SENSITIVITY-WEIGHTED RECONSTRUCTION MITIGATES PRIORITY DILUTION IN LLM PRUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

One-shot pruning compresses large language models by matching sparse module outputs to dense counterparts. Existing reconstruction objectives assign unit weight to every calibration token occurrence, despite differences in how local reconstruction errors affect the global loss. We define occurrence sensitivity as the squared norm of the loss gradient at the module output. Across 11 models, the most sensitive 10% of occurrences account for a median of 63% of the sensitivity mass. This concentration exposes Uniform reconstruction to *priority dilution*: abundant low-sensitivity occurrences can reduce the reconstruction priority of more sensitive ones. We propose *Sensitivity-Weighted Reconstruction* (SWR), which converts raw occurrence sensitivities into normalized, clipped reconstruction weights and inserts them into existing solvers. With fixed pruning masks, adding the least sensitive half at unit weight raises perplexity in 8 of 11 models; replacing those unit weights with SWR weights reduces this degradation in 10 of 11. Across five pruning methods and 11 models at 70% unstructured sparsity, SWR lowers perplexity in 51 of 55 configurations, with a 3.0% median relative reduction, and improves zero-shot accuracy in 51 of 55 configurations. Further analysis finds that high-sensitivity occurrences are more strongly associated with the token being processed for K/V and generally with the token being predicted next for Q/O/FFN.

1 INTRODUCTION

Large language models (LLMs) have substantial parameter and compute requirements (Grattafiori et al., 2024; Touvron et al., 2023b; Yang et al., 2024; 2025a; Jiang et al., 2023; Zhang et al., 2022). One-shot pruning provides a direct path to efficiency by selecting low-impact weights for removal (Yu et al., 2022; Benbaki et al., 2023; Frantar & Alistarh, 2022). Some methods also update retained weights after selecting a pruning mask (Frantar & Alistarh, 2023; Meng et al., 2024). These procedures use a few passes over a small calibration set to record module statistics, typically input activations (Sun et al., 2024; Frantar & Alistarh, 2023) and gradients (Das et al., 2023).

Rather than updating the whole network jointly, one-shot methods use these statistics to optimize one weight matrix at a time (Frantar & Alistarh, 2023; Meng et al., 2024). They adjust retained weights so that each sparse module matches the corresponding dense module output. We call each token at a specific position in a calibration sequence a *token occurrence* and index all such occurrences by $t = 1, \ldots, T$. Let $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ be the dense weight matrix. For occurrence t, $x_t \in \mathbb{R}^{d_{\text{in}}}$ is the input activation received by this module, and Wx_t is the dense module output for that occurrence. For a binary pruning mask $M \in \{0,1\}^{d_{\text{out}} \times d_{\text{in}}}$, let $\mathcal{S}_M$ denote the subspace of matrices whose support is restricted to M. A standard reconstruction objective is (Frantar & Alistarh, 2023; Meng et al., 2024):

$$\min_{\widehat{W}} \sum_{t=1}^{T} \|(\widehat{W} - W)\, x_t\|_2^2 \qquad \text{s.t.} \quad \widehat{W} \in \mathcal{S}_M. \tag{1}$$

We refer to this as the *Uniform* objective. It gives equal-sized output residuals the same penalty at every occurrence; in joint methods, this reconstruction criterion also guides mask selection.

Related work has shown that token occurrences need not be equally important. Token-weighted training assigns unequal importance to tokens (Lin et al., 2024b; Gu et al., 2020); loss-aware quantization weights reconstruction errors by their effects on the loss (Li et al., 2021; Kim et al.,

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.

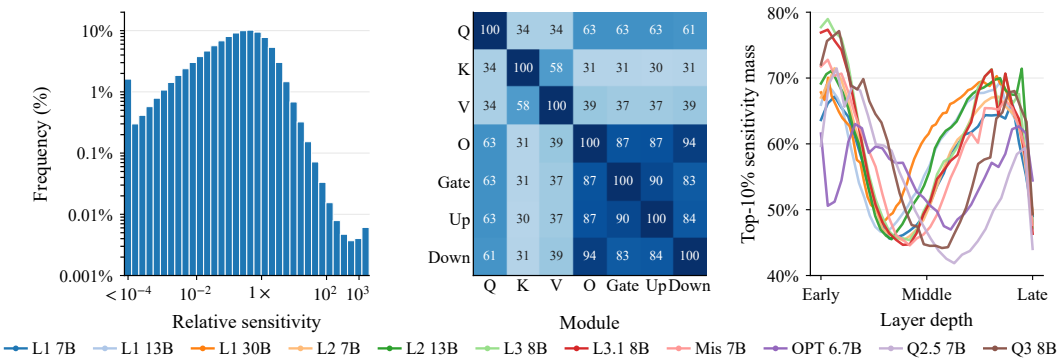


Figure 1: **Occurrence sensitivity is concentrated and module-specific. Left:** Sensitivity histograms after per-sequence, per-module mean normalization, averaged with equal weight across models. The right tail spans several orders of magnitude. This pattern is consistent across all 11 models. **Middle:** Each cell reports the overlap between the sets containing the most sensitive 10% of occurrences for the row and column module types; 100% means identical sets. **Right:** At each layer depth, the y-axis is the percentage of total sensitivity carried by the most sensitive 10% of occurrences.

2025); and autoregressive generation depends disproportionately on a subset of occurrences (Zhang et al., 2023; Liu et al., 2023). One-shot pruning, however, usually gives every calibration occurrence equal reconstruction priority. Because sparsity prevents exact reconstruction, some local error is unavoidable and can accumulate across layers (Huang et al., 2025). The question is therefore not only which weights to retain, but how to allocate limited reconstruction capacity across calibration occurrences to limit the resulting increase in global loss.

We analyze this allocation problem through a second-order expansion of the global next-token loss. Appendix C.1 bounds a block-diagonal empirical Fisher surrogate by a weighted sum of squared reconstruction errors. The coefficient for each occurrence is the squared norm of the loss gradient at the module output; we call it *occurrence sensitivity*. For each empirical Fisher block, this coefficient is the smallest scalar isotropic majorant, although it can be loose for residuals not aligned with the gradient. Across 11 models, the most sensitive 10% of occurrences carry a median of 63% of the sensitivity mass (Appendix B.2). Figure 1 shows partial overlap across modules and changes in sensitivity concentration with layer depth.

Because occurrence sensitivity is concentrated while the Uniform objective assigns unit weight to every occurrence, abundant low-sensitivity occurrences can reduce the reconstruction priority of more sensitive ones. We call this *priority dilution*.

We propose *Sensitivity-Weighted Reconstruction* (SWR) and integrate it into the local objective.

$$\min_{\widehat{W}} \sum_{t=1}^{T} \rho_t \|(\widehat{W} - W)\, x_t\|_2^2 \qquad \text{s.t.} \quad \widehat{W} \in \mathcal{S}_M, \tag{2}$$

where we construct ρ_t separately for each module by normalizing and clipping occurrence sensitivities within every calibration sequence. The analysis motivates raw occurrence sensitivity; normalization and clipping stabilize its use in sparse reconstruction. SWR retains all calibration occurrences and changes only their reconstruction weights. For methods that only produce pruning masks, we keep the mask fixed and perform Optimal Brain Surgeon (OBS) reconstruction (Hassibi & Stork, 1992) using SWR weights. For joint pruning methods, SWR replaces the Uniform objective within the existing procedure, guiding both mask selection and retained-weight updates (Section 3.3).

Section 5 tests priority dilution using a fixed Wanda pruning mask at 70% sparsity. Starting with SWR weights on the more sensitive half, adding the least sensitive half at unit weight raises perplexity (PPL) in 8 of 11 models. Replacing those unit weights with SWR weights mitigates the increase in 10 of 11. The full SWR-weighted set remains close to excluding that half.

Overall, the contributions of our work are as follows:

- We identify *priority dilution*: uniform weighting can reduce the reconstruction priority of sensitive occurrences. Controlled comparisons show that adding low-ranked occurrences at unit weight can worsen PPL, while reweighting those occurrences mitigates this degradation.
- We derive module-specific occurrence sensitivity from a second-order empirical Fisher surrogate and construct SWR through per-sequence normalization and magnitude clipping. Across the evaluated configurations, SWR generally lowers perplexity and improves zero-shot accuracy across diverse models and pruning methods.
- We reveal module-specific structure in occurrence sensitivity across 11 evaluated models: high-sensitivity occurrences are more strongly associated with the token being processed for K/V and generally with the token being predicted next for Q/O/FFN.

2 Related Work

One-Shot Pruning of LLMs. Pruning methods differ in pruning mask selection and retained-weight reconstruction. Wanda scores weights by magnitude and input activation norm (Sun et al., 2024), RIA incorporates importance across rows and columns (Zhang et al., 2024a), and GBLM uses calibration gradients (Das et al., 2023). DSnoT refines masks through prune-and-grow updates (Zhang et al., 2024b), while Wanda++ combines block gradient scoring with regional weight optimization (Yang et al., 2025b). GBLM and Wanda++ use gradients to score weights for mask selection, not to weight reconstruction. Joint methods select masks while updating retained weights: SparseGPT uses column-wise OBS updates (Frantar & Alistarh, 2023; Hassibi & Stork, 1992; Frantar & Alistarh, 2022), and ALPS optimizes sparse reconstruction with ADMM (Meng et al., 2024). OWL and AlphaPruning address layer sparsity allocation (Yin et al., 2023; Lu et al., 2024), while other work studies calibration-data selection and reconstruction overfitting (Ji et al., 2025; Shin et al., 2024). WoodFisher estimates an inverse empirical Fisher in parameter space (Singh & Alistarh, 2020), while LLM Surgeon uses Kronecker-factored target-loss curvature (van der Ouderaa et al., 2024). Both change the curvature model and the solver; SWR changes only the reconstruction weights.

Loss-Aware Reconstruction. Loss-aware reconstruction uses end-loss information to prioritize compression errors. BRECQ weights block reconstruction with diagonal Fisher information (Li et al., 2021), while FWSVD uses parameter-wise Fisher information for low-rank factorization (Hsu et al., 2022). For LLM quantization, GuidedQuant constructs loss-guided Gram matrices from output-channel gradients and shares them within channel groups (Kim et al., 2025). SWR shares this principle but differs in how it constructs reconstruction weights through per-sequence normalization and clipping and how it uses them under pruning masks; in joint methods, the weights also enter mask selection. Section 3.1 relates its raw one-group Gram matrix to GuidedQuant; Section 3.2 evaluates the candidate statistics, their transformation, and output-channel grouping.

Occurrence Heterogeneity and Reweighting in Compression. Occurrence-specific treatment spans training and inference. Rho-1 selects tokens for the language-modeling loss (Lin et al., 2024b), token-adaptive training assigns per-occurrence loss weights (Gu et al., 2020), and focal loss reweights examples by difficulty (Lin et al., 2017). H2O and Scissorhands prioritize inference-time KV-cache entries (Zhang et al., 2023; Liu et al., 2023). Attribution studies characterize context-dependent token influence (Hao et al., 2021; Modarressi et al., 2022), while interpretability work describes feed-forward layers as key–value memories for next-token prediction (Geva et al., 2021; 2022). In compression, AWQ and LLM.int8() protect salient channels using activations (Lin et al., 2024a; Dettmers et al., 2022), while SpQR isolates outlier weights (Dettmers et al., 2024).

3 Sensitivity-Weighted Reconstruction

3.1 Sensitivity from the Global Loss

Let $\Delta W = \widehat{W} - W$ and $e_t = \Delta W x_t$ be the reconstruction residual at occurrence t, with t indexing calibration occurrences across sequences. Let $\mathcal{L}$ denote the global mean next-token loss. Writing $y_t = W x_t$ for the dense module output, let $g_t = \nabla_{y_t} \mathcal{L}$ denote its gradient, evaluated at the dense

model. The aggregate first-order term satisfies

$$\sum_t g_t^\top e_t = \langle \nabla_W \mathcal{L}, \Delta W \rangle_F.$$

Here $\langle A, B \rangle_F = \mathrm{Tr}(A^\top B)$ is the Frobenius inner product. We neglect this term under the near-stationarity approximation $\nabla_W \mathcal{L} \approx 0$ for the dense model; this does not require the individual g_t to vanish. We further omit curvature interactions across occurrences, including those through attention, and replace each remaining output-Hessian block by the empirical-Fisher surrogate $g_t g_t^\top$ (Hassibi & Stork, 1992; Frantar & Alistarh, 2022; Singh & Alistarh, 2020; Li et al., 2021; Kunstner et al., 2019). These approximations yield the quadratic surrogate $\mathcal{Q}(\Delta W) := \frac{1}{2}\sum_{t=1}^{T}(g_t^\top e_t)^2$ for the loss change (Appendix C.1), which couples output channels. The reconstruction solvers used here are row-separable, so we seek one scalar majorant per occurrence, shared across output rows. Defining the *occurrence sensitivity* as $a_t = \|g_t\|_2^2$ gives

$$\mathcal{Q}(\Delta W) \le \frac{1}{2}\sum_{t=1}^{T} a_t \|e_t\|_2^2 = \frac{1}{2}\,\mathrm{Tr}\left[\Delta W \left(\sum_{t=1}^{T} a_t x_t x_t^\top\right) \Delta W^\top\right]. \tag{3}$$

Since $a_t = \lambda_{\max}(g_t g_t^\top)$, where $\lambda_{\max}$ denotes the largest eigenvalue, it is the smallest isotropic majorant of the empirical Fisher block, and the Gram matrix $\sum_t a_t x_t x_t^\top$ is shared by all output rows. The bound is tight only when e_t is parallel to g_t and may be loose for other residual directions.

Constructing reconstruction weights. Raw sensitivities span several orders of magnitude (Figure 1). SWR converts them into bounded reconstruction weights by normalizing within each calibration sequence and module and then clipping. Writing $\bar{a}$ for the mean sensitivity within the calibration sequence and module containing occurrence t, we define

$$\rho_t = \begin{cases} \mathrm{clip}\big(a_t/\bar{a};\, 0.2,\, 5\big), & \bar{a} > 0, \\ 1, & \bar{a} = 0. \end{cases} \tag{4}$$

Normalization preserves ordering within each sequence and removes differences in mean sensitivity across sequences. It changes relative contributions and is not a static rescaling of the raw objective. The weights have mean one before clipping; the lower bound keeps every weight positive, and the upper bound limits outliers. The majorization in Equation 3 applies to raw occurrence sensitivity, not to the transformed reconstruction weights, which Section 3.2 evaluates empirically.

GuidedQuant (Kim et al., 2025) also uses global-loss gradients to weight local reconstruction, but applies separate weighted Gram matrices to output-channel groups in quantization. SWR instead converts the gradient signal into normalized, clipped reconstruction weights for pruning. It uses one Gram matrix per module by default, and joint methods also use the weights during mask selection (Section 3.3). Section 3.2 compares occurrence sensitivity with alternative statistics and evaluates normalization, clipping, and output-channel grouping.

3.2 Choosing and Validating Reconstruction Weights

We examine three design questions: which candidate statistic supplies useful reconstruction weights, how to transform raw occurrence sensitivities, and what finer output-channel grouping adds.

Table 1 evaluates candidate statistics on fixed Wanda pruning masks at 70% sparsity, with the same calibration data, OBS solver, and clipping parameters. The activation norm $\|x_t\|_2^2$ measures feature magnitude without global gradient information, whereas prediction loss provides one scalar per occurrence, shared across all modules. The shuffled control permutes each module's sensitivities within sequences, preserving the marginal distribution but disrupting occurrence alignment. Occurrence sensitivity achieves the lowest PPL on 10 of 11 models and improves over Uniform on 10 of 11, compared with 6 each for activation norm and next-token loss and 3 for the shuffled control. Because the shuffled control keeps the weight distribution, the benefit depends on occurrence alignment.

Transforming raw occurrence sensitivities. With the same Wanda masks, calibration data, and OBS reconstruction, we compare Raw, Mean-one, and Mean-one + clip (SWR; Appendix C.2). The number of models with lower PPL than Uniform rises from 6 of 11 for Raw to 8 for Mean-one and

Table 1: WikiText-2 test PPL for candidate statistics with one fixed Wanda pruning mask and OBS reconstruction.

Candidate statistic	L1-7B	L1-13B	L1-30B	L2-7B	L2-13B	L3-8B	L3.1-8B	Mis-7B	OPT-6.7B	Q2.5-7B	Q3-8B
Uniform	40.49	29.00	14.91	37.89	28.51	63.80	62.27	38.74	50.65	29.58	37.53
Activation norm	41.20	29.21	**14.77**	37.40	28.89	63.71	60.78	37.84	51.08	29.01	37.92
Next-token loss	40.06	29.37	15.22	38.12	28.57	63.58	62.25	39.22	50.17	29.27	37.10
Shuffled sensitivity	40.42	29.26	14.94	38.05	28.74	67.07	61.56	38.75	51.25	29.46	37.59
SWR	**38.33**	**28.38**	15.01	**37.28**	**28.29**	**63.01**	**60.69**	**37.66**	**49.25**	**28.61**	**36.44**

10 for SWR, and SWR improves over Raw in 9 of 11 models. For 11 models with Wanda, GBLM, and DSnoT masks, each of the 12 clipping intervals improves at least 30 of 33 configurations, and the default $[0.2, 5]$ improves 31 of 33 (Appendix C.2).

Output-channel grouping. We compare SWR with finer output-channel grouping following GuidedQuant (Kim et al., 2025). We evaluate one, two, and four groups on four models with OBS on fixed Wanda pruning masks and with SparseGPT. Each group normalizes and clips sensitivities and constructs a Gram matrix (Appendix D). Relative to one group, four groups lower WikiText-2 test PPL by at most 0.6% in seven of eight settings and by 5.6% for SparseGPT on Llama-3.1 8B, while requiring $4.02\times$ and $3.16\times$ the one-group reconstruction time for OBS with a Wanda mask and SparseGPT, respectively, on Llama-2 7B. We therefore retain one Gram matrix per module.

3.3 SWR in Existing Pruning Methods

Algorithm 1 in the appendix outlines the complete procedure. A single forward and backward sweep over the calibration set evaluates the dense model and caches occurrence sensitivities for all target modules. SWR replaces the Uniform Gram matrix with

$$\sum_t x_t x_t^\top \quad \longrightarrow \quad G_{\mathrm{SWR}} := \sum_t \rho_t x_t x_t^\top. \tag{5}$$

Wanda (Sun et al., 2024), GBLM (Das et al., 2023), and DSnoT (Zhang et al., 2024b) generate pruning masks without retained-weight reconstruction. We apply OBS reconstruction (Hassibi & Stork, 1992; Frantar & Alistarh, 2023) after mask generation while holding the pruning masks and all other settings fixed, so Uniform and SWR differ only in the Gram matrix. SparseGPT (Frantar & Alistarh, 2023) and ALPS (Meng et al., 2024) jointly select the pruning mask and reconstruct retained weights. We inject G_{SWR} into their existing update rules, so the reconstruction weights can affect both steps. Appendix E reports the sensitivity-computation cost and the conditions for reusing the cache across pruning configurations.

4 Experiments

4.1 Experimental setup

Models, Datasets, and Pruning Methods. We evaluate SWR on 11 models: LLaMA-1 7B/13B/30B (Touvron et al., 2023a), Llama-2 7B/13B (Touvron et al., 2023b), Llama-3 and Llama-3.1 8B (Grattafiori et al., 2024), Mistral 7B (Jiang et al., 2023), OPT 6.7B (Zhang et al., 2022), and Qwen2.5-7B and Qwen3-8B (Yang et al., 2024; 2025a). We evaluate language modeling using perplexity on WikiText-2 test (Merity et al., 2016), C4 validation (Raffel et al., 2020), and PTB test (Marcus et al., 1994). For zero-shot evaluation, we report accuracy on seven benchmarks: BoolQ (Clark et al., 2019), PIQA (Bisk et al., 2020), HellaSwag (Zellers et al., 2019), WinoGrande (Sakaguchi et al., 2021), ARC-Easy (Boratko et al., 2018), ARC-Challenge (Boratko et al., 2018), and OpenBookQA (Mihaylov et al., 2018). We prune each linear module using Wanda (Sun et al., 2024), GBLM (Das et al., 2023), DSnoT (Zhang et al., 2024b), SparseGPT (Frantar & Alistarh, 2023), or ALPS (Meng et al., 2024). The main operating point is 70% unstructured sparsity. Additional experiments cover 50–80% unstructured sparsity and 2:4 and 4:8 structured sparsity. Unless stated otherwise, calibration uses 128 sampled C4 train sequences of 2,048 tokens. Models use their native tokenizers. Each table defines its boldface comparison.

Table 2: WikiText-2 test PPL ↓ at 70% unstructured sparsity.

