-family contrast: Qwen2.5 vs. Qwen3 toggles per-head QK normalization (Henry et al., 2020).

Pre-registration. Pilots at *query-head* granularity on Llama had shown a strong proxy anti-signal ($3\times$ worse than random); Qwen3 the opposite. We hypothesized that *QK normalization determines the proxy's sign* and registered predictions before measurement (`predictions.json`, timestamped). We disclose a registration error the test itself exposed: the rule was derived from query-head pilots but registered for a group-granularity experiment; the regimes differ, and the registration prevented retrofitting.

Results. Joint agreement is **1/6**; the within-family toggle fails outright — Qwen2.5, *without* QK-norm, shows the largest positive proxy effect. Across all backbones, the estimation-based traffic/magnitude-style proxies remain weakly correlated with interventional importance ($|\rho_s| < 0.26$), and the gauge-invariant Taylor/Fisher controls reach larger but still modest correlations (up to $|\rho_s|$=0.50; §3, App. A); yet their practical consequence swings from -2.24 (helpful) to $+2.05$ (harmful) with no architectural predictor we tested. The query-head anti-signal on Llama is real but granularity-specific. Where causal selection is informative it is decisively so: on OLMo-2, causal top-k reaches ΔNLL 1.39 vs. 4.18 (random) and 6.22 (proxy). We report the mechanism as open.

Measurement's own pitfall (Mistral). The one "successful" prediction is the table's anomaly: causal top-k *collapses* on Mistral ($+7.27$ vs. $+0.58$ random). Its solo landscape is nearly flat (attention max 0.064 vs. 0.114 Llama, 0.478 OLMo-2; median 0.001; 32% negative). Two controls localize the failure: bottom-k collapses symmetrically ($+5.94$; the scores are noise, not anti-signal),

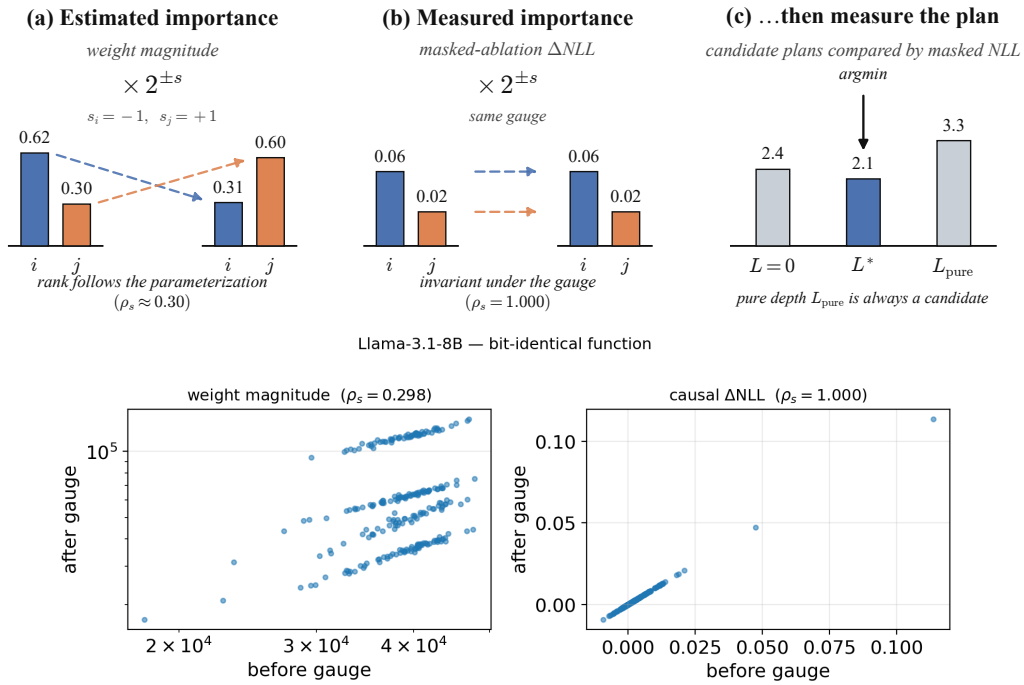

Figure 1: **Function-preserving reparameterization destroys magnitude importance and leaves causal importance untouched.** *Top (a–c):* the thesis in schematic form — magnitude rankings ride the gauge while causal scores are invariant, so MAPS measures both unit scores and the plan itself (candidates compared by masked NLL, §5). *Bottom:* Each point is one KV-group unit of Llama-3.1-8B, before vs. after a random power-of-two gauge under which the network is bit-identical. *(d):* magnitude scores split into four bands — precisely the gauge exponents $\{\frac{1}{4}, \frac{1}{2}, 2, 4\}$ ($\rho_s = 0.298$). *(e):* causal ΔNLL on the diagonal ($\rho_s = 1.000$).

backbone	QK-norm	ρ_s(traffic)	ρ_s(Taylor)	ρ_s(Fisher)	A−B (ΔNLL)	prediction
Llama-3.1-8B	absent	+0.01	+0.01	−0.03	**−0.67**	✗ (proxy helped)
Qwen3-8B	present	**−0.05**	+0.02	−0.05	−0.19	✗ (sign)
Qwen2.5-7B	absent	**+0.17**	+0.19	+0.17	**−2.24**	✗✗
Mistral-7B-v0.3	absent	−0.06	+0.09	+0.25	+0.06	✓✓ (see text)
OLMo-2-7B	present	**−0.06**	+0.25	+0.50	**+2.05**	✗✗
Gemma-3-1B	present	−0.04	−0.19	+0.27	−0.27	✗ (sign)

Table 2: **Pre-registered predictions vs. outcomes.** Joint agreement 1/6; the hypothesis is falsified. Negative A−B means the proxy helped. Caveats: Gemma-3-1B has one KV group per layer (attention arm degenerate; FFN and correlation carry the row); OLMo-2 is full MHA (n_{kv}=32). Taylor and Fisher are *exactly* gauge-invariant (rank survival > 0.9999 on Llama — exact up to floating-point ties — and ≥ 0.91 under QK-norm on Qwen3; App. A) yet remain weak everywhere ($|\rho_s| \leq 0.50$); behaviorally, a Taylor top-k keep-set on Llama degrades NLL to 8.73 vs. 2.37 random. Wanda-style values here use our pair-summed aggregation, which we prove invariant in App. A; the standard row-wise aggregation survives the gauge only at ρ_s=0.25. Full per-proxy grids and all arm values in Appendix A.

and a single mid-budget score refresh restores safety (+0.64). Solo scores are exact measurements of single-unit ablations; on flat landscapes one-shot extreme selection compounds interactions solo scores cannot see — *selection must itself be measured* (§5). This reclassifies Mistral's "correct prediction" as coincidence. The QK-norm hypothesis and the six-backbone map were pre-registered; the Mistral refresh diagnostic was added *after* the registered failure and is post-hoc localization, not part of the confirmatory test (App. F).

