LOST IN PRUNING: SILENT MULTILINGUAL COLLAPSE AND TRAINING-FREE RECOVERY FOR ONE-SHOT MOE PRUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

One-shot expert pruning shrinks the resident memory of sparse mixture-of-experts (MoE) models without retraining and is validated on benchmarks that resemble its calibration data. We show that this validation misses a silent multilingual collapse: a 25%-pruned Qwen3-30B-A3B keeps code near baseline and English perplexity within $\times 1.6$, yet answers Korean prompts in Chinese. The damage is a loss of language access rather than of knowledge alone: on 2,800 parallel MMLU-ProX items the pruned model loses 56 points in Korean but 15 in English, and still answers 74% of its Korean losses correctly in English. Fluency, knowledge, answer extraction, and translation direction fail differently; on this model generative MMLU-ProX also falls to near zero in Japanese and French, generation into an uncovered language degrades far more than translation out of it, and the collapse recurs across six MoE architectures and four scoring criteria, implicating calibration coverage rather than any single scoring rule. Controlled expert exchanges at fixed compression, each designed before its checkpoints were built, probe this account directly: restoring 28% of the experts on which the default mask and a multi-view coverage mask disagree closes about 80% of the Korean NLL gap between the two masks and restores Korean-script responses, but only 11% of the KMMLU gap. This language-before-knowledge ordering recurs on a 256-expert model, where it was predicted before the run, and is absent in a model whose default pruning does not collapse Korean generation. Building on this analysis, Cross-Observation Expert Retention (COER) recomposes cached per-domain calibration views on CPU without raw text or retraining; at matched compression and observation budgets (three seeds on Qwen3) it restores target-language generation under all four criteria, generalizes to five languages without labeled calibration examples in those languages, and recovers 34–77% of lost HAERAE accuracy under gate weighting, with rank-union retaining code at or above the Qwen3 default.

1 INTRODUCTION

Sparse MoE models decouple per-token computation from total capacity (Jiang et al., 2024; Qwen Team, 2025; Zhipu AI, 2025), yet serving still stores every expert. One-shot expert pruning removes the experts a calibration set ranks as unimportant, without retraining: on Qwen3-30B-A3B, removing 25% and 50% of experts cuts resident weight memory by 24% and 47% at unchanged per-token compute and, where KV-cache capacity binds, serves a saturating workload $1.57\times$ and $1.94\times$ faster (two runs within 3%; Appendix G.1). The method is validated on benchmarks that resemble its calibration data (Lasby et al., 2026; Liu et al., 2026b).

That validation is blind to what it does not measure. The reference-pruned model keeps code near baseline and English perplexity within $\times 1.6$ yet answers a Korean prompt in fluent Chinese, and on identical questions posed in both languages it loses 56 points in Korean against 15 in English while still answering most of its Korean losses correctly in English. The damage is not a uniform loss of quality but a disproportionate loss of access through the uncovered language, beside a smaller English loss that perplexity and code scores hide. We characterize this failure across six MoE families and four scoring criteria, diagnose how calibration coverage under-scores experts needed by uncovered languages, and then intervene: exchanging individual experts at fixed compression

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.

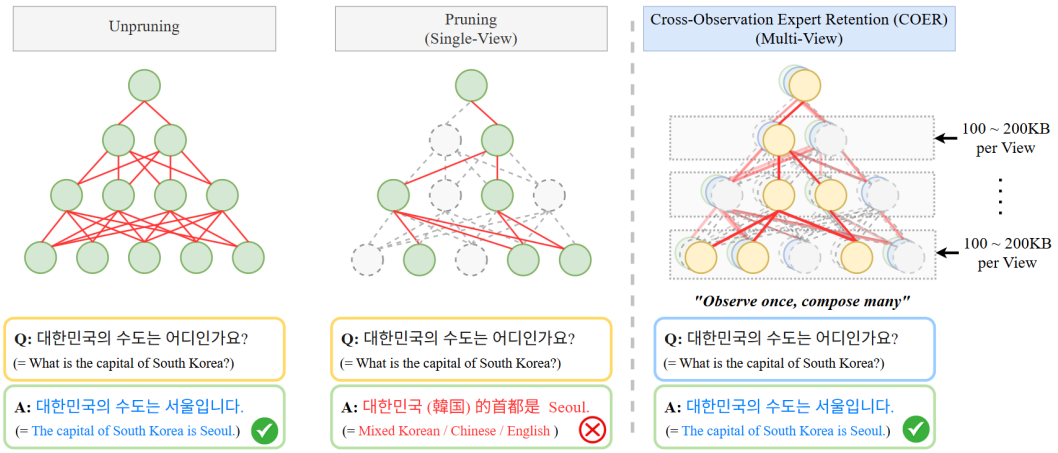


Figure 1: Silent multilingual collapse under one-shot expert pruning (Qwen3-30B, 25%). With single-view code calibration (middle), the pruned model answers a Korean question in mixed Korean/Chinese/English while code stays near baseline and English perplexity within $\times 1.6$. Composing cached calibration views (COER, right) restores the Korean answer at the same compression.

shows that a few well-chosen restorations return the language while benchmark accuracy needs most of them. Composing cached calibration views (COER) is the training-free remedy this analysis yields.

Contributions. First, we establish distinct fluency, knowledge, extraction, and translation failures across six MoE architectures (nine languages, three seeds, and three compression ratios on Qwen3) and show on 2,800 parallel items that the pruned model loses far more in Korean than in English yet answers most of its Korean losses in English—a loss of target-language access beside a smaller English loss that perplexity hides. Second, teacher-forced and source-ablation tests across three families trace the failure to underestimated target-side conditional saliency and reveal distinct fluency and knowledge regimes. Third, controlled expert exchanges at fixed compression, each design committed before its checkpoints were built, attribute most of the Korean likelihood gain to which experts are *restored* rather than removed, and show language returning after a small exchange while benchmark accuracy needs most of it—an ordering observed in the two tested models whose default mask collapses the language (Qwen3-30B, Qwen3.6) and not in one that keeps it (ERNIE). Fourth, the analysis yields a training-free remedy, COER, which recomposes cached 100–200 KB per-domain calibration views on CPU without raw text. At matched compression and observation budget, it restores target-language generation under all four criteria, including five languages without labeled calibration examples in those languages; controls show that observation diversity, not the composition rule, is the dominant factor given a diverse pool.

2 Related Work

One-shot MoE expert pruning. REAP scores experts by conditional gate-weighted activation norms and reports near-lossless English and code compression (Lasby et al., 2026); related work studies expert scoring, including Shapley- and coalition-based importance (Liu et al., 2026a; Zhang, 2026; Huang et al., 2025; Hyeon and Do, 2026; Goel et al., 2026), layer-adaptive pruning rates (Bai et al., 2025), calibration-free criteria (Liu et al., 2026b), pruning and merging (Li et al., 2025; Cao et al., 2024; Muzio et al., 2024; Gu et al., 2025), and distillation (Muralidharan et al., 2024). Multilingual pruning studies target dense weights (Choenni and Titov, 2025; Zeng et al., 2024; Kurz et al., 2026) or translation-specific MoE subnetworks (Koishekenov et al., 2023; Martin et al., 2026); none characterizes the fluency, knowledge, extraction, and directional failures of general-purpose expert pruning outside its calibration distribution. Prior work studies expert specialization and interpretation (Zoph et al., 2022; Dai et al., 2024; Lo et al., 2024; Herbst et al., 2026); we instead exchange experts at fixed compression to attribute recovered capabilities to the restored set.

Coverage-aware and recovery methods. Lang-Prune adds language-aware importance to dense pruning (Liang et al., 2026), TB-Coverage protects top experts per generic English corpus (Zeng et al., 2026), and Data-Aware Calibration mixes calibration data (Anon., 2026a); distillation pipelines (Bercovich et al., 2025; 2026; Zhang et al., 2026) require recovery training, and channel-level compression is complementary (Ding et al., 2026). We share the coverage premise but compose cached observations without raw data under matched budgets. Coalition-aware scoring and adaptive per-layer rates change which experts are removed, not which languages calibration observes (Appendix C.1).

3 PRELIMINARIES

One-shot expert pruning and saliency criteria. A sparse MoE layer routes each token to the top-k of N experts. One-shot pruning scores experts on calibration observations and removes the lowest-ranked per layer. Our reference criterion, REAP, scores expert j by the conditional mean $S_j = \mathrm{E}_{x \in X_j}[\, g_j(x) \, \| f_j(x) \| \,]$ over tokens X_j selecting j (Lasby et al., 2026). Criterion controls replace only this function with frequency, MAN, or MSAN (Liu et al., 2026a); all other stages remain fixed (Fig. 2, gray upper path). We correct sigmoid-router reproduction (on GLM, top-k over raw logits matches the true expert set on only 27.5% of tokens; Qwen3 is unaffected; Appendix A).

Metrics tied to failure modes. Metrics follow the capability affected. Fluency and language consistency use held-out Korean Wikipedia perplexity normalized to each unpruned model (koPPL) and output-script fractions that detect language switching (Hangul: share of generated characters in Korean script). Knowledge (throughout, multiple-choice accuracy, not stored facts) uses KMMLU (four-option, 0.25 chance), KoBEST (five understanding tasks), and HAE-RAE Bench (HAERAE; Korea-specific knowledge and vocabulary) (Son et al., 2024; Jang et al., 2022; Son et al., 2023); HAERAE-history isolates historical knowledge. MMLU-ProX zero-shot generative exact match is decomposed into valid-answer extraction and accuracy conditional on extraction (Xuan et al., 2025). Bidirectional FLORES-200 chrF++ probes target-language generation and an operational comprehension proxy (NLLB Team, 2022). HumanEval+ and MBPP+ measure code retention with EvalPlus-hardened greedy pass@1 (Chen et al., 2021; Austin et al., 2021; Liu et al., 2023).

Models. We evaluate six sparse MoE architectures spanning 8–256 routed experts per layer and 21B–236B parameters: Qwen3-30B-A3B, Qwen3.6-35B-A3B, GLM-4.5-Air, ERNIE-4.5-21B, Mixtral-8x7B, and DeepSeek-V2 (Qwen Team, 2025; 2026; Zhipu AI, 2025; Baidu, 2025; Jiang et al., 2024; DeepSeek-AI, 2024). Qwen3-30B is the primary controlled setting (three calibration seeds, three compression ratios); DeepSeek-V2 uses a reduced 256-batch protocol (Table 1).

Calibration data and baselines. Calibration views use evol-codealpaca, KoAlpaca, Aya, and Korean Wikipedia (Luo et al., 2023; Lee, 2023; Singh et al., 2024). The released single-view default is evol-codealpaca-v1. We use lm-evaluation-harness where applicable (Gao et al., 2024). Four interchangeable criteria, frequency, REAP, MAN, and MSAN, share the same pruning ratio, selection, and surgery (Lasby et al., 2026; Liu et al., 2026a). Direct comparisons include TB-Coverage, AIMER, and a resident-parameter-matched dense reference (Zeng et al., 2026; Liu et al., 2026b; Mistral AI, 2025).

4 SILENT MULTILINGUAL COLLAPSE AND FAILURE MODES

Beyond its calibration distribution, standard one-shot expert pruning fails severely and invisibly to calibration-aligned evaluation, in distinct modes that are most pronounced in the finest-grained MoEs studied here (Table 1).

Experimental protocol. The reference setup uses 25% pruning and about 2.1M calibration tokens (Section 6: 40% and 50%). The primary Qwen3 comparison uses seeds 42–44 (prune-set Jaccard 0.937–0.939, benchmark std at most 2.6 points); other configurations are single-seed where marked.

Korean case study (Qwen3-30B). Korean perplexity rises to $\times 14.4 \pm 3.8$ (English control $\times 1.6$); generation switches to fluent Chinese (Hangul 79.6%$\rightarrow$1.3%); KMMLU and HAERAE drop 17.6

Model (experts/layer)	koPPL ratio	Δ HAERAE	Δ KMMLU	Δ HumanEval+	Δ MMLU-ProX ko (gen.)
Qwen3-30B (128)	$\times 14.4 \pm 3.8$	-28.1 ± 1.5	-17.6 ± 0.3	-3.2 ± 0.3	-58.0
GLM-4.5-Air 106B (128)	$\times 5.5$	-32.7	-16.8	-0.7	-38.6
ERNIE-4.5-21B (64)	$\times 1.9$	-18.5	-9.2	-0.6	-16.9
Mixtral-8x7B (8)	$\times 1.25$	-9.1	-8.0	-7.3	-9.4
DeepSeek-V2 236B (160)	$\times 2.4$	-28.3	-18.7	-2.5	–
Qwen3.6-35B (256)	$\times 8.89$	-14.0	-5.5	-0.6	-71.1

Table 1: Cross-family damage after 25% pruning with the default English-code calibration: task columns are Pruned minus Unpruned (points), koPPL is Pruned/Unpruned on held-out wiki. Qwen3-30B is mean±std over seeds 42–44; other rows are single runs. DeepSeek-V2 uses a reduced 256-batch calibration; its always-on shared experts buffer perplexity but not knowledge, and its generative ProX-ko baseline sits near the ten-way chance floor (7.5%), so that column is uninformative (−−).

and 28.1 points; French and Japanese perplexity rise $\times 21.6$ and $\times 7.9$ while fallback Chinese rises only $\times 2.3$; generative MMLU-ProX for Korean, Japanese, and French falls to 0.0–0.5%.

The collapse occurs under all four scoring criteria: frequency yields $\times 8.0$ perplexity, MAN and MSAN $\times 4.6$ with Hangul below 7%, Chinese fallback, and knowledge at or below the gate-weighted reference (Liu et al., 2026a). No expert has zero calibration frequency; affected experts are active but under-scored (Appendix Table S11).

Two collapse modes (extraction vs. knowledge). MMLU-ProX strict accuracy equals extraction rate times conditional accuracy. Qwen3's near-zero score is a format-language collapse (extraction $0.80 \rightarrow 0.003$, outputs in Chinese script), so knowledge is unmeasurable; ERNIE keeps extraction ($0.84 \rightarrow 0.79$) but loses conditional accuracy ($0.48 \rightarrow 0.31$), an in-format accuracy loss rather than a language switch. Qwen3.6 reproduces the dissociation (Korean/Japanese ProX 0.713/0.704 → 0.002/0.000; en→ko retains 6.5%, ko→en 97%).

Generation vs. comprehension, across languages. On 400 FLORES-200 sentences per direction, default pruning degrades translation into four uncovered languages (en→X retention 10/20/34/79% for ko/ja/fr/zh) far more than out of it (X→en retention 76/84/88/93%). This pattern holds for every language in Appendix Table S7. Korean and Japanese switch to Chinese script; Hebrew receives English and Chinese. The fallback varies, but output no longer matches the prompt language.

On 2,800 parallel MMLU-ProX items (identical questions and answer letters in English and Korean), a rebuilt default model loses 56 points in Korean but 15 in English (paired 95% CIs within ±2; two more seeds: 44/57 and 14/15), still answering 74% of its Korean losses correctly in English; English MMLU drops 10.5 points under the code-only default but 3.6 under Korean-only calibration, so a dedicated English view is not what preserves English accuracy here (Appendix C.5).

Severity and expert granularity. Multiple-choice knowledge damage grows from the 8e model (9.1 HAERAE points) through 64e (18.5) to the 128e and 160e models (28–33; Table 1), and the two 128e models differ 3.5× in size yet suffer comparable damage, so granularity separates these families better than scale. The trend is not monotone at the finest pool: the 256e Qwen3.6, with an always-on shared expert and stronger multilingual pretraining, loses 14.0 HAERAE points while its generative collapse is the most severe (ProX-ko -71.1), and DeepSeek-V2's shared experts hold perplexity to $\times 2.4$ while knowledge falls as far (an association across co-varying families, not a causal claim). Mixtral has the least stable prune set yet the smallest drops (Appendix Figs. S4 and S7).

5 Diagnosing the Collapse: Underestimated Saliency and Dual Recovery Regimes

Why does a saliency estimate that is faithful in-domain delete another language's ability? Default pruning removes 27–30% of Korean routing mass per layer in all four models (Jaccard against Korean calibration 0.149–0.330). Varying the Korean share of a fixed-budget code:Korean calibration

mixture (plain mixtures, not COER; the 95:5 and 90:10 points confirmed by re-observation, Sec. 6) reveals two recovery regimes (Appendix Fig. S1): a 5% share cuts perplexity from $\times 17.8$ (seed 42) to $\times 1.50$ while preserving code—97.0% of the perplexity gap, 86% in log space—but closes 19.8% of the knowledge gap: HAERAE-history stays at 0.309 against the unpruned 0.633 and reaches only 0.335 and 0.495 at 10% and 50% (seeds 43 and 44: 89%/15% and 97%/5%; Fig. S1). Removed routing mass on knowledge text correlates with damage across six natural calibrations ($r = -0.863$, $\rho = -0.829$, $n = 6$; exact two-sided permutation $p = 0.058$, $n = 6$, exploratory), yet optimizing it directly fails to close the gap—a Goodhart result.