Method		L1-7B	L1-13B	L1-30B	L2-7B	L2-13B	L3-8B	L3.1-8B	Mis-7B	OPT-6.7B	Q2.5-7B	Q3-8B
Wanda	Mask only	83.64	54.95	17.37	74.96	45.76	122.85	124.37	61.40	165.70	73.94	62.00
	Uniform	40.49	29.00	**14.91**	37.89	28.51	63.80	62.27	38.74	50.65	29.58	37.53
	SWR	**38.33**	**28.38**	15.01	**37.28**	**28.29**	**63.01**	**60.69**	**37.66**	**49.25**	**28.61**	**36.44**
GBLM	Mask only	99.18	48.49	17.57	96.20	45.98	177.08	163.22	62.08	51.53	73.80	70.65
	Uniform	66.43	26.48	15.75	40.52	28.88	104.18	100.98	41.34	36.85	28.09	38.45
	SWR	**39.98**	**25.52**	**14.83**	**38.78**	**25.96**	**80.89**	**71.49**	**39.73**	**36.37**	**27.18**	**37.28**
DSnoT	Mask only	61.78	44.43	16.09	54.06	33.95	114.17	119.43	61.43	41.92	67.03	65.57
	Uniform	39.75	28.71	**14.91**	37.15	27.99	61.13	61.52	38.83	36.42	29.39	37.74
	SWR	**37.14**	**28.14**	14.98	**36.55**	**27.76**	**60.88**	**59.94**	**37.67**	**35.50**	**28.37**	**36.64**
SparseGPT	Native	26.84	19.91	13.16	28.35	20.11	44.88	43.24	**22.32**	21.12	20.94	**25.39**
	SWR	**26.26**	**19.65**	**13.06**	**26.16**	**19.09**	**39.86**	**39.23**	23.11	**20.08**	**19.46**	25.90
ALPS	Native	18.87	16.27	11.55	19.59	14.53	30.62	30.35	17.32	19.18	16.43	20.86
	SWR	**18.11**	**16.03**	**11.10**	**19.45**	**14.01**	**27.40**	**27.28**	**16.98**	**17.62**	**15.82**	**20.22**

Table 3: Seven-task mean zero-shot accuracy ↑ (%) at 70% unstructured sparsity.

Method		L1-7B	L1-13B	L1-30B	L2-7B	L2-13B	L3-8B	L3.1-8B	Mis-7B	OPT-6.7B	Q2.5-7B	Q3-8B
Wanda	Uniform	38.57	40.99	52.00	37.59	40.95	36.62	37.35	38.84	40.90	44.01	43.51
	SWR	**38.82**	**41.34**	**52.09**	**37.78**	**41.07**	**37.21**	**37.86**	**39.03**	**41.28**	**44.30**	**43.64**
GBLM	Uniform	37.89	43.31	50.23	38.49	42.65	36.40	36.92	40.67	43.71	45.66	44.50
	SWR	**40.39**	**43.62**	**50.96**	**38.60**	**43.68**	**38.18**	**38.58**	**40.84**	**44.02**	**46.09**	**44.53**
DSnoT	Uniform	38.40	41.09	52.02	37.54	41.19	37.01	37.21	38.84	43.19	44.06	**43.43**
	SWR	**38.75**	**41.66**	**52.24**	**37.79**	**41.32**	**37.36**	**37.61**	38.84	**43.42**	**44.60**	43.41
SparseGPT	Native	42.44	45.40	52.21	41.93	47.67	43.30	43.07	**45.70**	46.87	50.73	**50.68**
	SWR	**43.47**	**46.18**	**52.34**	**42.67**	**48.41**	**43.64**	**44.11**	45.31	**47.69**	**51.79**	50.52
ALPS	Native	45.89	49.19	55.44	45.48	51.74	46.55	46.09	47.49	48.89	51.65	52.58
	SWR	**46.39**	**49.65**	**55.45**	**46.06**	**52.33**	**47.30**	**46.80**	**47.71**	**49.55**	**53.42**	**53.22**

4.2 Main results

Language modeling. At 70% unstructured sparsity, SWR lowers WikiText-2 test perplexity in 51 of 55 configurations, with a 3.0% median relative reduction (Table 2); the exceptions are Wanda and DSnoT on LLaMA-1 30B and SparseGPT on Mistral 7B and Qwen3 8B. On C4 validation and PTB test, SWR lowers PPL in 46 and 40 of 55 configurations (Appendix F). Per-method median reductions are 2.5% for Wanda, 4.3% for GBLM, and 2.5% for DSnoT with fixed pruning masks, and 4.9% for SparseGPT (9 of 11 models) and 3.7% for ALPS (all 11) with joint mask selection. On Llama-3 8B, PPL decreases from 30.62 to 27.40 for ALPS and from 44.88 to 39.86 for SparseGPT. An extension to Qwen2.5-32B improves all five evaluated pruning methods by 0.4–2.9% (Appendix F.3).

Zero-shot tasks. SWR also improves seven-task mean accuracy in 51 of 55 configurations (Table 3), with a median change of +0.38 percentage points and a maximum increase of 2.50 percentage points for GBLM on LLaMA-1 7B. The absolute median change is modest; the count summarizes the direction of the three-draw means. Appendix G reports per-task results.

Calibration-source transfer. We calibrate on the training splits of C4, PTB, or WikiText-2 and evaluate on all three evaluation corpora. Table 4 shows Qwen2.5-7B with Wanda at 70% sparsity; Appendix J reports the full grid over Qwen2.5-7B and Llama-2 7B with Wanda and GBLM, where SWR lowers PPL in all 12 same-corpus and 20 of 24 cross-corpus configurations, with median reductions of 2.01% and 1.30%. The four regressions occur under cross-corpus evaluation: three on PTB; the other is 0.02% on C4.

Table 4: Calibration-source transfer on Qwen2.5-7B with Wanda pruning masks. Uniform → SWR PPL.

Calibration source	WikiText-2 test	C4 validation	PTB test
C4 train	29.58 → **28.61**	40.34 → **39.89**	43.21 → **42.40**
PTB train	29.44 → **29.15**	45.03 → **44.84**	27.10 → **26.53**
WikiText-2 train	23.40 → **22.90**	44.70 → **44.69**	42.93 → **42.67**

Sparsity levels and structured sparsity. From 50% to 70% unstructured sparsity, the median relative PPL improvement increases for each of the four evaluated methods (Table 5; Appendix H). For Wanda and DSnoT, the median improvement is close to zero at 50% and reaches approximately 2.5% at 70%. At 80%, the pattern is method-dependent: SparseGPT improves all 11 models, whereas

Table 5: WikiText-2 test PPL across sparsity levels. Each cell gives the number of models (of 11) on which SWR lowers PPL and the median relative reduction against Uniform (Native for SparseGPT).

Method	$S = 0.5$	$S = 0.6$	$S = 0.7$	$S = 0.8$
Wanda	6/11 (0.10%)	9/11 (0.48%)	10/11 (2.54%)	7/11 (1.05%)
GBLM	9/11 (0.43%)	11/11 (1.18%)	11/11 (4.29%)	9/11 (6.56%)
DSnoT	7/11 (0.14%)	9/11 (0.37%)	10/11 (2.53%)	6/11 (0.15%)
SparseGPT	10/11 (1.33%)	10/11 (2.05%)	9/11 (4.92%)	11/11 (4.26%)

Table 6: WikiText-2 test PPL ↓ under 2:4 structured sparsity.

Method		L1-7B	L1-13B	L1-30B	L2-7B	L2-13B	L3-8B	L3.1-8B	Mis-7B	OPT-6.7B	Q2.5-7B	Q3-8B
Wanda	Uniform	11.22	9.67	7.18	11.81	9.10	**18.62**	17.78	**11.31**	15.16	12.09	15.47
	SWR	**11.04**	**9.60**	**7.16**	**11.74**	**9.05**	18.67	**17.72**	11.34	**15.08**	**12.05**	**15.32**
SparseGPT	Native	10.95	9.07	7.15	10.95	8.75	**15.98**	15.98	10.53	**14.07**	11.19	14.48
	SWR	**10.55**	**8.91**	**7.12**	**10.88**	**8.62**	16.19	**15.92**	**10.42**	14.08	**10.96**	**14.30**

Wanda, GBLM, and DSnoT lower PPL on 7, 9, and 6 models, respectively. Exact 2:4 sparsity imposes strict positional constraints on retained weights within every output row. Table 6 evaluates all 11 models using Wanda and SparseGPT. SWR reduces perplexity in 9 of 11 models under both solvers, by at most 1.6% (Wanda) and 3.7% (SparseGPT). Appendix I reports 4:8 structured sparsity results.

Additional analyses. SWR yields lower PPL in 18 of 22 comparisons under a fixed OWL layer sparsity profile (Yin et al., 2023) (WikiText-2; Appendix L). On Llama-2 7B, SWR increases the post-setup runtime of Wanda with OBS by 17.7% relative to Uniform and approximately doubles peak allocated GPU memory; Appendix E details the cost and the conditions for reusing cached sensitivities. After matched LoRA recovery, SWR yields lower WikiText-2 PPL in 17 of 20 comparisons across four pruning methods and five models, with a 0.37% median relative advantage (Appendix K).

5 Why Does Weighting Help?

We examine three questions: whether sensitivity helps only through occurrence selection, whether its values improve reconstruction after selection, and whether more uniformly weighted calibration data closes the gap. The diagnostics use fixed Wanda pruning masks at 70% sparsity, OBS reconstruction on C4 train text, and WikiText-2 test PPL, with settings identical to those in Section 3.3.

Module-specific weights add information beyond a shared ranking. We average sensitivities normalized within each sequence across reconstructed modules to form one global ranking and select nested top fractions $q \in \{0.1, 0.2, \ldots, 1.0\}$. Uniform and SWR use identical occurrences at each fraction. SWR weights use the complete calibration pool; excluded values are zeroed without renormalizing or reclipping retained weights (Appendix M). The following statistics cover the nine proper subsets with $q < 1$. Across 11 models and these nine fractions, SWR lowers PPL against subset unit weighting in 92 of 99 configurations (median reduction, 1.76%). Against full Uniform, selection alone improves 71 of 99 configurations (median reduction 0.28%), versus 82 of 99 (2.35%) with SWR. Selection peaks at $q = 20\%$ (0.51%; Figure 2(a)), while retaining SWR weights yields larger, more stable gains as subsets grow (Figure 2(b)). Thus, the ranking identifies useful occurrences, while module-specific reconstruction weights add information beyond it.

More calibration data does not remove the gap. We vary the nested calibration budget n_{cal} from 16 to 1,024 sequences with fixed Wanda masks, using identical nested subsets for both configurations. Figure 2(c) plots four models. Appendix O reports PPL results for all 11 models and wall time and peak GPU memory for Llama-2 7B. At matched budgets, SWR has lower PPL than Uniform on 3 of 11 models at $n_{\text{cal}} = 16$, 9 at $n_{\text{cal}} = 32$, 10 at $n_{\text{cal}} = 64$ and 128, and all 11 from $n_{\text{cal}} = 256$ onward. SWR at $n_{\text{cal}} = 256$ also outperforms Uniform at $n_{\text{cal}} = 512$ on all 11 models and at $n_{\text{cal}} = 1{,}024$ on 9 of 11. Thus, SWR can outperform Uniform with fewer sequences. On Llama-2 7B, SWR at $n_{\text{cal}} = 256$ lowers PPL from 37.38 to 36.71 (1.8%) versus Uniform at $n_{\text{cal}} = 512$, using 17.5% less post-setup time but 23.4% more peak GPU memory. Against Uniform at $n_{\text{cal}} = 1{,}024$ (37.35 PPL), SWR remains lower, using 51.1% less time and 19.3% less peak GPU memory. Within

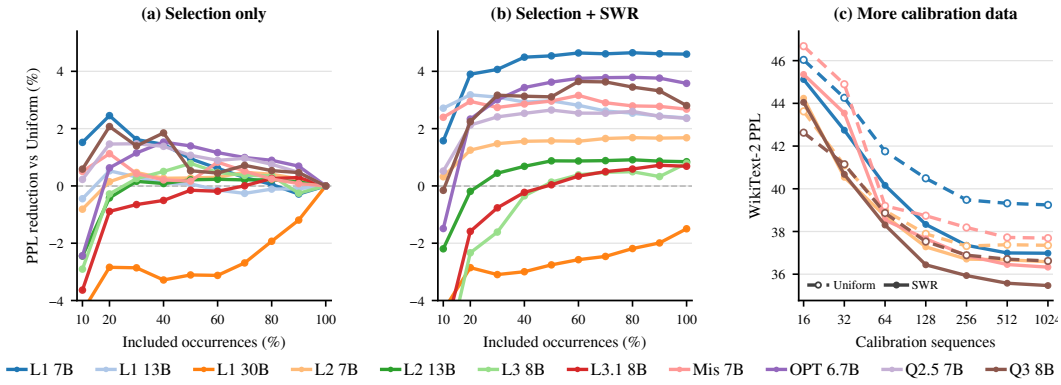


Figure 2: **Module-specific reconstruction weights add information beyond a shared occurrence ranking and additional calibration data.** (a,b) WikiText-2 test PPL reduction relative to Uniform reconstruction on all calibration occurrences. For each retained fraction, both panels use the same sensitivity-ranked subset: (a) assigns unit weights, whereas (b) retains the SWR weights. Each curve denotes one model. (c) WikiText-2 test PPL as the number of calibration sequences increases, using fixed Wanda pruning masks. Colors identify models using the shared legend.

this range, increasing the calibration set improves both configurations but does not remove SWR's matched-budget advantage under the same relative damping rule (Section 3.3).

Region-wise tests of priority dilution. Table 7 compares inclusion and reweighting of three regions under the same sensitivity ranking. Reweighting replaces one region's unit weights with SWR weights; occurrence inclusion, the mask, and all other weights remain fixed.

Reweighting the Top 20% (Rows 1–2) improves 10 of 11 models (median relative PPL reduction, 1.03%). Adding the Bottom 50% at unit weight to an SWR-weighted top half (Rows 4–5) degrades 8 of 11 models (median PPL increase, 0.97%). Reweighting this added half (Rows 5–6) improves 10 of 11 models (median reduction, 1.06%). Compared with excluding the Bottom 50% (Row 4), retaining it with SWR weights (Row 6) improves only 5 of 11 models and has a median relative PPL increase of 0.03%. These results support *priority dilution*: adding low-ranked occurrences at unit weight can reduce the reconstruction priority of more sensitive occurrences. In this experiment, reweighting mitigates the degradation, but does not reliably improve over excluding the same half. Appendix M.2 reports the full results.

Table 7: **Region-wise tests of priority dilution.** Sensitivity-ranked regions receive weights 0 (excluded), 1 (unit), or SWR weights ρ. The Gain column reports the median relative PPL reduction from the immediately preceding row; positive values indicate performance improvement. The Improved column reports the number of evaluated models with lower PPL values.

	Region weights			Versus previous row	
Row	Top 20%	Next 30%	Bottom 50%	Gain (%)	Improved
1	1	0	0	–	–
2	ρ	0	0	+1.03	10/11
3	ρ	1	0	+0.13	6/11
4	ρ	ρ	0	+0.60	10/11
5	ρ	ρ	1	−0.97	3/11
6	ρ	ρ	ρ	+1.06	10/11

6 Module-Specific Lexical Structure

SWR computes separate reconstruction weights for each module. We ask which tokens are associated with high-sensitivity occurrences and whether sharing one array per layer is sufficient (Figure 3).

We first ask which token categories occur immediately after high-sensitivity occurrences. In a causal language model, a module output at position t contributes to predicting the token at $t+1$, so we label occurrence t by the following token's category. If the occurrence at *is* in *The capital is Paris* belongs to the most sensitive 20%, we count it under proper nouns because the following token is *Paris*. Sensitivity is measured at *is*, not at *Paris*.

For each model, we normalize occurrence sensitivities per sequence, average them across modules, and select the most sensitive 20%. For each token category, Figure 3(a) reports its frequency among

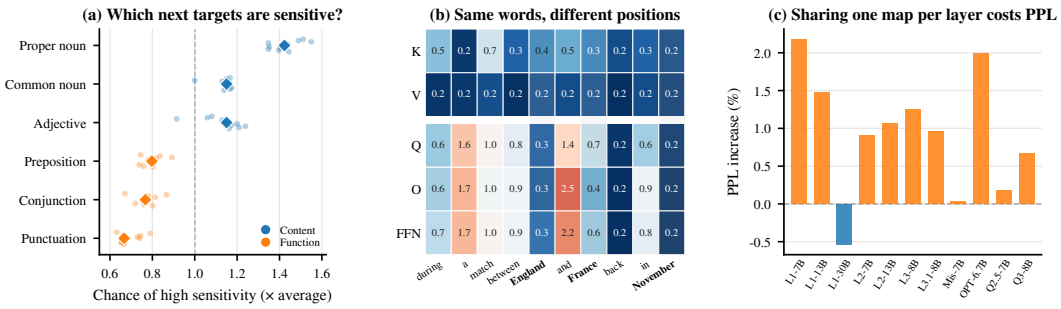


Figure 3: **Occurrence sensitivity varies with token category and module type.** (a) Frequency of each category among tokens immediately following the most sensitive 20% of occurrences, divided by its position-matched expectation. Circles are models and diamonds are medians. (b) Reconstruction weights in one selected Llama-2 7B C4 span, averaged across layers. Q/O/FFN weights are higher immediately before the proper nouns, while K/V change little; aggregate results are in Appendix P.2. (c) Relative PPL change after replacing module-specific weights with one shared array per layer.

the following tokens divided by its position-matched expectation. A ratio of one means no enrichment; values above or below one mean over- or underrepresentation. Across models, the median ratio is $1.42\times$ for proper nouns, $1.15\times$ for common nouns and adjectives, at most $0.80\times$ for prepositions, conjunctions, and punctuation, and about $1.01\times$ for verbs and adverbs.

Proper nouns show the largest enrichment, so we compare reconstruction weights across module types. For a noun at position t, the module output at $t-1$ predicts it, while at t the noun is being processed. Across 202,960 proper-noun occurrences, we compare both positions, averaging within models and then equally across models. Weights are higher at $t-1$ for Q (1.146 versus 0.788), O (1.193 versus 0.901), and FFN (1.127 versus 0.888). K is nearly unchanged (0.600 versus 0.604), while V is slightly lower at the noun (0.302 versus 0.287). Thus, Q/O/FFN place more weight before a proper noun, whereas K/V show no aggregate peak at the noun. Figure 3(b) provides a selected example; Appendices P and P.2 report the protocol and aggregate results.

Beyond proper nouns, we compare the token being processed with the token being predicted next across occurrences. For each module type, a cross-fitted logistic regression predicts whether held-out occurrences belong to its most sensitive 20%. Removing either token identity measures the decrease in AUC after controlling for position and both tokens' frequencies and byte lengths (Appendix P.3). The token being processed provides more predictive information for K/V in all 11 models; the token being predicted next provides more for Q/O in 8 of 11 and FFN in 10 of 11. Prior work links FFNs to next-token prediction (Geva et al., 2021; 2022), without comparing these associations across module types. We compare five module types and 11 models. These associations are descriptive. A reconstruction ablation tests module-specific reconstruction weights. With fixed Wanda pruning masks at 70% sparsity, replacing module-specific weights with their within-layer average increases PPL in 10 of 11 models by a median of 1.0%; LLaMA-1 30B is the exception (Figure 3(c); Appendix P.4).

7 Conclusion

We study how one-shot pruning distributes reconstruction priority across calibration occurrences. Uniform reconstruction treats all occurrences equally, although their local errors can affect the global loss differently. This can cause *priority dilution*: low-sensitivity occurrences can reduce the priority assigned to sensitive ones. Sensitivity-Weighted Reconstruction (SWR) mitigates this problem by using reconstruction weights to form the Gram matrix of existing solvers without changing their update rules. Controlled interventions separate occurrence inclusion from weighting, and evaluations across models and pruning methods show that this substitution generally improves sparse-model quality. Within the measured range, more calibration data does not eliminate the gap between Uniform and SWR. We also find module-specific lexical structure in occurrence sensitivity, while sharing reconstruction weights across modules in a layer usually increases PPL. Pruning masks therefore determine which weights remain, whereas reconstruction weights determine which reconstruction errors receive priority. Future one-shot pruning methods can optimize these two decisions jointly.

REFERENCES

Riade Benbaki, Wenyu Chen, Xiang Meng, Hussein Hazimeh, Natalia Ponomareva, Zhe Zhao, and Rahul Mazumder. Fast as chita: Neural network pruning with combinatorial optimization. In *International Conference on Machine Learning*, pp. 2031–2049. PMLR, 2023.

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.

Michael Boratko, Harshit Padigela, Divyendra Mikkilineni, Pritish Yuvraj, Rajarshi Das, Andrew McCallum, Maria Chang, Achille Fokoue-Nkoutche, Pavan Kapanipathi, Nicholas Mattei, et al. A systematic classification of knowledge, reasoning, and context within the arc dataset. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pp. 60–70, 2018.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, 2019.

