TOKENIZATION INVARIANCE IS LEARNED LATE AND BROKEN AT THE HEAD

Anonymous authors
Paper under double-blind review

ABSTRACT

Splitting the input of a subword language model into single characters raises its bits per character five- to thirteen-fold, which is commonly taken to mean that the model is tied to the tokenization it was trained on. We argue that the failure is largely confined to the output head, and that what the trunk has learned survives re-tokenization on a schedule and with a specificity that current accounts do not predict. Probing fifteen pretraining checkpoints of OLMo-2-7B with frozen linear probes and matched random-direction control probes, we find that the model solves a word-order task within its first 9 billion tokens, but that the probe direction becomes reliably readable on character-split input, after a per-layer mean correction, only at about 630 billion tokens, roughly seventy times later. This invariance is specific to the task direction, increases with scale within OLMo-2 (one anomalous final checkpoint is flagged), and has not appeared by the end of two shorter Pythia runs. In released models it is graded in the fraction of tokens split and specific to the construct: no semantic-content probe in four models transfers more selectively than random directions in a way that survives a change of pooling rule, so re-tokenization spares positional structure in strong trunks rather than content. On the generation side, every model continues character-split prefixes at 1.3 to 1.8 times its canonical bits per character but emits character-split text at 5 to 13 times, and restricting the output distribution to character tokens removes at most a quarter of that collapse. Retraining the output head, 6.5% of the parameters, over the frozen trunk brings Llama-3.1-8B to within 0.63 bits per character of canonical; a capacity-matched adaptation of the trunk, run as a registered control, repairs less and does not preserve canonical reading. The part of the head repair attributable to the trunk tracks probe-measured invariance across eight models (exact one-sided Spearman $\rho = 0.83$, $p = 0.008$, registered in advance). Twenty predictions were registered before their data were collected; twelve failed, and every outcome is reported.

1 INTRODUCTION

Every input to a subword language model passes through a tokenizer, one point in a large design space of segmentations (Mielke et al., 2021). The usual assumption is that a trained model and its tokenizer cannot be separated, and the strongest evidence for this assumption is behavioral. When the input of Llama-3.1-8B (Grattafiori et al., 2024) is split into single characters, its bits per character rise from 1.31 to 16.63; among the eight models we measure (Table 5, Appendix D), even the mildest case, GPT-2 (Radford et al., 2019), degrades nearly five-fold. Whether the assumption holds decides whether a trained model's tokenizer can be replaced without retraining the model, a question that tokenizer-transfer and byte-level work (Minixhofer et al., 2024; 2025a) treats as open. We argue that the behavioral evidence mislocates the failure. The representations that support task performance survive re-segmentation, they do so late in pretraining and only for some constructs, and the residual failure sits in the output head, the final projection from hidden states to token logits, which in Llama-3.1-8B holds 6.5% of the parameters.

Our measurements address four gaps in prior designs. Behavioral studies of non-canonical tokenization (Zheng et al., 2025; Chai et al., 2024) do not observe representations, and representational studies of segmentation geometry (Jain et al., 2026) do not separate task-specific from generic robustness. Neither line of work compares when invariance is acquired with when task competence

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.