A target-side Korean token-prediction deficit. Similar gate mass is removed from Korean tokens in generation and comprehension (Appendix Table S8); generation-minus-comprehension differences are $+0.034$, -0.007, and -0.015 for Qwen3, GLM, and ERNIE. Allocation volume cannot explain the directional gap.

Feeding the gold target throughout FLORES evaluation removes free-running error propagation; on Qwen3, pruning still raises Korean-target NLL by 2.77 nats per token (95% CI $[2.64, 2.90]$) but English-target NLL by only 0.42 ($[0.30, 0.54]$). The first target token rises by 0.80 and later terciles by 2.61, 2.81, and 2.78—a flat profile that rules out compounding as the main explanation.

Because the two directions change both role and target language, the comparison localizes a target-side Korean deficit without isolating an internal pathway; source ablation (shuffling the Korean source) retains 0.90 (95% CI $[0.82, 0.98]$) of the English-target source sensitivity after pruning.

The signature replicates on ERNIE and GLM: pruning raises en→ko target NLL by 0.85 and 1.25 nats but ko→en by only 0.27 and 0.47, with disjoint confidence intervals, flat positional profiles, and 90–110% retained source sensitivity across all three families. These results motivate retaining experts supported by complementary calibration views rather than a single in-domain saliency estimate.

6 A Training-Free Remedy: COER at Fixed Compression

We therefore propose Cross-Observation Expert Retention (COER), a criterion-agnostic framework that composes cached domain views into one expert-selection decision while keeping the criterion, pruning ratio, observation budget, and surgery fixed (Fig. 2): mixture synthesis reconstructs weighted empirical unions from cached statistics, rank-union protects experts salient in any view, and the controls below separate the two: given a diverse pool the composition rule matters little, whereas with only code and Korean views the rule decides third-language recovery (French ProX-lite 0.532 for rank-union vs. 0.000 for the mixture).

Mixture prescription. A 50:50 code:Korean calibration keeps code near baseline (HumanEval+ 0.876 ± 0.014 vs. 0.896) and recovers 57–77% of HAERAE damage on Qwen3 and GLM and 34% on ERNIE, whereas Korean-only calibration loses up to 40.8 code points; the two-view mixture leaves a third-language hole (Qwen3 French 0.000).

Closed-form mixture synthesis (exact over cached observations). For the gate-weighted criterion saliency is a routed-token conditional mean, so any mixture's saliency follows in closed form from per-view sufficient statistics. For view d and expert j, store routed-token count $n_{dj} = \sum_t \mathbf{1}[j \in \mathrm{TopK}(t)]$ and the gate-weighted activation-norm sum $u_{dj} = \sum_{t:\, j \in \mathrm{TopK}(t)} g_j(t)\, \|f_j(t)\|_2$. For view weights $\alpha_d \geq 0$, the score on their empirical union is $S_j(\alpha) = \big(\sum_d \alpha_d\, u_{dj}\big) / \big(\sum_d \alpha_d\, n_{dj}\big)$. Here α_d weights view d's routed tokens in the intended mixture: equal-budget views with $\alpha_d{=}1$ reproduce the honest mixture's statistics exactly, and for views of unequal size $\alpha_d \propto$ (target share of d)/(routed tokens of d); an expert routed in no view scores 0 and is pruned first. We prune the smallest layerwise scores. The identity follows by weighting each view's numerator and denominator in the routed-token conditional mean. Against independently sampled, re-observed mixtures, synthesized prune sets reach Jaccard 0.970–0.981 (two views) and 0.937 (four views), and benchmarks differ by at most 1.2 points for two-view mixtures and 1.7 for the four-view one (French ProX-lite 0.551 vs. 0.534). It transfers to ERNIE and GLM. MAN and MSAN replace the numerator with $\sum \|f\|_2$ and $\sum \|f\|_2^2$ (Appendix E).

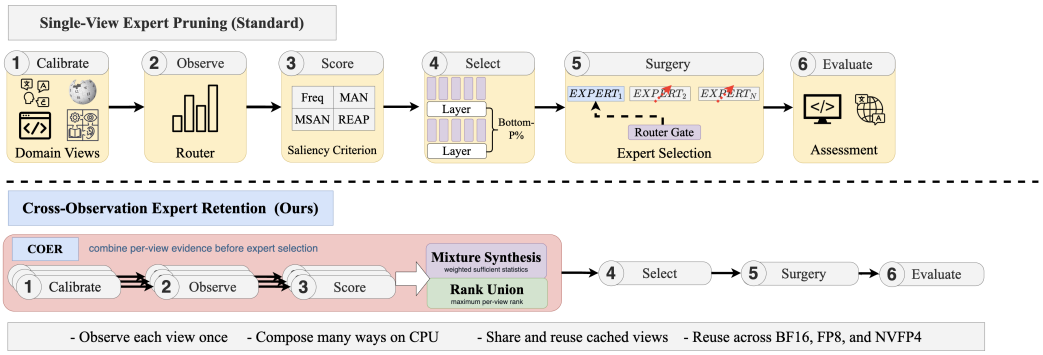


Figure 2: Unified comparison of standard single-view expert pruning and COER. Standard pruning passes one view through an interchangeable saliency criterion and then applies selection, surgery, and evaluation. COER observes multiple views independently and inserts one of two alternative composition rules before the unchanged selection step. Mixture synthesis reconstructs a weighted virtual mixture from cached sufficient statistics, whereas rank-union uses each expert's best per-view saliency rank. The 100–200 KB views are observed once, shared, and recomposed on CPU, and can be reused across BF16, FP8, and NVFP4 without re-observation.

Rank-union over diverse observations. Within each layer ℓ, let $r_{\ell dj}$ be expert j's saliency rank under view d (larger is more salient); rank-union prunes the experts with the smallest $R_{\ell j} = \max_d r_{\ell dj}$, so one high rank cannot be averaged away (Appendix Fig. S2). Two views use evol-code and KoAlpaca; four add Aya and Korean Wikipedia, receiving 512 and 256 batches per view at the matched 1024-batch budget. Rank-union matches mixture fluency and code; on KMMLU it reaches 0.505 at the matched 1024 batches (seed 42) vs. 0.530 for the mixture (0.511 ± 0.002 over three seeds at 2048 batches), exceeds the default on HumanEval+ (0.880 ± 0.003 vs. 0.864 ± 0.003, three seeds), and restores en→ko to 28.7 and en→ja to 26.0 chrF++ (mixture 29.7/28.1; default 3.2/5.7). On Qwen3.6 the four-view pool restores en→ko to 99% of baseline with code at or above baseline.

The four-view pool also closes the third-language hole: French MMLU-ProX-lite 0.551 at 1024 batches (seed 42) vs. 0.000 for the code+Korean mixture at the same budget (0.551 ± 0.009 over three seeds at 3328 batches; Table 2), en→fr 63.0 chrF++ (94% of baseline) vs. the 22.5 default, and 88–93% generation retention for Hebrew, Arabic, Hindi, Russian, and Swahili vs. 3–18% by default—Hebrew although no view targets it (three-seed std $\leq$2.0). A leave-language-out control removes the Aya rows labeled French, Hindi, Russian, Swahili, or Arabic (61 of 71 categories remain). Across three held-out observations, generation into all five languages retains 87–93% of baseline, French ProX is 0.540 ± 0.018, Korean perplexity $\times 1.16$, and HumanEval+ 0.882 ± 0.003: recovery does not require labeled examples of the evaluated languages.

Transferring the remedy to gate-free criteria. Applied to MAN and MSAN (Liu et al., 2026a), COER cuts Korean perplexity from $\times 4.6$ to $\times 1.2$–$\times 1.4$, raises Hangul from at most 7% to 92–93%, restores en→ko to about 90% of baseline, and retains code, but KMMLU stays at 0.28–0.31 versus $0.40 \rightarrow 0.53$ under gate weighting. Across three seeds and four criteria, composition restores generation, activation-based scores retain code, only gate weighting recovers knowledge, and frequency sacrifices code (Appendix Tables S11–S12).

Same-pool control. In the four-view configuration the French recovery is a property of the pool: a token-balanced mixture of the same four observations reaches 0.539 (re-observed four-way mixture 0.534) against the rank-union's 0.551, with KMMLU 0.491 vs. 0.482, koPPL $\times 1.11$ vs. $\times 1.17$, and HumanEval+ 0.854 vs. 0.878 (Table 2); with this pool the rule mainly trades code for Korean fluency, and only with code and Korean alone does it decide French (rank-union 0.532, mixture 0.000). The gain is not budget-driven, four rules over identical observations differ by about four points (Appendix Table S2), and exact minimax over removed mass improves fluency but lowers KMMLU by 6–9 points—routing mass predicts fluency better than knowledge.

Qwen3-30B setting	koPPL ratio	HumanEval+ score	KMMLU score	MMLU-ProX-lite fr (gen.) score
Unpruned baseline	1.00	0.896	0.577	0.636
25% pruned, 1024-batch obs.:				
default setup (REAP)	14.4	0.864	0.401	0.000
mixture 50:50	1.22	0.876	0.530	0.000
rank-union 2-view	1.30	0.884	0.505	0.532
rank-union 4-view	1.17	0.878	0.482	0.551
mixture 4-view (same observations)	1.11	0.854	0.491	0.539

Table 2: Absolute scores for Qwen3-30B remedies (koPPL normalized to unpruned); lower rows share 25% compression and a matched 1024-batch budget. Default and 50:50 mixture: three-seed means; rank-union rows: seed 42; mixture 4-view: the same four seed-42 256-batch views as rank-union 4-view, composed as a token-balanced mixture (Blackwell node, where rank-union 4-view reproduces French 0.551). French ProX-lite: zero-shot generative exact match on 588 items. Other budgets and the 2048-batch 2-view run: Appendix Table S1.

Robustness to compression ratio. Qwen3 collapse worsens with compression (default koPPL $\times 17.8 \rightarrow \times 38 \rightarrow \times 131$ at 25/40/50%, seed 42); mixture and rank-union hold it to $\times 2.0$ and $\times 1.9$ at 50%, retain 5–18 KoBEST and HAERAE points, and keep HumanEval+ ≥ 0.854, while KMMLU converges near 0.31 for every evaluated method—a floor not tested against a matched-compression oracle (Appendix Table S9, Fig. S6).

Data, rules, and a dense reference. A TB-Coverage-style adaptation in our observation space with a generic-corpus setting modelled on the original (English C4 + English Wikipedia in place of its WikiText-2 + C4; no Korean) protects Korean fluency (koPPL $\times 1.9$) but not KMMLU (0.370, below the default) and loses code (HumanEval+ 0.695): in the views tested, the fluency half behaves as a code-versus-natural-language domain effect that the generic English views repair, whereas KMMLU recovered only with target-language views (HAERAE improves under generic English too). The rule adds little beyond the data: a TB-Coverage re-implementation on the code+Korean observations trails rank-union by 1–4 points and costs 9.2 code points in the generic setting (Table S2). A resident-parameter-matched dense reference, Mistral Small 3.1 (Mistral AI, 2025), trails leave-language-out rank-union on code and Korean multiple-choice knowledge at about 7$\times$ more activated parameters but leads on generative axes (Appendix Table S10). In practice: keep the deployment view (dropping it costs 2–4 HumanEval+ points over three seeds), add a broad multilingual view (dropping Aya halves French translation), and add the target if known (KoAlpaca: 1–2 KMMLU points; Appendices C.3, C.4). Weight-only FP8/NVFP4 leave selection, collapse, and recovery unchanged; the 38%/21% packed-weight figures are projections, not measured footprints (Appendix I).

7 Anatomy of Recovery: Which Experts, in What Order

Section 5 supports the mechanism with correlational evidence and Section 6 measures COER on whole checkpoints; we now intervene directly, exchanging individual experts at fixed compression, to ask which retained experts carry the recovery and in what order capabilities return. Section 5 leaves two readings—removed experts carried Korean-specific function, or the survivors' routing was disturbed; within this fixed-compression design and candidate set, the 2×2 exchange below shows that the Korean likelihood axes are more sensitive to which experts are restored than to which are removed, while script share, KMMLU, and code also respond to the removed set; it does not by itself exclude routing disturbance. The answer is the expert-level counterpart of Section 5's dual regime: a small, well-chosen restoration repairs generation, and knowledge returns only after most of the disagreement set is restored.

Expert-exchange design at fixed compression. Start from a seed's default mask B and its matched-budget four-view rank-union C (Section 6); both keep the same number of experts per layer, so the restore candidates $P_\ell = \text{pruned}_B \cap \text{kept}_C$ and removal candidates $Q_\ell = \text{kept}_B \cap \text{pruned}_C$ have equal size. An exchange of dose m_ℓ restores m_ℓ experts from P_ℓ and removes m_ℓ from Q_ℓ, so every condition keeps 96 of 128 experts per layer. Doses follow $m_\ell(t) = 4 + \lceil t\,(|P_\ell| - 4) \rceil$, $t \in \{0, \frac{1}{3}, \frac{2}{3}, 1\}$, i.e. 28, 54, 78, and 100% of $\sum_\ell |P_\ell|$ (t=1 is C). The removal order is shared (Q_ℓ ascending in C-score); the *restore* order varies: **C-rank** (descending rank-union score), **B-saliency** and **B-frequency** (from B's own 1024-batch observation), and random orders fixed before

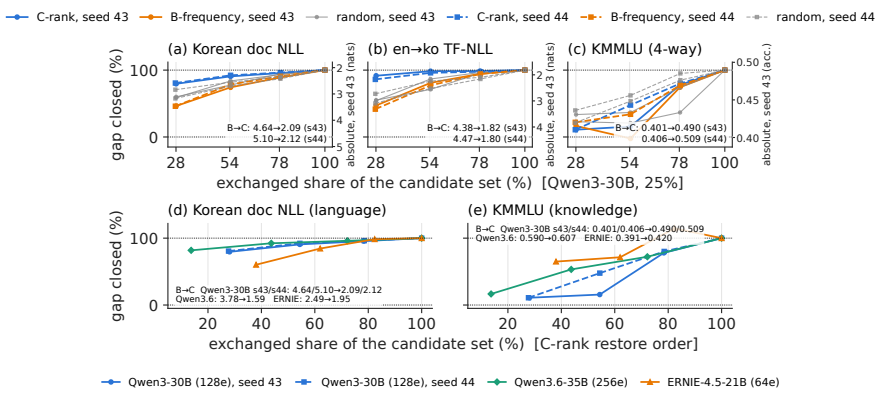


Figure 3: **(a–c)** Language and knowledge recover on different schedules (Qwen3-30B, 25%; seed 43 solid, 44 dashed): share of the $B \to C$ gap closed vs. exchanged share of the candidate set; right axes give the seed-43 absolute scale and the boxes list both seeds' $B \to C$ endpoints, so a closed gap is not recovery to the unpruned model. At 28% C-rank has closed 80/81% of Korean NLL and 92/86% of en→ko but 10.9/10.9% of KMMLU, which reaches 78/80% at the 78% dose. **(d–e)** C-rank curves for three families: Qwen3-30B and Qwen3.6, whose default masks destroy Korean generation, recover language before KMMLU; ERNIE, whose default mask keeps Korean, does not.

any build. Checkpoints come from the unchanged surgery, verified tensor-by-tensor against the original weights (14,259 tensors, zero mismatches); each design was committed before its checkpoints were built. We report the share of the $B \to C$ gap closed on Korean document NLL, en→ko teacher-forced NLL, ProX-ko-lite accuracy and extraction, Korean-script share (≥80% Hangul), KMMLU, and HumanEval+ (averaging units in Appendix K).

Restore selection dominates the Korean likelihood gain. A 2×2 cross-exchange (seed 42, m=4) swaps the restore and removal sets between C-rank and each of three random orders. Swapping the *restore* set accounts for most of the targeted-vs-random gap (Korean perplexity 13.7/15.5/6.0 of 14.2/16.8/6.3 points; token-weighted en→ko excess NLL 0.75–1.27 nats). With the C-rank restore fixed, changing the removal set moves perplexity by at most 0.24 and en→ko by at most 0.035 nats; with a random restore fixed the perplexity change reaches 1.38. Removal also shifts script share, KMMLU (±1–2 points), and code (up to 7 HumanEval+ problems) in some quartets; the two Korean likelihood axes, however, follow which experts are restored.