Rocktim Jyoti Das, Mingjie Sun, Liqun Ma, and Zhiqiang Shen. Beyond size: How gradients shape pruning decisions in large language models. *arXiv preprint arXiv:2311.04902*, 2023.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

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

Elias Frantar and Dan Alistarh. Optimal brain compression: A framework for accurate post-training quantization and pruning. *Advances in Neural Information Processing Systems*, 35:4475–4488, 2022.

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.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 5484–5495, 2021.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 30–45, 2022.

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.

Shuhao Gu, Jinchao Zhang, Fandong Meng, Yang Feng, Wanying Xie, Jie Zhou, and Dong Yu. Token-level adaptive training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1035–1046, 2020.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Self-attention attribution: Interpreting information interactions inside transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 12963–12971, 2021.

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

Yen-Chang Hsu, Ting Hua, Sungen Chang, Qian Lou, Yilin Shen, and Hongxia Jin. Language model compression with weighted low-rank factorization. *arXiv preprint arXiv:2207.00112*, 2022.

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 *The Tenth International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Weizhong Huang, Yuxin Zhang, Xiawu Zheng, Fei Chao, and Rongrong Ji. Determining layer-wise sparsity for large language models through a theoretical perspective. *arXiv preprint arXiv:2502.14770*, 2025.

Yixin Ji, Yang Xiang, Juntao Li, Qingrong Xia, Ping Li, Xinyu Duan, Zhefeng Wang, and Min Zhang. Beware of calibration data for pruning large language models. In *International Conference on Learning Representations*, volume 2025, pp. 18604–18620, 2025.

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

Jinuk Kim, Marwa El Halabi, Wonpyo Park, Clemens JS Schaefer, Deokjae Lee, Yeonhong Park, Jae W. Lee, and Hyun Oh Song. GuidedQuant: Large language model quantization via exploiting end loss guidance. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30011–30037, 2025. URL https://proceedings.mlr.press/v267/kim25d.html.

Frederik Kunstner, Lukas Balles, and Philipp Hennig. Limitations of the empirical fisher approximation for natural gradient descent. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yuhang Li, Ruihao Gong, Xu Tan, Yang Yang, Peng Hu, Qi Zhang, Fengwei Yu, Wei Wang, and Shi Gu. BRECQ: Pushing the limit of post-training quantization by block reconstruction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=POWv6hDd9XH.

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, 2024a.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision*, 2017.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, et al. Rho-1: Not all tokens are what you need. *arXiv preprint arXiv:2404.07965*, 2024b.

Zichang Liu, Aditya Desai, Fangshuo Liao, Weitao Wang, Victor Xie, Zhaozhuo Xu, Anastasios Kyrillidis, and Anshumali Shrivastava. Scissorhands: Exploiting the persistence of importance hypothesis for llm kv cache compression at test time. *Advances in Neural Information Processing Systems*, 36:52342–52364, 2023.

Haiquan Lu, Yefan Zhou, Shiwei Liu, Zhangyang Wang, Michael W Mahoney, and Yaoqing Yang. Alphapruning: Using heavy-tailed self regularization theory for improved layer-wise pruning of large language models. *Advances in neural information processing systems*, 37:9117–9152, 2024.

Mitch Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. The penn treebank: Annotating predicate argument structure. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994.

Xiang Meng, Kayhan Behdin, Haoyue Wang, and Rahul Mazumder. Alps: Improved optimization for highly sparse one-shot pruning for large language models. *Advances in neural information processing systems*, 37:37594–37625, 2024.

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

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2381–2391, 2018.

Ali Modarressi, Mohsen Fayyaz, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. Globenc: Quantifying global token attribution by incorporating the whole encoder layer in transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 258–271, 2022.

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.

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.

Sungbin Shin, Wonpyo Park, Jaeho Lee, and Namhoon Lee. Rethinking pruning large language models: Benefits and pitfalls of reconstruction error minimization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1182–1191, 2024.

Sidak Pal Singh and Dan Alistarh. WoodFisher: Efficient second-order approximation for neural network compression. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023a. URL https://arxiv.org/abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Tycho F. A. van der Ouderaa, Markus Nagel, Mart van Baalen, and Tijmen Blankevoort. The LLM surgeon. In *International Conference on Learning Representations*, 2024.

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

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*, 2025a.

Yifan Yang, Kai Zhen, Bhavana Ganesh, Aram Galstyan, Goeric Huybrechts, Markus Müller, Jonas M Kübler, Rupak Vignesh Swaminathan, Athanasios Mouchtaris, Sravan Babu Bodapati, et al. Wanda++: Pruning large language models via regional gradients. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4321–4333, 2025b.

Lu Yin, You Wu, Zhenyu Zhang, Cheng-Yu Hsieh, Yaqing Wang, Yiling Jia, Gen Li, Ajay Jaiswal, Mykola Pechenizkiy, Yi Liang, et al. Outlier weighed layerwise sparsity (OWL): A missing secret sauce for pruning LLMs to high sparsity. *arXiv preprint arXiv:2310.05175*, 2023.

Xin Yu, Thiago Serra, Srikumar Ramalingam, and Shandian Zhe. The combinatorial brain surgeon: Pruning weights that cancel one another in neural networks. In *International Conference on Machine Learning*, pp. 25668–25683. PMLR, 2022.

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.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Yingtao Zhang, Haoli Bai, Haokun Lin, Jialin Zhao, Lu Hou, and Carlo Vittorio Cannistraci. Plug-and-play: An efficient post-training pruning method for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 50490–50508, 2024a.

Yuxin Zhang, Lirui Zhao, Mingbao Lin, Sun Yunyun, Yiwu Yao, Xingjia Han, Jared Tanner, Shiwei Liu, and Rongrong Ji. Dynamic sparse no training: Training-free fine-tuning for sparse LLMs. In *International Conference on Learning Representations*, volume 2024, pp. 249–264, 2024b.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, et al. H2O: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in neural information processing systems*, 36:34661–34710, 2023.

APPENDIX MAP

Table 8 lists the additional results and implementation details for each part of the main paper.

Table 8: Additional material for each part of the main paper.

Main paper	Additional material
Section 1, Figure 1	Appendix B: sensitivity distribution, module overlap, and variation across depth.
Section 3	Appendices C and C.4: reconstruction weight construction, numerical details, and comparisons with candidate statistics.
Efficiency measurements	Appendix E: wall time, peak memory, scaling, and reuse of occurrence sensitivity.
Section 4.2, Table 2	Appendix F: C4 and PTB PPL, three reconstruction configurations, and solver details.
Section 4.2, Table 3	Appendix G: results for all seven zero-shot tasks.
Section 4.2, Table 5	Appendix H: complete results at 50–80% unstructured sparsity.
Section 4.2, Table 6	Appendix I: complete 4:8 results for four pruning methods.
Section 4.2, Table 4	Appendix J: complete source and evaluation dataset results.
Section 4.2	Appendix K: LoRA recovery results and training settings.
OWL layer allocation	Appendix L: results on three PPL corpora.
Section 5, Figure 2, and Table 7	Appendices M and N: occurrence selection and residual allocation results.
Section 5, Figure 2	Appendix O: calibration scaling for all 11 models.
Section 6, Figure 3	Appendix P: lexical categories, position analysis, layer-shared reconstruction weights, and the static lexical prior.

A ALGORITHM

Algorithm 1 summarizes the complete SWR procedure. SWR computes occurrence sensitivity once, constructs reconstruction weights for each module, and supplies the weighted Gram matrix to the pruning solver.

Here $\mathcal{P}.\mathrm{Mask}$ denotes the method's native pruning mask routine and receives its required calibration statistics.

B ADDITIONAL RESULTS AND DETAILS FOR FIGURE 1

Figure 1 measures the distribution of occurrence sensitivity, its overlap between module types, and its variation across depth. Across all 11 models, the most sensitive 1% and 10% of occurrences account for 25.1% and 63.1% of the total sensitivity. The overlap between the most sensitive occurrence sets of two module types ranges from 30.0 to 94.3%. Sensitivity concentration also varies with depth, with a median range of 26.2 percentage points within each model. These results motivate module-specific reconstruction weights for each occurrence.

B.1 EXPERIMENTAL SETUP AND UNIT OF ANALYSIS

We evaluate the same 11 model checkpoints used in the main experiments, spanning LLaMA, Qwen, Mistral, and OPT. Each calibration draw supplies $n = 128$ unpadded sequences of length $L = 2{,}048$ from C4 train. For sequence s, position ℓ, and linear module m, a dense forward and backward pass

Algorithm 1: SWR

Input: dense model; calibration set $\mathcal{C}$; pruning method $\mathcal{P}$; target sparsity s
Output: pruned model

Compute occurrence sensitivities for all modules on the dense model and cache them; this requires one forward and backward pass per calibration sequence;
foreach *module with dense weights W, in $\mathcal{P}$'s order* **do**
  $x_1, \ldots, x_T \leftarrow$ current inputs of the module;
  $\rho_1, \ldots, \rho_T \leftarrow$ construct the module's reconstruction weights from its cached occurrence sensitivities using Equation 4;
  $G_{\mathrm{SWR}} \leftarrow \sum_{t=1}^{T} \rho_t \, x_t x_t^\top$;
  if *$\mathcal{P}$ jointly selects and reconstructs* **then**
    $\widehat{W} \leftarrow \mathcal{P}.\mathrm{Joint}(W, \{x_t\}, G_{\mathrm{SWR}}, s)$;
  else
    $M \leftarrow \mathcal{P}.\mathrm{Mask}(W, \mathcal{C}, s)$;
    $\widehat{W} \leftarrow \text{OBS-Reconstruct}(W, M, G_{\mathrm{SWR}})$;
  replace W by $\widehat{W}$ in the model;
return the pruned model;

records the local squared gradient norm

$$a_{m,s,\ell} = \left\| \frac{\partial \mathcal{L}_s}{\partial y_{m,s,\ell}} \right\|_2^2, \quad s = 1, \ldots, n, \quad \ell = 1, \ldots, L, \tag{6}$$

where $y_{m,s,\ell}$ is the module output at occurrence (s, ℓ) and $\mathcal{L}_s$ is the mean next-token loss. The squared norm spans all output channels.

Each module yields a matrix of occurrence sensitivities with shape $128 \times 2{,}048$. Across the evaluated models, this diagnostic covers 396 Transformer layers and 2,740 linear modules. Figure 1 uses these nonnegative, unnormalized values without SWR clipping. Panel (a) normalizes them only to compare scales. The other panels use the unnormalized values.

Because models use native tokenizers, we do not enforce token identity across models. We compute positional comparisons within the same model, sequence, and layer. Every model contributes equally to the aggregate. Panel (b) uses the ten models that share the same seven module types. We exclude OPT because it lacks a gated FFN.

B.2 Sensitivity distribution and concentration

For Figure 1(a) we divide each sequence and module vector by its mean before binning all 2,048 values. We group exact zeros and values below 10^{-4} into the labelled leftmost bar. Remaining values occupy bins of equal width on a logarithmic scale from 10^{-4} to the upper bound of 2,048. Within each model we compute the arithmetic frequency across all vectors to construct a normalized histogram. We then average the 11 model histograms arithmetically, so every model contributes equally to the display.

To quantify this concentration, we compute the cumulative fraction of occurrence sensitivity within each sequence and module. We then take medians across sequences, modules, and models. The most sensitive 1, 5, 10, and 20% of occurrences account for 25.1%, 48.9%, 63.1%, and 78.3% of the total sensitivity, respectively. Table 9 reports the corresponding result for each model.

B.3 Variation across modules

For each model, layer, sequence, and module, we select the most sensitive 10% of occurrences. We store these 205 positions in a set $\mathcal{I}_{m,s}$. For two distinct modules m and m', the overlap is

$$o(\mathcal{I}_{m,s}, \mathcal{I}_{m',s}) = \frac{|\mathcal{I}_{m,s} \cap \mathcal{I}_{m',s}|}{205}. \tag{7}$$

We divide by 205 because all sets have the same cardinality. A value of 100% indicates identical selections, whereas 0% indicates disjoint sets. At cutoff ties, the smaller position index comes first. We compare sets from the same model, layer, and calibration sequence.

Table 9: Sensitivity concentration for each model in Figure 1(a). The mass rows report the fraction of occurrence sensitivity accounted for by the most sensitive occurrences.

Statistic	L1-7B	L1-13B	L1-30B	L2-7B	L2-13B	L3-8B	L3.1-8B	Mis-7B	OPT-6.7B	Q2.5-7B	Q3-8B
Module cells	224	280	420	224	280	224	224	224	192	196	252
Top 1% mass	24.1%	25.1%	25.5%	24.9%	25.7%	27.8%	26.1%	24.5%	23.7%	23.7%	30.6%
Top 5% mass	47.3%	48.9%	50.1%	48.9%	51.1%	50.9%	50.2%	48.3%	43.4%	46.1%	53.5%
Top 10% mass	61.3%	63.1%	64.8%	63.1%	64.9%	64.0%	63.3%	62.4%	55.5%	59.0%	65.5%
Top 20% mass	77.3%	78.7%	80.3%	78.3%	79.1%	78.9%	78.9%	77.4%	70.9%	73.7%	78.2%

Table 10: Sensitivity concentration across layers for each model in Figure 1(c). Values are the sensitivity mass of the most sensitive 10% at the first, minimum, maximum, and last layers.

Statistic	L1-7B	L1-13B	L1-30B	L2-7B	L2-13B	L3-8B	L3.1-8B	Mis-7B	OPT-6.7B	Q2.5-7B	Q3-8B
Early layer	63.6%	65.9%	67.8%	67.0%	69.2%	77.7%	76.9%	71.7%	61.6%	59.7%	72.1%
Minimum	45.6%	46.5%	48.3%	45.3%	45.5%	45.4%	44.7%	44.6%	47.0%	41.9%	44.2%
Maximum	67.2%	69.4%	70.3%	71.5%	71.4%	79.0%	77.3%	72.8%	63.0%	71.5%	77.1%
Late layer	49.0%	47.1%	49.5%	49.8%	47.7%	47.3%	46.3%	47.7%	54.4%	44.0%	49.3%

For each combination of two module types, we first take the median over 128 sequences within a layer. We then take the median across layers within a model and the arithmetic mean over the ten compatible models. Figure 1(b) shows the resulting 7×7 matrix. Its diagonal is 100% by definition. Among the 21 off-diagonal module-type combinations, overlap ranges from 30.0% for K and Up to 94.3% for O and Down. The median is 58.2%. Some module types therefore select similar sensitive occurrences, while others differ. A single ranking shared by all modules would remove these differences.

B.4 Variation across depth

Table 10 reports the first, minimum, maximum, and last sensitivity concentration values for each model. At each layer, we select the 205 largest unnormalized sensitivities for each sequence and module. We divide their sum by the total over all 2,048 occurrences. We then compute the median over 128 sequences and across all recorded modules in that layer. The layer index is divided by $L - 1$ so that models of different depths share the same fractional axis. Across the 11 models, the range within each model is 16.0–33.6 percentage points, with a median of 26.2 points. Several models have high concentration in early layers, lower concentration in middle layers, and higher concentration near the final layers. The magnitude and location of these changes differ across models.

C Reconstruction Weight Construction and Numerical Safeguards

This section gives the theoretical and practical construction of SWR. We first derive occurrence sensitivity. We then normalize the reconstruction weights to mean one within each sequence and clip them. Finally, we show how the weights enter standard pruning methods and compare occurrence sensitivity with candidate statistics.

C.1 Theoretical derivation of occurrence sensitivity

Consider one module and index its T calibration occurrences by t across all sequences and positions. Let $y_t = Wx_t$ denote the dense module output. The matrix perturbation is $\Delta W = \widehat{W} - W$, and the reconstruction residual is $e_t = \Delta W x_t$. We define $g_t = \nabla_{y_t}\mathcal{L}$ as the gradient of the global loss $\mathcal{L}$ with respect to this output. All gradients are evaluated at the dense model. The resulting sensitivities are held fixed during sequential pruning. Later module inputs may differ after earlier modules are pruned, so this is a dense-model approximation. A second order Taylor expansion with respect to the

stacked outputs gives

$$\Delta\mathcal{L} = \sum_{t=1}^{T} g_t^\top e_t + \frac{1}{2}\,\mathrm{vec}(E)^\top \mathcal{H}\,\mathrm{vec}(E) + O(\|E\|_F^3), \tag{8}$$

where E stacks all reconstruction residuals. The matrix $\mathcal{H}$ is the exact Hessian of $\mathcal{L}$ with respect to the stacked module outputs. Let $\mathcal{H}_{ts}$ denote the block that couples occurrences t and s. The quadratic term separates into self terms and cross terms:

$$\mathrm{vec}(E)^\top \mathcal{H}\,\mathrm{vec}(E) = \sum_{t=1}^{T} e_t^\top \mathcal{H}_{tt}\, e_t + \sum_{t=1}^{T}\sum_{\substack{s=1 \\ s\neq t}}^{T} e_t^\top \mathcal{H}_{ts}\, e_s. \tag{9}$$

Because $e_t = \Delta W x_t$ we have

$$\sum_{t=1}^{T} g_t^\top e_t = \sum_{t=1}^{T} g_t^\top \Delta W x_t = \mathrm{Tr}\Big(\sum_{t=1}^{T} x_t g_t^\top \Delta W\Big) = \mathrm{Tr}\left((\nabla_W \mathcal{L})^\top \Delta W\right), \tag{10}$$

using $\nabla_W \mathcal{L} = \sum_{t=1}^{T} g_t x_t^\top$. Thus, the first order contributions from all occurrences sum to one inner product between the calibration gradient and the matrix perturbation. We assume that the dense model is approximately stationary on the calibration distribution, so $\nabla_W \mathcal{L} \approx 0$. Under this assumption, the aggregate first order term vanishes. For a finite C4 calibration draw, this is an approximation; the derivation therefore motivates a surrogate rather than an identity.

Directly modeling the full Hessian $\mathcal{H}$ is intractable. The cross terms in Equation 9 contain dense interactions across occurrences. We therefore omit them and retain only the diagonal blocks $\mathcal{H}_{tt} = \nabla^2_{y_t}\mathcal{L}$. This gives

$$\Delta\mathcal{L} \approx \frac{1}{2}\sum_{t=1}^{T} e_t^\top \mathcal{H}_{tt}\, e_t. \tag{11}$$

We further approximate each block by its empirical Fisher surrogate (Kunstner et al., 2019),

$$\mathcal{H}_{tt} \approx g_t g_t^\top, \tag{12}$$

which gives

$$\Delta\mathcal{L} \approx \frac{1}{2}\sum_{t=1}^{T} \left(g_t^\top e_t\right)^2. \tag{13}$$

Equation 13 retains an anisotropic quadratic form aligned with the loss gradient. It is an empirical Fisher surrogate, not the exact Hessian curvature. The form couples output dimensions and is incompatible with the shared Gram matrix and independent row updates used by standard pruning methods. We therefore apply the spectral bound

$$g_t g_t^\top \preceq \|g_t\|_2^2 I \tag{14}$$

which implies (equivalently, by Cauchy–Schwarz)

$$\left(g_t^\top e_t\right)^2 \leq \|g_t\|_2^2 \|e_t\|_2^2. \tag{15}$$

Therefore we obtain

$$\frac{1}{2}\sum_{t=1}^{T} \left(g_t^\top e_t\right)^2 \leq \frac{1}{2}\sum_{t=1}^{T} a_t \|e_t\|_2^2 \qquad a_t := \|g_t\|_2^2. \tag{16}$$

The scalar a_t follows directly from the spectral bound. If a scalar b_t must satisfy

$$e_t^\top g_t g_t^\top e_t \leq b_t \|e_t\|_2^2 \tag{17}$$

for every possible reconstruction residual e_t, then

$$b_t \geq \lambda_{\max}(g_t g_t^\top) = \|g_t\|_2^2. \tag{18}$$

Table 11: **Clipping range ablation with fixed pruning masks.** WikiText-2 test PPL ↓ for 11 models.