Algorithm 1 MAPS (single hyperparameter ρ; fixed constants disclosed: $K{=}6$ prompts, candidate window $\{0, L^*{-}1, L^*, L^*{+}1, L_{\text{pure}}\}$)

1: **Discover** units from the architecture: KV groups, FFN groups, layers (MoE experts: Appendix C).
2: **Score** each unit by masked-ablation ΔNLL on K calibration prompts; a layer's mask is a block skip — *exactly* the removal it prices.
3: **Groups:** within every surviving stage keep the top-k by causal score at a uniform residual keep ratio (≥ 1 per stage).
4: **Layers by measurement:** rank layers by causal score per parameter; with L^* the predicted (score-sum) optimum and $L_{\text{pure}}{=}\text{round}((1{-}\rho)n_{\text{layers}})$, evaluate the *full candidate plans* for $L \in \{0, L^*{-}1, L^*, L^*{+}1, L_{\text{pure}}\}$ by masked NLL; keep the measured argmin.
5: **Materialize:** slice groups, drop layers, renumber indices. No gradients, no canonicalization, no recovery.

4 Fixing the Gauge Does Not Rescue the Proxies

If magnitude fails because the gauge is arbitrary, canonicalize it: for every group, rescale so $\|W_Q^{(g)}\|_F{=}\|W_K^{(g)}\|_F$, $\|W_V^{(g)}\|_F{=}\|W_O^{(:,g)}\|_F$, and per FFN group $\|W_{\text{up}}\|_F{=}\|W_{\text{down}}\|_F$ — the attention/SwiGLU analogue of path-norm balancing. One practical subtlety: canonicalization must transform *affine* parameters with the gauge (Qwen2.5 carries attention biases; weight-only rescaling broke the function, ΔNLL $+1.38$; bias-aware rescaling restores drift ≤ 0.004 on all six backbones). **Result:** every proxy–causal correlation moves by less than 0.02 (Llama magnitude $0.181 \to 0.177$; OLMo-2 $0.150 \to 0.150$; Qwen2.5 traffic $0.171 \to 0.155$; Gemma Wanda $0.254 \to 0.254$; full grid Appendix A). The canonical representative is as uninformative as the arbitrary one: the failure is deeper than the gauge, and the last cheap escape is closed.

5 MAPS: Measuring the Cut, Not Estimating It

Scoring (Algorithm 1, step 2) costs one batched forward per unit (Llama-3.1-8B: 576 group $+\,32$ layer probes, $\approx$5 min); plan construction and materialization take $\approx$1 min (Table 5: 39–62 s measured). The design encodes §3's lessons: removal is *spread* across stages, depth enters only when a measured comparison prefers it, and L_{pure} guarantees the pure-depth regime is always examined.

Allocation is as important as the signal. Holding the causal scores fixed, we compare four allocators (Table 3, Figure 2). Feeding the same unit scores to four allocators, allocation alone changes zero-shot MC by over 20 points on three of four backbones (Llama-8B $\approx$28, Llama-3B 24, OLMo-32B $\approx$22; Qwen-14B 12.5), and the global per-param greedy — even with its one mid-budget remeasurement — trails *random* at the shallowest budget on three of four backbones (Llama-3B is the exception, 0.516 vs. 0.504): concentrating removals compounds interactions solo scores cannot see. MAPS leads every allocator on every backbone at $\rho \in \{0.875, 0.75\}$. At the deepest budget the picture is honest rather than uniform: on the 64-layer 32B at $\rho{=}0.625$ the measured probe-NLL criterion prefers width while downstream MC prefers depth (layer-drop 0.659 vs. 0.548) — probe-metric choice can bind (§8). Plans may differ across calibration draws at deep budgets (the hub's three draws drop $1/6/10$ layers at $\rho{=}0.625$); evaluation is deterministic, so we report mean$\pm$range over draws as the calibration-sensitivity axis. Distinct measured plans can also tie downstream: on Qwen3-14B at $\rho{=}0.75$, a 1-layer and a 6-layer plan (selected under different candidate windows) reach identical MC (0.647) — the plan landscape has plateaus.

Score-swap ablation: the allocator carries the transfer. Holding the full measured allocator fixed (candidate probing, plan argmin, guard) and swapping only the unit-score family — causal ΔNLL (MAPS) vs. gauge-invariant Taylor (Molchanov et al., 2019; Ma et al., 2023) ("Taylor-final") — separates the two contributions. On the hub the two families are statistically tied on matched draws ($+0.4$, -0.0, $+0.6$ at $\rho{=}0.875/0.75/0.625$; App. E); the draw-mean advantage of interventional scores at $\rho \leq 0.75$ (Table 4) comes from Taylor-final's far larger calibration-draw variance (0.442–0.617 vs. 0.584–0.600): interventional scores are more *reliable* across draws, not better on a given draw. On Mistral-7B, whose solo scores are flat (§3), the ordering *reverses* at every budget ($0.727/0.654/0.493$ vs. $0.712/0.604/0.445$; paired $+1.5$ $[+0.8, +2.2]$, $+5.0$ $[+4.1, +6.0]$;

allocator (same scores)	Llama-8B	Llama-3B	Qwen-14B	OLMo-32B
MAPS (spread + measured layers)	**0.704–0.731**	**0.646**	**0.734**	**0.757**
greedy (global per-param, refresh)	0.493–0.513	0.516	0.634	0.541
group (spread only)	0.612–0.650	0.574	0.688	0.737
layer-drop (depth only)	0.438	0.406	0.609	0.741
random	0.615	0.504	0.656	0.722
dense	*0.745*	*0.679*	*0.749*	*0.760*

Table 3: **Allocation ablation at** ρ=0.875 (zero-shot MC average; identical causal scores). Ranges are min–max over three disjoint calibration draws; single values denote draw-independent methods under this protocol. Full curves in Figure 2; all budgets in Appendix D.

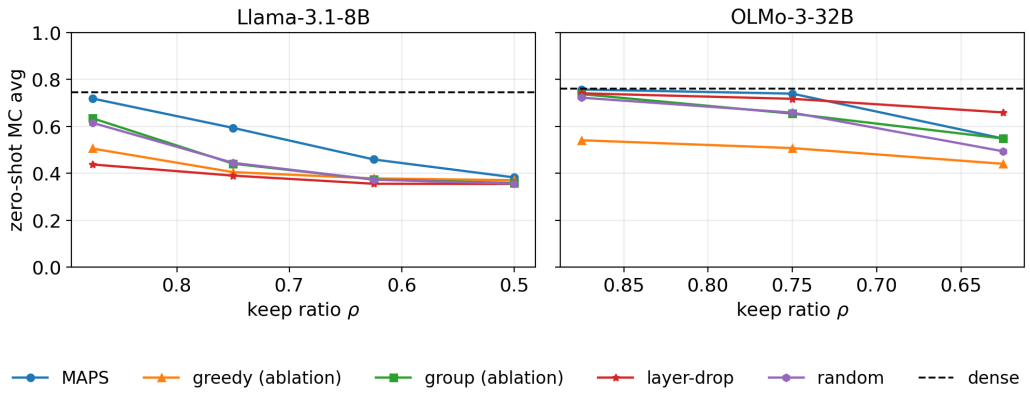

Figure 2: **One score set, four allocators** (Llama-3.1-8B, OLMo-3-32B). Global per-param greedy (orange) trails random at shallow budgets on both backbones; spreading with measured depth selection (blue, MAPS) dominates. Allocation policy, not the importance signal, separates the best and worst structured pruners here.