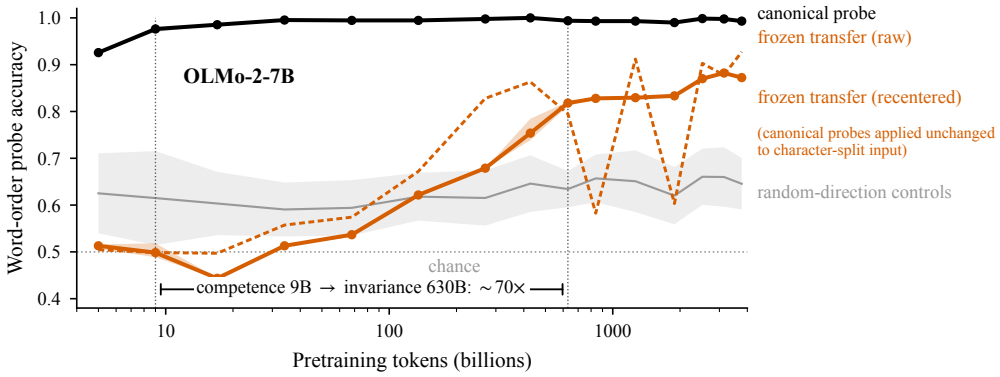

Figure 1: Competence and invariance are acquired on different timescales. Word-order probe accuracy (intact vs. word-shuffled sentences) for OLMo-2-7B at fifteen checkpoints spanning 5 to 3,775 billion pretraining tokens (log axis). Black: probes fitted and evaluated on canonical input (the model's own tokenization). Solid vermilion: recentered frozen transfer (the same probes applied unchanged to character-split input, in which every character is encoded as its own token, after the per-layer mean correction of Section 2). Dashed vermilion: frozen transfer without recentering. Gray line and band: mean and ± 1 SD of the recentered transfer of twenty matched random-direction control probes per checkpoint (Section 2). Bands on the solid curves: range over three probe-fitting seeds. Dotted horizontal line: chance. Dotted vertical lines and the bracket mark the competence point (9 billion tokens), the invariance plateau (630 billion tokens), and their ratio (rules in Section 3).

is acquired, and neither locates the behavioral collapse by intervention. Closing two of these gaps changes qualitative conclusions (Section 2): without matched controls, generic drift of the representation space is mistaken for task robustness, and without a per-layer translation correction, a probe applied unchanged across tokenizations can lose its signal and appear as total loss.

We report three main findings. First, tokenization invariance is learned on its own schedule. OLMo-2-7B (Team OLMo et al., 2025) reaches competence on a word-order task within 9 billion pretraining tokens but holds the same probe direction stably under character-split input only from about 630 billion tokens, roughly seventy times later on the recentered measure and thirty times later on raw transfer (Figure 1). Within OLMo-2 the final level of invariance increases with scale (the 13B endpoint is anomalous, Section 3), and two Pythia models (Biderman et al., 2023) have not acquired it by the end of their runs (Section 3). Second, what survives is positional structure. Word-order retention rises from 0.08 to 0.87 across four models and, in OLMo-2-7B, clears the matched controls under both pooling rules and at every re-tokenization dose; semantic-content and acceptability probes clear theirs in no cell that survives a change of pooling rule (Section 4). Third, the behavioral collapse is chiefly a failure of writing, and it can be attributed among the output support, the trunk's features, and the head's mapping. Where the trunk (all layers below the output head) is strong only the mapping is broken: retraining 6.5% of Llama-3.1-8B's parameters over its frozen trunk restores generation to within 0.63 bits per character of canonical, and the part of the repair attributable to the trunk rises with the trunk's measured invariance across eight models. A capacity-matched adaptation of the trunk, run as a registered control, repairs less and loses canonical reading, an outcome the registration had not anticipated (Section 5).

Twenty quantitative predictions were registered before the data that tested them were collected; twelve failed, and Section 6 reports each outcome.

2 Measurement Design and Controls

Our measurements compare representations of the same text under two segmentations. Both design choices, where to compare and against what baseline, can invalidate probing claims (Belinkov, 2022), so we state them explicitly.

Substrate and conditions. In a causal model the hidden state at position t depends on the prefix, and two segmentations agree on their prefixes exactly at shared segment boundaries. We therefore mean-pool hidden states over the end-of-word positions that both segmentations share (alignment rule in Appendix C), giving one vector per text and layer; comparing on a character grid would confound positional mismatch with representational change. The *canonical* condition encodes text with the model's own tokenizer. The *character-split* condition encodes every character as its own token, drawn from the model's own vocabulary, so no unknown tokens arise. Behavior is measured in bits per character (bpc), which is comparable across segmentations because the character count is fixed while token counts vary (Ahia et al., 2023; Chatzi et al., 2025).