(a) Wanda

Model	Uniform	$l=0$			$l=0.01$			$l=0.1$			$l=0.2$		
		$u=5$	10	20	$u=5$	10	20	$u=5$	10	20	$\mathbf{u=5}$	10	20
L1-7B	40.49	38.41	38.45	38.51	38.38	38.43	38.49	38.30	38.34	38.40	38.33	38.35	38.40
L1-13B	29.00	27.56	27.50	27.53	27.65	27.56	27.56	28.20	28.04	27.92	28.38	28.24	28.10
L1-30B	14.91	15.09	15.10	15.11	15.11	15.13	15.14	15.07	15.09	15.10	15.01	15.03	15.04
L2-7B	37.89	36.71	36.51	36.43	36.91	36.70	36.53	37.21	37.14	37.06	37.28	37.25	37.22
L2-13B	28.51	28.23	28.24	28.25	28.18	28.18	28.21	28.22	28.20	28.21	28.29	28.27	28.27
L3-8B	63.80	63.39	63.62	63.67	63.27	63.42	63.53	63.12	63.26	63.16	63.01	63.19	63.11
L3.1-8B	62.27	60.32	60.61	61.00	60.40	60.64	61.08	60.57	60.75	61.18	60.69	60.87	61.26
Mis-7B	38.74	37.50	37.49	37.63	37.43	37.45	37.57	37.50	37.43	37.51	37.66	37.50	37.59
OPT-6.7B	50.65	48.96	48.76	48.62	48.94	48.74	48.60	49.08	48.86	48.69	49.25	49.00	48.80
Q2.5-7B	29.58	28.50	28.45	28.45	28.49	28.46	28.44	28.55	28.49	28.48	28.61	28.56	28.54
Q3-8B	37.53	36.08	36.08	36.44	36.08	36.09	36.45	36.18	36.18	36.40	36.44	36.22	36.61

(b) GBLM

Model	Uniform	$l=0$			$l=0.01$			$l=0.1$			$l=0.2$		
		$u=5$	10	20	$u=5$	10	20	$u=5$	10	20	$\mathbf{u=5}$	10	20
L1-7B	66.43	40.19	40.19	40.33	40.23	40.25	40.39	40.13	40.20	40.36	39.98	40.07	40.27
L1-13B	26.48	25.61	25.59	25.63	25.60	25.59	25.64	25.56	25.57	25.64	25.52	25.52	25.58
L1-30B	15.75	14.78	14.81	14.83	14.80	14.82	14.85	14.83	14.85	14.88	14.83	14.84	14.86
L2-7B	40.52	39.12	39.14	39.11	39.07	39.10	39.06	38.84	38.92	38.90	38.78	38.85	38.84
L2-13B	28.88	26.03	26.09	26.18	25.99	26.06	26.15	25.96	26.02	26.09	25.96	26.00	26.06
L3-8B	104.18	79.26	79.32	79.51	79.90	79.75	79.92	80.24	80.11	79.96	80.89	80.06	79.81
L3.1-8B	100.98	69.62	69.49	69.84	69.91	69.56	69.93	70.46	69.60	69.69	71.49	70.11	69.77
Mis-7B	41.34	38.94	38.87	38.80	39.04	38.91	38.89	39.41	39.19	39.22	39.73	39.47	39.41
OPT-6.7B	36.85	36.26	36.26	36.26	36.26	36.27	36.27	36.30	36.31	36.31	36.37	36.37	36.36
Q2.5-7B	28.09	27.12	27.03	27.05	27.13	27.02	27.07	27.15	27.08	27.08	27.18	27.17	27.11
Q3-8B	38.45	37.28	37.21	37.32	37.27	37.22	37.38	37.27	37.22	37.30	37.28	37.38	37.14

(c) DSnoT

Model	Uniform	$l=0$			$l=0.01$			$l=0.1$			$l=0.2$		
		$u=5$	10	20	$u=5$	10	20	$u=5$	10	20	$\mathbf{u=5}$	10	20
L1-7B	39.75	37.18	37.22	37.31	37.16	37.21	37.31	37.09	37.13	37.23	37.14	37.15	37.22
L1-13B	28.71	27.36	27.30	27.35	27.46	27.36	27.37	27.99	27.82	27.71	28.14	28.01	27.89
L1-30B	14.91	15.04	15.04	15.04	15.07	15.07	15.08	15.04	15.04	15.05	14.98	14.99	15.00
L2-7B	37.15	36.16	36.00	35.96	36.31	36.12	35.99	36.49	36.44	36.37	36.55	36.53	36.51
L2-13B	27.99	27.71	27.73	27.79	27.67	27.70	27.77	27.71	27.73	27.76	27.76	27.77	27.80
L3-8B	61.13	61.39	61.64	61.82	61.26	61.61	61.71	61.05	61.31	61.49	60.88	61.21	61.33
L3.1-8B	61.52	59.76	59.92	60.19	59.81	59.91	60.28	59.81	60.01	60.26	59.94	60.03	60.27
Mis-7B	38.83	37.52	37.46	37.61	37.47	37.44	37.59	37.54	37.43	37.49	37.67	37.49	37.50
OPT-6.7B	36.42	35.52	35.46	35.41	35.51	35.46	35.41	35.49	35.44	35.38	35.50	35.45	35.38
Q2.5-7B	29.39	28.28	28.20	28.18	28.27	28.22	28.17	28.31	28.25	28.22	28.37	28.31	28.28
Q3-8B	37.74	36.41	36.33	36.49	36.46	36.31	36.61	36.53	36.55	36.53	36.64	36.60	36.57

Thus $a_t = \|g_t\|_2^2$ is the smallest scalar isotropic majorant of the empirical Fisher surrogate for each occurrence. This form uses the same shared Gram matrix structure as the pruning methods. It remains an approximation to the exact loss curvature. The bound is tight when e_t is parallel to g_t and can be loose for other residual directions.

Substituting $e_t = \Delta W x_t$ into Equation 16 yields

$$\sum_{t=1}^{T} a_t \|\Delta W x_t\|_2^2 = \mathrm{Tr}\left[\Delta W \left(\sum_{t=1}^{T} a_t x_t x_t^\top\right) \Delta W^\top\right] \tag{19}$$

Thus, the occurrence sensitivity surrogate has the standard reconstruction form. It replaces the Uniform Gram matrix $\sum_t x_t x_t^\top$ with $\sum_t a_t x_t x_t^\top$.

This derivation uses three approximations. First, the dense model is assumed to be approximately stationary. Second, the output Hessian is approximated by its diagonal blocks. Third, the exact curvature is replaced by an empirical Fisher approximation. The final step from Equation 13 to Equation 16 is an exact spectral majorization of that approximation. SWR is therefore compatible with pruning methods that use a shared Gram matrix, but it is not a strict upper bound on the exact finite loss change. This property applies to the occurrence sensitivity a_t in Equation 16. Normalization and clipping do not retain it. Appendix C.2 evaluates the resulting reconstruction weight ρ_t empirically.

Table 12: **Normalization ablation with fixed pruning masks.** WikiText-2 test PPL ↓ for 11 models. Bold marks the lowest value within each model.

Setting	L1-7B	L1-13B	L1-30B	L2-7B	L2-13B	L3-8B	L3.1-8B	Mis-7B	OPT-6.7B	Q2.5-7B	Q3-8B
Raw	39.16	**27.91**	15.57	37.30	28.44	68.68	64.47	39.24	49.80	**28.40**	38.45
Mean-one	38.80	28.18	15.47	37.35	28.48	66.28	61.75	38.76	**48.38**	28.43	37.01
Mean-one + clip	**38.33**	28.38	**15.01**	**37.28**	**28.29**	**63.01**	**60.69**	**37.66**	49.25	28.61	**36.44**

C.2 Normalization and clipping

Equation 16 defines occurrence sensitivity a_t. Normalization to mean one within each sequence and clipping do not retain the theoretical bound, but they improve numerical stability.

For one sequence and module, we collect the L unnormalized occurrence sensitivities in $\mathbf{a} = (a_1 \dots a_L)$. We construct the reconstruction weights as

$$\bar{a} = \frac{1}{L}\sum_{t=1}^{L} a_t \qquad \boldsymbol{\rho} = \begin{cases} \mathrm{clip}(\mathbf{a}/\bar{a}; 0.2, 5) & \bar{a} > 0 \\ \mathbf{1} & \bar{a} = 0. \end{cases} \tag{20}$$

We apply this transformation independently to every sequence and module. Normalization to mean one does not change the ordering within that sequence and module. It also prevents a sequence with larger absolute gradients from receiving excessive reconstruction weight. We then clip the normalized values to limit the effect of extreme reconstruction weights on conditioning. If the mean sensitivity is zero, we use a vector of ones.

We evaluate every combination of lower bounds $\{0, 0.01, 0.1, 0.2\}$ and upper bounds $\{5, 10, 20\}$. The ablation covers all 11 models and the three methods with fixed pruning masks: Wanda, GBLM, and DSnoT. It contains 33 evaluations per clipping range. Every comparison fixes the pruning mask, calibration set, sparsity, and reconstruction method. We exclude methods that select the pruning mask during reconstruction because the reconstruction weights can also change mask selection. Every tested range outperforms Uniform in at least 29 of the 33 cases. We use $[0.2, 5]$ as a fixed numerical range rather than a tuned optimum. It improves 31 of 33 cases. The exceptions are Wanda and DSnoT on LLaMA-1 30B, as shown in Table 11.

Table 12 evaluates the two numerical steps in Equation 4 separately. It uses the Wanda pruning mask at 70% sparsity. Raw uses occurrence sensitivity a_t directly as the reconstruction weight, without normalization or clipping. Mean-one divides a_t by its mean $\bar{a}$ over the same sequence and module. Mean-one + clip also clips the result to $[0.2, 5]$ and is the SWR configuration.

Raw is better than Uniform in 6 of 11 models but is worse by up to 7.6% on Llama-3 8B. A few occurrences with very large a_t can receive most of the reconstruction weight. Mean-one lowers PPL relative to Raw in 7 of 11 models, by up to 4.2% on Llama-3.1 8B. Its four increases are all below 1.0%. Clipping lowers PPL relative to Mean-one in 8 of 11 models, by a median of 1.2% and up to 5.2% on Llama-3 8B. Together, normalization and clipping lower PPL relative to Raw by a median of 2.2% and up to 9.0%. They also increase the number of models that improve over Uniform from 6 to 10.

C.3 Weighted objective and fixed experimental settings

For one module, we concatenate all T calibration inputs into $X \in \mathbb{R}^{d_{\text{in}} \times T}$. We define the diagonal matrix of reconstruction weights as $D = \mathrm{diag}(\rho_1, \dots, \rho_T)$ and the matrix perturbation as $\Delta W = \widehat{W} - W$. The local weighted objective is

$$\sum_t \rho_t \|\Delta W x_t\|_2^2 = \|\Delta W X D^{1/2}\|_F^2 = \mathrm{Tr}\big(\Delta W X D X^\top \Delta W^\top\big). \tag{21}$$

SWR replaces the Uniform Gram matrix $XX^\top$ with

$$G_{\text{SWR}} = XDX^\top = \sum_t \rho_t x_t x_t^\top. \tag{22}$$

Scaling each calibration input by $\sqrt{\rho_t}$ is an implementation that accumulates this matrix. Both the objective and G_{SWR} are linear in ρ_t. SWR retains the pruning solver's update rule and changes its Gram

matrix. In experiments with fixed pruning masks, Uniform and SWR use the same dense weights, pruning mask, calibration set, block size, and sequential OBS update order. Both configurations apply the same relative damping rule

$$G' = G + \lambda I \qquad \lambda = 0.01 \ \mathrm{mean}(\mathrm{diag}\, G) \tag{23}$$

where G is the Gram matrix for the configuration: $XX^\top$ for Uniform and G_{SWR} for SWR. SWR changes the diagonal of G, so the scalar λ differs between configurations. The relative damping rule remains fixed. The reconstruction weights are the only experimental change.

C.4 Comparison with candidate statistics

Section C.1 derives occurrence sensitivity from the loss gradient. This experiment compares it with candidate statistics while holding all other components fixed. We test five configurations across 11 models, for 55 matched evaluations. Every configuration uses the same Wanda pruning mask at 70% sparsity, calibration set, relative damping rule, and OBS solver. Uniform assigns unit reconstruction weight to every occurrence. The remaining configurations apply the same weight construction to four candidate statistics. For module m and occurrence t, the four candidate statistics are

$$\begin{aligned}
\text{SWR} \quad & s_{m,t} = \left\| \nabla_{y_{m,t}} \mathcal{L} \right\|_2^2 \\
\text{Activation Norm} \quad & s_{m,t} = \| x_{m,t} \|_2^2 \\
\text{Next-token loss} \quad & s_{m,t} = \ell_t \\
\text{Shuffled} \quad & s_{m,t} = \left\| \nabla_{y_{m,\pi_m(t)}} \mathcal{L} \right\|_2^2
\end{aligned}$$

where $x_{m,t}$ and $y_{m,t}$ are the input and output of module m at occurrence t. The SWR statistic is the occurrence sensitivity a_t from Appendix C.1, evaluated separately for each module. The next-token loss ℓ_t is the loss term for occurrence t and is shared across modules. We draw the permutation π_m independently for each sequence and module. Shuffling preserves the distribution of unnormalized occurrence sensitivity but changes its alignment with occurrences. Every candidate statistic uses the same normalization to mean one, clipping, and reconstruction procedure as SWR.

Activation norm tests a feature magnitude available from a forward pass. Next token loss measures prediction difficulty without distinguishing modules. Shuffling tests whether the sensitivity distribution is useful without its alignment to occurrences.

Table 1 reports the results. SWR outperforms Uniform on 10 of the 11 models. LLaMA-1 30B is the exception. The shuffled control improves only 3 of 11 models. The distribution of occurrence sensitivity is therefore insufficient without the correct occurrence alignment. Activation norm and next-token loss are also less consistent than occurrence sensitivity.

D Grouped Reconstruction Weights

Grouped SWR splits the d_{out} output channels of a module into g contiguous groups of equal size. For group $\mathcal{G}$ and occurrence t, the grouped sensitivity is the sum of squared output-gradient entries, $a_{\mathcal{G},t} = \sum_{i \in \mathcal{G}} g_{t,i}^2$. For $g = 1$, this is a_t. Each group is normalized and clipped independently by Equation 4. It receives a Gram matrix $\sum_t \rho_{\mathcal{G},t}\, x_t x_t^\top$ with the same relative damping rule, and its rows are reconstructed with that matrix. The remaining procedure is unchanged. A module therefore requires g Gram matrix inverses instead of one. All values of g reuse the same forward and backward pass. Table 13 reports PPL on three evaluation sets for $g \in \{1, 2, 4\}$ and four models. It evaluates reconstruction on a fixed Wanda pruning mask and on SparseGPT with joint mask selection and reconstruction. The $g = 1$ rows are SWR.

Increasing g can lower PPL, but its additional benefit is usually small relative to its cost. On WikiText-2 test, $g = 4$ is lower than $g = 1$ in all eight configurations, but the relative reduction is at most 0.6% in seven of them. SparseGPT on Llama-3.1 8B is the exception, with a 5.6% reduction. On C4 validation, the reductions range from 0.3% to 2.6%. On PTB test, $g = 4$ is lower in seven configurations but increases PPL for Wanda on Llama-2 7B. Moreover, $g = 2$ is worse than $g = 1$ for SparseGPT on three of the four models on WikiText-2 test.

The computational increase is substantial. On Llama-2 7B, reconstruction at $g = 2$ and $g = 4$ takes $2.2\times$ and $4.0\times$ the $g = 1$ time with the Wanda pruning mask, and $1.7\times$ and $3.2\times$ for SparseGPT. The

Table 13: Grouped reconstruction weights at 70% unstructured sparsity. PPL ↓ on three evaluation corpora; $g = 1$ is SWR.

Model	Method	WikiText-2 test			C4 validation			PTB test		
		$g = 1$	$g = 2$	$g = 4$	$g = 1$	$g = 2$	$g = 4$	$g = 1$	$g = 2$	$g = 4$
L2-7B	Wanda	37.28	37.25	**37.16**	47.02	46.91	**46.77**	**257.61**	264.25	267.97
	SparseGPT	26.16	26.55	**26.13**	29.38	29.20	**29.08**	1709.58	**1498.27**	1525.07
L3.1-8B	Wanda	60.69	60.50	**60.39**	90.32	90.16	**90.09**	77.31	77.05	**76.73**
	SparseGPT	39.23	38.73	**37.05**	48.47	48.57	**47.21**	55.27	51.36	**49.63**
Q2.5-7B	Wanda	28.61	28.61	**28.53**	39.89	39.85	**39.72**	42.40	42.16	**41.97**
	SparseGPT	19.46	19.54	**19.43**	27.18	27.04	**26.99**	28.19	27.75	**27.50**
OPT-6.7B	Wanda	49.25	49.21	**48.96**	49.75	49.69	**49.50**	55.31	55.36	**55.13**
	SparseGPT	20.08	20.09	**19.97**	21.94	21.88	**21.87**	27.48	**26.69**	27.24

Table 14: Cost of grouped reconstruction weights on Llama-2 7B.

	Measured reconstruction time ratio			Complete SWR time (s)		
Pruning method	$g = 1$	$g = 2$	$g = 4$	$g = 1$	$g = 2$	$g = 4$
Wanda	1.00×	2.16×	4.02×	348.3	578.8	949.8
SparseGPT	1.00×	1.67×	3.16×	244.1	377.2	673.2

estimated complete time at $g = 4$ increases from 348.3 to 949.8 seconds for Wanda and from 244.1 to 673.2 seconds for SparseGPT (Table 14). We therefore use $g = 1$ as the default: it retains one Gram matrix per module and provides the best measured balance between PPL and reconstruction cost. This is also the structure used by OBS, SparseGPT, and ALPS. Before normalization and clipping, its occurrence sensitivity is the scalar spectral majorizer derived in Section 3.1. The deployed reconstruction weights do not preserve this bound. Larger g remains available when its additional cost is acceptable.

E Post-Training Cost, Memory, and Reuse

This section reports all cost measurements. The benchmark over five pruning methods uses Llama-2 7B, 70% sparsity, FP16 weights, and one fixed C4 train calibration draw of $128 \times 2{,}048$ tokens. For each method, its standard and SWR configurations are measured once on the same GPU. Model and tokenizer loading, calibration construction, and WikiText-2 test evaluation are excluded from every timing and peak memory window. CUDA is synchronized at each boundary. Table 15 reports wall time and peak memory for all five methods. These measurements quantify computational overhead.

Table 15: Cost after setup on Llama-2 7B at 70% sparsity on one NVIDIA RTX PRO 6000 Blackwell GPU. Fixed mask rows include pruning mask construction and Uniform OBS. Rows for methods that select the pruning mask during reconstruction report both operations. Each SWR total includes its separately measured sensitivity pass. Memory is peak PyTorch CUDA allocation.

	Wall time (s)			Peak GPU memory (GiB)		
Pruning method	Uniform / Native	SWR	Δ	Uniform / Native	SWR	Δ
Wanda	295.9	348.3	+52.5	18.58	37.34	+18.76
GBLM	748.0	765.5	+17.5	37.34	37.34	+0.00
DSnoT	309.5	358.6	+49.1	19.97	37.34	+17.38
SparseGPT	196.1	244.1	+48.0	18.15	37.34	+19.19
ALPS	1814.2	1888.7	+74.5	20.56	37.34	+16.78

For Wanda, GBLM, and DSnoT, the reported Uniform procedures include pruning mask construction and Uniform OBS. SWR constructs the same pruning mask, computes occurrence sensitivity, and performs SWR OBS. SparseGPT and ALPS select the pruning mask during reconstruction, so reconstruction weights may alter both steps. These totals describe different complete procedures. Methods with fixed pruning masks use a separate pruning mask pass. SparseGPT and ALPS select the

Table 16: SparseGPT post-setup runtime by stage and model size. All times exclude model and tokenizer loading, calibration construction, the precomputed layer sparsity profile, and evaluation. The SWR reconstruction row uses cached occurrence sensitivities, while the SWR total covers sensitivity computation and reconstruction. Peak memory is the maximum PyTorch CUDA allocation in GiB.

Stage / quantity	Llama-2 7B	Llama-3.1 8B	Llama-2 13B
Sensitivity time (s)	45.34	53.18	85.12
Native SparseGPT time (s)	196.10	248.94	377.24
SWR SparseGPT time, ρ cached (s)	198.77	250.95	378.86
SWR total time (s)	244.11	304.13	463.98
SWR peak allocation (GiB)	37.34	44.75	67.14

pruning mask during reconstruction. For these measurements, every SWR configuration recomputes occurrence sensitivity. The sensitivity stages take 44.54–45.69 seconds, reach 37.34 GiB of allocated memory, and produce the same occurrence sensitivities. For GBLM, the SWR run includes a separate 45.5 s sensitivity pass, but its independently timed pruning mask construction was 31.3 s faster than in the Uniform run, reducing the observed total increase to 17.5 s. In normal use, these values are computed once for a compatible dense model and calibration setup. They can then be reused across pruning methods, masks, sparsity levels, and clipping ranges. GBLM uses gradients during its native pruning mask construction and already reaches the same peak. For the other four methods, the sensitivity pass determines the SWR peak.

E.1 Runtime across model scales

Table 16 reports post-setup SparseGPT runtime on three model sizes. All columns exclude model and tokenizer loading, calibration construction, the precomputed layer sparsity profile, and PPL evaluation. Sensitivity and reconstruction use separate timing windows; the SWR total covers both stages.