App. D). No score family is universal; both orderings sit far above their estimation-only top-k counterparts, so the measured plan, not the score, is the load-bearing, transferable component.

6 Application: No-Recovery Pruning Across Four Backbones

Protocol. Every cell is prune $\rightarrow$ immediate evaluation: WikiText-2 PPL (4096-token windows) (Merity et al., 2017), a five-task zero-shot MC composite, an unweighted macro-average over (ARC-E/C (Clark et al., 2018), HellaSwag (Zellers et al., 2019), WinoGrande (Sakaguchi et al., 2021), PIQA (Bisk et al., 2020)) (per-task scores in App. G), GSM8K 8-shot (Cobbe et al., 2021), and on the hub MMLU 5-shot (Hendrycks et al., 2020) at $\rho \in \{1.0, 0.875, 0.625\}$. Evaluation is deterministic for a fixed pruned model; draw ranges describe pipeline repeatability under prompt selection, not finite-sample uncertainty. A paired bootstrap over evaluation examples (App. E) gives $\pm$0.7–1.1-point 95% intervals: every MAPS–SLEB and Taylor-final–MAPS gap excludes zero; the hub score-swap pairs are ties ($+0.4$, -0.0, $+0.6$). Rankings are stable in prompt count (Spearman $0.90/0.89$ for K=4/12 vs. 6; Appendix E); every pruned model passes a structural probe gate; all evaluations use the lm-eval harness (Gao et al., 2021).

Baselines in two tiers. *Signal isolation* (our materialization, score swapped): magnitude, Wanda-style, random. *External systems*: official SparseGPT via `llmcompressor` (unstructured 50%, measured as a cross-regime reference), Torch-Pruning Taylor restricted to MLP channels at budget-equivalent ratio (its dependency graph does not support GQA-constrained attention), ShortGPT-style layer-drop; SliceGPT is discussed from its paper (official implementation lacks these backbones); TP-magnitude is omitted to avoid duplicating the signal-isolation row. At 14B and 32B the official implementation is unavailable in the evaluated release (its sequential tracer fails across these archi-

method	MC avg by ρ 0.875	0.75	0.625	0.5	at ρ=0.875 GSM8K	PPL	MMLU
MAPS	**0.704–0.731**	**0.584–0.600**	0.444–0.479	0.372–0.401	**0.216–0.468**	8.5–9.3	0.548–0.653
SLEB (measured depth)	0.671–0.675	0.568–0.612	**0.473–0.484**	**0.391–0.392**	0.152–0.250	12.0–12.9	–
Taylor-final (swap ablation)	0.687–0.733	0.442–0.617	0.371–0.438	0.358–0.374	0.254–0.498	8.9–13.2	–
FLAP-style (An et al., 2024)	0.626–0.635	0.527–0.538	0.400–0.418	0.355–0.359	0.102–0.134	17.6–17.8	–
Taylor top-k (group)	0.620–0.657	0.361–0.473	0.360–0.387	0.358–0.361	0.020–0.104	14.8–17.8	–
TP-Taylor (MLP)	0.662–0.667	0.532–0.543	0.417–0.424	0.367–0.375	0.278–0.320	16.8–18.4	–
Wanda-style	0.606–0.630	0.446–0.493	0.384–0.396	0.357–0.366	0.082–0.106	17.1–19.5	–
magnitude	0.619	0.432	0.369	0.358	0.026	21.4	–
random	0.615	0.445	0.373	0.356	0.038	18.0	–
group (ablation)	0.612–0.650	0.417–0.465	0.366–0.383	0.358–0.362	0.042–0.072	16.7–26.3	–
greedy (ablation)	0.493–0.513	0.402–0.408	0.373–0.384	0.365–0.376	0.002–0.022	55.6–136.0	0.255–0.270
layer-drop	0.438	0.390	0.356	0.354–0.356	0.008	95.0	–
SparseGPT (unstr. 50%)†		0.650–0.651 (cross-regime)			0.360–0.372	16.0–16.1	–
dense		*0.745*			*0.514*	*7.7*	*0.654*

Table 4: **Hub matrix (Llama-3.1-8B), no recovery.** Ranges are min–max over three disjoint calibration draws (draw means and per-draw scores in App. E); single values denote draw-independent methods under this protocol. Best per MC column (draw-mean) in bold; at $\rho \leq 0.625$ the best cells belong to SLEB — both leaders are measurement-based. †Unstructured 50% is a different budget axis (no tensor-shape change), measured over the same three calibration draws as a cross-regime reference; deployment consequences in Table 5. The quality-viable deployment window under no recovery is $\rho \geq 0.75$; at ρ=0.625 every structured method pays double-digit MC costs.

tectures, and the basic pipeline's simultaneous Hessians exceed 80 GB), so measured rows are the 3B spoke (Appendix D) and the 8B hub.

Results (Figure 3, Table 4). The comparison resolves into a *regime map*. At $\rho \in \{0.875, 0.75\}$ MAPS is the strongest structured method in every cell on all four backbones. At the deepest budget (ρ=0.625) both exception cells go to *depth-only* removal: on the hub the fully-measured depth search SLEB (Song et al., 2024) edges MAPS (0.473–0.484 vs. 0.444–0.479), and on OLMo-3-32B — the documented dissociation cell, where dropping layers is simply the right axis — even ShortGPT-style layer-drop suffices (0.659; SLEB 0.651, MAPS 0.548). MAPS holds the 3B and 14B spokes, taking 10 of 12 cells overall (the hub ρ=0.75 cell is the least robust: $+1.5$ $[+0.6, +2.5]$ on the matched draw, but not on every draw; App. E), and no score-estimation method (magnitude, Wanda, Taylor, FLAP) takes a single cell anywhere. We read this as dose–response: MAPS measures five candidate plans, SLEB measures every layer at every step; where depth-only removal suffices, more measurement wins, and where width must be cut (3B, 14B, moderate budgets), SLEB cannot follow. Estimation-based rows (magnitude, Wanda-style, Taylor, FLAP-style) trail the measured pair throughout; the backbone spread at ρ=0.875 (MAPS 0.646–0.757) makes backbone choice an explicit, reported variable of the comparison rather than a hidden one. On OLMo-3-32B (Olmo et al., 2025), at a nominal keep-ratio budget of ρ=0.875, the pruned model retains 99.5% of dense MC (0.7570 vs. 0.7605) and 99.8% of GSM8K (0.820 vs. 0.822); on Qwen3-14B GSM8K survives at 0.900 (dense 0.924). Two regularities cut across the matrix: *generation dies first* — GSM8K collapses toward zero for baselines and deep budgets while MC degrades gracefully, so "accuracy retention" claims depend heavily on task type (Liang et al., 2022) — and MMLU tracks the MC composite. The cross-regime reference is itself informative: official SparseGPT at 50% unstructured costs 9.4 MC points and doubles perplexity on the hub (0.650 vs. 0.745; PPL 16.1 vs. 7.7), and 8.9 points on the 3B spoke — in our no-recovery evaluation on the tested $\leq$8B backbones, the near-lossless 50% behavior reported at larger scales is not reproduced in the no-recovery protocol.