Which restore order? At the smallest dose (seeds 43 and 44) C-rank beats B-saliency, B-frequency, and five random restores on Korean document NLL and en→ko NLL in all 14 arm×seed comparisons per axis (CIs exclude zero, also on fresh documents and pairs) and is the only order that brings the Korean response language back (0.80/0.76 of responses in Korean script vs. at most 0.49). Scored on B's own 1024-batch observation, the B-only selectors stay 0.84–1.03 nats behind on Korean NLL and 1.14–1.60 on en→ko (Appendix Table S19). KMMLU reverses: among the m=4 exchange arms C-rank has the lowest observed KMMLU in both seeds (-0.50 and -1.14 points against the 1024-batch B-frequency control, p=.052 and 3.9×10^{-6}), above B and far below C.

The recovery schedule. Figure 3 traces the four doses. At 28% of the candidate set C-rank has closed 105/106% of the response-language gap (above the full C), 92/86% of en→ko NLL, 80/81% of Korean NLL, and 65/72% of ProX-ko accuracy, but 10.9/10.9% of KMMLU, which reaches 78/80% at the 78% dose. Both seeds show the same ordering—NLL and response language before KMMLU—with different intermediate KMMLU trajectories (15.6 vs. 47.7% at 54%) and no consistent HumanEval+ trend (139–146 of 164 across the 26 distinct policy–dose checkpoints). The m=4 KMMLU deficit is thus a difference in curve shape rather than a fixed trade-off—the knowledge has not arrived yet; against B-frequency it reverses by 54% ($+1.59/+1.41$ points, $p < 10^{-8}$), against random orders it is seed-dependent at 78%. Free generation confirms the likelihood ordering: on 600 fresh FLORES sentences en→ko chrF++ rises B 3.2/3.5 → B-frequency 9.0/8.2 →

C-rank 24.8/23.9 $\rightarrow$ C 29.6/29.7 in both seeds. B-frequency shows why one score misclassifies these models: at 28% it recovers 68–74% of ProX-ko accuracy with 0% of the Korean response language.

Compression ratio. At 50% compression (28 checkpoints, 9,651 tensors verified each) the language half is unchanged—the smallest dose (24%) closes 78.0/79.5% of Korean NLL, 83.9/85.2% of en→ko, and 96/107% of the response language—while every KMMLU value lies in 0.209–0.362 (chance 0.25, unpruned 0.577), so the $B \rightarrow C$ knowledge gap cannot be fractioned: Section 6's floor, reached by every restoration order. B-frequency does not restore the response language at this ratio.

Two more families: the schedule tracks the collapse regime. We ported the design unchanged to ERNIE-4.5-21B (64 experts) and, after committing the hypothesis, to Qwen3.6-35B (256 experts). On ERNIE the ordering does *not* appear: at the smallest dose (38%) C-rank closes 60.2% of Korean NLL, 76.5% of en→ko, and 65.0% of KMMLU: document NLL and KMMLU are within five points and en→ko leads KMMLU by 11.5, far less separated than on Qwen3 (Fig. 3d–e; Appendix Table S22). The endpoints suggest why: ERNIE's default mask keeps 88% of responses in Korean script (Qwen3-30B: 1%), its NLL gap is 4–5 times smaller (0.54 vs. 2.54 nats), and full C lowers script share, extraction, and HumanEval+ (145→137), so the four-view rank-union is not cost-free here (Section 8). For Qwen3.6 we predicted, before the exchange runs, that its default mask—ProX-ko accuracy 0.0017, no Korean-script responses, yet KMMLU 0.590 vs. 0.607 for C—should show the ordering again. It does, more sharply: at 13.7% of candidates C-rank closes 82–99% of every language axis (Korean NLL 81.8%, en→ko 95.7%, ProX-ko accuracy 99.2%) but 16.6% of KMMLU; its $B \rightarrow C$ KMMLU gap is only 1.7 points, so this fraction is not a precise rate. Across the three tested families at 25%, early language recovery relative to the KMMLU gap appears in the two whose default masks collapse Korean generation and is absent on ERNIE—consistent with the collapse-regime reading, though family, expert count, and smallest dose (38/28/13.7%) vary together, so expert count is not isolated. C-rank improves both Korean NLL axes over every tested comparator at each partial dose in all three families (CIs exclude zero; all policies coincide at full exchange), with effects an order of magnitude smaller on ERNIE; two 50% Korean-NLL comparisons are unresolved.

8 Discussion and Limitations

Implications for practice. Attention-only LoRA on the same Korean corpus (two calibration seeds, 62–65 GPU-minutes) restores Korean script but leaves koPPL at 13.9–14.6 (COER 7.7), recovers 0–3 knowledge points (COER 10–12), and lowers HumanEval+ by 9–11 points, so it does not match COER's joint recovery and code retention (Appendix C.6).

Scope and limitations. Korean is the primary case study (eight further languages on translation or generative knowledge); the granularity trend is cross-family, not causal. Code preservation at or above the default is a Qwen result (on ERNIE the four-view rank-union costs 8 HumanEval+ problems and lowers extraction). The exchange analysis defines its candidate set through C, uses single seeds for ERNIE and Qwen3.6, and lacks Qwen3.6's code axis; FP8/NVFP4 are weight-only fake quantization on one seed.

9 Conclusion

One-shot expert pruning silently removes access to languages absent from calibration: the pruned model still answers in English most of what it can no longer answer in Korean, while calibration-aligned validation stays healthy. Expert exchanges at fixed compression reveal a distinct recovery order: a few restored experts return the language, whereas most of the disagreement set is needed for benchmark accuracy, and in the tested families this ordering appears only where the default mask collapsed the language. Composing cached calibration views (COER) turns this analysis into a training-free remedy. Near-lossless MoE pruning therefore needs certification that measures language access and knowledge separately.

Reproducibility statement

Appendix B gives the full protocol (calibration packing, seeds, decoding, metrics, hyperparameters). Appendix L records the provenance freeze of the reference protocol: hashes of the reference checkpoints and result files; expert masks reconstructed from cached observations and verified by exact weight equality against every checkpoint (14,259 tensors per Qwen3-30B checkpoint); an observer reproducibility bridge showing that re-collected observations match the original ones on 48/48 layers; and a re-run of the reference protocol on a fixed evaluation set that reproduces the deterministic axes to three decimals. Every expert-exchange design in Section 7 was committed before its new checkpoints were built; verification records and build manifests accompany the checkpoints, and those for the reference and pilot checkpoints are included in the repository (Appendix K). All numbers in tables and figures are generated from the result files by scripts. We release the observe–prune–evaluate pipeline, the 100–200 KB cached views, the exchange designs, and the analysis scripts; composition and the closed-form synthesis check run on CPU in seconds.

Ethics statement

This work compresses publicly released MoE language models with public datasets and benchmarks; no human subjects or private data are involved. Its central finding is safety-relevant: standard one-shot pruning can silently remove a model's competence in languages absent from calibration while calibration-aligned benchmarks stay healthy, which would disproportionately affect speakers of those languages if such models were deployed. We report the failure, its mechanism, and a training-free remedy, and argue that compressed models should be certified with multilingual, knowledge-aware evaluation.

AI use statement

In this work, we used generative AI tools (Anthropic Claude and OpenAI Codex) to assist with writing and editing the manuscript, with implementing evaluation and analysis scripts, and with orchestrating experiment runs under author-specified protocols. The method, the experimental design, and all scientific claims are the authors' own; every AI-assisted script, number, and passage of text was reviewed and verified by the authors. AI assistants also proposed candidate experiments and analyses, which the authors selected, directed, and approved; no experimental data were generated by AI. We take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

References

Lasby, M.; Lazarevich, I.; Sinnadurai, N.; Lie, S.; Ioannou, Y.; and Thangarasa, V. 2026. REAP the Experts: Why Pruning Prevails for One-Shot MoE Compression. ICLR 2026. arXiv:2510.13999.

Liang, J.; Zhang, S.; Zhang, M.; Yang, H.; and Wang, B. 2026. Lang-Prune: Unlocking Fair and Powerful Pruning for Multilingual Large Language Models. OpenReview R9BUpZJ2fF.

Zeng, Y.; et al. 2026. Generic Expert Coverage for Pruning Sparse MoE Language Models. arXiv:2607.01710.

Compressing Large MoE Models via Efficient Pruning and Data-Aware Calibration. ICLR 2026 submission, OpenReview dJy6z9peC7.

Ding, Y.; et al. 2026. Attribution-Guided and Coverage-Maximized Pruning for Structural MoE Compression. arXiv:2606.18304.

Liu, Z.; Zhang, J.; Ma, Z.; Chen, G.; and Yuan, X. 2026. How to Score Experts for One-Shot MoE Expert Pruning: A Unified Formulation and Selection Principle. arXiv:2606.15716.

Hyeon, S.; and Do, J. 2026. Is Retraining-Free Enough? The Necessity of Router Calibration for Efficient MoE Compression. arXiv:2603.02217.

Zhang, Y. 2026. SHAPE: Coalition-Aware Expert Pruning for Sparse Mixture-of-Experts LLMs. arXiv:2606.09886.

Huang, W.; Zhang, Y.; Zheng, X.; Chao, F.; Ji, R.; and Cao, L. 2025. Discovering Important Experts for Mixture-of-Experts Models Pruning Through a Theoretical Perspective. In *Advances in Neural Information Processing Systems 38*.

Bai, S.; Li, H.; Zhang, J.; Hong, Z.; and Guo, S. 2025. DiEP: Adaptive Mixture-of-Experts Compression through Differentiable Expert Pruning. In *Advances in Neural Information Processing Systems 38*. arXiv:2509.16105.

Goel, P.; Maheshwari, A.; and Chakraborty, T. 2026. It Takes a MAESTRO To Prune Bad Experts. arXiv:2607.08601.

Liu, Z.; Chen, G.; Tang, S.; Shen, Y.; Wang, H.; and Yuan, X. 2026. AIMER: Calibration-Free Task-Agnostic MoE Expert Pruning. arXiv:2603.18492.

Bercovich, A.; et al. 2025. Puzzle: Distillation-Based NAS for Inference-Optimized LLMs. arXiv:2411.19146.

Bercovich, A.; et al. 2026. Nemotron-Labs-3-Puzzle-75B-A9B. arXiv:2607.04371.

Li, K.; Yang, Z.; Zhou, Z.; Xue, F.; Jiang, Z.; and Wang, W. 2025. HEAPr: Hessian-based Efficient Atomic Expert Pruning in Output Space. arXiv:2509.22299.

Cao, M.; Li, G.; Ji, J.; et al. 2024. Condense, Don't Just Prune: Enhancing Efficiency and Performance in MoE Layer Pruning. arXiv:2412.00069.

Gu, H.; Li, W.; Li, L.; et al. 2025. Delta Decompression for MoE-based LLMs Compression. arXiv:2502.17298.

Muzio, A.; et al. 2024. SEER-MoE. arXiv:2404.05089.

Muralidharan, S.; et al. 2024. Minitron. arXiv:2407.14679.

Jiang, A. Q.; et al. 2024. Mixtral of Experts. arXiv:2401.04088.

Qwen Team. 2025. Qwen3 Technical Report. arXiv:2505.09388.

Qwen Team. 2026. Qwen3.6-35B-A3B. Hugging Face model card, `Qwen/Qwen3.6-35B-A3B`.

Mistral AI. 2025. Mistral Small 3.1. Hugging Face model card, `mistralai/Mistral-Small-3.1-24B-Instruct-2503`.

Zhipu AI. 2025. GLM-4.5 Technical Report.

Baidu ERNIE Team. 2025. ERNIE 4.5 Technical Report.

DeepSeek-AI. 2024. DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model. arXiv:2405.04434.

Son, G.; et al. 2024. KMMLU: Measuring Massive Multitask Language Understanding in Korean. arXiv:2402.11548.

Jang, M.; et al. 2022. KoBEST. arXiv:2204.04541.

Son, G.; et al. 2023. HAE-RAE Bench. arXiv:2309.02706.

Xuan, W.; Yang, R.; Qi, H.; et al. 2025. MMLU-ProX: A Multilingual Benchmark for Advanced Large Language Model Evaluation. arXiv:2503.10497.

Liu, J.; et al. 2023. EvalPlus. arXiv:2305.01210.

Chen, M.; et al. 2021. HumanEval. arXiv:2107.03374.

Austin, J.; et al. 2021. MBPP. arXiv:2108.07732.

Luo, Z.; et al. 2023. WizardCoder / Evol-Instruct. arXiv:2306.08568.

Singh, S.; et al. 2024. Aya Dataset. arXiv:2402.06619.

Lee, J. (beomi). 2023. KoAlpaca. github.com/Beomi/KoAlpaca.

Gao, L.; et al. 2024. lm-evaluation-harness. Zenodo/EleutherAI.

Kurz, S.; et al. 2026. On the Limitations of Language-targeted Pruning. TACL.

Choenni, R.; and Titov, I. 2025. M-Wanda. EMNLP.

Zeng, H.; et al. 2024. Multilingual Brain Surgeon. arXiv:2404.04748.

Martin, L. O.; Bandarkar, L.; and Peng, N. 2026. Extracting Small Translation Specialists from LLMs by Aggressively Pruning Experts. arXiv:2605.28042.

Koishekenov, Y.; Berard, A.; and Nikoulina, V. 2023. Memory-efficient NLLB-200. ACL.

NLLB Team. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation (FLORES-200). arXiv:2207.04672.

Zhang, Q.; Yuan, Q.; Lin, H.; Lu, Y.; Han, X.; Sun, L.; et al. 2026. ShortOPD: Recovering Pruned LLMs with Short-to-Long On-Policy Distillation. arXiv:2607.13124.

Zoph, B.; Bello, I.; Kumar, S.; Du, N.; Huang, Y.; Dean, J.; Shazeer, N.; and Fedus, W. 2022. ST-MoE: Designing Stable and Transferable Sparse Expert Models. arXiv:2202.08906.

Dai, D.; Deng, C.; Zhao, C.; et al. 2024. DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models. ACL 2024. arXiv:2401.06066.

Lo, K. M.; Huang, Z.; Qiu, Z.; Wang, Z.; and Fu, J. 2024. A Closer Look into Mixture-of-Experts in Large Language Models. arXiv:2406.18219.

Herbst, J.; Wermter, S.; and Lee, J. H. 2026. The Expert Strikes Back: Interpreting Mixture-of-Experts Language Models at Expert Level. arXiv:2604.02178.

This appendix provides architecture-specific routing corrections (App. A), full experimental details (App. B), extended remedy and combination-rule tables and two controls stated in full (App. C), additional figures and tables (App. D), criterion robustness (App. E), routing exposure by domain (App. F), compute and memory cost (App. G), hyperparameters (App. H), weight-only FP8 and NVFP4 compatibility (App. I), agent-domain calibration (App. J), the expert-exchange designs and full tables (App. K), and the provenance freeze (App. L). Reviewers are not obliged to consult it; all claims essential to the paper's evaluation are stated in the main body.

A Sigmoid-Router Implementation Corrections

REAP scores an expert by the conditional gate-weighted activation norm $S_j = \mathrm{E}_{x \in X_j}[\, g_j(x) \, \| f_j(x) \| \,]$, so both the conditioning set X_j (which tokens select j) and the gate g_j must match the model's true router. The reference observer approximates routing by topk over the raw logits with softmax gates. This is exact for Qwen3 (softmax top-k with `norm_topk_prob`), but wrong for sigmoid-router models:

- **GLM-4.5-Air.** The true router applies a sigmoid, an additive score-correction bias, and group-limited top-k; gates are the sigmoid scores renormalized over the selected set and scaled by a routed-scaling factor. Measured over 45 MoE layers on 32 texts, the raw-logit approximation reproduces the true top-8 expert set on only **27.5% of tokens** (mean per-token Jaccard 0.785), with a $\sim$50% median relative gate error even when the selected set agrees. Under such disagreement, REAP is computed from the wrong X_j and the wrong g_j.
- **ERNIE-4.5-21B.** The modeling file bundled with the reference code *instantiates* an expert score-correction bias (`moe_statics`) but never applies it during routing (top-k is taken over the raw logits), unlike the official in-tree implementation, which selects $\mathrm{topk}(\text{bias} + \text{softmax probs})$. We use the official in-tree routing.