The cached SWR solve differs from Native by 1.36%, 0.81%, and 0.43% for the three models. The cached sensitivities occupy 224 MiB for the 7B and 8B models and 280 MiB for the 13B model. These values are the total size of the CPU tensors in memory. They exclude Python container overhead, GPU residency, and serialized file size. The cached tensors are much smaller than the temporary autograd peak. Unnormalized occurrence sensitivities depend on the dense model and calibration computation, not on the pruning method, mask, target sparsity, or clipping range. We therefore compute them once for each model and fixed calibration setup, then reuse them across compatible methods and sparsity levels. Reuse requires the same model, tokenizer, ordered calibration sequences, sequence length, reconstructed modules, and sensitivity definition. Cached values allow normalization and clipping to change without another backward pass.

Cached occurrence sensitivities are used only during pruning. The final sparse model does not load them, so SWR adds no parameter, persistent state, or inference operation. The measured cost is incurred once for a fixed post-training setup, not for each inference request.

F Additional PPL Results

For a matched comparison, relative PPL reduction is $100(P_{\text{ref}} - P_{\text{variant}})/P_{\text{ref}}$. The reference is Uniform when the pruning mask is fixed. It is Native when mask selection and reconstruction are performed jointly, unless another configuration is stated. Medians include all evaluated configurations, including regressions. Accuracy differences are in percentage points. Counts of improved configurations describe the direction of the reported scores rather than statistical significance.

Table 2 presents the main WikiText-2 test results at 70% sparsity. This appendix complements that table in two ways. First, we report C4 validation and PTB test results for Wanda, GBLM, and DSnoT with fixed pruning masks. Second, we provide a controlled reconstruction comparison for SparseGPT. We also report C4 validation and PTB test results for methods that select the pruning mask during reconstruction.

Table 17: WikiText-2 test PPL ↓ and variability at 70% unstructured sparsity. Central values reproduce Table 2; ± values report the population standard deviation across three calibration draws.

Model	Configuration	Wanda	GBLM	DSnoT	SparseGPT	ALPS
LLaMA-1 7B	Uniform/Native	40.49 ± 1.10	66.43 ± 3.43	39.75 ± 0.61	26.84 ± 0.47	18.87 ± 0.61
	SWR	$\mathbf{38.33} \pm 0.73$	$\mathbf{39.98} \pm 0.64$	$\mathbf{37.14} \pm 0.46$	$\mathbf{26.26} \pm 0.93$	$\mathbf{18.11} \pm 0.66$
LLaMA-1 13B	Uniform/Native	29.00 ± 0.51	26.48 ± 0.55	28.71 ± 0.84	19.91 ± 0.62	16.27 ± 0.35
	SWR	$\mathbf{28.38} \pm 0.50$	$\mathbf{25.52} \pm 0.39$	$\mathbf{28.14} \pm 0.74$	$\mathbf{19.65} \pm 0.50$	$\mathbf{16.03} \pm 0.22$
LLaMA-1 30B	Uniform/Native	$\mathbf{14.91} \pm 0.46$	15.75 ± 0.36	$\mathbf{14.91} \pm 0.42$	13.16 ± 0.35	11.55 ± 0.26
	SWR	15.01 ± 0.43	$\mathbf{14.83} \pm 0.38$	14.98 ± 0.42	$\mathbf{13.06} \pm 0.12$	$\mathbf{11.10} \pm 0.26$
Llama-2 7B	Uniform/Native	37.89 ± 0.40	40.52 ± 0.86	37.15 ± 0.44	28.35 ± 0.82	19.59 ± 0.20
	SWR	$\mathbf{37.28} \pm 0.28$	$\mathbf{38.78} \pm 0.65$	$\mathbf{36.55} \pm 0.43$	$\mathbf{26.16} \pm 0.40$	$\mathbf{19.45} \pm 0.44$
Llama-2 13B	Uniform/Native	28.51 ± 0.42	28.88 ± 0.77	27.99 ± 0.43	20.11 ± 0.38	14.53 ± 0.41
	SWR	$\mathbf{28.29} \pm 0.36$	$\mathbf{25.96} \pm 0.77$	$\mathbf{27.76} \pm 0.37$	$\mathbf{19.09} \pm 0.43$	$\mathbf{14.01} \pm 0.37$
Llama-3 8B	Uniform/Native	63.80 ± 0.93	104.18 ± 9.89	61.13 ± 1.16	44.88 ± 0.98	30.62 ± 1.70
	SWR	$\mathbf{63.01} \pm 0.68$	$\mathbf{80.89} \pm 3.85$	$\mathbf{60.88} \pm 0.74$	$\mathbf{39.86} \pm 1.96$	$\mathbf{27.40} \pm 1.67$
Llama-3.1 8B	Uniform/Native	62.27 ± 1.71	100.98 ± 4.19	61.52 ± 1.68	43.24 ± 1.29	30.35 ± 1.51
	SWR	$\mathbf{60.69} \pm 0.86$	$\mathbf{71.49} \pm 4.07$	$\mathbf{59.94} \pm 0.98$	$\mathbf{39.23} \pm 1.11$	$\mathbf{27.28} \pm 1.04$
Mistral 7B	Uniform/Native	38.74 ± 0.68	41.34 ± 0.55	38.83 ± 0.35	$\mathbf{22.32} \pm 0.02$	17.32 ± 0.18
	SWR	$\mathbf{37.66} \pm 0.74$	$\mathbf{39.73} \pm 0.91$	$\mathbf{37.67} \pm 0.52$	23.11 ± 0.48	$\mathbf{16.98} \pm 0.10$
OPT 6.7B	Uniform/Native	50.65 ± 1.34	36.85 ± 1.13	36.42 ± 0.49	21.12 ± 0.43	19.18 ± 0.72
	SWR	$\mathbf{49.25} \pm 1.19$	$\mathbf{36.37} \pm 1.13$	$\mathbf{35.50} \pm 0.62$	$\mathbf{20.08} \pm 0.40$	$\mathbf{17.62} \pm 0.53$
Qwen2.5 7B	Uniform/Native	29.58 ± 0.29	28.09 ± 0.32	29.39 ± 0.34	20.94 ± 0.37	16.43 ± 0.24
	SWR	$\mathbf{28.61} \pm 0.08$	$\mathbf{27.18} \pm 0.13$	$\mathbf{28.37} \pm 0.17$	$\mathbf{19.46} \pm 0.39$	$\mathbf{15.82} \pm 0.13$
Qwen3 8B	Uniform/Native	37.53 ± 0.76	38.45 ± 0.58	37.74 ± 0.83	$\mathbf{25.39} \pm 0.94$	20.86 ± 0.89
	SWR	$\mathbf{36.44} \pm 0.42$	$\mathbf{37.28} \pm 0.41$	$\mathbf{36.64} \pm 0.45$	25.90 ± 0.88	$\mathbf{20.22} \pm 0.72$

Table 17 adds the observed variability across three calibration draws to the main WikiText-2 test values. Mask only configurations are omitted because they are not part of the matched reconstruction comparison over three calibration draws.

F.1 PPL ON ADDITIONAL EVALUATION SETS WITH FIXED PRUNING MASKS

The following tables report C4 validation and PTB test results for Wanda, GBLM, and DSnoT. These evaluation sets test whether improvements on WikiText-2 test also occur on other text corpora.

For all three pruning methods, we distinguish three configurations: *Mask only*, *Uniform*, and *SWR*, matching the terminology in Table 2. Mask only is the pruned model before reconstruction. Uniform applies sequential OBS reconstruction to the same fixed pruning mask with unit reconstruction weights. SWR uses the same pruning mask and reconstruction procedure with weights derived from occurrence sensitivity. The comparison between Mask only and Uniform measures the effect of OBS reconstruction. The comparison between Uniform and SWR measures the effect of the reconstruction weights. Changes from Mask only to Uniform are therefore not attributed to SWR. Tables 18 and 19 report all three configurations on C4 validation and PTB test, respectively.

F.2 SPARSEGPT AND ALPS RECONSTRUCTION

SparseGPT and ALPS differ from the experiments with fixed pruning masks above. Both methods select the pruning mask during reconstruction. Table 2 therefore compares their complete *Native* and *SWR* configurations rather than reconstruction with fixed pruning masks.

For SparseGPT, we additionally decompose its joint mask selection and reconstruction, as shown in Table 20. *Native* is standard SparseGPT. The other three configurations first obtain the native SparseGPT pruning mask for the same model and calibration draw. *SWR mask* reruns mask selection with SWR and reconstructs the retained weights with the Uniform Gram matrix. *SWR recon.* keeps the native pruning mask fixed and reconstructs with SWR. *Full SWR* applies SWR to both steps. All four configurations hold the dense model, calibration draw, sparsity, and solver schedule fixed. Changing only mask selection improves 7 of 11 models (median 1.94%) and changing only reconstruction

Table 18: C4 validation PPL ↓ with fixed pruning masks. Bold compares Uniform and SWR only.

Method		L1-7B	L1-13B	L1-30B	L2-7B	L2-13B	L3-8B	L3.1-8B	Mis-7B	OPT-6.7B	Q2.5-7B	Q3-8B
Wanda	Mask only	84.56	53.51	18.35	79.31	57.29	166.96	178.38	73.29	279.45	84.11	87.94
	Uniform	44.46	32.58	**16.93**	47.28	35.78	91.02	91.45	**46.55**	50.16	40.34	51.57
	SWR	**43.24**	**32.20**	17.05	**47.02**	**35.58**	**90.82**	**90.32**	46.63	**49.75**	**39.89**	**50.90**
GBLM	Mask only	94.35	44.35	18.89	94.17	63.46	183.84	188.95	72.96	52.31	85.19	98.18
	Uniform	56.54	26.68	17.76	47.00	32.97	97.96	105.78	**45.66**	37.28	38.18	54.54
	SWR	**39.37**	**26.29**	**16.94**	**44.86**	**31.08**	**90.17**	**90.09**	45.77	**37.15**	**37.38**	**53.74**
DSnoT	Mask only	76.70	54.16	17.11	83.70	57.97	161.35	179.83	72.56	50.07	78.27	91.61
	Uniform	44.16	32.25	**16.71**	46.76	35.26	**88.10**	90.77	**46.87**	35.90	39.85	52.06
	SWR	**42.28**	**31.92**	16.80	**46.35**	**35.06**	88.52	**89.79**	46.90	**35.70**	**39.45**	**51.09**

Table 19: PTB test PPL ↓ with fixed pruning masks. Bold compares Uniform and SWR only.

Method		L1-7B	L1-13B	L1-30B	L2-7B	L2-13B	L3-8B	L3.1-8B	Mis-7B	OPT-6.7B	Q2.5-7B	Q3-8B
Wanda	Mask only	583.72	335.52	99.84	390.67	373.92	152.56	150.74	217.64	176.40	106.67	96.93
	Uniform	307.72	**192.68**	63.05	**242.35**	228.95	84.01	78.72	**143.11**	57.08	43.21	59.34
	SWR	**298.19**	197.94	**62.89**	257.61	**227.18**	**83.07**	**77.31**	145.92	**55.31**	**42.40**	**58.19**
GBLM	Mask only	614.19	456.12	115.85	464.78	377.24	256.59	236.64	227.72	75.53	98.79	117.85
	Uniform	428.26	196.24	72.36	**265.82**	217.96	158.20	143.09	**154.67**	46.65	39.55	66.02
	SWR	**332.80**	**194.80**	**64.15**	269.78	**211.81**	**127.75**	**108.90**	157.61	**45.63**	**39.23**	**64.01**
DSnoT	Mask only	559.49	307.98	87.14	389.28	348.09	148.21	135.07	211.74	58.73	99.41	101.96
	Uniform	307.94	**187.24**	62.46	**235.22**	224.06	82.21	77.24	**142.36**	42.93	42.54	59.41
	SWR	**292.55**	192.50	**62.10**	251.88	**220.65**	**81.23**	**75.56**	144.20	**41.42**	**42.14**	**58.17**

improves 6 of 11 (median 0.30%). Neither component alone accounts for most of the gain. Full SWR improves on both individual components in almost every case, 11 of 11 over SWR mask and 9 of 11 over SWR recon., with a median reduction of 4.92% against Native, beyond either component's individual median. Combining both changes produces most of the measured gain. For Mistral 7B and Qwen3 8B, Full SWR is worse than Native. In both cases, the mask-only change accounts for most of the regression, while the reconstruction-only change is smaller. Tables 21 and 22 report the corresponding C4 validation and PTB test results.

Table 20: SparseGPT decomposition of mask selection and reconstruction on WikiText-2 test PPL ↓ at 70% unstructured sparsity. Native is standard SparseGPT; SWR mask changes only mask selection; SWR recon. changes only reconstruction.

Method		L1-7B	L1-13B	L1-30B	L2-7B	L2-13B	L3-8B	L3.1-8B	Mis-7B	OPT-6.7B	Q2.5-7B	Q3-8B
SparseGPT	Native	26.84	19.91	13.16	28.35	20.11	44.88	43.24	**22.32**	21.12	20.94	**25.39**
	SWR mask	26.50	20.25	13.35	26.43	19.72	42.04	40.75	23.59	20.67	20.36	26.76
	SWR recon.	26.30	19.85	13.17	28.67	20.34	43.39	42.53	22.36	20.48	20.57	25.44
	Full SWR	**26.26**	**19.65**	**13.06**	**26.16**	**19.09**	**39.86**	**39.23**	23.11	**20.08**	**19.46**	25.90

F.3 QWEN2.5-32B EXTENSION

We extend WikiText-2 test PPL evaluation to Qwen2.5-32B at 70% unstructured sparsity. For Wanda, GBLM, and DSnoT, the comparison holds the native pruning mask fixed and changes only the OBS reconstruction weights. For SparseGPT and ALPS, we compare the complete Native and SWR configurations. Table 23 reports all five method comparisons. SWR lowers PPL in every setting.

G COMPLETE ZERO-SHOT RESULTS

This section provides the complete results for Table 3 and Section 4.2. We use accuracy on BoolQ (Clark et al., 2019), PIQA (Bisk et al., 2020), HellaSwag (Zellers et al., 2019), Wino-

Table 21: SparseGPT decomposition on C4 validation PPL ↓. The middle configuration changes reconstruction only.

Method		L1-7B	L1-13B	L1-30B	L2-7B	L2-13B	L3-8B	L3.1-8B	Mis-7B	OPT-6.7B	Q2.5-7B	Q3-8B
SparseGPT	Native	28.29	22.53	**15.28**	31.65	23.73	52.01	51.89	**27.99**	22.22	27.75	**32.89**
	SWR recon.	27.88	22.54	15.36	31.96	23.88	51.12	51.12	28.18	22.02	27.90	33.07
	SWR	**27.26**	**22.09**	15.34	**29.38**	**22.80**	**48.53**	**48.47**	29.19	**21.94**	**27.18**	32.91
ALPS	Native	19.32	15.75	12.28	20.82	16.48	35.71	36.14	21.60	19.50	21.73	26.76
	SWR	**18.85**	**15.60**	**12.23**	**19.84**	**15.94**	**34.75**	**34.23**	**21.02**	**19.05**	**21.14**	**26.67**

Table 22: SparseGPT decomposition on PTB test PPL ↓. The middle configuration changes reconstruction only.

Method		L1-7B	L1-13B	L1-30B	L2-7B	L2-13B	L3-8B	L3.1-8B	Mis-7B	OPT-6.7B	Q2.5-7B	Q3-8B
SparseGPT	Native	286.98	138.26	60.38	**1526.88**	261.55	69.05	63.61	**111.30**	28.83	29.51	**38.30**
	SWR recon.	282.56	**137.42**	59.22	1599.82	256.61	67.13	63.63	112.46	28.06	29.48	38.99
	SWR	**248.96**	138.67	**57.82**	1709.58	**217.46**	**60.93**	**55.27**	123.45	**27.48**	**28.19**	41.76
ALPS	Native	213.46	104.92	46.50	485.16	143.87	50.56	46.01	**76.94**	23.16	**22.37**	**31.60**
	SWR	**179.39**	**99.41**	**46.38**	**389.46**	**125.42**	**45.25**	**44.92**	79.30	**21.45**	22.44	33.29

Grande (Sakaguchi et al., 2021), ARC-Easy (Boratko et al., 2018), ARC-Challenge (Boratko et al., 2018), and OpenBookQA (Mihaylov et al., 2018). Tables 25, 26, 27, 28, and 29 report all seven tasks for each pruning method. SWR attains the highest mean accuracy in most settings. Table 24 reports the observed variability of the main mean accuracies across the same three calibration draws.

H Performance at Different Sparsity Levels

The main experiments use 70% unstructured sparsity. To test whether the benefit of SWR is specific to this sparsity level, we additionally evaluate four sparsity levels, $S \in \{0.5, 0.6, 0.7, 0.8\}$, across the same 11 models. For Wanda, GBLM, and DSnoT, we compare SWR with Uniform on the same fixed pruning mask. For SparseGPT, we compare SWR with Native. Table 30 reports the complete WikiText-2 test PPL results. Overall, the effect of SWR persists across sparsity levels, although the magnitude and consistency of the gains vary with both the pruning method and sparsity.

I Exact 4:8 Sparsity

Under unstructured sparsity, each row can retain weights at any positions. We next test whether SWR remains useful when the pruning mask must satisfy a local 4:8 constraint. Here four is the number of retained positions and eight is the group width. For every output row, each eligible matrix is partitioned from input column zero into disjoint groups of eight. The resulting pruning mask contains exactly four pruned and four retained positions per group. No incomplete group is padded. All evaluated matrices have input widths divisible by eight. The pruning mask therefore has exactly 50% sparsity. Each cell uses the common C4 train calibration protocol: 128 sequences of 2,048 tokens sampled by document. Evaluation uses all complete chunks of 2,048 tokens from the concatenated WikiText-2 test set.

For Wanda, GBLM, and DSnoT, each comparison uses the same 4:8 pruning mask produced by that pruning method and compares Uniform OBS with SWR OBS. For SparseGPT, the comparison is Native versus SWR. The pruning mask satisfies the exact 4:8 pattern in every configuration. Table 31 reports all 44 method–model comparisons. SWR lowers PPL in 8/11, 10/11, 7/11, and 10/11 comparisons for Wanda, GBLM, DSnoT, and SparseGPT, respectively.

Table 23: Qwen2.5-32B WikiText-2 test PPL ↓ at 70% unstructured sparsity. Bold marks the lowest PPL within each method block.

Method	Configuration	WikiText-2 test PPL
Wanda	Uniform	14.04
	SWR	**13.92**
GBLM	Uniform	13.65
	SWR	**13.59**
DSnoT	Uniform	14.08
	SWR	**13.95**
SparseGPT	Native	11.20
	SWR recon.	10.97
	SWR	**10.87**
ALPS	Native	9.86
	SWR	**9.74**

Table 24: Mean zero-shot accuracy over seven tasks ↑ (%) and variability at 70% unstructured sparsity. ± values report the population standard deviation across three calibration draws.