Deployment cost (Table 5). Structured causal pruning buys real resources: at ρ=0.625 on Llama-3.1-8B, parameters -24%, KV-cache bytes/token -27%, peak VRAM -24%, with one-shot pruning in $\approx$1 minute from cached scores. The reference row is the point: official unstructured SparseGPT at 50% changes no tensor shape — **0% memory saved and no decode speedup** absent sparse kernels (measured: 150.7 vs. 153.0 tok/s). Small-batch decode latency is governed by depth, not width (layer-drop $+52\%$ tokens/s at ρ=0.625; width $+1$–3%): MAPS's allocator trades width (memory, batch capacity) against depth (latency) as the budget deepens.

Calibration transfer (Table 13). At ρ=0.875, general-text (WikiText) and task (GSM8K) calibration converge to the *same plan*; at deeper budgets they diverge, and general-text calibration preserves

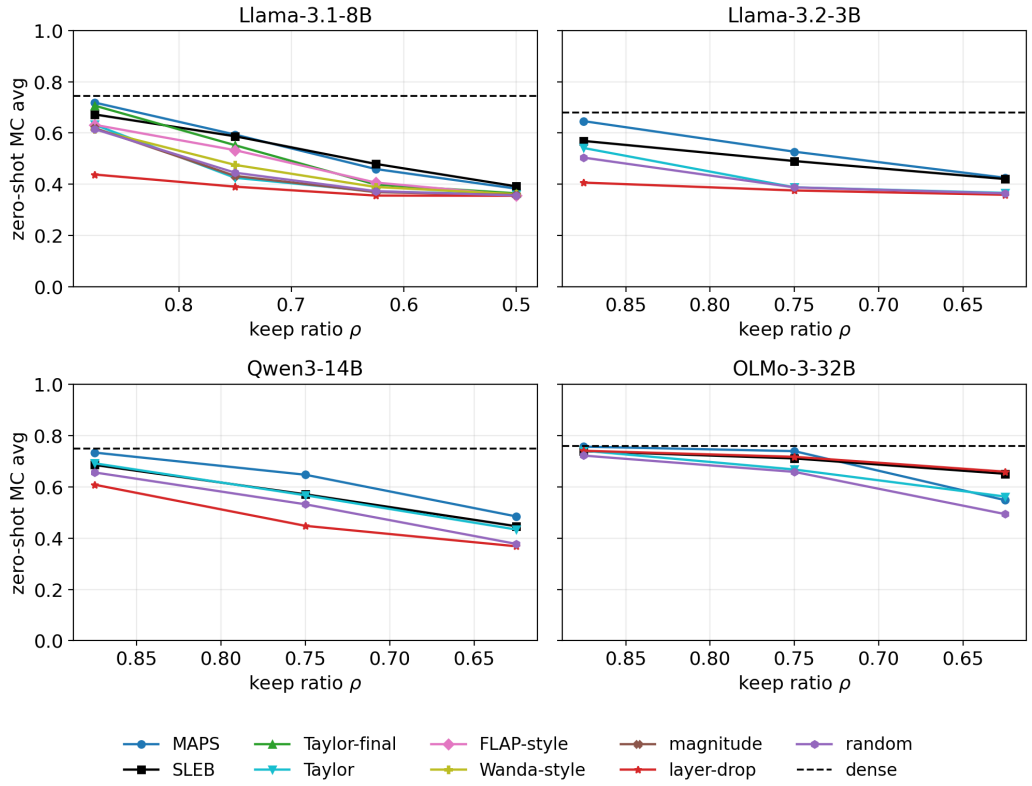

Figure 3: **No-recovery collapse curves on four backbones** (zero-shot MC vs. ρ). ρ=0.9 is omitted for group-granular methods (8-unit stages cannot resolve it). MAPS leads every cell at $\rho \geq 0.75$; the deepest OLMo-3-32B cell is the documented probe-metric dissociation (§5); a task-aware probe metric is a drop-in mitigation we leave unexplored to preserve pre-registration fidelity.

method@ρ	params (B)	KV B/tok	tok/s (bs8)	peak VRAM	prune time
dense	8.03	131,072	153.0	15.5 GB	–
MAPS@0.875	7.81	126,976	157.2	15.1 GB	62 s
MAPS@0.75	6.97	111,104	154.8	13.5 GB	39 s
MAPS@0.625	**6.12**	**95,232**	154.1	**11.8 GB**	43 s
layer-drop@0.625	5.41	81,920	**232.2**	10.5 GB	2.5 s
SparseGPT (unstr. 50%)	8.03	131,072	150.7	15.5 GB	584 s

Table 5: **Deployment cost** (Llama-3.1-8B, A100, greedy decode, 128 new tokens). Unstructured sparsity changes no shape: memory savings 0% and no decode speedup without sparse kernels; width converts budget into memory and batch capacity, depth into latency. ρ is a nominal unit-keep budget: the hub's materialized reduction is $2.7/13.2/23.8\%$ at ρ=$0.875/0.75/0.625$ ($8.03 \rightarrow 7.81/6.97/6.12$B; layer-drop 10.9% at 0.875); 32B counts were not logged, so its budgets are nominal.

the general MC composite better (0.513 vs. 0.444 at ρ=0.625). Calibration should match the deployment distribution; mismatch cost grows with depth, and the measured answer costs minutes per cell in this protocol.

7 Related Work

Importance-based structured pruning. Magnitude and Taylor criteria descend from CNN-era pruning (Han et al., 2015; Wen et al., 2016; He et al., 2018; Blalock et al., 2020) and anchor LLM-Pruner-style systems (Ma et al., 2023); our Torch-Pruning row instantiates the same Taylor criterion via its official dependency-graph implementation (Fang et al., 2023). Wanda's $|W| \cdot \|x\|$ (Sun et al., 2024) is, as we prove, invariant to the gauges of §2 — and, as we measure, still uncorrelated with causal effect: gauge dependence is sufficient but not necessary for proxy failure. SparseGPT (Frantar & Alistarh, 2023) reconstructs weights in the unstructured regime; we run its official implementation as a measured cross-regime reference (its 50% accuracy cost is substantial at $\leq$8B scale; Table 4). SliceGPT (Ashkboos et al., 2024) is discussed from its paper. Depth pruning (Fan et al., 2019; Men et al., 2025) appears as a baseline and *inside* MAPS as the L_{pure} candidate; SLEB (Song et al., 2024) is its fully-measured form and, in our taxonomy, MAPS's closest relative — the two measured methods bracket the top of every comparison in §6. First-order Taylor scoring (Molchanov et al., 2019; Ma et al., 2023) and activation-fluctuation scoring (An et al., 2024) are the estimation families our E4 sweep tests directly: gauge-invariant, and still uninformative. Pruning with continued pretraining (Xia et al., 2024) targets a different budget regime and is out of scope for a no-recovery study. **Reparameterization invariance.** That QK and OV circuits are defined only up to invertible transforms is established in interpretability (Elhage et al., 2021); Proposition 1 is the diagonal per-group instance, sharpened into a bit-exact falsifier and aimed at pruning scores. Rescaling-invariance critiques have precedent in generalization research (Salimans & Kingma, 2016; Neyshabur et al., 2015; Dinh et al., 2017); our §4 canonicalization is the attention/SwiGLU analogue of weight balancing, and the new observation is that even the canonical representative carries no causal information here. We view our contribution as turning a known symmetry into a bit-exact falsifier and mapping its empirical consequences for pruning practice, not as new mathematics. **Causal importance and evaluation practice.** Ablation is interpretability's oldest tool (Michel et al., 2019; Meng et al., 2022); we systematize it as pruning currency with cost accounting, show where solo ablations mislead, and show recovery fine-tuning (Liu et al., 2018; Frankle & Carbin, 2018) can compress and reorder method gaps (Appendix B) — motivating the no-recovery primary regime.