Frozen and refitted probes. A linear probe (Alain & Bengio, 2016; Conneau et al., 2018a) fitted on canonical representations and applied unchanged to character-split representations tests whether task information survives *in place*, with the same direction and the same threshold. A probe refitted on character-split representations tests whether the information survives at all (Voita & Titov, 2020; Elazar et al., 2021). Fitting on clean text and evaluating on perturbed input follows Zhuo et al. (2025). Probes are logistic classifiers fitted in float64 with an internal validation split for the ridge parameter and three fitting seeds, reported as min-to-max bands; bootstrap confidence intervals are stored with each result file (Appendix C). We report frozen *transfer*, the accuracy of the canonical probe on character-split input, and *retention*, the fraction of the probe's canonical headroom (accuracy above chance) that survives frozen transfer. Throughout, *invariance* refers to a task probe's recentered frozen transfer (the Recentering paragraph below), *retention* to its normalized form, which we use when comparing models whose canonical accuracies differ, and a *strong* or *weak* trunk to high or low word-order retention.

Selectivity controls. A frozen probe can also transfer for reasons unrelated to the task, since every direction survives if the representation cloud barely moves between segmentations. Following the logic of control tasks (Hewitt & Liang, 2019), each cell (one model and task pair) is accompanied by twenty control probes fitted on dichotomies defined by random directions in canonical representation space (median splits with the 20% of texts nearest the boundary dropped; Appendix C), and task transfer is reported against the spread of these control transfers, which we call the control envelope, rather than against chance.

Recentering. Frozen-probe transfer between segmentations can be destroyed by a per-layer translation: the character-split cloud lies elsewhere in representation space, a fixed threshold falls on the wrong side of it, and accuracy drops to the floor of a constant prediction. Subtracting the character-split mean and adding the canonical mean restores the comparison; because the probe is linear, this recalibrates its intercept and leaves its direction unchanged. Recentered values describe geometry exposed by the analysis; the model does not apply this correction on its own, and Section 5 shows that it does not repair generation. We report raw and recentered transfer side by side.

Layer selection and registration. Reported statistics use the layer selected on a class-stratified half of the evaluation set and are evaluated on the disjoint other half (split construction and its integrity record in Appendix B). Because probing studies are easy to steer after the fact (Cockburn et al., 2020), the project adopted preregistration (Nosek et al., 2018; Chambers & Tzavella, 2022; van Miltenburg et al., 2021; Bertinetto et al., 2021): quantitative thresholds were fixed and timestamped before their metrics were computed. Section 6 reports all twenty registered predictions.

3 Acquisition of invariance during pretraining

If robustness to re-tokenization were a by-product of learning a task, the two would appear together during pretraining. We tested this on OLMo-2-7B at fifteen checkpoints spanning 5 to 3,775 billion tokens, each verified by shard checksum (Appendix B), using a word-order task: distinguishing sentences drawn from SST-2 (Socher et al., 2013) from word-shuffled counterparts. Shuffling preserves lexical content and destroys the positional binding that character splitting disrupts, and unlike standard NLI benchmarks (Sinha et al., 2021; Pham et al., 2021) the task is read nearly perfectly by canonical probes. We use competence as shorthand for canonical linear decodability and date it at the first checkpoint whose canonical probe accuracy reaches 0.95, the invariance plateau at the first