Model	Configuration	Wanda	GBLM	DSnoT	SparseGPT	ALPS
LLaMA-1 7B	Uniform/Native	38.57 ± 0.11	37.89 ± 0.35	38.40 ± 0.03	42.44 ± 0.49	45.89 ± 0.27
	SWR	**38.82** ± 0.07	**40.39** ± 0.27	**38.75** ± 0.05	**43.47** ± 0.19	**46.39** ± 0.36
LLaMA-1 13B	Uniform/Native	40.99 ± 0.19	43.31 ± 0.13	41.09 ± 0.25	45.40 ± 0.08	49.19 ± 0.13
	SWR	**41.34** ± 0.17	**43.62** ± 0.18	**41.66** ± 0.30	**46.18** ± 0.29	**49.65** ± 0.08
LLaMA-1 30B	Uniform/Native	52.00 ± 0.17	50.23 ± 0.15	52.02 ± 0.27	52.21 ± 0.25	55.44 ± 0.45
	SWR	**52.09** ± 0.15	**50.96** ± 0.19	**52.24** ± 0.18	**52.34** ± 0.29	**55.45** ± 0.26
Llama-2 7B	Uniform/Native	37.59 ± 0.13	38.49 ± 0.12	37.54 ± 0.08	41.93 ± 0.33	45.48 ± 0.26
	SWR	**37.78** ± 0.17	**38.60** ± 0.18	**37.79** ± 0.21	**42.67** ± 0.28	**46.06** ± 0.28
Llama-2 13B	Uniform/Native	40.95 ± 0.13	42.65 ± 0.31	41.19 ± 0.18	47.67 ± 0.26	51.74 ± 0.44
	SWR	**41.07** ± 0.10	**43.68** ± 0.24	**41.32** ± 0.17	**48.41** ± 0.32	**52.33** ± 0.08
Llama-3 8B	Uniform/Native	36.62 ± 0.27	36.40 ± 0.64	37.01 ± 0.16	43.30 ± 0.25	46.55 ± 0.34
	SWR	**37.21** ± 0.31	**38.18** ± 0.18	**37.36** ± 0.05	**43.64** ± 0.12	**47.30** ± 0.45
Llama-3.1 8B	Uniform/Native	37.35 ± 0.26	36.92 ± 0.38	37.21 ± 0.26	43.07 ± 0.31	46.09 ± 0.16
	SWR	**37.86** ± 0.27	**38.58** ± 0.17	**37.61** ± 0.22	**44.11** ± 0.24	**46.80** ± 0.23
Mistral 7B	Uniform/Native	38.84 ± 0.07	40.67 ± 0.06	38.84 ± 0.03	**45.70** ± 0.33	47.49 ± 0.68
	SWR	**39.03** ± 0.24	**40.84** ± 0.16	38.84 ± 0.11	45.31 ± 0.25	**47.71** ± 0.06
OPT 6.7B	Uniform/Native	40.90 ± 0.11	43.71 ± 0.18	43.19 ± 0.01	46.87 ± 0.09	48.89 ± 0.20
	SWR	**41.28** ± 0.24	**44.02** ± 0.15	**43.42** ± 0.08	**47.69** ± 0.04	**49.55** ± 0.23
Qwen2.5 7B	Uniform/Native	44.01 ± 0.08	45.66 ± 0.07	44.06 ± 0.16	50.73 ± 0.31	51.65 ± 0.33
	SWR	**44.30** ± 0.09	**46.09** ± 0.01	**44.60** ± 0.17	**51.79** ± 0.53	**53.42** ± 0.53
Qwen3 8B	Uniform/Native	43.51 ± 0.15	44.50 ± 0.24	**43.43** ± 0.11	**50.68** ± 0.40	52.58 ± 0.05
	SWR	**43.64** ± 0.15	**44.53** ± 0.21	43.41 ± 0.04	50.52 ± 0.19	**53.22** ± 0.58

J Calibration Source Transfer

This section expands the view in Table 4. For both Wanda and GBLM, every comparison uses the same pruning mask for Uniform OBS and SWR at 70% sparsity. Calibration uses 128 sequences of 2,048 tokens from C4 train, PTB train, or WikiText-2 train. Each sparse model is evaluated on WikiText-2 test, the first 512 complete chunks of C4 validation, and all complete chunks of 2,048 tokens from the local PTB test split. Results are shown in Table 32.

K LoRA Recovery after Pruning

We test whether LoRA recovery (Hu et al., 2022) preserves the ordering between Uniform and SWR. The experiment covers LLaMA-1 7B, Llama-2 7B, Llama-3.1 8B, Qwen2.5-7B, and Qwen3 8B at 70% sparsity. We use fixed pruning masks for Wanda, GBLM, and DSnoT, and Full SWR for SparseGPT. Evaluation uses WikiText-2 test.

Table 25: Zero-shot accuracy ↑ (%) for each task with Wanda.

Model	Configuration	BoolQ	PIQA	HellaSwag	WinoGrande	ARC-E	ARC-C	OBQA	Mean
L1-7B	Uniform	62.13	58.60	**30.02**	52.80	32.42	20.34	13.67	38.57
	SWR	**62.23**	**58.65**	30.01	**53.67**	**32.79**	**20.39**	**14.00**	**38.82**
L1-13B	Uniform	62.17	62.99	32.46	**53.25**	40.10	20.99	15.00	40.99
	SWR	62.17	**63.60**	**32.75**	53.12	**41.33**	**21.33**	**15.07**	**41.34**
L1-30B	Uniform	64.61	**72.22**	**43.19**	**66.51**	62.01	31.68	**23.80**	52.00
	SWR	**65.26**	71.82	42.95	66.30	**63.01**	**32.08**	23.20	**52.09**
L2-7B	Uniform	62.02	57.83	**29.00**	**51.01**	30.35	**19.31**	13.60	37.59
	SWR	62.02	**58.16**	28.99	50.91	**31.06**	19.28	**14.07**	**37.78**
L2-13B	Uniform	62.09	**61.41**	**31.20**	56.14	**41.26**	20.36	**14.20**	40.95
	SWR	**62.11**	61.10	31.09	**57.30**	41.25	**20.82**	13.80	**41.07**
L3-8B	Uniform	57.53	56.67	28.09	50.38	33.68	**17.01**	13.00	36.62
	SWR	**60.15**	**56.96**	28.09	**50.83**	**34.55**	16.81	**13.07**	**37.21**
L3.1-8B	Uniform	60.63	56.75	28.39	50.86	33.95	**17.24**	13.60	37.35
	SWR	**61.87**	**56.86**	**28.48**	**51.62**	**34.72**	16.98	**14.47**	**37.86**
Mis-7B	Uniform	62.39	58.45	30.48	51.70	36.53	18.17	**14.13**	38.84
	SWR	**62.51**	**58.54**	**30.55**	**53.04**	**36.62**	**18.26**	13.67	**39.03**
OPT-6.7B	Uniform	**62.18**	63.53	30.78	**52.51**	43.95	19.51	13.87	40.90
	SWR	62.17	**64.36**	**31.16**	51.70	**45.37**	**19.54**	**14.67**	**41.28**
Q2.5-7B	Uniform	62.13	**64.38**	33.42	56.77	52.43	21.53	17.40	44.01
	SWR	**62.17**	64.15	**33.56**	**57.70**	**52.78**	**22.10**	**17.67**	**44.30**
Q3-8B	Uniform	65.36	62.21	**32.54**	**55.56**	**50.07**	21.87	16.93	43.51
	SWR	**66.94**	**62.89**	32.45	54.06	49.66	**22.44**	**17.07**	**43.64**

All configurations use the same training protocol. We train rank 8 LoRA adapters for one epoch on the first 30,000 tokenized blocks of 1,024 tokens from the first official C4 training shard. Adapters target the query and value projections, with scaling 16 and dropout 0.05. We use AdamW, batch size one, gradient accumulation over 128 steps, learning rate 10^{-4}, a linear schedule with five warmup steps, and no weight decay. Sparse base parameters remain fixed.

Table 33 reports WikiText-2 test PPL after identical LoRA training. SWR remains lower in 17 of 20 method–model comparisons, with a median relative advantage of 0.37%. LoRA reduces the difference but usually does not reverse the ordering.

L Compatibility with OWL Layer Sparsity Allocation

OWL (Yin et al., 2023) sets sparsity separately for each layer. SWR sets reconstruction weights separately for each occurrence and module. We combine the two methods by fixing the OWL layer sparsity profile within each comparison and changing only the reconstruction weights. This experiment does not compare OWL with uniform layer sparsity.

We use one fixed OWL layer sparsity profile per model. We use the repository default $\lambda = 0.08$, an outlier threshold of 5, and a mean layer sparsity of 70%. Calibration uses 128 C4 train sequences of length 2,048. For OWL + Wanda, a single native Wanda pruning mask is shared by Uniform OBS and SWR OBS reconstruction. For OWL + SparseGPT, we compare Native and SWR, so the reconstruction weights may change both mask selection and reconstruction. We evaluate WikiText-2 test, the first 512 complete chunks of C4 validation, and all complete chunks of PTB test.

Table 34 reports every model and both configurations. SWR lowers PPL in 18/22 WikiText-2 test, 18/22 C4 validation, and 16/22 PTB test comparisons. Thus, SWR can be combined with nonuniform layer allocation.

Table 26: Zero-shot accuracy ↑ (%) for each task with GBLM.

Model	Configuration	BoolQ	PIQA	HellaSwag	WinoGrande	ARC-E	ARC-C	OBQA	Mean
L1-7B	Uniform	59.40	59.07	29.74	52.28	31.80	19.88	13.07	37.89
	SWR	**62.14**	**60.88**	**31.60**	**55.49**	**37.08**	**20.48**	**15.07**	**40.39**
L1-13B	Uniform	62.17	65.56	35.37	56.67	45.55	22.04	15.80	43.31
	SWR	62.17	**66.21**	**35.66**	**56.91**	**46.25**	**22.27**	**15.87**	**43.62**
L1-30B	Uniform	63.76	**71.29**	41.65	62.51	61.13	30.23	21.07	50.23
	SWR	**64.16**	71.16	**42.54**	**65.27**	**61.15**	**30.80**	**21.67**	**50.96**
L2-7B	Uniform	62.17	59.32	29.74	**51.99**	33.74	**18.37**	**14.13**	38.49
	SWR	62.17	**60.07**	**30.16**	51.46	**34.55**	18.17	13.60	**38.60**
L2-13B	Uniform	62.27	63.29	33.19	54.72	46.37	22.10	**16.60**	42.65
	SWR	**62.51**	**64.04**	**33.58**	**55.88**	**49.62**	**23.63**	16.47	**43.68**
L3-8B	Uniform	50.89	58.05	28.62	**50.99**	35.44	17.06	**13.73**	36.40
	SWR	**60.18**	**59.19**	**28.91**	50.59	**37.63**	**17.15**	13.60	**38.18**
L3.1-8B	Uniform	53.44	58.09	28.46	52.01	35.58	**17.29**	13.60	36.92
	SWR	**59.69**	**59.30**	**29.13**	**53.20**	**37.58**	17.06	**14.13**	**38.58**
Mis-7B	Uniform	62.44	60.95	**32.27**	**54.28**	40.70	**19.25**	**14.80**	40.67
	SWR	**62.54**	**61.64**	32.14	53.67	**42.30**	19.20	14.40	**40.84**
OPT-6.7B	Uniform	**62.19**	66.43	34.01	54.80	49.72	**20.99**	17.80	43.71
	SWR	62.17	**66.70**	**34.44**	**55.01**	**50.59**	20.73	**18.47**	**44.02**
Q2.5-7B	Uniform	62.97	**65.67**	35.05	57.72	56.16	24.29	17.73	45.66
	SWR	**63.33**	65.40	**35.37**	**58.80**	**56.82**	**24.32**	**18.60**	**46.09**
Q3-8B	Uniform	68.06	63.26	32.42	**56.04**	51.68	22.41	17.60	44.50
	SWR	**68.44**	**63.28**	**32.45**	55.64	**51.85**	**22.44**	17.60	**44.53**

M Occurrence Selection, Weighting, and the Top 20% Peak

This section gives the protocols for Figure 2(a,b) and Table 7. The first experiment fixes one occurrence sensitivity ranking and compares Uniform and SWR on exactly the same selected occurrences. The second experiment uses the same ranking and changes the weights of the Top 20%, Next 30%, and Bottom 50% one region at a time. All experiments use the same Wanda pruning mask at 70% sparsity, the same C4 train calibration data with $(128 \times 2{,}048)$ tokens per calibration draw, and the same reconstruction solver and damping rule.

M.1 Occurrence subsets ranked by sensitivity

Figure 2(a,b) separates two choices: which calibration occurrences are used and which reconstruction weights they receive. Each model contains $T = 128 \times 2{,}048 = 262{,}144$ calibration occurrences. To obtain one ranking, we normalize each module's occurrence sensitivities to mean one within each sequence, average them across reconstructed linear modules in float64 without clipping, and rank all T occurrences. Ties are resolved by the original occurrence index. This ranking is shared by all modules only in this analysis. SWR itself still computes module-specific reconstruction weights.

For $q \in \{0.1, 0.2, \ldots, 1.0\}$, we select the first $\mathrm{round}(qT)$ occurrences in this ranking. The selected sets are nested, so every larger set contains all occurrences from the smaller sets. The same selected set is used in every reconstructed module. For each value of q, we compare:

- **Unit subset:** weight 1 for selected occurrences and 0 otherwise;
- **SWR subset:** the SWR reconstruction weight for each selected occurrence and module, and 0 otherwise.

For the SWR subset, we first compute the SWR reconstruction weights and then set the weights outside the selected set to zero. We do not renormalize or clip the nonzero weights again. Both

Table 27: Zero-shot accuracy ↑ (%) for each task with DSnoT.

Model	Configuration	BoolQ	PIQA	HellaSwag	WinoGrande	ARC-E	ARC-C	OBQA	Mean
L1-7B	Uniform	62.22	58.63	30.01	52.51	31.99	20.51	**12.93**	38.40
	SWR	**62.29**	**59.68**	**30.38**	**52.80**	**32.74**	**20.82**	12.53	**38.75**
L1-13B	Uniform	62.17	63.49	32.63	53.80	40.39	20.51	14.67	41.09
	SWR	62.17	**63.66**	**33.04**	**55.41**	**41.67**	**20.90**	**14.80**	**41.66**
L1-30B	Uniform	64.88	72.36	**43.38**	66.75	62.33	31.57	**22.87**	52.02
	SWR	**65.96**	72.36	43.08	**66.85**	**62.37**	**32.25**	22.80	**52.24**
L2-7B	Uniform	**62.13**	57.78	29.10	**50.43**	30.72	18.83	13.80	37.54
	SWR	62.11	**58.27**	**29.26**	50.36	**30.89**	**19.54**	**14.13**	**37.79**
L2-13B	Uniform	62.15	61.48	31.22	**56.22**	41.71	**20.53**	15.00	41.19
	SWR	**62.20**	**62.30**	**31.46**	55.49	**42.21**	20.22	**15.33**	**41.32**
L3-8B	Uniform	59.40	**57.04**	**28.24**	50.38	**33.73**	**17.04**	13.27	37.01
	SWR	**60.76**	56.64	28.20	**52.09**	33.29	16.98	**13.53**	**37.36**
L3.1-8B	Uniform	60.91	56.98	28.29	50.46	33.77	16.92	13.13	37.21
	SWR	**61.28**	**57.02**	**28.31**	**50.75**	**34.26**	**17.83**	**13.80**	**37.61**
Mis-7B	Uniform	**62.34**	**58.38**	30.48	51.88	**36.11**	18.15	14.53	38.84
	SWR	62.23	58.11	**30.53**	**52.01**	35.98	**18.52**	14.53	38.84
OPT-6.7B	Uniform	62.17	**66.10**	33.73	**54.64**	48.36	20.42	16.93	43.19
	SWR	62.17	65.61	**33.91**	54.54	**49.45**	**20.56**	**17.67**	**43.42**
Q2.5-7B	Uniform	62.13	64.04	33.72	56.96	52.76	21.47	17.33	44.06
	SWR	**62.17**	**64.53**	**34.05**	**57.54**	**53.58**	**21.93**	**18.40**	**44.60**
Q3-8B	Uniform	**65.43**	62.04	32.40	55.83	49.71	**21.99**	**16.60**	**43.43**
	SWR	64.71	**62.30**	**32.54**	**56.99**	**50.29**	20.82	16.20	43.41

configurations start from the dense model and use the same Wanda pruning mask, OBS solver, calibration order, and WikiText-2 test evaluation. Therefore, within each comparison, the selected occurrences are identical and only their reconstruction weights differ. The same stored occurrence sensitivities and rankings are reused for all selected fractions, so no additional backward pass is required.

Across the 99 intermediate settings, the Unit subset improves over Uniform using all occurrences in 71 cases, with a median PPL reduction of 0.28%. The SWR subset improves over Uniform in 82 cases, with a median reduction of 2.35%. When Uniform and SWR use exactly the same selected occurrences, the SWR subset is better in 92 of 99 comparisons, with a median relative PPL reduction of 1.76%.

The SWR subset has lower PPL more consistently as the selected fraction increases. It is better in 9/11 models at 10%, 10/11 at 20–50%, all 11 at 60–80%, and 10/11 at 90%. At 50%, the SWR subset is better than Uniform in 10/11 models, with a median reduction of 2.72%. The 50% SWR result is also slightly better than SWR using all occurrences in 6/11 models, with a median relative PPL reduction of 0.03%. Thus, occurrence sensitivity provides useful information beyond the ranking alone, and using more occurrences does not always reduce PPL. Table 35 reports results for every model and retained fraction.

M.2 Testing priority dilution one region at a time

This experiment tests why performance declines after the Top 20% in Figure 2(a). We use the same occurrence sensitivity ranking from Appendix M.1 and divide the occurrences into the Top 20%, Next 30%, and Bottom 50%. The same three regions are used in every reconstructed module. All configurations use the same dense model, Wanda pruning mask at 70% sparsity, OBS solver, calibration order, and WikiText-2 test evaluation.

Table 28: Zero-shot accuracy ↑ (%) for each task with SparseGPT.

Model	Configuration	BoolQ	PIQA	HellaSwag	WinoGrande	ARC-E	ARC-C	OBQA	Mean
L1-7B	Native	64.61	63.08	33.73	55.09	40.46	**23.58**	16.53	42.44
	SWR	**64.83**	**65.02**	**34.74**	**56.83**	**42.68**	23.55	**16.67**	**43.47**
L1-13B	Native	**62.65**	67.14	36.76	60.27	46.84	**25.26**	18.87	45.40
	SWR	62.39	**67.25**	**37.59**	**62.04**	**49.33**	25.17	**19.47**	**46.18**
L1-30B	Native	**65.97**	72.29	44.13	67.30	61.08	**31.60**	23.07	52.21
	SWR	65.54	**72.36**	**44.22**	**68.51**	**61.15**	31.23	**23.40**	**52.34**
L2-7B	Native	62.90	62.69	33.27	**57.33**	38.87	21.39	17.07	41.93
	SWR	**63.03**	**63.55**	**34.35**	56.27	**41.20**	**22.70**	**17.60**	**42.67**
L2-13B	Native	67.98	66.61	36.87	61.40	53.51	25.91	21.40	47.67
	SWR	**68.26**	**66.92**	**37.22**	**61.72**	**56.65**	**26.28**	**21.80**	**48.41**
L3-8B	Native	**66.56**	62.73	**34.15**	56.01	45.20	**21.59**	**16.87**	43.30
	SWR	66.48	**64.42**	33.79	**56.35**	**46.46**	21.50	16.47	**43.64**
L3.1-8B	Native	**67.21**	61.64	34.21	56.77	43.81	20.99	16.87	43.07
	SWR	65.99	**64.36**	**34.23**	**57.85**	**47.60**	**21.67**	**17.07**	**44.11**
Mis-7B	Native	**65.23**	**65.27**	**36.56**	**59.56**	**52.10**	23.49	**17.67**	**45.70**
	SWR	65.08	65.23	36.34	57.54	51.98	**23.81**	17.20	45.31
OPT-6.7B	Native	**62.22**	69.46	38.90	58.91	54.49	23.69	20.40	46.87
	SWR	62.20	**69.86**	**39.71**	**59.83**	**55.68**	**24.49**	**22.07**	**47.69**
Q2.5-7B	Native	**72.60**	68.57	39.92	62.38	61.66	29.27	20.73	50.73
	SWR	71.47	**69.86**	**40.41**	**63.22**	**63.76**	**30.80**	**23.00**	**51.79**
Q3-8B	Native	73.19	67.52	**38.75**	**63.14**	60.37	**30.46**	21.33	**50.68**
	SWR	**73.21**	**68.23**	38.44	62.19	**60.77**	29.18	**21.60**	50.52

The six steps in Table 7 change one region at a time. Steps 1 and 2 use only the Top 20%, first at weight 1 and then with SWR reconstruction weights. Steps 3 and 4 add the Next 30%, first at weight 1 and then with SWR reconstruction weights. Steps 5 and 6 repeat this comparison for the Bottom 50%. Each reweighting step therefore changes one region from weight 1 to its original SWR reconstruction weights. The selected occurrences, pruning mask, reconstruction method, and weights of all other regions remain unchanged. No new sensitivity computation, renormalization, or additional clipping is used. Table 36 reports the complete results.

For the Top 20%, changing the weights from 1 to SWR improves 10/11 models, with a median relative PPL reduction of 1.03%. After fixing the Top 20% at its SWR reconstruction weights, adding the Next 30% at weight 1 is nearly neutral. It improves 6/11 models with a median reduction of 0.13%. Changing the same Next 30% to SWR reconstruction weights improves 10/11 models, with a median reduction of 0.60%.

Adding the Bottom 50% at weight 1 worsens 8/11 models and increases median PPL by 0.97%. Changing only that region to SWR reconstruction weights improves 10/11 models, with a median reduction of 1.06%. Using all occurrences with SWR is not always better than excluding the Bottom 50%: Step 6 is better than Step 4 in only 5/11 models, with a median relative PPL increase of 0.03%.

N Reconstruction Error Allocation on Separate Evaluation Data

This section provides the analysis on data not used for reconstruction in Section 5. It measures how reconstruction weights change the distribution of squared reconstruction error while the pruning mask remains fixed.

Table 29: Zero-shot accuracy ↑ (%) for each task with ALPS.