8 Limitations

The mechanism is open: we falsified our own normalization hypothesis and do not know what makes a proxy helpful on Qwen2.5 and harmful on OLMo-2. **Probe-metric dissociation**: MAPS optimizes measured probe NLL; on the deepest 32B cell the NLL-optimal plan is width-heavy while downstream MC prefers depth (0.548 vs. 0.659) — criterion choice can bind, and the dissociation itself deserves study. **Fixed constants**: the candidate window and $K{=}6$ are fixed; plans vary across calibration draws at deep budgets ($1/6/10$ layers), reported as mean$\pm$range. **Residual estimation and the deep-budget frontier**: MAPS measures only five candidate plans; its unit scores remain solo estimates with test–retest reliability $\bar{r}{=}0.55$ at $K{=}6$ (App. E). The $\rho{=}0.625$ cells where SLEB's exhaustive depth measurement wins locate exactly this residual: scaling the number of *measured* candidates (score refresh per stage, larger plan sets) is the natural extension, and the dose–response pattern in §6 predicts it pays. **Evaluation breadth**: the suite is MC-centric (five zero-shot tasks, GSM8K, WikiText, partial MMLU); generation-heavy and long-context breadth is future work. **MoE scope**: OLMoE experts are pruned by router mass, not ablation, and do not beat random without recovery. **Regime and scale**: claims concern no-recovery pruning of decoder-only text models up to 32B; the gauge argument is untested beyond text.

Reproducibility Statement

All results are reproducible from the artifact provided as supplementary material (code, cached interventional scores, every result cell including failures, and a checksummed manifest): three shared modules plus click-to-run per-experiment notebooks; every (model, method, ρ, calibration) cell caches its result JSON and resumes idempotently; every pruned model passes a structural self-test ladder before entering the matrix; cached interventional scores (from which every plan is rebuilt deterministically; each cell records its plan hash and dropped layers), the aggregate CSV, the pre-registration file (written and timestamped before any E2 measurement), and the failure ledger ship together. Appendix F documents the pipeline, pre-registration boundaries, and operational pitfalls.

Ethics Statement

This work studies evaluation methodology and compression of publicly released language models on public benchmarks; it introduces no new data collection, human subjects, or dual-use capability beyond making existing models cheaper to serve. Pruning can amplify a model's existing failure modes under aggressive budgets; our task-type analysis (generation collapses before multiple choice) is a caution for deployment, not a recommendation.

AI Use Statement

In this work we used generative AI tools (large language models) for: implementing and debugging the experiment pipeline and baselines, designing and critiquing the experimental protocol, interpreting results, helping formalize and draft the exposition of the theoretical statements, drafting and editing the manuscript, generating figures from cached results, identifying literature and formatting references, and building the automated consistency checks that compare every reported number against the released result files. We did not use generative AI tools to generate data. All AI-assisted code was executed by the authors and its outputs verified against cached results; every reported number was re-derived from the released artifact; theoretical claims and their proofs were checked by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

Joshua Ainslie, James Lee-Thorp, Michiel De Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 4895–4901, 2023.

Yongqi An, Xu Zhao, Tao Yu, Ming Tang, and Jinqiao Wang. Fluctuation-based adaptive structured pruning for large language models. In *AAAI Conference on Artificial Intelligence*, volume 38, pp. 10865–10873, 2024.

Saleh Ashkboos, Maximilian Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. SliceGPT: Compress large language models by deleting rows and columns. In *International Conference on Learning Representations*, volume 2024, pp. 11682–11701, 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.

Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Guttag. What is the state of neural network pruning? *Proceedings of machine learning and systems*, 2:129–146, 2020.

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.

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

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pp. 1019–1028. PMLR, 2017.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Angela Fan, Edouard Grave, and Armand Joulin. Reducing transformer depth on demand with structured dropout. *arXiv preprint arXiv:1909.11556*, 2019.

Gongfan Fang, Xinyin Ma, Mingli Song, Michael Bi Mi, and Xinchao Wang. DepGraph: Towards any structural pruning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16091–16101. IEEE, 2023.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

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.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, et al. A framework for few-shot language model evaluation. *Zenodo*, 2021.

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.

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.

Yihui He, Ji Lin, Zhijian Liu, Hanrui Wang, Li-Jia Li, and Song Han. AMC: Automl for model compression and acceleration on mobile devices. In *European conference on computer vision*, pp. 815–832. Springer, 2018.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4246–4253, 2020.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. *arXiv preprint arXiv:1810.05270*, 2018.

Xudong Lu, Qi Liu, Yuhui Xu, Aojun Zhou, Siyuan Huang, Bo Zhang, Junchi Yan, and Hongsheng Li. Not all experts are equal: Efficient expert pruning and skipping for mixture-of-experts large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6159–6172, 2024.

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.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in neural information processing systems*, 35:17359–17372, 2022.

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

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32, 2019.

Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11256–11264. IEEE, 2019.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, et al. OLMoE: Open mixture-of-experts language models. In *International Conference on Learning Representations*, volume 2025, pp. 62061–62121, 2025.

Behnam Neyshabur, Russ R Salakhutdinov, and Nati Srebro. Path-sgd: Path-normalized optimization in deep neural networks. *Advances in neural information processing systems*, 28, 2015.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

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.

Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, 29, 2016.

Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Jiwon Song, Kyungseok Oh, Taesu Kim, Hyungjun Kim, Yulhwa Kim, and Jae-Joon Kim. SLEB: Streamlining LLMs through redundancy verification and elimination of transformer blocks. In *International Conference on Machine Learning (ICML)*, pp. 46136–46155, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Mingjie Sun, Zhuang Liu, Anna Bair, and Zico Kolter. A simple and effective pruning approach for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 4942–4964, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Learning structured sparsity in deep neural networks. *Advances in neural information processing systems*, 29, 2016.

Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. Sheared LLaMA: Accelerating language model pre-training via structured pruning. In *International Conference on Learning Representations*, volume 2024, pp. 5385–5409, 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.