We contribute a per-architecture routing-reproduction registry (`qwen3`, `glm4moe`, `ernie4_5_intree`) that reproduces each model's true expert selection and gate values; a unit test matches the ERNIE reproduction to the in-tree reference exactly. All sigmoid-router results in this paper (GLM, ERNIE) use it. Because our GLM numbers therefore differ in provenance from the reference observer, GLM results should be read in this corrected-routing context rather than compared directly to reference-observer GLM numbers. A latent defect — the router's expert-count attribute not being updated after the expert ModuleList is sliced — is also fixed.

B EXPERIMENTAL DETAILS

Calibration observations. One observation *batch* is a single packed sequence of `model_max_length`$= 2048$ tokens: calibration documents are concatenated (with a one-token separator whose attention is masked at document boundaries, so statistics are not contaminated across documents). Runs use batch size 1 and 1024 batches, i.e. $\approx$2.1M calibration tokens (1024×2048). A composite calibration splits this budget across sources by a per-source batch count; for example:

Arm	Composition (batches per source)
default	evol-codealpaca-v1: 1024
Korean-only	KoAlpaca: 1024
mixture 50:50	evol-codealpaca: 512 + KoAlpaca: 512
honest four-way	code: 256 + KoAlpaca: 256 + Aya: 256 + ko-wiki: 256

Rank-union operates on separate per-domain observations. Primary matched-budget comparisons fix the total at 1024 batches: two views receive 512 batches each, and four views receive 256 each. Extended full-budget variants use 2048 total batches for two views and 3328 for four views. The observation shuffling seed equals the pruning seed.

Pruning. Experts are removed layerwise-uniformly at 25% (also 40/50% for the ratio study), taking the lowest score under the criterion being tested. REAP is the default criterion; frequency, MAN, and MSAN use the same selection and surgery. When the model renormalizes its top-k gates (`norm_topk_prob`; true for Qwen3) the surviving experts' router weights are renormalized after slicing. Qwen3 uses softmax top-k routing directly; GLM-4.5-Air and ERNIE-4.5 use the faithful routing reproduction of App. A. Closed-form mixture synthesis composes per-domain observations offline via $S_\alpha = (\alpha f_A S_A + (1-\alpha) f_B S_B)/(\alpha f_A + (1-\alpha) f_B)$; synthesized mixtures were validated against honestly re-collected mixed observations (prune-set Jaccard 0.94–0.98). The transfer to other families is validated by honest re-observation: at 50:50, closed-form synthesis matches an honestly re-observed mixture at prune-set Jaccard 0.974 (ERNIE) and 0.970 (GLM), Pearson ≥ 0.9997 — the same band as Qwen3. The four-view synthesis is validated the same way: against an honestly re-observed four-way mixture, closed-form MIX4 reaches prune-set Jaccard 0.948 (ERNIE) and 0.921 (GLM), Pearson ≥ 0.998.

COER procedure. COER takes a checkpoint, a layerwise pruning ratio, one saliency criterion, and cached observations from one or more domains. First, it computes the chosen per-expert score independently in every view. Second, it composes the views before selection. Mixture synthesis combines stored routed counts and numerator sums to reproduce the requested empirical mixture. Rank-union ranks experts separately within each layer and view, assigns $R_{\ell j} = \max_d r_{\ell d j}$ with larger ranks denoting greater saliency, and prunes the smallest $R_{\ell j}$. Third, the unchanged selector removes the same number of experts per layer, and the unchanged surgery slices expert modules and router rows. The output is a standard pruned checkpoint; no recovery training is performed.

Code and reproducibility. We will publicly release the full observe–prune–evaluate pipeline (pruning with any criterion, mixture synthesis, or rank-union) together with the analysis scripts. Because COER composes cached per-domain views rather than raw data, its central claim is verifiable without a GPU: from the minimal per-domain sufficient statistics (the REAP conditional saliency and routed counts, $\approx$74 KB per view — the 100–200 KB artifacts referenced in the main text), a bundled check reconstructs both the closed-form mixture and the rank-union decision and matches honestly re-observed mixtures (REAP-score correlation ≥ 0.997; prune-set Jaccard 0.94–0.98 for mixtures) in seconds on CPU. Rank-union is a distinct mass-insensitive rule (prune-set Jaccard ≈ 0.65 against the mixture), so its equivalence to data mixing is at the benchmark level (same-pool control) rather than prune-set identity.

Perplexity (fluency). The `korean_ppl` protocol computes token-weighted perplexity over the first 128 documents of `wikimedia/wikipedia 20231101.{ko,en}` (held out from calibration; streaming, seed 42), each truncated to 2048 tokens; English wiki is a control. All perplexities

are reported as ratios to each model's own unpruned baseline, since the collapsed regime is high-variance in absolute terms. The multilingual sweep (Fig. S3) uses a different document sampling and therefore different absolute magnitudes; only ratios are compared.

Free-form generation. Greedy decoding (`do_sample=False`), `max_new_tokens=` 256, chat template with `enable_thinking=False` for Qwen3. Output script mix (Hangul / Han / Latin fractions) is measured to detect language switching.

Translation (FLORES-200). `Muennighoff/flores200` `devtest`, first $n = 400$ sentences per direction; language codes `kor_Hang`, `jpn_Jpan`, `fra_Latn`, `zho_Hans`, and `eng_Latn`. Served with vLLM, temperature 0 (greedy), `max_tokens=` 512, zero-shot chat prompt (*"Translate the following {src} sentence into {tgt}. Output only the translation, nothing else."*). chrF++ is `sacrebleu corpus_chrf` with `word_order=` 2; BLEU is `corpus_bleu`. En→X is target-language generation; X→en is source-conditioned English translation.

Generative knowledge (MMLU-ProX). lm-evaluation-harness MMLU-ProX tasks, zero-shot (`num_fewshot=` 0), generative chain-of-thought with letter extraction (A–J). The full MMLU-ProX task is used for the model-family damage comparisons in Table 1; the `lite` subset (588 items per language) is used only for the decomposition and remedy screens where explicitly marked. Strict `exact_match` is decomposed as extraction rate × conditional accuracy (accuracy given a recovered letter).

Multiple-choice knowledge. KMMLU, KoBEST, and HAERAE, zero-shot (`num_fewshot=` 0) via lm-evaluation-harness; we report accuracy with harness standard errors, and mean±std over seeds 42–44 for the primary Qwen3 arms.

Code. HumanEval(+) and MBPP(+) under EvalPlus, greedy pass@1, served via vLLM.

Teacher-forced NLL. FLORES `devtest` first 200 en/ko pairs; per-token negative log-likelihood of the *gold* target under teacher forcing (both directions), which removes free-running error propagation. We report the first-target-token excess separately from the early/mid/late positional terciles, with sentence-level bootstrap 95% CIs (10,000 resamples, seed 0). Teacher-forced NLL and source-ablation are evaluated directly on the saved pruned checkpoints and do not invoke the observer-side router reconstruction used during saliency collection. An ERNIE-4.5 replication of the full protocol gives en→ko excess 0.849 $[0.810, 0.889]$ vs. ko→en 0.272 $[0.211, 0.335]$ (first token 0.32; terciles $0.86/0.87/0.81$, flat). A GLM-4.5-Air replication gives en→ko excess 1.252 $[1.166, 1.342]$ vs. ko→en 0.470 $[0.395, 0.544]$ (first token 0.54; terciles $1.31/1.30/1.19$, flat). The directional deficit and its positional signature thus reproduce on two further families; the absolute en→ko excess follows the same Qwen3 > GLM > ERNIE ordering as the observed Korean fluency collapse.

Source ablation. ko→en target NLL under three source conditions: *correct*; *token-shuffled* (the Korean source tokens randomly permuted — meaning destroyed, re-tokenized length within $\sim$1 token of the original); and *empty*. The source-use signal is $\mathrm{SIG} = \mathrm{NLL}(\text{ablated}) - \mathrm{NLL}(\text{correct})$; we compute a paired bootstrap over the shared 200 sentences (100,000 resamples, seed 0). The retained fraction is $\mathrm{SIG}_{\text{pruned}}/\mathrm{SIG}_{\text{baseline}}$. On ERNIE, the token-shuffled SIG is 1.76 before and 1.93 after pruning, a retained-sensitivity ratio of 1.10 $[1.06, 1.15]$; on GLM it is 1.57 before and 1.50 after, a ratio of 0.95 $[0.91, 1.00]$. Source conditioning therefore remains substantial on both further families.

Gate-mass diagnostics. Per-layer gate-weighted routing mass removed from Korean tokens in the generation role (en→ko continuation positions) versus the reading role (ko→en prompt positions), using the same architecture-faithful router reproduction as the observer.

Seeds. The primary Qwen3 three-arm comparison uses seeds 42–44 (prune-set Jaccard 0.937–0.939); other configurations are single-seed unless noted. Engineered and synthesized observations use distinct directory seeds to avoid cache collisions.

C Extended Remedies and Combination Rules

Two controls summarized in Section 6, stated in full.

Data-versus-rule decomposition. A TB-Coverage-style adaptation in our observation space with a generic-corpus setting modelled on the original (English C4 + English Wikipedia in place of the original's WikiText-2 + C4; no Korean or code) protects Korean fluency (koPPL $\times 1.9$) but not KMMLU (0.370, below even the default) and loses code (HumanEval+ 0.695). Plain REAP on the same data gives koPPL $\times 2.1$ and code 0.787. Thus, natural-language data supplies almost all fluency protection; round-robin adds $\times 2.1 \rightarrow \times 1.9$ at a 9.2-point code cost. Our target-aware observations dominate both on every axis. Generic English views substantially repair Korean fluency and improve HAERAE (0.502 vs. 0.397) but leave KMMLU below the default (0.370 vs. 0.401); these comparisons do not isolate language coverage from domain and observation effects. Table S2 reports the full rule and calibration-free comparison.

A resident-parameter-matched dense reference. At comparable resident parameters, leave-language-out rank-union activates 3.3B rather than 24B parameters per token and exceeds dense Mistral Small 3.1 (Mistral AI, 2025) on code and Korean multiple-choice knowledge. The dense model leads on the remaining multilingual axes, indicating that COER provides a distinct compute–quality trade-off rather than uniformly dominating the dense alternative (Appendix Table S10).

Qwen3, 25%	budget	koPPL×	HEval+	KMMLU	HAERAE
Korean-only	1024	1.14	0.752	0.521	0.624
rank-union 2-view eq	1024	1.30	0.884	0.505	0.562
rank-union 2-view (3-seed)	2048	1.30	0.880	0.511	0.562
rank-union 4-view	3328	1.16	0.866	0.484	0.547

Table S1: Extended remedy variants omitted from the main-paper remedy table (Table 2) for space, with the HAERAE knowledge column. The 2-view 3-seed row (2048 batches) has HAERAE 0.562 ± 0.006 and French 0.518 ± 0.004. The same-pool fairness control (a data mixture over the identical four domains) is discussed in the main paper's rank-union section.

Method	koPPL×	HEval+	MBPP+	KMMLU	HAERAE
TB-Coverage rule (our re-implementation, B=40)	1.37	0.878	0.717	0.497	0.526
mean-rank	1.30	0.890	0.728	0.504	0.538
round-robin	1.26	0.878	0.733	0.515	0.560
max-rank (ours, 3-seed)	1.30	0.880	0.730	0.511	0.562
max-rank, raw (ablation)	1.25	0.872	0.735	0.470	0.545
minimax-mass exact (WVEP)[‡]	1.26	0.866	0.725	0.449	0.570
AIMER (calib.-free)[†]	1.59	0.854	0.664	0.423	0.465

Table S2: Under *identical* code+ko observations, the four coverage-family rules (TB-Coverage, mean-rank, round-robin, max-rank) differ by at most about four points (HAERAE 0.526–0.562). TB-Coverage is our reproduction of the generic-coverage round-robin rule (arXiv:2607.01710). [‡]WVEP solves the worst-view removed-saliency-mass objective *exactly* (per-layer 0–1 MILP: $\min_{|P|=k} \max_d \sum_{e \in P} q_{de}$ with q_{de} the per-domain-normalized mass $f_{de} S_{de}$; under this additive proxy the worst mixture over the whole domain simplex is attained at a single-domain vertex, so worst-mixture robustness reduces to worst-view protection). It dominates every rule on that proxy (worst-view removed mass 0.053 vs. 0.129 for max-rank) and improves fluency, yet lowers KMMLU (0.449 vs. 0.511 for max-rank; a 4-view variant reaches the best fluency of all methods, koPPL $\times 1.12$, at the largest KMMLU cost, 0.425, and -4.5pt code) while HAERAE remains comparable (0.570). Even exact optimization of the mass objective fails to protect KMMLU — the sharpest Goodhart failure of mass diagnostics: routing mass predicts fluency more reliably than knowledge retention. [†]AIMER (arXiv:2603.18492) is *not* part of this identical-observation comparison: it is calibration-free (weight-only ℓ_1/ℓ_2, no observations), listed as a reference point. It avoids language collapse (ko PPL $\times 1.6$) but protects neither knowledge (KMMLU 0.42 $\approx$ default-setup REAP) nor code (MBPP+ 0.66).

C.1 Relation to coalition-aware, layer-adaptive, and language-aware pruning

Three families of prior work change different parts of the pipeline, and none was run here. *Coalition-aware importance* (SHAPE (Zhang, 2026); Shapley-MoE (Huang et al., 2025)) replaces the per-expert score with a Shapley-style contribution estimated from Monte-Carlo forward passes over expert subsets; it still evaluates those subsets on the calibration distribution, so an uncovered language is invisible to it for the same reason it is invisible to REAP. *Layer-adaptive allocation* (DiEP (Bai et al., 2025)) learns a non-uniform per-layer pruning rate by gradient-based relaxation on calibration data; our criteria use a uniform ratio, so DiEP's gain and COER's are additive in principle—this is a hypothesis, not a result. *Language-aware importance* (Lang-Prune (Liang et al., 2026)) aggregates per-language importance for dense weights; it does not transfer to expert units directly. The closest expert-level analogue we tested is the raw-versus-rank aggregation ablation in Table S2 (raw max keeps fluency and French, rank normalization adds about four KMMLU points); it is an ablation of the aggregation principle, not a reproduction of Lang-Prune. What this paper verifies is confined to observation coverage under fixed per-expert scores and a fixed ratio.

C.2 Capability-by-calibration summary

Table S3 places the calibrations discussed in Sections 5–6 on one set of axes with their budgets and seeds. Rows differ in budget, seed count and rule, so only rows with matching labels are directly comparable; no statistical Pareto ordering is asserted. In the tested controls, generic English natural-language calibration substantially protects Korean fluency but does not recover Korean knowledge scores, and target-aware calibration improves the latter.

calibration	rule	budget (batches)	seeds	koPPL×	KMMLU	HAERAE	HumanEval+	ProX-lite fr
code default (evol-codealpaca)	REAP	1024	42–44	14.4	0.401	0.397	0.864	0.000
generic English (C4 + en-wiki)	TB-Coverage rule, B=40	1024	42	1.9	0.370	0.502	0.695	–
Korean only (KoAlpaca)	REAP	1024	42	1.13	0.528	0.628	0.732	–
code + Korean mixture 50:50	synthesis	1024	42–44	1.22	0.530	0.556	0.876	0.000
code + Korean	rank-union (2 views)	1024	42	1.30	0.505	0.562	0.884	0.532
code + Korean + Aya + ko-wiki	rank-union (4 views)	1024	42	1.17	0.482	0.547	0.878	0.551

Table S3: Same axes, explicit labels: Qwen3-30B-A3B at 25%. koPPL is relative to unpruned (6.32); ProX-lite fr is zero-shot generative exact match (588 items); "–" = not measured. Multi-seed rows are means. The generic-English row uses our re-implementation of the TB-Coverage rule.

C.3 Choosing views without a downstream benchmark

The controls imply a procedure that needs no target-language evaluation. Keep the deployment view the default already uses (it protects the deployment domain); add one broad multilingual instruction view (Aya here), which protects fluency and third-language access for languages it never labels (leave-language-out, Section 6); add the target language only if it is known, which is what recovers target knowledge (dual regime); and add a knowledge-dense target view only for that purpose (C-v2 did not improve HAERAE). A two-view mixture restores the target but leaves a third-language hole (French 0.000) that the two-view rank-union (0.532) and four-view pools close (Table 2); 256 batches per view suffice (matched-budget rows, Table S1); and once the pool is diverse the composition rule is secondary (Table S2). The leave-one-view-out ablation in Appendix C.4 shows which view carries which axis.