Model	Configuration	BoolQ	PIQA	HellaSwag	WinoGrande	ARC-E	ARC-C	OBQA	Mean
L1-7B	Native	**65.42**	66.97	37.41	58.67	47.74	**26.34**	18.67	45.89
	SWR	64.80	**67.03**	**37.93**	**60.69**	**49.54**	26.02	**18.73**	**46.39**
L1-13B	Native	**63.13**	70.69	41.47	64.11	53.52	27.84	**23.60**	49.19
	SWR	62.72	**71.49**	**42.26**	**64.33**	**56.36**	**28.24**	22.13	**49.65**
L1-30B	Native	67.61	**74.86**	**48.55**	**70.09**	**65.36**	35.13	**26.47**	55.44
	SWR	**70.37**	74.21	48.41	69.93	65.19	**35.58**	24.47	**55.45**
L2-7B	Native	**62.63**	66.49	**37.44**	**60.04**	47.00	24.66	20.07	45.48
	SWR	62.32	**68.01**	37.36	59.98	**48.78**	**25.51**	**20.47**	**46.06**
L2-13B	Native	71.31	70.35	41.28	65.67	59.75	28.56	**25.27**	51.74
	SWR	**71.83**	**71.06**	**42.17**	**65.98**	**60.94**	**30.03**	24.27	**52.33**
L3-8B	Native	70.42	65.14	36.82	61.14	51.19	22.58	18.53	46.55
	SWR	**70.64**	**67.25**	**37.09**	**61.25**	**52.99**	**23.04**	**18.87**	**47.30**
L3.1-8B	Native	**69.11**	65.23	37.02	60.59	49.37	22.30	**19.00**	46.09
	SWR	68.78	**67.68**	**37.54**	**61.09**	**50.84**	**22.95**	18.73	**46.80**
Mis-7B	Native	**65.06**	**67.28**	38.67	**62.04**	55.20	24.97	**19.20**	47.49
	SWR	64.86	67.25	**38.94**	60.62	**57.74**	**26.11**	18.47	**47.71**
OPT-6.7B	Native	62.22	71.58	41.50	60.48	57.74	25.80	22.93	48.89
	SWR	**62.23**	**71.76**	**41.76**	**61.80**	**58.46**	**25.85**	**25.00**	**49.55**
Q2.5-7B	Native	69.23	69.93	41.93	65.32	63.55	29.95	21.67	51.65
	SWR	**73.46**	**71.22**	**43.29**	**66.06**	**65.03**	**31.23**	**23.67**	**53.42**
Q3-8B	Native	73.59	**70.13**	**41.46**	63.14	**63.17**	**32.76**	23.80	52.58
	SWR	**79.02**	69.31	41.05	**64.33**	62.58	32.25	**24.00**	**53.22**

N.1 Protocol and sensitivity deciles

For each of the 11 models, we compare Uniform and SWR using the same fixed Wanda pruning mask and identical reconstruction solver settings. The reconstruction weights are the only intended difference. We evaluate both configurations on a disjoint draw of 32 C4 train sequences. Each sequence is defined by a fixed C4 train document ID and token offset. Every model tokenizes the same document with its own tokenizer, applies the same integer offset, and retains up to 2,048 tokens. The models therefore share the document and token index, but not necessarily the exact character span. For each sequence, one dense model forward and backward pass records the module inputs used to measure local reconstruction residuals. It also records the module output gradients used to compute occurrence sensitivity.

Within each module, we rank these occurrences by occurrence sensitivity and divide them into ten bins of nearly equal size, $B_1, \ldots, B_{10}$. The least sensitive 10% of occurrences form B_1, and the most sensitive 10% form B_{10}. We use the same squared module output gradient norm as SWR, but omit normalization to mean one and clipping because only the ranking is needed. These sensitivities and residuals are used only for analysis. They do not update the reconstructed model, change the pruning mask, or select hyperparameters.

N.2 Residual share

For module m, let $x_{m,t}$ be the dense model input for an evaluation occurrence t. For $a \in \{\mathrm{U}, \mathrm{S}\}$, denoting Uniform and SWR, the local reconstruction residual is

$$e^a_{m,t} = (\widehat{W}^a_m - W_m)x_{m,t}.$$

Both configurations are evaluated on the same dense model inputs. Uniform and SWR optimize different local objectives: Uniform minimizes the unweighted sum of squared residuals, whereas SWR minimizes its sensitivity-weighted form. Absolute unweighted residuals are therefore aligned

Table 30: Performance at different sparsity levels. WikiText-2 test PPL (↓) for 11 models at $S \in \{0.5, 0.6, 0.7, 0.8\}$. Wanda, GBLM, and DSnoT additionally report the Mask only configuration before reconstruction.

Method	Sparsity	L1-7B	L1-13B	L1-30B	L2-7B	L2-13B	L3-8B	L3.1-8B	Mis-7B	OPT-6.7B	Q2.5-7B	Q3-8B
Wanda	Mask only	7.26	6.15	5.25	6.92	5.97	9.84	9.66	6.55	11.97	8.62	11.35
	Uniform $S = 0.5$	7.32	6.24	5.23	6.99	6.01	9.60	9.51	6.61	11.75	8.43	11.63
	SWR	7.33	6.23	5.24	7.00	6.02	9.59	9.50	6.60	11.71	8.45	11.60
	Mask only	10.70	8.77	6.56	10.77	8.39	23.41	21.81	11.28	15.20	14.05	16.08
	Uniform $S = 0.6$	10.68	9.08	6.65	11.01	8.60	18.98	18.18	10.73	14.46	11.76	15.27
	SWR	10.54	9.02	6.67	10.96	8.61	18.84	18.12	10.70	14.39	11.62	15.09
	Mask only	83.64	54.95	17.37	74.96	45.76	122.85	124.37	61.40	165.70	73.94	62.00
	Uniform $S = 0.7$	40.49	29.00	14.91	37.89	28.51	63.80	62.27	38.74	50.65	29.58	37.53
	SWR	38.33	28.38	15.01	37.28	28.29	63.01	60.69	37.66	49.25	28.61	36.44
	Mask only	5504.09	4709.67	2734.32	2012.56	1040.86	859.39	927.67	319.20	3544.56	796.24	334.42
	Uniform $S = 0.8$	968.47	481.84	140.71	286.49	211.56	327.77	292.14	169.18	1868.52	196.97	166.65
	SWR	1001.27	482.14	135.61	288.85	209.33	334.84	288.00	162.42	1861.58	185.15	156.23
GBLM	Mask only	7.21	6.14	5.12	6.92	5.91	9.41	9.31	6.51	12.04	8.47	11.61
	Uniform $S = 0.5$	7.50	6.24	5.16	7.04	6.04	9.33	9.30	6.56	11.82	8.33	11.56
	SWR	7.24	6.21	5.14	7.01	5.96	9.22	9.16	6.57	11.79	8.34	11.55
	Mask only	11.27	8.43	6.53	11.59	8.39	21.47	21.17	11.08	15.74	13.02	16.72
	Uniform $S = 0.6$	13.17	8.84	6.76	11.62	8.96	18.09	18.08	10.57	15.02	11.29	15.30
	SWR	11.03	8.79	6.68	11.44	8.60	17.86	17.50	10.56	14.94	11.27	15.17
	Mask only	99.18	48.49	17.57	96.20	45.98	177.08	163.22	62.08	51.53	73.80	70.65
	Uniform $S = 0.7$	66.43	26.48	15.75	40.52	28.88	104.18	100.98	41.34	36.85	28.09	38.45
	SWR	39.98	25.52	14.83	38.78	25.96	80.89	71.49	39.73	36.37	27.18	37.28
	Mask only	5677.89	6009.53	5580.25	2254.70	1846.46	1636.17	1513.60	447.47	7835.69	1101.66	511.31
	Uniform $S = 0.8$	912.01	469.38	117.68	408.97	303.94	458.50	476.71	198.85	867.89	166.18	176.97
	SWR	710.75	477.78	101.39	393.49	248.18	488.88	399.95	196.47	794.31	155.28	173.11
DSnoT	Mask only	7.28	6.16	5.26	6.92	5.97	9.81	9.66	6.56	11.96	8.60	11.39
	Uniform $S = 0.5$	7.36	6.24	5.23	7.02	6.02	9.60	9.53	6.59	11.74	8.43	11.61
	SWR	7.35	6.23	5.24	7.00	6.01	9.58	9.52	6.61	11.70	8.44	11.66
	Mask only	10.57	8.76	6.54	10.74	8.40	22.98	21.51	11.36	15.02	14.03	16.31
	Uniform $S = 0.6$	10.66	9.09	6.64	11.01	8.59	18.80	18.15	10.73	14.52	11.76	15.33
	SWR	10.51	9.04	6.66	10.97	8.60	18.73	18.09	10.71	14.45	11.64	15.10
	Mask only	61.78	44.43	16.09	54.06	33.95	114.17	119.43	61.43	41.92	67.03	65.57
	Uniform $S = 0.7$	39.75	28.71	14.91	37.15	27.99	61.13	61.52	38.83	36.42	29.39	37.74
	SWR	37.14	28.14	14.98	36.55	27.76	60.88	59.94	37.67	35.50	28.37	36.64
	Mask only	4143.56	3516.38	1182.64	1108.72	253.46	945.77	853.20	327.75	7561.86	485.28	355.18
	Uniform $S = 0.8$	1298.34	408.75	147.58	301.84	195.22	319.29	336.93	171.76	5557.05	180.28	163.09
	SWR	1304.30	414.14	144.26	301.40	200.61	327.72	331.53	165.57	6088.44	162.33	154.56
SparseGPT	Native $S = 0.5$	7.23	6.21	5.32	7.00	6.02	9.40	9.34	6.48	11.61	8.43	11.63
	SWR $S = 0.5$	7.11	6.14	5.16	6.87	5.94	9.19	9.14	6.49	11.51	8.37	11.61
	Native $S = 0.6$	10.46	8.47	6.68	10.22	8.27	15.23	15.51	9.43	13.44	10.84	14.26
	SWR $S = 0.6$	9.96	8.26	6.59	10.01	8.01	14.98	14.93	9.45	13.24	10.61	14.19
	Native $S = 0.7$	26.84	19.91	13.16	28.35	20.11	44.88	43.24	22.32	21.12	20.94	25.39
	SWR $S = 0.7$	26.26	19.65	13.06	26.16	19.09	39.86	39.23	23.11	20.08	19.46	25.90
	Native $S = 0.8$	178.12	102.18	56.18	120.12	92.43	185.65	185.19	97.27	98.30	86.91	85.30
	SWR $S = 0.8$	156.68	89.13	54.55	94.02	88.49	185.51	163.85	94.96	94.76	66.98	84.30

Table 31: Complete WikiText-2 test PPL ↓ results under exact 4:8 structured sparsity. Bold marks the lower value in each matched comparison.

Method		L1-7B	L1-13B	L1-30B	L2-7B	L2-13B	L3-8B	L3.1-8B	Mis-7B	OPT-6.7B	Q2.5-7B	Q3-8B
Wanda	Uniform	8.670	7.552	6.081	8.767	**7.098**	13.206	**12.902**	**8.406**	13.002	9.860	13.299
	SWR	**8.622**	**7.538**	**6.058**	**8.723**	7.104	**13.182**	12.920	8.421	**12.962**	**9.856**	**13.081**
GBLM	Uniform	9.469	7.484	6.093	9.118	7.391	12.992	12.993	**8.295**	13.278	9.709	13.220
	SWR	**8.745**	**7.466**	**6.040**	**9.065**	**7.196**	**12.853**	**12.712**	8.315	**13.213**	**9.697**	**13.105**
DSnoT	Uniform	8.658	7.552	**6.060**	8.753	**7.085**	13.255	**12.938**	**8.422**	13.109	9.866	13.298
	SWR	**8.613**	**7.541**	6.081	**8.708**	7.097	**13.221**	12.973	8.439	**13.063**	**9.862**	**13.105**
SparseGPT	Native	8.603	7.417	6.181	8.498	7.038	12.282	12.213	8.083	**12.522**	9.539	12.660
	SWR	**8.406**	**7.298**	**6.032**	**8.381**	**6.904**	**12.021**	**11.975**	**8.063**	12.528	**9.458**	**12.559**

Table 32: Calibration source transfer across models. Calibration uses the row corpus and evaluation uses the column corpus. Uniform → SWR PPL (lower is better); bold marks the lower value.

Model	Pruning method	Calibration source	WikiText-2 test	C4 validation	PTB test
Qwen2.5-7B	Wanda	C4 train	29.58 → **28.61**	40.34 → **39.89**	43.21 → **42.40**
		PTB train	29.44 → **29.15**	45.03 → **44.84**	27.10 → **26.53**
		WikiText-2 train	23.40 → **22.90**	44.70 → **44.69**	42.93 → **42.67**
	GBLM	C4 train	28.09 → **27.18**	38.18 → **37.38**	39.55 → **39.23**
		PTB train	27.90 → **27.74**	42.79 → **41.85**	25.39 → **24.85**
		WikiText-2 train	21.54 → **20.93**	41.95 → **41.77**	**39.46** → 40.08
Llama-2-7B	Wanda	C4 train	37.89 → **37.28**	47.28 → **47.02**	**242.35** → 257.61
		PTB train	43.63 → **42.49**	62.87 → **61.51**	169.09 → **165.92**
		WikiText-2 train	28.81 → **28.78**	**52.07** → 52.08	258.15 → **250.69**
	GBLM	C4 train	40.52 → **38.78**	47.00 → **44.86**	**265.82** → 269.78
		PTB train	41.46 → **40.50**	63.42 → **61.76**	182.97 → **181.89**
		WikiText-2 train	26.02 → **25.52**	51.73 → **50.74**	242.78 → **241.34**

with the Uniform objective but not the SWR objective, and their comparison mixes total residual scale with its distribution across sensitivity bins. We instead normalize each bin by the total residual of the same configuration to isolate this distribution. This statistic does not test whether SWR lowers the absolute residual in every bin.

For sensitivity bin $B_{m,q}$, define its share of the module's total squared reconstruction error as

$$R^a_{m,q} = \frac{\sum_{t \in B_{m,q}} \|e^a_{m,t}\|_2^2}{\sum_{j=1}^{10} \sum_{t \in B_{m,j}} \|e^a_{m,t}\|_2^2}, \tag{24}$$

$$\Delta_q = \frac{100}{M} \sum_{m=1}^{M} \left(R^{\mathrm{S}}_{m,q} - R^{\mathrm{U}}_{m,q} \right). \tag{25}$$

We normalize within each module before averaging, so every module contributes equally regardless of its width or residual scale. Δ_q is measured in percentage points: a negative value means that bin q contains a smaller share of the residual under SWR, while a positive value means a larger share. For example, $\Delta_q = -0.10$ denotes a reduction of 0.10 percentage points in that bin's residual share, not a 10% reduction. Because $\sum_q \Delta_q = 0$, this quantity measures the redistribution of squared reconstruction error across sensitivity bins rather than a change in total error. Figure 4 reports this difference in residual share.

N.3 Complete Wanda results for 11 models

All 11 Wanda curves are strictly decreasing from the least sensitive to the most sensitive decile. The first five deciles are positive in every model and the last five are negative in every model. The median curve is $(0.1021, 0.0652, 0.0459, 0.0262, 0.0094, -0.0069, -0.0234, -0.0418, -0.0676, -0.1072)$ percentage points. For the most sensitive decile, the shift for each model ranges from -0.1665 to -0.0855 points and is negative in all 11 models. At the module level, the same direction for the most sensitive decile holds in 2,738 of 2,740 modules (99.93%). The median difference between the least and most sensitive deciles is 0.2108 points, with a range of 0.1593–0.3095. Together, these results show a consistent reduction in the share of residual error assigned to sensitive occurrences.

Table 33: **Uniform versus SWR after matched LoRA recovery at 70% unstructured sparsity.** Each cell reports WikiText-2 test PPL after LoRA (lower is better). Both sides use the same data, initialization, optimization schedule, and number of updates. For SparseGPT, the reference configuration is Native rather than Uniform. Bold marks the lower value in each matched comparison.

Pruning method	Configuration	L1-7B	L2-7B	L3.1-8B	Qwen2.5-7B	Qwen3-8B
Wanda	Uniform	**12.495**	12.877	**20.359**	14.220	16.566
	SWR	12.497	**12.827**	20.388	**14.148**	**16.289**
GBLM	Uniform	12.056	11.611	19.461	13.320	15.499
	SWR	**11.713**	**11.575**	**18.467**	**13.272**	**15.482**
DSnoT	Uniform	12.217	12.773	20.333	14.213	15.999
	SWR	**12.186**	**12.699**	**20.271**	**14.179**	**15.936**
SparseGPT	Native	11.430	**11.098**	17.342	12.597	14.565
	SWR	**11.145**	11.195	**17.103**	**12.453**	**14.554**

Table 34: Compatibility with OWL layer sparsity allocation at 70% global unstructured sparsity (PPL ↓). Every row uses the same OWL layer sparsity profile; "Uniform" assigns unit reconstruction weight to every occurrence. Bold marks the lower PPL in each matched comparison.

Evaluation set	Pruning method	Reconstruction weights	L1-7B	L1-13B	L1-30B	L2-7B	L2-13B	L3-8B	L3.1-8B	Mis-7B	OPT-6.7B	Q2.5-7B	Q3-8B
WikiText-2 test	OWL + Wanda	Uniform	22.04	16.01	11.02	23.75	**16.71**	50.15	45.02	27.78	36.88	28.18	40.67
		SWR	**21.47**	**15.89**	**10.91**	**23.69**	16.72	**49.72**	**44.82**	**27.51**	**36.26**	**27.89**	**39.58**
	OWL + SparseGPT	Native	19.73	13.99	**10.37**	20.39	14.53	35.93	34.54	**18.67**	21.68	19.89	**26.51**
		SWR	**18.32**	**13.95**	10.39	**19.68**	**14.12**	**34.53**	**31.89**	19.53	**20.64**	**19.25**	26.91
C4 validation	OWL + Wanda	Uniform	24.11	18.09	**13.43**	28.30	20.45	68.43	61.80	35.49	37.37	39.93	**55.50**
		SWR	**23.97**	**18.08**	13.53	**28.24**	**20.40**	**68.37**	**61.14**	**35.25**	**37.15**	**39.77**	55.55
	OWL + SparseGPT	Native	21.38	16.06	12.77	22.90	17.22	41.29	40.45	**23.44**	22.11	27.87	**34.17**
		SWR	**19.84**	**15.94**	**12.72**	**22.26**	**16.61**	**39.65**	**38.98**	24.31	**21.83**	**27.12**	34.56
PTB test	OWL + Wanda	Uniform	198.78	109.84	47.53	**155.60**	**140.71**	76.97	73.93	**120.75**	44.33	41.87	**58.06**
		SWR	**190.47**	**109.65**	**47.15**	156.90	142.42	**75.55**	**70.84**	123.13	**43.91**	**41.84**	58.15
	OWL + SparseGPT	Native	215.39	94.56	47.43	943.03	170.60	59.26	56.57	**103.23**	31.16	31.45	**39.25**
		SWR	**165.53**	**84.66**	**46.77**	**647.10**	**144.43**	**58.71**	**51.73**	110.75	**29.46**	**30.37**	41.95

O CALIBRATION SET SCALING

Figure 2(c) shows four representative models. Here we report the complete results for all 11 models. For each model, we construct nested C4 train calibration sets with $n_{\text{cal}} \in \{16, 32, 64, 128, 256, 512, 1024\}$ sequences, each of length 2,048. A larger calibration set contains all sequences from the smaller sets plus additional sequences from the same ordered C4 train draw.

At every value of n_{cal}, Uniform and SWR use the same fixed Wanda pruning mask at 70% sparsity, the same reconstruction solver, and exactly the same calibration sequences. Both are evaluated on the full WikiText-2 test set. Therefore, the only quantities that change across columns are the number of calibration sequences and the reconstruction weights.

Table 37 shows that PPL generally decreases as the calibration set grows, although not every row is monotonic. The largest reductions occur between the smallest set sizes, while the changes become smaller at larger n_{cal}. Larger calibration sets therefore usually lower PPL for both configurations.

The comparison between Uniform and SWR becomes more consistent as the calibration set grows. SWR has lower PPL in 3/11 models at $n_{\text{cal}} = 16$, 9/11 at $n_{\text{cal}} = 32$, 10/11 at $n_{\text{cal}} = 64$ and 128, and all 11 models at $n_{\text{cal}} = 256$, 512, and 1,024. Across all seven set sizes, SWR is better in 65 of 77 matched comparisons.

The remaining exceptions occur at small calibration set sizes. L1-30B favors Uniform from $n_{\text{cal}} = 16$ through $n_{\text{cal}} = 128$, but favors SWR from $n_{\text{cal}} = 256$ onward. L2-13B favors Uniform at $n_{\text{cal}} = 16$ and $n_{\text{cal}} = 32$, and then favors SWR for every larger set size. Several other models favor Uniform only at $n_{\text{cal}} = 16$. Thus, a small calibration set can change the ordering between the two configurations,

Table 35: Complete WikiText-2 test PPL ↓ for nested subsets from the occurrence sensitivity ranking shared by all modules. At each retained fraction, Unit and SWR use exactly the same selected occurrences; bold marks the lower PPL.