A FULL PROXY–CAUSAL CORRELATIONS AND BEHAVIORAL ARMS

Table 6 reports, per backbone: Spearman correlations of each proxy family against causal ΔNLL over all group units, before → after balanced gauge canonicalization (§4; function drift ≤ 0.004 nats after bias-aware rescaling), and the three behavioral arms of §3 at ρ=0.75: (A) traffic top-k, (B) random, (C) causal top-k.

backbone	ρ_s vs. causal (before → canonical) traffic	magnitude	Wanda	ΔNLL arms A	B	C
Llama-3.1-8B	$+0.01 \to -0.00$	$+0.18 \to +0.18$	$+0.15 \to +0.15$	1.70	2.37	1.08
Qwen3-8B	$-0.05 \to -0.05$	$+0.09 \to +0.09$	$+0.08 \to +0.08$	0.46	0.65	0.56
Qwen2.5-7B	$+0.17 \to +0.15$	$+0.09 \to +0.09$	$+0.18 \to +0.18$	0.69	2.93	0.48
Mistral-7B-v0.3	$-0.06 \to -0.07$	$+0.24 \to +0.24$	$+0.06 \to +0.06$	0.64	0.58	7.27
OLMo-2-7B	$-0.06 \to -0.06$	$+0.15 \to +0.15$	$+0.19 \to +0.19$	6.22	4.18	1.39
Gemma-3-1B	$-0.04 \to -0.02$	$-0.23 \to -0.23$	$+0.25 \to +0.25$	0.86	1.13	0.49

Table 6: Correlations never exceed $|\rho_s|$=0.26 before or after canonicalization; behavioral consequences swing in both directions; Mistral's causal arm is the flat-landscape collapse of §3.5 (a single refresh restores $+0.64$).

Gauge-invariant proxies (E4). Table 7 extends the map to Taylor ($|\sum_i w_i g_i|$ per unit), empirical Fisher ($\sum_i (w_i g_i)^2$), and standard-aggregation Wanda. Both gradient proxies are exactly gauge-invariant on Llama (rank survival > 0.9999, exact up to floating-point ties) and ε-invariant under QK-norm (Qwen3 ≥ 0.91) — the ε in Prop. 2 is thus empirically ≤ 0.09 in rank terms for this class — yet their causal correlations stay in $[-0.20, +0.50]$ across all six backbones, and their top-k keep-sets are behaviorally unsafe: on Llama the Taylor arm degrades ΔNLL to 8.73 against 2.37 for random and 1.08 for causal selection.

Which Wanda is invariant. Our main-text "Wanda-style" aggregates $\sum_{i \in \text{unit}} \|X_i\|_2 \, |W_{\cdot i}|$ over *paired* in/out channels of each gauge orbit, so every factor α cancels ($\|\alpha X\| \, |W/\alpha| = \|X\| \, |W|$): survival 1.000. The standard row-wise form aggregates unpaired rows and rides the gauge exactly like magnitude — measured survival ρ_s=0.253 on Llama. Reported Wanda numbers elsewhere in the paper use the invariant form; the standard form is strictly worse under reparameterization.

backbone	ρ_s(Tay)	ρ_s(Fis)	ρ_s(W-std)	ΔNLL Tay	Fis	random	gauge surv. (T/F/W)
Llama-3.1-8B	$+0.01$	-0.03	$+0.18$	8.73	1.20	–	1.000/1.000/0.253
Qwen3-8B	$+0.02$	-0.05	$+0.01$	0.79	1.06	–	0.911/0.997/0.935
Qwen2.5-7B	$+0.19$	$+0.17$	$+0.10$	3.22	1.12	–	–
Mistral-7B-v0.3	$+0.09$	$+0.25$	$+0.14$	0.52	0.34	–	–
OLMo-2-7B	$+0.25$	$+0.50$	$+0.33$	2.97	2.02	–	–
Gemma-3-1B	-0.19	$+0.27$	$+0.18$	1.28	0.63	–	–

Table 7: **E4: invariant proxies across six backbones.** Spearman vs. causal ΔNLL; behavioral arms are mean ΔNLL of top-k keep-sets (k matched); gauge survival on the two 8B backbones (exact on Llama, QK-norm on Qwen3).

B PERSISTENCE UNDER RECOVERY FINE-TUNING

We attach LoRA adapters (rank 16, 800 steps, WikiText-103 text) to every $\rho \in \{0.875, 0.75\}$ hub checkpoint of six methods and re-evaluate zero-shot MC — a deliberately method-agnostic, domain-mismatched recovery that favors no ranking.

GSM8K does not survive this protocol for *any* method (accuracies 0.00–0.27): text-only LoRA overwrites the answer format. Task-matched recovery would be required to compare generation post-recovery; we report the MC comparison and flag this as a protocol limitation rather than sweeping it into a favorable cell.

method (+LoRA)	ρ=0.875	ρ=0.75
MAPS	0.708	0.642
SLEB	0.677	0.624
Wanda-style	0.661	0.578
random	0.668	0.574
Taylor top-k	0.671	0.430
layer-drop	0.598	0.524

Table 8: **Recovery compresses gaps but preserves the ordering.** Zero-shot MC after identical LoRA recovery. At ρ=0.75: MAPS 0.642 > SLEB 0.624 > Wanda 0.578 $\approx$ random 0.574 > layer-drop 0.524 > Taylor 0.430; at ρ=0.875: MAPS 0.708 > SLEB 0.677 > Taylor 0.671 $\approx$ random 0.668 > Wanda 0.661 > layer-drop 0.598. Both post-recovery orderings coincide with the no-recovery orderings of Table 4 (draw d_0): the measured pair stays on top, and no estimation baseline overtakes any measured method after recovery.

C MoE EXPERT PRUNING (OLMoE)

Expert units of OLMoE-1B-7B (Muennighoff et al., 2025) are selected by routed token mass (causal expert ablation requires router-internal intervention and is future work; cf. Lu et al., 2024); the top-k_ρ experts by routed mass are kept per layer. Table 9 reports zero-shot MC across budgets.

OLMoE-1B-7B (zero-shot MC)	ρ=0.9	0.875	0.7	0.625
MAPS (router-mass experts)	0.680	0.661	0.608	0.569
greedy (abl.)	–	0.661	–	0.569
random	**0.686**	**0.668**	**0.625**	0.568
dense		0.702		

Table 9: Router-mass expert selection does not beat random at any tested budget without recovery (random leads by $+0.006$ to $+0.018$ MC at $\rho \geq 0.7$; tie at 0.625) — consistent with the paper's thesis that routed mass is traffic, not measurement. The greedy ablation coincides with MAPS where run (0.661/0.569) because expert stages admit only per-layer top-k: there is no allocation freedom for the allocator to exploit. Grid points follow the appendix runner's schedule (expert stages have width 64, so ρ=0.9 is resolvable).

D FULL SPOKE RESULTS

Table 10 reports zero-shot MC by budget, with GSM8K and WikiText-2 PPL at ρ=0.875. The 3B spoke additionally ran $\rho \in \{0.9, 0.8, 0.7\}$ (released CSV). SparseGPT rows exist where the official implementation runs (§6).

Diagnostic backbones and the Mistral grid. The three map-only backbones (§3) confirm the pattern out of sample at ρ=0.875 (left), and the full Mistral grid (right) realizes the score-swap reversal of §5: with the identical measured allocator, Taylor scores beat causal solo scores at every budget on this flat-landscape backbone — and both beat every estimation-only row.