C.4 Leave-one-view-out ablation

Four 256-batch views of Qwen3-30B-A3B (code, KoAlpaca, Aya, Korean Wikipedia; seed 42) give the four-view rank-union (1024 batches in total), the four three-view unions that each drop one view (768 batches each), and the two-view code+KoAlpaca union (512 batches), all pruned at 25% with the unchanged surgery and evaluated under the Section 6 protocol (ProX-fr on the first 43 items per subject, 602 items; FLORES on 400 sentences per direction). Dropping a view also lowers the observation budget, so we add a budget-matched control: the four-view union observed at 192 batches per view (768 in total). It performs similarly to the 1024-batch union on every axis (koPPL 7.32 vs. 7.39, KMMLU 0.484 vs. 0.482, HAERAE 0.569 vs. 0.567, en→fr 63.3 vs. 62.7, HumanEval+

0.890 vs. 0.884; no equivalence test). A same-observation control (the four seed-42 views composed as a token-balanced mixture instead of a rank-union) lands within 1 point of the rank-union on K-MMLU, KoBEST, HAERAE, and French, with lower Korean perplexity (7.03 vs. 7.39) and 3 fewer HumanEval+ points (0.854 vs. 0.884), so with this pool the composition rule mainly moves the code–fluency balance. At a matched total observation budget, different view compositions therefore protect different capabilities, indicating that total budget alone does not explain the leave-one-view-out results; removing a view also changes how the remaining budget is distributed, so the effects below are effects of view composition, not of view identity in isolation.

view set	batches	koPPL	Hangul	KMMLU	KoBEST	HAERAE	ProX-fr	en→ko	en→fr	ko→en	HumanEval+	MBPP+
all four views	1024	7.39	0.993	0.482	0.555	0.567	0.550	28.8	62.7	52.6	0.884	0.746
all four views, 192 batches each	768	7.32	0.982	0.484	0.556	0.569	0.556	29.2	63.3	52.5	0.890	0.741
all four views, token-balanced mixture (same observations)	1024	7.03	0.978	0.491	0.565	0.562	0.545	30.0	63.2	54.3	0.854	0.735
− code	768	7.01	0.996	0.507	0.559	0.598	0.583	29.8	63.4	53.7	0.848	0.688
− KoAlpaca	768	7.35	0.985	0.462	0.562	0.559	0.553	28.7	62.9	52.2	0.878	0.728
− Aya	768	7.31	0.980	0.500	0.559	0.569	0.508	28.9	27.7	52.4	0.878	0.720
− ko-wiki	768	8.15	0.996	0.486	0.557	0.513	0.545	28.2	62.7	52.0	0.860	0.738
code + KoAlpaca only	512	8.26	0.962	0.507	0.563	0.567	0.533	28.0	26.1	51.8	0.884	0.733
all four views (seed 43)	1024	7.40	0.997	0.489	0.559	0.543	0.543	28.7	63.4	53.0	0.878	0.733
− code (seed 43)	768	7.02	0.997	0.504	0.561	0.621	0.561	29.7	63.4	53.9	0.860	0.722
− KoAlpaca (seed 43)	768	7.41	0.983	0.466	0.551	0.522	0.548	28.2	63.0	52.6	0.878	0.743
− Aya (seed 43)	768	7.37	0.991	0.510	0.566	0.572	0.550	29.0	26.2	52.8	0.878	0.728
− ko-wiki (seed 43)	768	8.19	0.984	0.495	0.563	0.523	0.553	29.0	63.5	53.2	0.872	0.725
code + KoAlpaca only (seed 43)	512	8.08	0.984	0.496	0.567	0.546	0.532	29.0	27.6	53.0	0.878	0.720
all four views (seed 44)	1024	7.54	0.987	0.509	0.558	0.596	0.551	28.9	62.8	52.5	0.860	0.730
− code (seed 44)	768	7.01	0.998	0.496	0.564	0.614	0.566	29.8	63.5	53.8	0.817	0.677
− KoAlpaca (seed 44)	768	7.46	0.976	0.498	0.552	0.563	0.532	28.1	62.5	52.0	0.890	0.733
− Aya (seed 44)	768	7.47	0.971	0.506	0.563	0.580	0.513	28.5	27.9	52.4	0.878	0.746
− ko-wiki (seed 44)	768	8.14	0.965	0.516	0.563	0.545	0.543	28.6	62.6	52.1	0.860	0.717
code + KoAlpaca only (seed 44)	512	8.22	0.981	0.518	0.567	0.566	0.548	28.9	26.9	51.9	0.884	0.730

Table S4: Leave-one-view-out at fixed 25% compression (rank-union). Upper block: observation seed 42, including the budget-matched four-view control at 192 batches per view and a same-observation control that replaces rank-union by a token-balanced mixture of the same four views; lower blocks: independent sets of four 256-batch observations at seeds 43 and 44. koPPL is absolute (unpruned 6.32); Hangul is the generation-only Korean-script share; FLORES columns are chrF++.

Reading (within the 768-batch drop variants). Different view compositions trade capabilities against each other; the clearest repeated effects are French translation without Aya and Korean perplexity without Korean Wikipedia, whereas the KoAlpaca–KMMLU assignment is weaker and not unique at seed 44. Dropping the code view lowers HumanEval+ to 0.848 and MBPP+ to 0.688 (the other drops: 0.860–0.878 and 0.720–0.738) while Korean perplexity, KMMLU, and HAERAE improve in the seed-42 comparison, so the deployment view protects the deployment domain and competes with Korean knowledge for the budget. Dropping KoAlpaca gives the lowest KMMLU (0.462 vs. 0.486–0.507) with fluency intact, because Aya and Korean Wikipedia still cover Korean. Dropping Aya more than halves French translation (en→fr 27.7 vs. 62.7–63.4 chrF++) and lowers ProX-fr to 0.508, so the broad multilingual view carries third-language generation and access. Dropping Korean Wikipedia gives the worst Korean perplexity (8.15 vs. 7.01–7.35) and the lowest HAERAE (0.513 vs. 0.559–0.598), the Korea-specific knowledge benchmark. The two-view union (512 batches) reproduces the Aya-drop pattern on French (26.1) with the highest perplexity (8.26). At the shared 768-batch budget the four-view union keeps every axis at once (French 63.3, KMMLU 0.484, koPPL 7.32, HumanEval+ 0.890) while each drop loses one. The evidence is not uniform in strength: the French collapse without Aya (more than 35 chrF++ points, both seeds) and the perplexity loss without Korean Wikipedia are large relative to seed variation, the KMMLU loss without KoAlpaca is 1–2 points, whereas the code cost of dropping the code view (2–4 HumanEval+ points) is of the order of the seed-to-seed differences. A second, independent set of four observations (seed 43; lower block) reproduces every assignment: dropping code again gives the lowest HumanEval+ (0.860 vs. 0.872–0.878) with the best Korean axes (koPPL 7.02, HAERAE 0.621); dropping KoAlpaca again gives the lowest KMMLU (0.466 vs. 0.495–0.510); dropping Aya again collapses French translation (26.2 vs. 63.0–63.5) while the two-view union does the same (27.6); dropping Korean Wikipedia again gives the worst perplexity (8.19) and, with the KoAlpaca drop, the lowest HAERAE (0.523/0.522). A third set (seed 44; lowest block) reproduces three of the four assignments: dropping code again gives the lowest HumanEval+ and MBPP+ (0.817/0.677 vs. 0.860–0.890/0.717–0.746) with the best Korean axes (koPPL 7.01, HAERAE 0.614); dropping Aya again collapses French translation (27.9 vs. 62.5–63.5; two-view union 26.9) and gives the lowest ProX-fr (0.513); dropping Korean Wikipedia again gives the worst perplexity among the three-view unions (8.14) and the lowest HAERAE (0.545). The KMMLU assignment is the weakest: dropping KoAlpaca lowers KMMLU by 1.1 points against the four-view union (0.498 vs. 0.509) and the code

drop is as low (0.496), whereas at seeds 42 and 43 the KoAlpaca drop was the lowest KMMLU by 1.9 and 2.3 points against the union. Over the three seeds the code cost of dropping the code view is 2–4 HumanEval+ points and the KMMLU cost of dropping KoAlpaca 1–2 points, both against the four-view union; the main text quotes these ranges. These observations motivate the view-selection heuristic in Appendix C.3; they do not certify it as an optimal procedure.

C.5 Same-item English/Korean knowledge probe

MMLU-ProX provides the same MMLU-Pro items in English and Korean. We take the first 200 items of each of the 14 subjects (2,800 items per language; selection fixed before any result was seen) and verify that every paired item carries the same answer letter in both languages (0 mismatches). Scoring is zero-shot generative exact match after answer extraction (lm-evaluation-harness `custom-extract`), vLLM, seed 42. Arms: unpruned; a default model rebuilt from a 256-batch code observation (seed 42, with rebuilds from seeds 43 and 44); the token-balanced 50:50 code:Korean mixture composed from cached observations (seed 610); the four-view rank-union over 256-batch views (Appendix C.4); and the Korean-only calibration (KoAlpaca, 1024 batches, the paper's B-ko arm). These are single-node rebuilds, not the 1024-batch endpoints B/C of Section 7. Full-set aggregates on all 11,759 items for unpruned/default/mixture are en 0.699/0.539/0.617 and ko 0.584/0.011/0.473. English MMLU (0-shot log-likelihood; different items, a script-independent English diagnostic rather than a paired KMMLU) is 0.778/0.673/0.731, 0.671/0.669 for the seed-43 and seed-44 defaults, 0.721 for the rank-union, and 0.742 for the Korean-only calibration.

	strict accuracy		extraction rate		acc. | extracted		Δ vs. unpruned (points, 95% CI)	
arm	en	ko	en	ko	en	ko	en	ko
unpruned	0.692	0.572	0.910	0.828	0.760	0.691	–	–
default (rebuilt, 256-batch)	0.539	0.011	0.846	0.018	0.637	0.588	−15.3 [−17.1, −13.5]	−56.1 [−57.9, −54.2]
COER mixture 50:50	0.617	0.473	0.853	0.765	0.724	0.618	−7.5 [−9.1, −5.9]	−9.9 [−11.6, −8.2]
default (rebuilt, 256-batch, seed 43)	0.549	0.131	0.866	0.224	0.634	0.588	−14.3 [−16.1, −12.4]	−44.0 [−46.0, −42.1]
default (rebuilt, 256-batch, seed 44)	0.543	0.000	0.861	0.000	0.631	–	−14.9 [−16.7, −13.0]	−57.2 [−59.0, −55.4]
COER rank-union (4 views, 256-batch each)	0.620	0.463	0.859	0.760	0.721	0.609	−7.2 [−8.8, −5.6]	−10.9 [−12.6, −9.1]
Korean-only calibration (KoAlpaca)	0.640	0.531	0.898	0.839	0.712	0.632	−5.2 [−6.7, −3.8]	−4.1 [−5.7, −2.5]

Table S5: Same 2,800 items in two languages (Qwen3-30B-A3B, 25%; one calibration seed per row, three seeds for the default). Conditional accuracies condition on each arm's own extracted subset and are not comparable across arms (the default model's Korean subset has 50 items). Paired bootstrap CIs over items, 10,000 resamples.

Reading. The default model's Korean strict-score collapse coincides with a collapse of valid-answer extraction (0.828 to 0.018) and with the script switching observed separately, whereas English strict accuracy and the separate MMLU likelihood diagnostic decline more moderately. Because extracted subsets differ across arms, the conditional accuracies are descriptive rather than a decomposition of knowledge loss. Of the 1,574 items the unpruned model answers correctly in Korean and the default model does not, the default model still answers 1,157 (73.5%) correctly in English; restricted to the 1,406 of them that the unpruned model also answers correctly in English, 1,078 (76.7%). The second default seed gives 890 of 1,290 (69.0%; restricted 822 of 1,136, 72.4%). Pairing the two drops item by item, the Korean drop exceeds the English drop by 40.8 points [38.3, 43.4] at seed 42 and 29.8 [27.2, 32.3] at seed 43; for the composed arms the two drops nearly coincide ($+2.5$ [0.2, 4.8] mixture, $+3.6$ [1.3, 5.9] rank-union). The Korean collapse is therefore mostly a loss of access through Korean rather than a matching loss of the ability to answer the same items in English; at the same time English knowledge is measurably reduced (15 points here, 10.5 on MMLU), which English perplexity ($\times 1.6$) and code scores do not reveal. Composition recovers most of both losses, and the four-view rank-union behaves like the mixture. Calibrating on Korean text alone (KoAlpaca, 1024 batches; no code and no dedicated English view) loses only 5 points in English and 4 in Korean: in this control, Korean-only calibration preserves substantially more English accuracy than the code-only default, so the benefit is not restricted to matching the evaluation language; language, domain, and observation budget are not independently isolated here. Two further default seeds lose 14 and 15 points in English and 44 and 57 in Korean: the English loss is stable across the three calibration seeds while the Korean loss varies with the depth of the script collapse (Korean extraction 0.018, 0.224, 0.000), yet remains three to four times the English loss. The probe measures

access through each language on identical items; it is not a language-independent measurement of knowledge.

Answer-candidate likelihood scoring. Scoring the same 2,800 items by the length-normalized log-likelihood of each option text (no generation, hence no extraction step) has a low ceiling on this model—unpruned 0.280 in English and 0.246 in Korean against a mean chance level of 0.112 (10 options on most items)—so it is a weak axis, and we use it only as a consistency check. Within it the default model's paired losses are nearly symmetric, -5.3 points [-6.6, -4.0] in English and -6.5 [-8.1, -4.9] in Korean (mixture: -3.5 [-4.8, -2.3] and -1.9 [-3.0, -0.8]; same items, 5,000 paired resamples). The 41-point Korean-over-English excess of the generative probe does not appear when the model is not required to produce a Korean answer, which is consistent with a response-format or access contribution to the collapse, but the low unpruned performance also limits this probe's sensitivity to pruning damage; it does not measure stored knowledge either.

C.6 Recovery fine-tuning versus COER

A natural alternative to composing observations is to fine-tune the default-pruned model on the target language. Table S6 compares, on one node under the Section 6 protocol, the default 25% Qwen3-30B (rebuilt from a 256-batch observation; it matches the paper's default on every benchmark within the seed band), the same model after LoRA fine-tuning on KoAlpaca—the corpus COER uses as its Korean view—with rank 16 on the attention projections (13.4 M trainable parameters, sequence 1024, batch 16, learning rate 10^{-4}, 250 and 500 steps on one Blackwell GPU), and the COER 50:50 mixture composed from cached observations. Fine-tuning restores Korean-script output but recovers little knowledge and costs code; extending training from 250 to 500 steps improves perplexity but not the joint trade-off (KMMLU, HAERAE, and code decline in this run), whereas composition recovers knowledge with code unchanged and no training. A second calibration seed (43; lower rows, 500 steps) repeats the pattern: koPPL $33.98 \rightarrow 13.85$, knowledge $+0.1$ to $+1.6$ points, HumanEval+ $0.860 \rightarrow 0.774$. Attention-only LoRA is a low-cost recovery baseline tested here; expert-targeted or full fine-tuning is a different experiment and may do better.

Qwen3-30B, 25%	GPU train	koPPL	Hangul frac.	KMMLU	KoBEST	HAERAE	HumanEval+	MBPP+
default pruned	–	86.81	0.009	0.409	0.475	0.425	0.866	0.728
+ LoRA, 250 steps	31.6 min	16.03	0.797	0.438	0.489	0.456	0.787	0.696
+ LoRA, 500 steps	62.1 min	14.60	0.780	0.433	0.492	0.437	0.756	0.685
COER mixture 50:50	0	7.74	0.785	0.527	0.573	0.540	0.866	0.728
unpruned	–	6.32	0.80	0.577	0.581	0.678	0.896	0.738
default pruned (seed 43)	–	33.98	0.034	0.397	0.489	0.437	0.860	0.722
+ LoRA, 500 steps (seed 43)	64.8 min	13.85	0.987	0.413	0.496	0.438	0.774	0.706

Table S6: The tested short attention-only LoRA runs do not match composition's joint Korean recovery and code retention (one run per arm; the lower rows repeat the default and its 500-step fine-tune at a second calibration seed; the mixture is token-balanced across a 256-batch code view and a 1024-batch Korean view).

D Additional Figures and Tables

Table S7 gives the bidirectional FLORES-200 results behind Section 4 and Table S8 the gate-mass and retention numbers behind Section 5; Fig. S2 illustrates the COER selection principle; Figs. S3–S5 show the per-language damage sweep, the granularity plot, and the third-language hole; Table S9 and Fig. S6 report the compression-ratio sweep; and Table S10 the resident-parameter-matched dense reference.