Included	Weights	L1-7B	L1-13B	L1-30B	L2-7B	L2-13B	L3-8B	L3.1-8B	Mis-7B	OPT-6.7B	Q2.5-7B	Q3-8B
10%	Unit weight	39.89	29.13	15.54	38.20	29.20	**65.62**	64.45	38.55	51.90	29.51	**37.31**
	SWR	**39.52**	**28.28**	**15.42**	**37.80**	**29.14**	67.55	**64.33**	**37.77**	**51.84**	**29.15**	37.54
20%	Unit weight	39.53	28.85	15.32	37.83	28.63	**63.98**	62.80	38.31	50.33	29.15	36.76
	SWR	**38.61**	**28.15**	**15.21**	**37.44**	**28.58**	64.97	**62.05**	**37.56**	**49.89**	**28.68**	**36.65**
30%	Unit weight	39.85	28.91	15.32	37.71	28.47	**63.64**	62.66	38.58	50.06	29.15	37.01
	SWR	**38.54**	**28.17**	**15.24**	**37.36**	**28.40**	64.52	**61.56**	**37.64**	**49.54**	**28.60**	**36.31**
40%	Unit weight	39.92	28.97	15.38	37.79	28.49	**63.48**	62.57	38.66	49.87	29.18	36.85
	SWR	**38.37**	**28.22**	**15.23**	**37.33**	**28.34**	63.72	**61.24**	**37.59**	**49.33**	**28.56**	**36.32**
50%	Unit weight	40.12	28.98	15.36	37.79	28.45	**63.31**	62.36	38.68	49.94	29.27	37.33
	SWR	**38.35**	**28.21**	**15.18**	**37.32**	**28.28**	63.42	**61.08**	**37.55**	**49.23**	**28.53**	**36.33**
60%	Unit weight	40.25	29.04	15.36	37.78	28.45	63.51	62.38	38.43	50.06	29.32	37.36
	SWR	**38.31**	**28.26**	**15.17**	**37.33**	**28.28**	**63.26**	**60.90**	**37.48**	**49.16**	**28.56**	**36.13**
70%	Unit weight	40.38	29.07	15.30	37.71	28.45	63.57	62.27	38.54	50.14	29.30	37.27
	SWR	**38.33**	**28.31**	**15.15**	**37.29**	**28.28**	**63.21**	**60.80**	**37.58**	**49.15**	**28.56**	**36.14**
80%	Unit weight	40.47	29.03	15.19	37.74	28.45	63.58	62.12	38.65	50.19	29.36	37.33
	SWR	**38.31**	**28.33**	**15.11**	**37.28**	**28.27**	**63.20**	**60.75**	**37.62**	**49.14**	**28.54**	**36.20**
90%	Unit weight	40.60	29.03	15.08	37.80	28.47	63.96	62.10	38.71	50.30	29.44	37.36
	SWR	**38.33**	**28.36**	15.08	**37.29**	**28.28**	**63.30**	**60.67**	**37.62**	**49.16**	**28.59**	**36.25**
100%	Uniform	40.49	29.00	**14.91**	37.89	28.51	63.80	62.27	38.74	50.65	29.58	37.53
	SWR	**38.33**	**28.38**	15.01	**37.28**	**28.29**	**63.01**	**60.69**	**37.66**	**49.25**	**28.61**	**36.44**

Table 36: Complete WikiText-2 test PPL (↓) for the six priority dilution configurations in Table 7. The three two-column groups correspond to the Top 20%, Next 30%, and Bottom 50%. Within each group, the left column includes the indicated region at unit weight 1, while the right column changes only that region to its original SWR reconstruction weights ρ; bold marks the lower PPL.

	Top 20%		Next 30%		Bottom 50%	
Model	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6
L1-7B	39.53	**38.61**	38.62	**38.35**	38.62	**38.33**
L1-13B	28.85	**28.15**	28.61	**28.21**	28.68	**28.38**
L1-30B	15.32	**15.21**	15.19	**15.18**	**14.82**	15.01
L2-7B	37.83	**37.44**	37.65	**37.32**	37.68	**37.28**
L2-13B	28.63	**28.58**	28.43	**28.28**	28.42	**28.29**
L3-8B	**63.98**	64.97	**63.12**	63.42	63.20	**63.01**
L3.1-8B	62.80	**62.05**	61.37	**61.08**	60.86	**60.69**
Mis-7B	38.31	**37.56**	37.88	**37.55**	38.17	**37.66**
OPT-6.7B	50.33	**49.89**	49.63	**49.23**	50.01	**49.25**
Q2.5-7B	29.15	**28.68**	28.70	**28.53**	28.92	**28.61**
Q3-8B	36.76	**36.65**	36.40	**36.33**	37.10	**36.44**

but this behavior largely disappears as more calibration sequences are used. The small-set reversals are consistent with noisier sensitivity estimates, but this experiment does not isolate that cause.

Increasing n_{cal} does not make Uniform approach SWR within the measured range. At the three largest set sizes, SWR is better for every model, even though both methods have already benefited substantially from the additional calibration data. This shows that the benefit of SWR is not simply a consequence of using too little calibration data: adding more equally weighted sequences improves Uniform, but does not remove the benefit of assigning different weights to occurrences.

Table 38 reports wall time and peak GPU memory for the calibration set scaling experiment of Section 5 (Llama-2 7B, Wanda pruning mask, 70% sparsity, one NVIDIA RTX PRO 6000 Blackwell GPU). Uniform includes pruning mask construction followed by Uniform OBS reconstruction. SWR also includes the sensitivity pass, measured once per n_{cal} and added to the same reconstruction. Peak memory is the maximum PyTorch CUDA allocation over the whole run.

The time difference grows with n_{cal}, from 10.2 s at $n_{\text{cal}} = 16$ to 391.0 s at $n_{\text{cal}} = 1{,}024$. The memory difference decreases from 18.76 GiB at $n_{\text{cal}} \leq 128$ to 0.20 GiB at $n_{\text{cal}} = 1{,}024$. Some

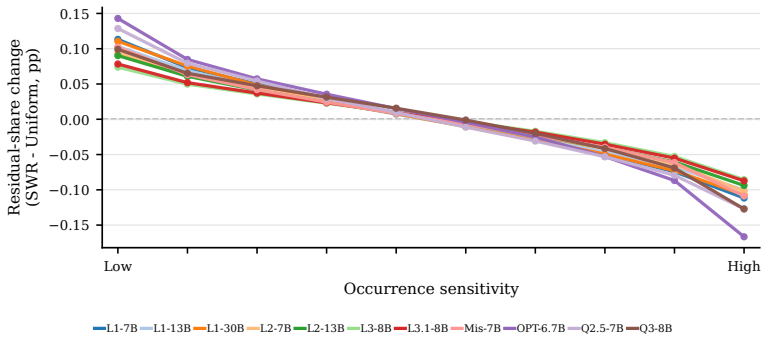


Figure 4: **SWR assigns a smaller share of squared reconstruction error to sensitive occurrences.** Evaluation occurrences are divided into ten groups of equal size within each module, from low to high sensitivity. Each curve denotes one model and reports the difference between SWR and Uniform in that group's share of the module's total squared reconstruction error. Values are averaged equally across modules. Negative values indicate a smaller residual share under SWR.

Table 37: Complete calibration set scaling results on WikiText-2 test PPL ↓. Each two-row block compares Uniform OBS and SWR on the same nested C4 train calibration set; bold marks the lower value.

Model	Configuration	16	32	64	128	256	512	1,024
L1-7B	Uniform	46.04	44.26	41.76	40.49	39.48	39.32	39.24
	SWR	**45.12**	**42.74**	**40.16**	**38.33**	**37.35**	**37.00**	**36.98**
L1-13B	Uniform	**34.31**	32.42	30.49	29.00	28.09	27.97	27.84
	SWR	34.90	**32.22**	**30.02**	**28.38**	**27.37**	**27.22**	**27.13**
L1-30B	Uniform	**17.26**	**16.57**	**15.76**	**14.91**	14.73	14.72	14.61
	SWR	17.77	16.63	15.78	15.01	**14.69**	**14.68**	**14.56**
L2-7B	Uniform	**43.63**	41.10	38.97	37.89	37.32	37.38	37.35
	SWR	44.24	**40.54**	**38.73**	**37.28**	**36.71**	**36.67**	**36.59**
L2-13B	Uniform	**32.62**	**31.15**	29.50	28.51	28.27	28.20	28.02
	SWR	33.70	31.34	**29.48**	**28.29**	**27.91**	**27.85**	**27.60**
L3-8B	Uniform	**79.87**	77.10	66.18	63.80	63.65	62.46	61.71
	SWR	80.36	**75.52**	**65.52**	**63.01**	**62.01**	**61.21**	**60.31**
L3.1-8B	Uniform	**72.36**	67.11	65.23	62.27	61.98	61.84	61.21
	SWR	72.50	**66.61**	**63.67**	**60.69**	**59.92**	**59.60**	**58.92**
Mis-7B	Uniform	46.68	44.90	39.20	38.74	38.19	37.72	37.69
	SWR	**45.36**	**43.54**	**38.54**	**37.66**	**36.87**	**36.45**	**36.33**
OPT-6.7B	Uniform	**58.55**	55.14	51.45	50.65	49.42	49.30	48.79
	SWR	59.62	**54.43**	**50.76**	**49.25**	**48.29**	**48.14**	**47.76**
Q2.5-7B	Uniform	34.33	31.93	30.53	29.58	29.07	29.04	28.91
	SWR	**34.31**	**31.45**	**29.57**	**28.61**	**28.05**	**28.00**	**27.89**
Q3-8B	Uniform	**42.63**	41.15	38.86	37.53	36.89	36.70	36.62
	SWR	44.05	**40.68**	**38.30**	**36.44**	**35.94**	**35.58**	**35.47**

peak memory values remain constant because the reported value is the maximum over independently measured stages. OBS memory does not exceed mask construction for Uniform until $n_{\text{cal}} = 256$ and does not exceed sensitivity computation for SWR until $n_{\text{cal}} = 1{,}024$.

P Lexical Structure of Occurrence Sensitivity

P.1 Occurrence sensitivity adjusted for position

Figure 3(a) uses the same 11 models and C4 train calibration occurrences as the main experiments. For each module and sequence, we divide the unnormalized occurrence sensitivity by its float64 mean over all 2,048 positions. We then average these values over reconstructed modules and select the most sensitive 20% of occurrences in each model.

Table 38: Cost after setup for calibration set scaling on Llama-2 7B at 70% sparsity on one NVIDIA RTX PRO 6000 Blackwell GPU. Each SWR total includes its independently measured sensitivity pass; memory is peak PyTorch CUDA allocation.

	Wall time (s)			Peak GPU memory (GiB)		
n_{cal}	Uniform	SWR	Δ	Uniform	SWR	Δ
16	210.0	220.2	+10.2	18.58	37.34	+18.76
32	220.9	238.6	+17.6	18.58	37.34	+18.76
64	247.7	274.9	+27.2	18.58	37.34	+18.76
128	295.9	348.3	+52.5	18.58	37.34	+18.76
256	397.8	498.6	+100.9	22.25	37.34	+15.10
512	604.6	801.2	+196.6	30.25	37.34	+7.10
1,024	1019.6	1410.6	+391.0	46.25	46.44	+0.20

Table 39: Frequency in the most sensitive 20% after adjustment for position. Values are ratios to the expectation at the same position and are medians over eligible models; 1.00 denotes average frequency.

Next target category	Global	Q	K	V	O	Gate	Up	Down
Proper noun	1.42	1.45	1.40	1.56	1.40	1.29	1.31	1.36
Common noun	1.15	1.14	0.89	0.87	1.18	1.17	1.18	1.19
Adjective	1.15	1.14	0.73	0.84	1.24	1.20	1.22	1.25
Verb	1.01	0.97	0.81	0.77	1.04	1.09	1.10	1.07
Adverb	1.01	0.97	0.87	0.85	1.10	1.09	1.11	1.06
Preposition	0.80	0.75	1.22	1.03	0.71	0.82	0.77	0.76
Conjunction	0.77	0.72	1.35	1.10	0.68	0.74	0.72	0.71
Pronoun	0.71	0.85	1.00	1.06	0.65	0.66	0.68	0.63
Determiner	0.63	0.67	0.91	1.14	0.58	0.62	0.60	0.62
Modal	0.68	0.74	0.86	0.65	0.70	0.72	0.69	0.66
Number	0.69	0.73	0.71	0.69	0.70	0.64	0.64	0.68
Punctuation	0.67	0.74	1.31	0.97	0.60	0.64	0.62	0.56

We analyze current positions 64–2,046 and label each occurrence by the token at position $t + 1$. The expected frequency in the most sensitive 20% is computed separately at each absolute position. The reported ratio is the observed frequency divided by the expectation at the same position. Categories are automatic Penn tags on English C4 train text. Proper noun denotes NNP or NNPS, not named entity recognition. Table 39 reports all categories and the results for each module type.

P.2 Reconstruction weights around nouns

For Figure 3(b), every module is normalized to mean one within each calibration sequence and clipped to $[0.2, 5]$. We average the weights across layers within K, V, Q, and O. The FFN row is the equal mean of Gate, Up, and Down. OPT uses Up and Down.

Figure 3(b) shows the selected example "during a match between England and France back in November." Q, O, and the FFN assign higher mean reconstruction weights one position before the proper nouns. K and V change little between the two positions. This example shows variation between module types. The aggregate analysis below tests the pattern over the full calibration set.

The aggregate analysis uses every noun at a word start whose offsets -3 through $+3$ remain inside positions 64–2,046. It contains 202,960 proper noun occurrences and 427,277 common noun occurrences across 11 models. We first average occurrences within each model and then give every model equal weight. For proper nouns, Q, O, and FFN have mean weights 1.146, 1.193, and 1.127 one position earlier, compared with 0.788, 0.901, and 0.888 at the noun. K is nearly unchanged (0.600 versus 0.604), and V is slightly lower at the noun (0.302 versus 0.287). For common nouns, K and V are higher at the noun (0.573 and 0.254) than one position earlier (0.449 and 0.247). Q, O, and FFN show the opposite pattern. Their values one position earlier are 0.935, 1.001, and 0.991, compared with 0.708, 0.726, and 0.753 at the noun.

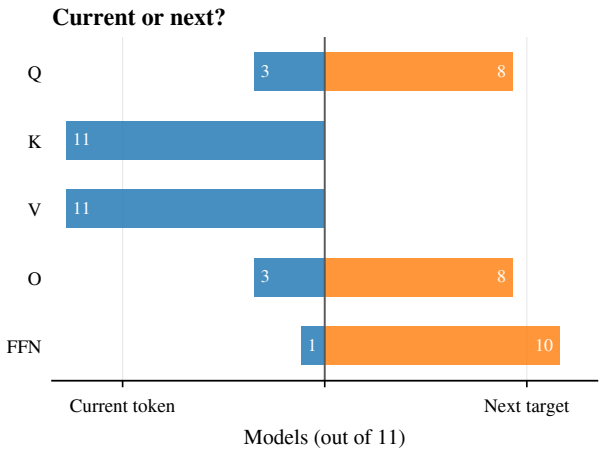


Figure 5: **Current-token and next-target identities differ in predictive value across module types.** Bars count the models for which either identity has the larger predictive increment on disjoint evaluation data after the stated controls.

Table 40: Predictive value of current-token and next-target identities for the most sensitive 20% of occurrences. Positive ΔAUC means that the next target has the larger predictive increment; negative values mean that the current token has the larger increment.

Module type	Current token	Next target	Median ΔAUC
Q	3/11	8/11	+0.010
K	11/11	0/11	-0.028
V	11/11	0/11	-0.015
O	3/11	8/11	+0.009
FFN	1/11	10/11	+0.010

P.3 Controlled analysis of current-token and next-target identities

Figure 5 asks whether current-token identity or next-target identity better predicts membership in the most sensitive 20% for each module type. We split sequences into two folds by sequence index. Token identity statistics for either fold are estimated only from the other fold; training features use a second split so no row encodes its own label. Token lookup rates use additive smoothing of 20, with the training prevalence for unseen identities. Logistic regression uses standardized features, L-BFGS, tolerance 10^{-7}, and 100 iterations. The controls are a cross-fitted absolute-position rate, normalized position and its square, current-token and next-target frequencies, and their byte lengths. We fit controls-only, current-token, next-target, and joint models, then average AUC over the two held-out folds. The reported value is the joint model's conditional AUC increment for the next target minus that for the current token. For FFN, we first average the available Gate, Up, and Down differences within each model.

Table 40 reports the controlled comparison. Current-token identity has the larger predictive increment for K and V in all 11 models. Next-target identity has the larger increment for the FFN in 10 of 11 models and for Q and O in 8 of 11. These are predictive associations rather than causal claims about module function. Lexical identity explains only part of occurrence sensitivity.

P.4 Layer-shared and module-specific reconstruction weights

Both configurations start from the same dense model and use the same Wanda pruning mask at 70% sparsity. They also use the same C4 sequences, OBS solver, damping, and WikiText-2 test evaluation. SWR retains module-specific reconstruction weights for each occurrence. The other configuration averages the SWR weights across reconstructed modules within each layer. Both configurations use the same normalization and $[0.2, 5]$ clipping.

Table 41: WikiText-2 test PPL ↓ under reconstruction with fixed pruning masks. The layer-shared configuration uses one set of reconstruction weights for every module in a layer; SWR retains module-specific reconstruction weights. Bold marks the lower value within each model.

Reconstruction weights	L1-7B	L1-13B	L1-30B	L2-7B	L2-13B	L3-8B	L3.1-8B	Mis-7B	OPT-6.7B	Q2.5-7B	Q3-8B
Layer-shared	39.15	28.79	**14.93**	37.61	28.58	63.79	61.26	37.68	50.23	28.66	36.68
SWR	**38.33**	**28.38**	15.01	**37.28**	**28.29**	**63.01**	**60.69**	**37.66**	**49.25**	**28.61**	**36.44**

Table 42: WikiText-2 test PPL ↓ for a static lexical prior with a fixed Wanda pruning mask. Values are arithmetic means over three calibration draws. SWR uses occurrence sensitivity from the target model. Bold marks the lowest PPL within each model.

Reconstruction weights	L1-7B	L1-13B	L1-30B	L2-7B	L2-13B	L3-8B	L3.1-8B	Mis-7B	OPT-6.7B	Q2.5-7B	Q3-8B
Uniform	40.49	29.00	14.91	37.89	28.51	63.80	62.27	38.74	50.65	29.58	37.53
Static prior	39.85	28.92	**14.79**	37.73	28.47	63.82	62.32	39.03	50.36	29.63	37.34
SWR	**38.33**	**28.38**	15.01	**37.28**	**28.29**	**63.01**	**60.69**	**37.66**	**49.25**	**28.61**	**36.44**

Figure 3(c) reports $100(P_{\text{shared}} - P_{\text{module}})/P_{\text{module}}$. Using layer-shared reconstruction weights increases PPL on 10 of 11 models, with a median increase of 1.0%. LLaMA-1 30B is the only exception. The only change is whether reconstruction weights remain module-specific. We do not test other averaging methods. Table 41 reports the corresponding PPL values.

P.5 Static lexical prior

We test whether the lexical structure above can provide reconstruction weights for a new model without computing occurrence sensitivity for that model. For each target model, the donor pool excludes the target itself and every model whose tokenized calibration inputs are exactly identical. Donor occurrence sensitivities are normalized within sequence and module and clipped to $[0.2, 5]$. For each module type and part-of-speech category, we first average within each donor model and then give donor models equal weight. Each target occurrence receives the current-token entry for K/V or the next-target entry for Q/O/FFN. Unmapped positions, positions 0–63, and the final position receive an initial weight of one. The target weights are then normalized within each sequence and clipped to $[0.2, 5]$.

All three configurations in Table 42 use the same fixed Wanda pruning mask at 70% sparsity, 128 C4 train calibration sequences, OBS reconstruction, and WikiText-2 test evaluation. Uniform assigns every occurrence unit weight. The static prior uses only the lookup estimated from donor models, whereas SWR uses occurrence sensitivity from the target model. The static prior does not use a backward pass to compute occurrence sensitivity on the target model. It still requires a calibration forward pass on the target model and OBS reconstruction. Every reported value is the arithmetic mean over three calibration draws.

The static prior lowers PPL relative to Uniform on 7 of 11 models, with a median relative reduction of 0.26%. SWR is lower than Uniform and the static prior on 10 of 11 models, with median reductions of 2.54% and 2.20%, respectively. Thus, the lexical analysis provides a static rule that can be used across models, but its benefit is small. Under this rule, most of the gap to SWR remains.

Q Large Language Models Usage in Our Paper

We gratefully acknowledge the use of large language models to assist in identifying relevant literature, improving the clarity and readability of the manuscript, and debugging code. All AI-assisted outputs were critically reviewed and, where appropriate, refined by the authors. Literature references were independently verified, and code modifications were tested. The authors ensure the accuracy and integrity of the final content.