Generation across budgets. GSM8K for the hub across all structured budgets — the task-type dependence of §6 in full:

Calibration $\times$ evaluation transfer.

E PROMPT-COUNT SENSITIVITY AND DETERMINISM

Re-scoring the hub with K=4 and K=12 calibration prompts reproduces the K=6 causal unit ranking at Spearman 0.896 and 0.889; plan-level choices are the sensitive quantity at deep budgets

Llama-3.2-3B	mc@.875	mc@.75	mc@.625	gsm@.875	ppl@.875
MAPS	0.646	0.527	0.425	0.186	11.743
greedy (abl.)	0.516	0.442	0.415	0.022	35.212
group (abl.)	0.574	0.469	0.380	0.056	26.370
layer-drop	0.406	0.376	0.359	0.004	77.683
random	0.504	0.388	0.364	0.012	68.550
SLEB (measured depth)	0.569	0.490	0.420	0.024	18.448
Taylor top-k (group)	0.541	0.388	0.366	0.012	22.359
SparseGPT (50% unstr.)		0.590		0.200	19.485
dense		0.679		0.266	9.776
Qwen3-14B	mc@.875	mc@.75	mc@.625	gsm@.875	ppl@.875
MAPS	0.734	0.647	0.485	0.900	11.587
greedy (abl.)	0.634	0.491	0.468	0.192	33.214
group (abl.)	0.688	0.535	0.400	0.500	16.838
layer-drop	0.609	0.448	0.368	0.528	18.413
random	0.656	0.532	0.378	0.566	17.838
SLEB (measured depth)	0.686	0.571	0.447	0.644	14.401
Taylor top-k (group)	0.692	0.567	0.434	0.306	15.793
dense		0.749		0.924	11.127
OLMo-3-32B	mc@.875	mc@.75	mc@.625	gsm@.875	ppl@.875
MAPS	0.757	0.739	0.548	0.820	5.981
greedy (abl.)	0.541	0.508	0.441	0.000	78.138
group (abl.)	0.737	0.654	0.548	0.594	8.443
layer-drop	0.741	0.717	0.659	0.796	7.040
random	0.722	0.659	0.494	0.542	8.645
SLEB (measured depth)	0.739	0.711	0.651	0.730	9.637
Taylor top-k (group)	0.741	0.668	0.562	0.598	7.639
dense		0.760		0.822	5.692

Table 10: Full spoke matrices. MAPS leads all structured methods at $\rho \geq 0.75$ on every backbone; at ρ=0.625 SLEB takes the 32B cell (the probe-metric dissociation of §5, where depth-only removal is optimal) while MAPS holds 3B and 14B.

ρ=0.875	MAPS	SLEB	dense
Qwen2.5-7B	0.707	0.635	0.721
OLMo-2-7B	0.715	0.708	0.754
Gemma-3-1B	0.597	0.554	0.613

Mistral-7B	mc@.875	mc@.75	mc@.625
MAPS	0.712	0.604	0.445
Taylor-final	0.727	0.654	0.493
SLEB	0.677	0.584	0.481
Taylor top-k	0.687	0.586	0.431
Wanda-style	0.668	0.542	0.422
magnitude	0.679	0.594	0.421
random	0.695	0.527	0.361
layer-drop	0.435	0.372	0.360

Table 11: Left: diagnostic backbones, no recovery. Right: full Mistral comparison. Taylor-final leads at every budget; MAPS is second at $\rho \geq 0.75$, and at the deepest budget the measured trio (Taylor-final, SLEB, MAPS) fills the top three. At $\rho \geq 0.75$ the third slot goes to random/magnitude — Mistral's flat shallow-budget landscape (§3) compresses those cells, which is exactly why solo-score selection was dangerous there.

GSM8K (hub)	ρ=0.875	0.75	0.625
MAPS	0.216–0.468	0.004–0.046	0.000–0.002
SLEB	0.152–0.250	0.006–0.040	0.000–0.002
Taylor-final	0.254–0.498	0.000–0.034	0.000–0.012
Wanda-style	0.082–0.106	0.000–0.002	0.000
random	0.038	0.000	0.000

Table 12: Generation collapses one budget earlier than MC for every method; measured selection delays but does not prevent it. Ranges are min–max over the three calibration draws.

calibration	ρ=0.875	0.75	0.625	0.5
GSM8K (task)	0.704	0.597	0.444	0.375
WikiText (general)	0.704	0.606	0.513	0.403

Table 13: **Calibration $\times$ evaluation transfer** (hub MAPS, MC average; draw d_0). Identical plans at ρ=0.875; growing divergence with depth.

and are exposed as mean±range over three disjoint draws throughout. Evaluation is deterministic by construction (log-likelihood scoring, greedy decoding); re-evaluating a fixed pruned model reproduces every reported metric identically, and each cell stores its evaluation-context marker (`eval_ctx`=4096).

Per-draw scores and draw-robustness of rankings. Table 14 lists the individual calibration draws ($d_0/d_1/d_2$) behind every hub range, with the draw mean in parentheses. Rankings in the main text are declared on draw means. They are not uniformly draw-robust for the closest pairs: at ρ=0.875 the best Taylor-final draw (0.733) exceeds the best MAPS draw (0.731), and at ρ=0.75 the best SLEB draw (0.612) exceeds every MAPS draw — so that cell's ranking is draw-fragile even though its matched-draw gap clears benchmark noise (Table 15). The wide-margin cells (ρ=0.875 vs. SLEB on every backbone; the deepest 32B cell) hold on every individual draw.

hub, MC $d_0/d_1/d_2$ (mean)	ρ=0.875	0.75	0.625	0.5
MAPS	0.704/0.720/0.731 (0.719)	0.597/0.584/0.600 (0.594)	0.444/0.454/0.479 (0.459)	0.375/0.372/0.401 (0.383)
SLEB	0.671/0.675/0.671 (0.672)	0.582/0.612/0.568 (0.587)	0.484/0.473/0.480 (0.479)	0.392/0.391/0.392 (0.392)
Taylor-final	0.701/0.687/0.733 (0.707)	0.597/0.442/0.617 (0.552)	0.438/0.371/0.387 (0.399)	0.374/0.358/0.361 (0.365)

Table 14: Per-draw zero-shot MC on the hub for the three measured methods; draw means in parentheses are the values on which main-text rankings are declared.