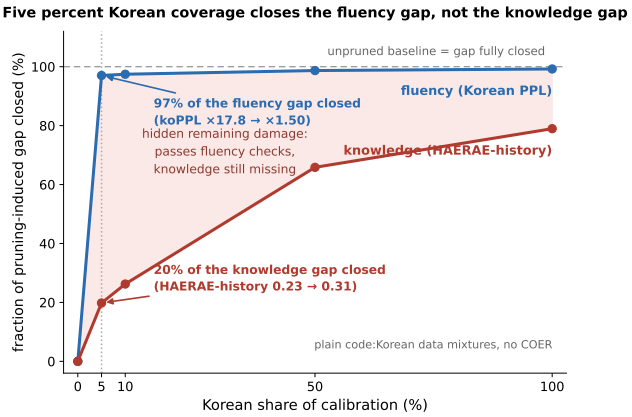


Figure S1: Dual regime under plain code:Korean calibration mixtures (Qwen3-30B, seed 42; standard gate-weighted pruning, no COER). Recovery is $100(m(\alpha) - m(0))/(m_{\text{base}} - m(0))$, signs reversed for perplexity: a 5% share closes 97.0% of the fluency gap in perplexity (86% in log-perplexity, $(\ln 17.8 - \ln 1.50)/\ln 17.8$; koPPL $\times 17.8 \rightarrow \times 1.50$) but 19.8% of the knowledge gap (HAERAE-history $0.229 \rightarrow 0.309$; baseline 0.633). The ordering holds at the other shares: 10% closes 97.6% (log 88%) of fluency and 26.2% of knowledge, 50% closes 98.7% and 65.8%. Two further calibration seeds (43 and 44; 256-batch views, 512 batches in total, so not budget-matched) preserve the ordering: the 5% mixture closes 88.7% (log 76.1%) and 97.1% (log 85.7%) of the fluency gap but 15.4% and 4.6% of the knowledge gap (koPPL $33.98 \rightarrow 9.45$ and $119.86 \rightarrow 9.62$ against the unpruned 6.32; HAERAE-history $0.287 \rightarrow 0.340$ and $0.287 \rightarrow 0.303$). A per-item bootstrap would not capture calibration-seed variance.

Direction	Baseline (no pruning)	En-code REAP (25% pruned)
en→ko (generation)	31.7	3.2
ko→en (comprehension)	56.0	42.7
en→ja (generation)	29.1	5.7
ja→en (comprehension)	55.2	46.6
en→fr (generation)	67.0	22.5
fr→en (comprehension)	67.3	59.2
en→zh (generation)	29.4	23.4
zh→en (comprehension)	57.3	53.2

Table S7: Bidirectional FLORES-200 translation for unpruned and 25% default-pruned Qwen3-30B (chrF++, 400 sentences per direction, seed 42). Directions from English measure target-language generation; reverse directions are operational comprehension proxies. Generation degrades far more in every language.

	gate mass removed		translation retention		
Model (experts)	gen	comp	en→ko	ko→en	gap
Qwen3 (128)	0.358	0.324	10%	76%	66 pt
GLM-Air (128)	0.324	0.331	33%	72%	39 pt
ERNIE (64)	0.297	0.312	43%	83%	40 pt

Table S8: Near-equal gate mass is removed from Korean tokens in both roles (left), yet translation into Korean degrades far more than out of it (right). Teacher-forced NLL supports a target-side deficit rather than aggregate mass imbalance or free-running error compounding.

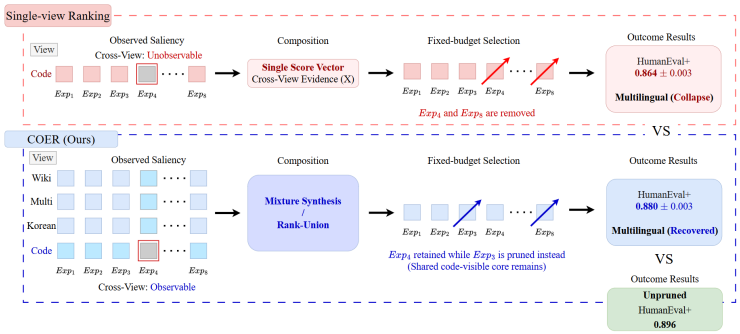


Figure S2: Controlled expert selection with the same criterion, pruning budget, and surgery. Single-view removes illustrative E4 because its cross-view value is unseen; COER retains E4, prunes E3 instead, and preserves the shared code-visible core. HumanEval+ values are measured Qwen3 rank-union results at 25% pruning over three seeds (unpruned: 0.896).

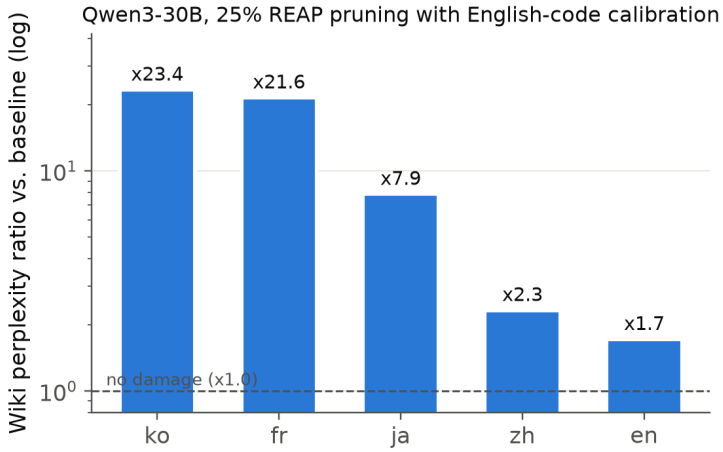


Figure S3: Per-language wiki-perplexity ratio after 25% default-setup pruning (Qwen3-30B). Document sampling differs from the main-paper `korean_ppl` protocol, so magnitudes differ; only ratios are compared.

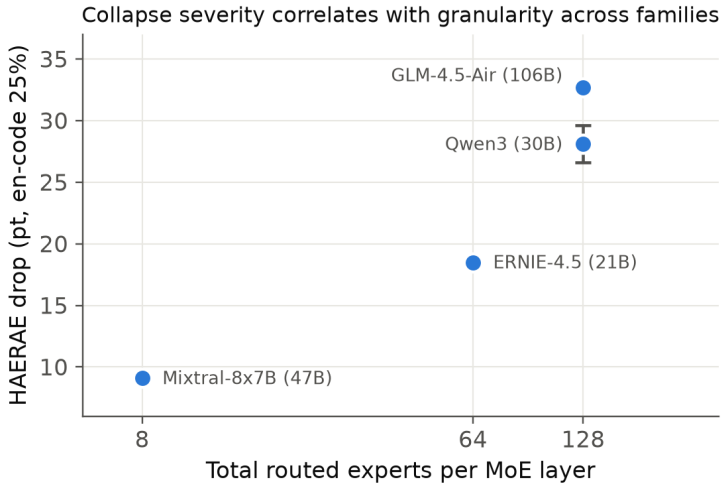


Figure S4: HAERAE damage vs. experts-per-layer; the two 128-expert models coincide despite a 3.5× size gap.

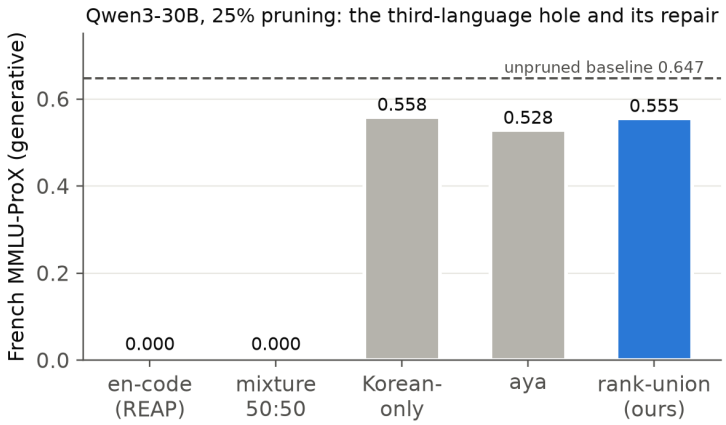


Figure S5: The third-language hole and its repair (MMLU-ProX-lite French, Qwen3-30B).

ratio (mem.)	method	koPPL×	HEval+	KMMLU	KoBEST	HAERAE
unpruned (61 GB)	—	1.00	0.896	0.577	0.581	0.678
25% (47 GB)	default	17.8	0.866	0.400	0.474	0.389
	mixture	1.22	0.884	0.531	0.571	0.572
	rank-union 4-view	1.16	0.866	0.484	0.547	0.547
40% (38 GB)	default	38.4	0.872	0.340	0.471	0.354
	mixture	1.6	0.872	0.448	0.560	0.551
	rank-union 4-view	1.5	0.884	0.397	0.546	0.490
50% (32 GB)	default	131	0.854	0.311	0.469	0.336
	mixture	2.0	0.884	0.312	0.557	0.511
	rank-union 4-view	1.9	0.854	0.304	0.522	0.387

Table S9: Compression-ratio robustness (Qwen3-30B, single seed 42; koPPL× vs. the unpruned baseline). Pruning $k\%$ of experts cuts resident weight memory by $24/38/47\%$ at $25/40/50\%$ ($61.1 \rightarrow 46.6/38.0/32.1$ GB) with unchanged per-token top-k compute. Under the default setup koPPL explodes with ratio ($\times 17.8 \rightarrow 38 \rightarrow 131$), whereas the mixture and four-view rank-union keep fluency ($\times 1.9$–2.0 at 50%) and code (HEval+ ≥ 0.854) viable throughout and hold KoBEST/HAERAE well above the default. Knowledge exhibits a shared high-compression floor: KMMLU converges to ≈ 0.31 for all methods and MMLU-ProX-lite French falls to $0.00/0.00/0.26$ (default/mixture/rank-union), so none of the evaluated calibration designs restores knowledge beyond this floor; whether a matched-compression oracle could is untested. Observation budgets: default/mixture 1024; four-view rank-union 3328.

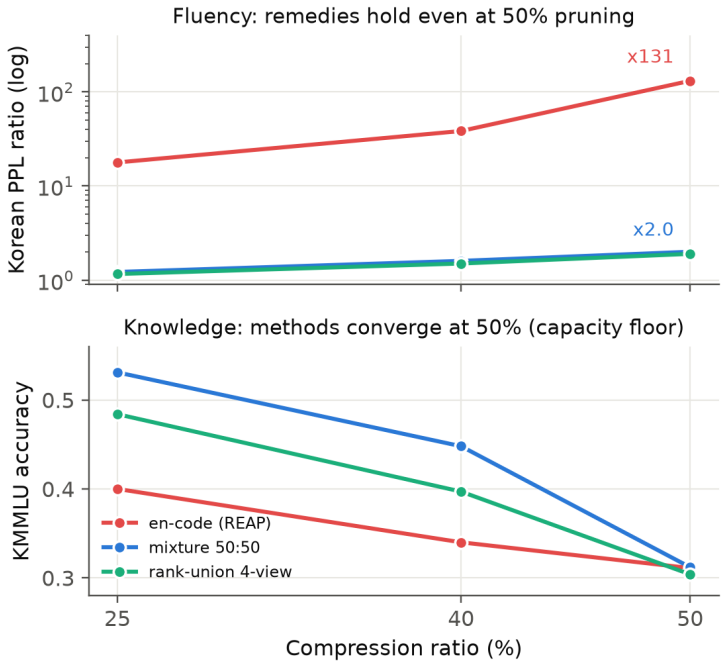


Figure S6: Compression-ratio robustness (Qwen3-30B).

Model	resident	activated	HEval+	MBPP+	KMMLU	KoBEST	HAERAE	ProX-ko	ProX-fr	en→ko
Mistral Small 3.1 (dense)	24B	24B	0.799	0.622	0.422	0.709	0.409	0.546	0.648	32.9
Qwen3-30B-A3B unpruned	30.5B	3.3B	0.896	0.738	0.577	0.581	0.678	0.585	0.636	31.7
default en-code pruned	23.3B	3.3B	0.864	0.727	0.400	0.474	0.389	0.00	0.000	3.2
rank-union, ours (LLO 4-view, 3 obs. seeds)	23.3B	3.3B	0.882	0.737	0.483	0.553	0.551	0.469	0.540	28.9

Table S10: Resident-parameter-matched dense reference (same lm-evaluation-harness; Mistral uses its own chat template). At 25% pruning the MoE holds 23.3B resident parameters (comparable resident-weight footprint to Mistral Small 3.1 24B) while MoE sparsity keeps 3.3B parameters activated per token, ≈7× fewer than the dense model. The comparison is mixed by axis: our rank-union pruned MoE leads on code (HumanEval+/MBPP+) and Korean multiple-choice knowledge (KMMLU/HAERAE) at ≈7× fewer activated parameters, while the dense model leads on KoBEST, generative-format Korean (MMLU-ProX-lite ko), French, and en→ko translation. Default en-code pruning falls below the dense reference on every Korean axis; the remedy brings the pruned MoE back into a competitive range. Unpruned Qwen3 is a reference (not resident-matched). The "ours" (LLO) row reports means over the same three observation seeds on every axis; Mistral and unpruned Qwen3 are fixed checkpoints, and other entries are point estimates unless stated otherwise. All ProX values are MMLU-ProX-lite (588 items). koPPL is omitted because the current perplexity harness does not support Mistral3's multimodal wrapper. HEval+/MBPP+ are greedy pass@1; EvalPlus "+" scores are not directly comparable to the base HumanEval/MBPP scores reported in the Mistral model card. en→ko is FLORES chrF++.

E Robustness Across Pruning Criteria

We test whether the multilingual collapse and its training-free remedy are specific to REAP's gate-weighted score or hold across scoring criteria. We reuse the unified formulation of Liu et al. (2026a), which parameterizes criteria as (b, α, β): REAP is $(1, 1, 1)$, frequency $(0, 0, 0)$, and their two proposed gate-free criteria are Mean Activation Norm MAN $(1, 0, 1) = \frac{1}{N_j} \sum_{t:\, j \in \mathrm{TopK}} \|f_{j,t}\|_2$ and Mean Squared Activation Norm MSAN $(1, 0, 2) = \frac{1}{N_j} \sum_t \|f_{j,t}\|_2^2$. We verified that our existing `ean_mean` statistic is *exactly* MAN (routed-token, gate-free, frequency-weighted online mean of $\|f\|_2$) and implemented MSAN as the analogous mean of squared norms; a re-observation confirms the invariant $\mathrm{MSAN}_j \geq \mathrm{MAN}_j^2$ (Jensen) at all 6144 layer-expert entries. All rows use the same Qwen3-30B, the same evol-code (English/code) calibration, and a 25% layer-uniform budget; MAN and MSAN are averaged over three seeds.

criterion	koPPL×	hangul%	han%	KMMLU	KoBEST	HAERAE	HEval+	MBPP+	en→ko
baseline (unpruned)	1.0	92	0	0.577	0.581	0.678	0.896	0.738	31.70
REAP $(1, 1, 1)$	17.8	1	88	0.400	0.474	0.389	0.866	0.728	3.25
frequency $(0, 0, 0)$	8.0	1	0	0.365	0.467	0.386	0.872	0.746	—
MAN $(1, 0, 1)$, 3 seeds	4.64	7	74	0.311	0.480	0.207	0.866	0.715	5.90
MSAN $(1, 0, 2)$, 3 seeds	4.64	5	81	0.304	0.479	0.205	0.870	0.728	5.54

Table S11: The multilingual collapse is robust across the four tested criteria. All four tested criteria severely damage Korean generation and knowledge under the same English/code calibration: perplexity blows up ($\times 4.6$–$\times 17.8$ over the unpruned baseline), free-form generation loses Korean (Korean script $\leq 7\%$; the fallback is criterion-dependent — REAP, MAN, and MSAN shift predominantly to Chinese-script output, whereas frequency shifts to Latin-script output), and Korean knowledge drops sharply; code (HEval+/MBPP+) is retained throughout. The gate-free MAN/MSAN scores collapse knowledge *at or below* REAP's level (HAERAE 0.21 vs. REAP 0.39) despite a milder perplexity blow-up, so the failure is driven by calibration coverage, not by REAP's gate weighting. koPPL× is relative to the unpruned baseline; hangul/han are output-script fractions over five fixed free-form Korean prompts; en→ko is FLORES chrF++. REAP and frequency are seed-42 point estimates; MAN and MSAN report means over observation seeds 42–44 (MAN koPPL× 4.64 ± 0.10, MSAN 4.64 ± 0.28).

variant	koPPL×	hangul%	han%	KMMLU	KoBEST	HAERAE	HEval+	MBPP+	en→ko
gate-free activation-based criteria									
MAN (pure)	4.6	7	74	0.311	0.480	0.207	0.866	0.715	5.90
MAN + mixture	1.25	92	1	0.273	0.519	0.230	0.868	0.709	29.81
MAN + rank-union	1.37	92	1	0.285	0.530	0.247	0.874	0.733	27.39
MSAN (pure)	4.6	5	81	0.304	0.479	0.205	0.870	0.728	5.54
MSAN + mixture	1.19	92	0	0.293	0.526	0.242	0.874	0.707	30.43
MSAN + rank-union	1.35	92	1	0.292	0.538	0.237	0.868	0.733	28.22
activation-free boundary (frequency criterion)									
frequency + mixture	1.23	92	0	0.293	0.523	0.283	0.593	0.497	29.24
frequency + rank-union	1.26	89	0	0.231	0.520	0.268	0.654	0.552	29.00

Table S12: Cross-criterion transfer and an activation-free boundary. COER restores Korean fluency and generation under the gate-free MAN and MSAN scores while retaining code, but the large KMMLU/HAERAE losses remain (unlike REAP, where the mixture lifts KMMLU $0.40 \rightarrow 0.53$). The bottom block applies the same compositions to frequency scoring. Compositions are offline from the same evol-code and KoAlpaca observations (no re-observation); pure-criterion collapse baselines are also in Table S11. All rows are means over three observation seeds (42–44); full per-seed values and standard deviations are provided in the artifact, and the qualitative hierarchy holds in every seed (generation recovers under all four criteria; frequency preserves neither code nor knowledge; knowledge recovers only under REAP's gate weighting). The two notable dispersions are frequency+mixture HEval+ (± 0.09) and frequency+rank-union hangul (± 4.5).