Finite-sample uncertainty: paired bootstrap. Draw ranges quantify calibration-prompt variance; they do not quantify benchmark sampling variance. For the headline cells we rebuilt each pruned model from its cached plan (plan hash and every reported metric reproduced exactly), logged per-example correctness, and resampled evaluation examples within each task (10^4 paired resamples, matched draw d_0, identical indices for both methods, unweighted five-task mean). Table 15 reports point differences with 95% percentile intervals and $P(\Delta \leq 0)$. Every MAPS–SLEB, MAPS–layer-drop, and Taylor-final–MAPS difference excludes zero; the hub score-swap pairs are ties at all three budgets; and the two depth-only removals at the deepest 32B cell are indistinguishable ($+0.8$ $[-0.1, +1.7]$). The hub ρ=0.75 MAPS–SLEB gap ($+1.5$) clears benchmark noise while remaining draw-fragile (above) — there, calibration-draw variance, not benchmark noise, is the binding uncertainty.

backbone	ρ	A − B	ΔMC (pt)	95% paired CI	$P(\Delta \leq 0)$
Llama-3.1-8B	0.875	MAPS − SLEB	+3.3	[+2.6, +4.0]	$< 10^{-4}$
Llama-3.1-8B	0.75	MAPS − SLEB	+1.5	[+0.6, +2.5]	0.0003
Llama-3.1-8B	0.625	SLEB − MAPS	+4.0	[+2.9, +5.1]	$< 10^{-4}$
Llama-3.1-8B	0.875	MAPS − Taylor-final	+0.4	[−0.4, +1.1]	0.163
Llama-3.1-8B	0.75	MAPS − Taylor-final	−0.0	[−1.0, +1.0]	0.526
Llama-3.1-8B	0.625	MAPS − Taylor-final	+0.6	[−0.5, +1.7]	0.144
Llama-3.2-3B	0.875	MAPS − SLEB	+7.7	[+6.8, +8.6]	$< 10^{-4}$
Qwen3-14B	0.875	MAPS − SLEB	+4.8	[+4.0, +5.6]	$< 10^{-4}$
OLMo-3-32B	0.875	MAPS − SLEB	+1.8	[+1.0, +2.5]	0.0001
OLMo-3-32B	0.875	MAPS − layer-drop	+1.6	[+1.0, +2.2]	$< 10^{-4}$
OLMo-3-32B	0.625	layer-drop − MAPS	+11.1	[+10.0, +12.2]	$< 10^{-4}$
OLMo-3-32B	0.625	SLEB − MAPS	+10.3	[+9.2, +11.4]	$< 10^{-4}$
OLMo-3-32B	0.625	layer-drop − SLEB	+0.8	[−0.1, +1.7]	0.040
Mistral-7B	0.875	Taylor-final − MAPS	+1.5	[+0.8, +2.2]	$< 10^{-4}$
Mistral-7B	0.75	Taylor-final − MAPS	+5.0	[+4.1, +6.0]	$< 10^{-4}$

Table 15: Paired bootstrap over evaluation examples for 15 headline pairs (draw d_0, 10^4 resamples).

Scaling the measurement (K=64). Re-running the deepest hub cell (ρ=0.625) with 64-prompt causal scoring collapses the plan's layer component to a single measured choice (two of three draws pick the identical plan) and tightens the draw range $\sim 2\times$ (0.435–0.451 vs. 0.444–0.479 at K=6) at comparable mean — consistent with plan variance, not score bias, dominating at depth. A Wanda-style variant with its recommended calibration regime (128 sequences) moves its own cells by $<$ 0.03 MC (0.653/0.505/0.393) and changes no conclusion.

Score reliability and attenuation. Across the three disjoint K=6 draws, causal unit scores have pairwise test–retest reliability $r \in [0.53, 0.57]$ ($\bar{r}$=0.552): individual scores are noisy solo estimates,

which is precisely why MAPS measures *plans* rather than trusting *units*. Correcting the proxy–causal correlations of Table 7 for this attenuation leaves them weak (Taylor 0.00, Fisher -0.05, Wanda-std $+0.25$ on the hub); partialling out depth adds at most $+0.17$. The proxies' failure is signal absence, not measurement noise.

F REPRODUCIBILITY LEDGER

Every experiment is a click-to-run notebook over three shared modules; each (model, method, ρ, calibration) cell caches its result and resumes idempotently; every pruned model passes a structural self-test ladder and a probe gate before evaluation; cached interventional scores (plans are rebuilt from them deterministically; plan hashes are recorded per cell), per-cell JSONs, and the pre-registration file are released (its timestamp precedes all E2 runs; one deviation is disclosed inline: the Mistral refresh of §3 was added *after* observing the collapse and is labeled as such rather than folded into the pre-registered protocol). Two operational lessons are recorded for practitioners: (i) accelerate-dispatched models are reference cycles — releasing them requires explicit garbage collection, or failed cells poison subsequent loads; (ii) in cached pipelines, *invalidation must erase metrics, not only status*, or incremental evaluation silently skips re-measurement. The failure ledger (orphaned and infeasible cells, all excluded by status) ships with the results.

G PER-TASK ZERO-SHOT SCORES (HUB)

The MC composite is an unweighted macro-average of five tasks. Table 16 reports every task for the primary hub cells (calibration draw d_0; acc_norm where lm-eval defines it, acc otherwise), so that the composite cannot hide task-specific failures.

method	ρ	ARC-E	ARC-C	HellaSwag	WinoGrande	PIQA	MC avg
dense	1.0	0.827	0.548	0.794	0.746	0.810	0.745
MAPS	0.875	0.774	0.488	0.754	0.710	0.797	0.704
SLEB	0.875	0.727	0.439	0.716	0.700	0.774	0.671
Taylor-final	0.875	0.781	0.506	0.763	0.674	0.779	0.701
Taylor top-k	0.875	0.656	0.383	0.651	0.661	0.746	0.620
FLAP-style	0.875	0.701	0.410	0.661	0.647	0.754	0.635
Wanda-style	0.875	0.654	0.386	0.611	0.636	0.746	0.607
magnitude	0.875	0.648	0.409	0.665	0.626	0.749	0.619
random	0.875	0.623	0.379	0.675	0.652	0.747	0.615
layer-drop	0.875	0.413	0.256	0.385	0.521	0.615	0.438
MAPS	0.75	0.641	0.369	0.624	0.624	0.726	0.597
SLEB	0.75	0.643	0.381	0.564	0.603	0.718	0.582
Taylor-final	0.75	0.638	0.386	0.632	0.594	0.737	0.597
Taylor top-k	0.75	0.264	0.258	0.260	0.489	0.532	0.361
FLAP-style	0.75	0.561	0.303	0.532	0.576	0.690	0.533
Wanda-style	0.75	0.431	0.237	0.381	0.563	0.616	0.446
magnitude	0.75	0.388	0.267	0.397	0.526	0.583	0.432
random	0.75	0.392	0.260	0.431	0.540	0.601	0.445
layer-drop	0.75	0.342	0.224	0.294	0.524	0.567	0.390
MAPS	0.625	0.415	0.251	0.408	0.529	0.619	0.444
SLEB	0.625	0.468	0.297	0.443	0.570	0.643	0.484
Taylor-final	0.625	0.425	0.253	0.389	0.512	0.612	0.438
Taylor top-k	0.625	0.260	0.264	0.260	0.488	0.530	0.360
FLAP-style	0.625	0.371	0.251	0.363	0.521	0.582	0.418
Wanda-style	0.625	0.342	0.235	0.307	0.528	0.566	0.396
magnitude	0.625	0.285	0.241	0.286	0.510	0.524	0.369
random	0.625	0.285	0.241	0.285	0.504	0.551	0.373
layer-drop	0.625	0.262	0.250	0.264	0.483	0.520	0.356

Table 16: Per-task zero-shot accuracy on Llama-3.1-8B (draw d_0), no recovery.