Activation-free boundary. Applying the same compositions to frequency scoring restores Korean fluency and generation (koPPL $\times 8.0 \rightarrow \times 1.23/\times 1.26$; hangul $1 \rightarrow 92/89\%$; en→ko

29.2/29.0) but sacrifices substantial code accuracy (HumanEval+ $0.872 \rightarrow 0.593/0.654$; MBPP+ $0.746 \rightarrow 0.497/0.552$; three-seed means) and does not recover knowledge. Thus observation diversity is sufficient to restore target-language generation even with routing counts alone, whereas code-preserving recovery in this Qwen3 experiment requires an activation-based criterion.

F Removed-Expert Routing Exposure by Domain

Figure S7 shows, per MoE layer, the fraction of each domain's routed expert assignments that fall on the experts removed by the default mask.

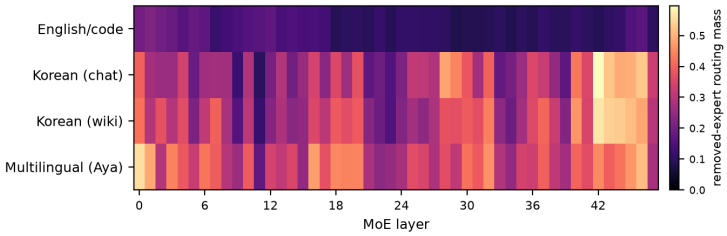


Figure S7: Language-dependent exposure to the default prune set (Qwen3-30B, 25%). Each cell is the fraction of a domain's routed expert assignments (expert-frequency count) that fall on the experts removed by the English/code default mask (bottom-25% REAP on evol-code), per MoE layer. English/code is uniformly low (mean 12.4%) — a domain discards its own least-used experts — whereas Korean chat/wiki and multilingual (Aya) inputs place 30–36% of their assignments on the removed experts, concentrated in the later layers. This *removed-expert routing-count exposure* does not identify the experts as Korean-exclusive; it provides a pre-pruning exposure diagnostic consistent with the disproportionate multilingual damage. Computed from cached observations (CPU).

G Compute and Memory Cost

Computing infrastructure. The main experiments ran on a single node with $8\times$ enterprise-class NVIDIA Hopper-generation GPUs (80 GB each; CUDA 13.0) with dual-socket server CPUs, under Linux. Software: Python 3.12, PyTorch 2.7.1 (CUDA 12.8 build), Transformers 4.55.0, vLLM 0.10.0, and lm-evaluation-harness 0.4.11 (EvalPlus 0.3.1 for code); exact versions are pinned in the code artifact's `requirements.txt`. A second node with $4\times$ enterprise-class NVIDIA Blackwell-generation GPUs (183 GB each; CUDA 13.0), also under Linux, hosted the NVFP4 compatibility experiments (App. I), the Qwen3.6 port, the agent-domain calibration study and its KoAlpaca-only control (Table S18), the serving-throughput measurements (App. G.1), and the single-node rebuild studies of Appendices C.4, C.5, and C.6, using the same pinned software stack (PyTorch 2.7.1 CUDA 12.8 build, which runs natively on Blackwell `sm_100`; lm-evaluation-harness 0.4.9.1 on this node). Each table compares arms evaluated on one node only.

The dominant cost is one-time calibration observation. Once per-view observations have been collected, COER's scoring, composition, and pruning overhead is negligible because it reuses stored per-view statistics. The expert-exchange experiments of Section 7 ran on the same Hopper node (checkpoint builds on CPU, 3–7 minutes each; evaluation about 35 minutes per model on one GPU), except the Qwen3.6 port, which used the isolated environment described in Appendix L.

stage or artifact	measured cost or size
resident weight memory	61.1 → 46.6 (-24%, 25%) → 32.1 GB (-47%, 50%)
observation (per view)	1024 batches, ≈ 4.2 h elapsed on 8× Hopper GPUs (≈34 GPU-h)
full cached view	6.4 MB for Qwen3; 1–13 MB across the four measured families
stripped pruning view	100–200 KB for Qwen3 (per-expert scalar statistics only)
COER scoring/composition	offline, CPU, $<$ 1 min (closed-form synthesis or rank-union)
prune + serialization	≈40 s (slice) + ≈38 s (save)
per-token active compute	unchanged (top-8 routing retained)

Table S13: Cost breakdown. Top-k routing and per-token expert FLOPs are unchanged. Serving throughput is measured in Appendix G.1; the measured benefit is the smaller resident-weight footprint (-24%/-47%), which can reduce serving-memory requirements or provide additional memory headroom: under identical vLLM startup-profiling settings on one 80 GB Hopper GPU (`gpu_memory_utilization`$=0.9$, bf16, `max_model_len`$=4096$, tensor parallel 1), 25% pruning raises the allocated KV-cache capacity from 95k to 243k tokens ($2.55\times$), equivalent in cache capacity to approximately 23 versus 59 full 4096-token sequences. These profiler-derived capacities are reproduced by the measured runs in Table S14. The composition overhead is negligible once the desired views are available; collecting each additional view costs approximately 4.2 elapsed hours on eight Hopper GPUs. Full view files are larger because they include $E \times E$ pairwise statistics used by expert-merging methods. COER pruning does not use those matrices, so they can be stripped before artifact release.

G.1 Measured serving throughput

Table S14 replaces the earlier profiler-only KV-cache figure with measured serving on one Blackwell GPU (183 GB; vLLM 0.10, tp=1, bf16, 4096-token context, identical engine settings within each budget). The 80 GB-class budget caps `gpu_memory_utilization` to about 73 GB, the setting of the Hopper profiling, and reproduces its 95k/243k-token capacities (96,144/243,664); it emulates an 80 GB card on a Blackwell GPU and is not a Hopper measurement. Pruning leaves top-k routing unchanged, so per-token compute is identical: the gain comes from KV-cache capacity and appears only where that capacity binds. Single-stream latency (batch 1, 1024→128) is 0.805/0.844/0.822 s for unpruned/25%/50%—no improvement. A second run reproduced every cell within 3% (saturating budget: 10,698/16,981/20,893 tokens/s, i.e. $1.59\times$ and $1.95\times$); the table reports the first run.

budget (workload)	model	KV tokens	conc. @4096	tok/s	vs. unpruned
∼73 GB (3584→512, 400 req.)	unpruned	96,144	23.5	10,903	1.00×
	25% pruned	243,664	59.5	17,138	1.57×
	50% pruned	391,184	95.5	21,112	1.94×
183 GB (1024→256, 200 req.)	unpruned	1,070,160	261.3	25,556	1.00×
	25% pruned	1,217,680	297.3	26,095	1.02×
	50% pruned	1,365,200	333.3	28,398	1.11×

Table S14: Serving throughput of Qwen3-30B-A3B after default 25%/50% expert pruning (run 1 of 2; run 2 agrees within 3%).

H Hyperparameters and Settings

Table S15 consolidates the final settings used across all experiments; per-configuration deviations (e.g., reduced batches for DeepSeek-V2 and FP8, the compression-ratio and mixture-share sweeps) are noted inline in the corresponding tables and are listed here for completeness.

component	final setting
Models (experts, top-k)	Qwen3-30B-A3B (128, 8); GLM-4.5-Air 106B (128, 8); ERNIE-4.5-21B-A3B (64, 6); Mixtral-8x7B (8, 2); DeepSeek-V2-Chat 236B (160 routed + 2 shared, 6)
Pruning criterion	REAP (default); frequency / MAN / MSAN (controls)
Compression ratio	0.25 (default); 0.40, 0.50 (robustness)
Layerwise rule	uniform per-layer removal of lowest-saliency experts
Router renormalization	on (models with `norm_topk_prob`, e.g. Qwen3)
Calibration batch size	1
Batches per view	1024 (256 for DeepSeek-V2, FP8, and NVFP4)
Packed sequence length	2048 tokens (≈2.1M tokens per view)
Calibration datasets	evol-codealpaca (default), KoAlpaca, Aya, ko-wiki
Seeds	42 (primary); 42–44 (three-seed); others single-seed
COER mixture weight	$\alpha = 0.5$ (50:50); sweep $\{0.95, 0.90\}$
COER rank-union	per-layer max-rank; M1 (2 views), M3 (4 views)
Matched-budget views	256 batches per view
FP8 rounding	`float8_e4m3fn`, per-channel; router/LM-head/embed in BF16
NVFP4 rounding	E2M1, 16-elem `e4m3` block scales, FP32 tensor scale; same exclusions
lm-eval tasks	0-shot, batch size auto, seed 42, chat template on KMMLU, KoBEST, HAERAE, MMLU-ProX-lite
Serving (vLLM)	BF16, gpu-mem-util 0.9, max-model-len 4096 (32768 for EvalPlus), tensor/expert parallel (model-dependent)
Code eval	EvalPlus greedy pass@1 (HumanEval+, MBPP+)
Translation	FLORES-200 devtest, 400 sentences, greedy, chrF++
MMLU-ProX	lite, exact-match (custom extraction)

Table S15: Consolidated final hyperparameters and settings across models, pruning, calibration, COER, quantization, and evaluation.

I WEIGHT-ONLY FP8 AND NVFP4 COMPATIBILITY

Method and scope. We round every eligible linear weight through two low-bit formats and evaluate the perturbed checkpoints. *FP8*: per-output-channel absolute-maximum scaling into `float8_e4m3fn` (mean relative weight error ≈0.5%). *NVFP4*: the Blackwell-native 4-bit floating-point format — E2M1 values $\{0, \pm 0.5, \pm 1, \pm 1.5, \pm 2, \pm 3, \pm 4, \pm 6\}$ with per-16-element `e4m3` block scales and an FP32 per-tensor scale (mean relative weight error ≈8.8%). In both formats the MoE router gate, language-model head, and embeddings remain in BF16; in total, 18,624 linear layers are rounded and 49 modules remain in BF16, identically for both formats. These are weight-only checkpoint perturbations (approximately W8A16 and W4A16), evaluated on Qwen3-30B-A3B with seed 42; FP8 on the Hopper node and NVFP4 on the Blackwell node (App. G). They do not evaluate activation quantization, real quantized-kernel execution (W8A8/W4A4), or runtime speed.

Selection integrity. The matched comparisons use the same 256-batch English-code calibration and change only the weight rounding. Table S16 shows that both formats change the prune set less than calibration-seed resampling: the margin shrinks monotonically with coarser rounding (FP8 0.989 > NVFP4 0.959 > the matched 256-batch seed band, 0.895), but even 4-bit rounding stays above the Jaccard ≥0.94 view-reuse gate. The recommended prune-then-quantize order therefore reuses BF16 views directly at both precisions, and the reverse quantize-then-prune order passes the same Jaccard integrity check used for cached-view synthesis.

selection diagnostic	FP8 vs. BF16	NVFP4 vs. BF16
prune-set Jaccard	0.989 (layer min. 0.939)	0.959 (layer min. 0.882)
matched seed-resampling band	0.895	0.895
saliency Pearson	0.9999 (layer min. 0.9996)	0.9991 (layer min. 0.9981)
different decisions	0.2 of 32 per layer	0.7 of 32 per layer
zero-frequency experts	0	0

Table S16: Selection integrity under weight-only FP8 and NVFP4 rounding, matched at 256 observation batches. Both formats are compared to the same Hopper-node 256-batch seed-resampling band, the mean over three seed pairs (42/43/44), 0.895 ± 0.009. Earlier drafts quoted per-format reference bands (0.937, a 1024-batch band; 0.888, a single Blackwell-node 256-batch pair); the budget-matched band above supersedes them, and the main text quotes it. Over seeds 42/43/44 the FP8-vs-BF16 prune-set Jaccard is 0.985–0.991 (mean 0.988 ± 0.003) and the NVFP4-vs-BF16 Jaccard is 0.959–0.962 (mean 0.960 ± 0.002); both stay above the band at every seed. Re-running the NVFP4 fake-quant on the Hopper node reproduces the earlier Blackwell-node measurement (0.959 at seed 42), confirming the weight-rounding diagnostic is hardware-independent. The tabulated per-format Jaccard and Pearson are the seed-42 point values.

Performance. Table S17 shows that both unpruned quantized checkpoints closely track BF16 on language metrics. For FP8 the only visible cost is HumanEval+ (-3.0 points, MBPP+ $+0.8$); for

checkpoint and pruning state	Korean PPL	English PPL	KMMLU	KoBEST	HAERAE	HumanEval+	MBPP+
BF16 unpruned	6.32	8.06	0.577	0.581	0.678	0.896	0.738
FP8 unpruned	6.31	8.03	0.581	0.576	0.665	0.866	0.746
NVFP4 unpruned	6.58	8.22	0.576	0.584	0.671	0.884	0.693
BF16 default single-view	112.7 (×17.8)	13.2	0.400	0.474	0.389	0.866	0.728
FP8 default single-view	102.3 (×16.2)	13.6	0.413	0.471	0.412	0.860	0.730
NVFP4 default single-view	92.8 (×14.1)	13.7	0.389	0.472	0.405	0.848	0.706
BF16 COER mixture (50:50)	7.72 (×1.22)	11.9	0.531	0.571	0.572	0.884	0.730
FP8 COER mixture (50:50)	7.68 (×1.22)	11.9	0.525	0.574	0.554	0.860	0.735
NVFP4 COER mixture (50:50)	7.96 (×1.21)	12.0	0.521	0.569	0.552	0.866	0.712

Table S17: BF16, weight-only FP8, and weight-only NVFP4 performance across unpruned, default-pruned, and COER-recovered states. Perplexity ratios are relative to each precision's own unpruned checkpoint. FP8 and NVFP4 rows use single-seed 256-batch observations; BF16 rows are the canonical project references. The matched 256-batch selection comparisons are reported separately in Table S16.

NVFP4, Korean perplexity rises 4% (6.58 vs. 6.32), Korean benchmarks stay within 1.3 points, and code is the only clearly affected axis (HumanEval+ -1.2, MBPP+ -4.5) — small for a 4-bit format. Default single-view pruning reproduces the multilingual collapse relative to each format's own unpruned checkpoint (Korean perplexity $\times 16.2$ FP8, $\times 14.1$ NVFP4; collapse-regime perplexity magnitudes are high-variance, and the benchmark deltas agree with BF16 within ≈ 2 points). The 50:50 COER mixture restores Korean perplexity to $\times 1.22$ (FP8) and $\times 1.21$ (NVFP4), retains code, and reproduces the BF16 recovery pattern at both precisions. We further reproduced the NVFP4 collapse and recovery on a different node (Hopper): koPPL and the MC-knowledge benchmarks (K-MMLU/KoBEST/HAERAE) are identical to the Blackwell-node values above (e.g. $\times 14.1/\times 1.21$ perplexity; HAERAE $0.407/0.552$), and code pass@1 agrees within generation variance (≤ 2.4 points), confirming these model-level results are hardware-independent (the fake-quant is deterministic software rounding). With physical low-bit weight storage, the measured 24% resident-weight reduction from 25% pruning composes to approximately 38% (FP8) and 21% (NVFP4, at ≈ 4.5 bits per parameter including block scales) of the unpruned BF16 weight bytes.

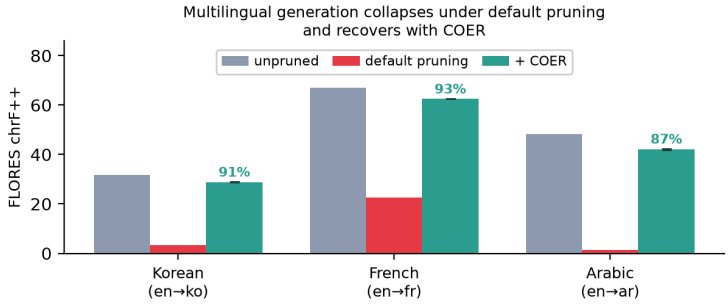


Figure S8: Collapse and recovery of generation into three languages (Qwen3-30B, FLORES chrF++). Default English/code pruning collapses generation into Korean, French, and Arabic (with output-script corruption — Korean shifts to Chinese, Arabic to Latin/Han); COER (four-view rank-union) restores 87–93% of baseline while keeping code near baseline (HumanEval+ 0.87). Unpruned and default-pruned values are fixed-checkpoint / seed-42 point estimates; COER bars report mean $\pm$ standard deviation over three leave-language-out observation seeds. Quantitative counterpart to the qualitative Figure 1 in the main text.

J AGENT-DOMAIN CALIBRATION (XLAM)

To test whether the collapse and its COER remedy are specific to the released English/code default rather than to calibration-language coverage in general, we repeat the pipeline on Qwen3-30B-A3B with an agent/tool-calling calibration (`Salesforce/xlam-function-calling-60k`; 1024 batches, 25% REAP, seed 42). Restoration observations are composed offline (mixture α=0.5;

two-view rank-union) from the xLAM and KoAlpaca views, without re-observation; a KoAlpaca-only arm (the same Korean observation used alone) is the control that separates composition from the Korean view's own effect. Tool-calling is measured with BFCL (Berkeley Function-Calling Leaderboard, v4) under a fixed prompt-mode handler applied identically to every arm, so only cross-arm differences are meaningful (absolute scores are not comparable to the public leaderboard). This is a single-seed reserve experiment.

Table S18 shows the same pattern as the code default. xLAM-only pruning keeps its own domain (BFCL within 1.2 points of unpruned) yet silently destroys Korean: free-form Hangul falls to 0.20, en→ko translation to 39% of unpruned, and HAERAE by 22.5 points; here the output degrades into a mixed English/Chinese script (Latin 0.44, Han 0.39) rather than the fluent Chinese substitution seen with the code default. COER (mixture or rank-union) restores fluency (Hangul ≈ 0.99), raises en→ko to 92–94% of unpruned, and recovers most knowledge (C-mix HAERAE -5.8), while tool-calling stays within -0.31 to $+1.44$ points of xLAM-only pruning. The KoAlpaca-only control recovers Korean at least as well (HAERAE 0.628) but is the lowest arm on BFCL (88.23/79.05) and loses 16 HumanEval+ points against the composed arms. Item-level paired bootstraps over the pooled AST items (1,150 non-live, 1,351 live; the table reports BFCL's category-averaged accuracies) place KoAlpaca-only -0.8 [-2.2, $+0.6$] and -1.3 [-2.9, $+0.4$] points below xLAM-only, C-mix $+1.8$ [$+0.5$, $+3.1$] above KoAlpaca-only on non-live and $+1.0$ [-0.5, $+2.7$] on live, and rank-union not separated from KoAlpaca-only on BFCL (a CI containing zero is not evidence of equivalence). On tool-calling only the mixture separates from the Korean view with confidence; on code the separation is large (HumanEval+ 0.866/0.872 vs. 0.707). The Korean view alone thus trades the deployment domain for Korean, and composition removes that trade on the code axis and, for the mixture, on non-live tool-calling. The collapse is therefore a property of calibration-language coverage, not of the code domain, and COER transfers to an agent domain while holding a domain-native benchmark near baseline.

axis	benchmark	Unpruned	xLAM-only	KoAlpaca-only	C-mix	rank-union
fluency	Korean PPL ↓	6.32	19.84	7.17	7.60	7.82
	gen. Hangul frac.	0.995	0.196	0.999	0.995	0.991
translation	FLORES en→ko chrF++	31.70	12.26	30.03	29.29	29.76
	FLORES ko→en chrF++	56.03	47.71	53.99	54.15	53.67
knowledge	KMMLU	0.577	0.449	0.528	0.533	0.505
	KoBEST	0.581	0.489	0.585	0.575	0.572
	HAERAE	0.678	0.453	0.628	0.620	0.522
tool-calling	BFCL Non-Live	90.17	89.02	88.23	90.46	88.71
	BFCL Live	81.50	80.31	79.05	80.09	80.01
code	HumanEval+	0.896	0.854	0.707	0.866	0.872
	MBPP+	0.738	0.704	0.611	0.701	0.688

Table S18: Agent-domain (xLAM tool-calling) calibration on Qwen3-30B-A3B, 25% REAP, single seed. xLAM-only pruning preserves tool-calling (BFCL) but collapses Korean (Hangul, translation, knowledge); COER mixture and rank-union restore Korean while keeping BFCL within -0.31 to $+1.44$ points of xLAM-only; the KoAlpaca-only control recovers Korean but is lowest on BFCL (paired CIs vs. xLAM-only include zero; C-mix is $+1.8$ [$+0.5$, $+3.1$] above it on non-live) and costs 16 HumanEval+ points. Restoration observations are offline-composed; BFCL uses a fixed prompt-mode handler (cross-arm deltas only).

K Exchange Framework: Designs, Verification, and Full Tables

Each design was fixed and committed before its new exchange checkpoints were built. Every checkpoint was produced by the unchanged pruning surgery from the original weights and verified tensor-by-tensor against them (expert, router, attention, norm, embedding, and head tensors); a checkpoint with any mismatch is marked failed and never evaluated. Table S21 lists the designs.

Dose axis. $m_\ell(t) = 4 + \lceil t(|P_\ell| - 4) \rceil$ per layer over the layer's own candidate set; $\sum_\ell |P_\ell|$ is 684/693 (Qwen3-30B seeds 43/44, 25%), 792/788 (50%), 284 (ERNIE), and 1,166 (Qwen3.6). Random restore orders keep the $m{=}4$ set as a prefix and order the remaining candidates once per layer with a fixed key, so the dose effect is separated from a redraw effect.

Qwen3-30B, 25%, m=4 (28% of candidates)	ko doc NLL ↓		en→ko TF-NLL ↓		≥80% Hangul		ProX-ko-lite acc.		KMMLU		HumanEval+	
restore policy (seed)	43	44	43	44	43	44	43	44	43	44	43	44
B: default mask (start)	4.64	5.10	4.38	4.47	0.010	0.000	0.146	0.007	0.401	0.406	0.854	0.866
C-rank restore	2.61	2.69	2.03	2.17	0.801	0.760	0.366	0.344	0.410	0.417	0.860	0.872
B-saliency restore (1024-batch obs.)	3.62	3.72	3.64	3.62	0.066	0.019	0.292	0.194	0.418	0.420	0.860	0.866
B-frequency restore (1024-batch obs.)	3.46	3.73	3.18	3.36	0.002	0.000	0.376	0.350	0.415	0.428	0.860	0.866
random restore (mean of 5)	3.15	3.28	2.97	3.09	0.113	0.210	0.292	0.281	0.426	0.430	0.856	0.856
C: matched four-view rank-union (end)	2.09	2.12	1.82	1.80	0.762	0.714	0.486	0.473	0.490	0.509	0.884	0.878

Table S19: Restore-order comparison at the smallest dose (m=4, 28% of candidates) on Qwen3-30B, seeds 43 and 44, all exchanges removing the same experts. The B-only selectors are scored on B's own 1024-batch observation (the corrected control; the first-design 256-batch values are in Table S20). C-rank leads every comparator on both Korean NLL axes (paired bootstrap CIs exclude zero, also on fresh documents and pairs) and is the only order that restores the Korean response language, yet has the lowest observed KMMLU of the exchange arms in both seeds ($-0.50/-1.14$ points vs. the 1024-batch B-frequency control, p=.052 and 3.9×10^{-6}), above B and far below C. Random rows average five fixed orders per seed; Table S20 lists each arm.

Qwen3-30B, 25%, m=4 (28% of candidates)	ko doc NLL ↓		en→ko TF-NLL ↓		≥80% Hangul		ProX-ko-lite acc.		KMMLU		HumanEval+	
restore policy (seed)	43	44	43	44	43	44	43	44	43	44	43	44
random order 1 (8101 / 8201)	3.14	3.38	2.99	3.13	0.000	0.022	0.228	0.360	0.421	0.428	0.854	0.866
random order 2 (8102 / 8202)	3.11	2.99	2.97	2.75	0.185	0.296	0.299	0.325	0.430	0.447	0.848	0.854
random order 3 (8103 / 8203)	3.16	3.79	2.96	3.59	0.270	0.012	0.272	0.236	0.429	0.431	0.866	0.854
random order 4 (8104 / 8204)	3.01	3.29	2.74	3.12	0.002	0.226	0.408	0.213	0.435	0.417	0.860	0.848
random order 5 (8105 / 8205)	3.32	2.98	3.16	2.86	0.109	0.493	0.252	0.270	0.417	0.429	0.854	0.860
B-saliency restore (256-batch obs., first design)	3.17	3.75	3.21	3.63	0.184	0.024	0.284	0.221	0.413	0.420	0.854	0.866
B-saliency restore (1024-batch obs., corrected)	3.62	3.72	3.64	3.62	0.066	0.019	0.292	0.194	0.418	0.420	0.860	0.866
B-frequency restore (256-batch obs., first design)	3.47	3.84	3.17	3.43	0.000	0.000	0.396	0.350	0.420	0.427	0.866	0.878
B-frequency restore (1024-batch obs., corrected)	3.46	3.73	3.18	3.36	0.002	0.000	0.376	0.350	0.415	0.428	0.860	0.866

Table S20: Individual restore arms behind the averages of Table S19: the five fixed random orders per seed and the B-only selectors under both observation budgets. The first design scored the B-only selectors on the 256-batch code view; the corrected design uses B's own 1024-batch observation, which is the comparator quoted in the text.

Corrected B-only control. The first matched-budget design scored the B-only selectors on a 256-batch code view although B itself uses its 1024-batch observation. All four masks differ and were rebuilt on the 1024-batch observation; the 256-batch results are kept and labeled. The corrected controls remain 0.84–1.03 nats behind C-rank on Korean document NLL and 1.14–1.60 on en→ko (all CIs exclude zero); frequency rankings barely move between budgets (40–42 of 48 layers keep the same top-4) while conditional-mean saliency rankings move substantially (14–27 of 48).

Statistics and averaging units. Dose curves use document-mean Korean NLL (117 held-out documents) and sentence-mean en→ko teacher-forced NLL (200 pairs); the seed-42 cross-exchange reports token-weighted en→ko excess NLL; ProX-ko denotes the 588-item Korean MMLU-ProX-lite task. Document- and pair-level paired bootstrap (10,000 resamples) for NLL differences; exact McNemar on the 588 ProX-lite items and 35,030 KMMLU items with per-item document hashes matched across runs (0 mismatches in every pairing). Whole-response Korean-script share is a descriptive proxy and is non-monotone in dose for C-rank; it is never used as a fluency or knowledge measure.

stage	model (experts, ratio)	seeds	conditions	checkpoints	tensors verified / mismatches
exchange pilot (m=4)	Qwen3-30B (128, 25%)	42	C-rank + 3 random	4	13,872 / 0
cross-exchange (2×2)	Qwen3-30B (128, 25%)	42	6 restore×remove swaps	6	14,259 / 0
selector + sensitivity	Qwen3-30B (128, 25%)	42	B-saliency, 5 random, L27 quartet	10	14,259 / 0
B-frequency selector	Qwen3-30B (128, 25%)	42	B-frequency restore	1	14,259 / 0
matched-budget replication	Qwen3-30B (128, 25%)	43, 44	B, C, C-rank, B-sal., B-freq., 5 random	18	14,259 / 0
recovery-path doses	Qwen3-30B (128, 25%)	43, 44	4 orders × 3 doses + 1024-batch B-only	20	14,259 / 0
50% compression doses	Qwen3-30B (128, 50%)	43, 44	4 orders × 3 doses + endpoints	28	9,651 / 0
ERNIE port	ERNIE-4.5-21B (64, 25%)	42	4 orders × 3 doses + endpoints	14	4,196 / 0
Qwen3.6 port	Qwen3.6-35B (256, 25%)	42	4 orders × 3 doses + endpoints	14	693 / 0

Table S21: Expert-exchange designs. Every design file records the explicit per-layer expert identities; the deterministic B rule reproduces the earlier default masks (48/48 layers on Qwen3-30B, 27/27 on ERNIE), and cutoff ties in C are resolved by a fixed identity rule and stored, never re-ranked. Verification counts are per checkpoint; the ERNIE model declares its 12 MTP tensors and the Qwen3.6 model its vision tower and MTP tensors explicitly rather than silently skipping them.

model (experts/layer)	smallest dose	ko doc NLL	en→ko TF	ProX-ko acc.	KMMLU	default mask: ≥80% Hangul
Qwen3-30B-A3B (128e), seeds 43/44	28%	79.5/80.7	91.6/86.1	64.5/72.3	10.9/10.9	0.010/0.000
Qwen3.6-35B-A3B (256e), seed 42	13.7%	81.8	95.7	99.2	16.6	0.000
ERNIE-4.5-21B-A3B (64e), seed 42	38.0%	60.2	76.5	53.3	65.0	0.881

Table S22: Share of the $B \to C$ gap closed by C-rank at each family's smallest dose: ≈70-point language-before-knowledge separation on Qwen3-30B, ≈80 on Qwen3.6, none on ERNIE, whose default mask keeps Korean (last column). Qwen3.6's KMMLU fraction divides a 1.7-point gap and is not a precise rate. ERNIE and Qwen3.6: one seed.

L Provenance Freeze and Reproduction of the Reference Protocol

Before any follow-up experiment, the reference protocol was frozen: the sha256 of the reference checkpoints, the model snapshot (`Qwen/Qwen3-30B-A3B` at revision `ad44e777`), the tokenizer and chat template, and every table-to-raw-file mapping were recorded. Expert masks for B (default) and C (four-view rank-union) were reconstructed from the cached observations and verified by exact weight equality against the saved checkpoints; the reconstruction exposed cutoff ties in five layers of C, so the canonical masks are the weight-derived ones, stored as explicit expert identities.

Reproduction of the reference protocol. On a fixed evaluation set the deterministic axes reproduce the reference values to three decimals (Korean perplexity 6.321 / 112.706 / 7.389 for the unpruned, default, and C models against 6.32 / 112.71 / 7.389; en→ko teacher-forced excess $+2.733$ against $+2.73$), and the generative axes agree within about one point (ProX-ko 0.578 / 0.003 / 0.449; HumanEval+ 0.884 / 0.866 / 0.872 against 0.896 / 0.866 / 0.878). Re-rendering the generative task in-task gives identical per-item outcomes under greedy decoding.

Observer bridge. A KoAlpaca observation re-collected months later with the current code reproduces the original 256-batch, seed-42 view exactly (`reap` and `expert_frequency` identical on 48/48 layers, 0 bottom-32 mismatches), so the matched-budget designs rest on the same observer definition as the reference results. The observation files (five Qwen3-30B and five Qwen3.6 views, 6.7–21.6 MB each) are released with their sha256; the exchange checkpoints themselves (4.5 TB) are regenerated on CPU in 3–7 minutes each from the design files and the original weights, and each carries an `exchange_manifest.json` with the design, builder, and surgery hashes.

Qwen3.6 environment. Qwen3.6 requires transformers 5 and vLLM 0.26 in an isolated environment. Three issues were fixed with the failing runs preserved: the pruned text-only checkpoint needs an explicit M-RoPE position interface in vLLM (exact for text-only input, where the three position channels coincide and the encoding reduces to one-dimensional RoPE); FLORES must be loaded from a local dump because `datasets` 5 rejects script datasets; and the EvalPlus harness could not load the model in this environment, so the code axis of the Qwen3.6 *exchange* checkpoints is reported as unmeasured rather than as a value (the Qwen3.6 code scores in Table 1 and Section 4 come from the earlier standard-surgery checkpoints with the flattened configuration described below, which EvalPlus did load). Reference Qwen3.6 checkpoints pruned earlier by the standard surgery carry a flattened text-only configuration without `mrope_section`, so they never take the M-RoPE

path; the exchange checkpoints keep the original configuration and do. For text-only evaluation the two paths are numerically identical.