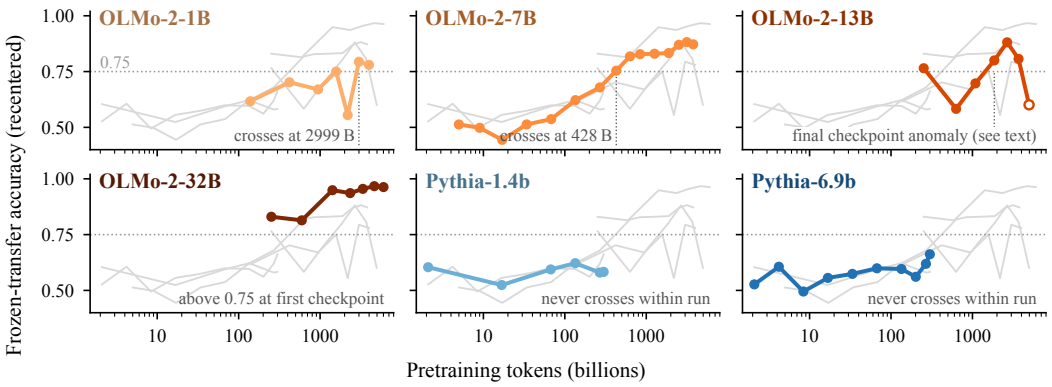

Figure 2: Acquisition of invariance across six pretraining runs: recentered frozen transfer (the probe fitted on canonical input, applied unchanged to character-split input, after the per-layer mean correction of Section 2) on the word-order task (intact vs. word-shuffled sentences) at the measured checkpoints, six to fifteen per run (Table 11). Gray lines repeat the other five runs for context; the dotted horizontal line marks the 0.75 criterion, and dotted vertical lines mark the sustained crossing where one occurs within the run (for OLMo-2-13B the unlabeled crossing at 1854 billion tokens holds only through its penultimate checkpoint). Within OLMo-2, final invariance increases with scale, and the 32B model is above the criterion at its first checkpoint; neither Pythia model crosses 0.75 within its run (Table 1). The final checkpoint of OLMo-2-13B (open marker) is out of line with its own trajectory (Section 5 and Appendix F).

checkpoint from which recentered transfer stays within 10% of its final value over two consecutive checkpoints, and a 0.75 crossing at the first of two consecutive checkpoints above 0.75 (Table 1).

Figure 1 shows the result. Canonical probe accuracy reaches 0.98 by 9 billion tokens and stays there. Recentered frozen transfer to character-split input starts at chance and dips briefly below it (0.44 at 17 billion tokens). It is still near chance at 68 billion tokens, reaches 0.82 to 0.83 by 630 billion tokens, does not fall back, and ends the run at 0.87. Invariance therefore arrives roughly seventy times later than decodability. The rise is significant (one-sided exact Spearman test, $p = 5 \times 10^{-5}$), while the matched random-direction controls stay between 0.59 and 0.66 across the whole run, so the growth is specific to the task direction. Raw frozen transfer shows the same late arrival more noisily: it first crosses 0.75 at 269 billion tokens, thirty times after competence, and then swings between 0.58 and 0.93 across neighboring checkpoints, because the per-layer translation that recentering removes differs between checkpoints and between the layers selected at each checkpoint (Appendix C).

Scale and lineage. We repeated the measurement on the Pythia-6.9b trajectory and on four further lineages whose predictions were registered in advance (OLMo-2-1B, 13B, and 32B, and Pythia-1.4b; Figure 2, with statistics in Table 1). Within OLMo-2, final invariance increases with scale: 0.78 at 1B, 0.87 at 7B, and 0.96 at 32B (0.79, 0.88, and 0.95 at matched budgets of 2999B, 3146B, and 3348B; Table 11). The 32B model is already above the 0.75 criterion at its first published checkpoint (0.83 at 252 billion tokens) and is still rising ($p = 0.006$). Neither Pythia model crosses 0.75 within its run, but the Pythia runs end near 300 billion tokens, before the 428 billion at which OLMo-2-7B first crossed. Pythia-1.4b ends flat at 0.58 (trend $p = 0.643$), and we had registered in advance that it would end at or below 0.6 (Table 2, row 5); Pythia-6.9b rises detectably ($p = 0.028$), but only from 0.53 to 0.66, a range that OLMo-2-7B had left behind by 269 billion tokens, while its control envelope rises faster, from 0.62 to 0.85, so the trend is not task-selective (Table 11). Data, tokenizer, and recipe differ between the families, so we make no causal attribution; the contrast shows that invariance is not automatic at these budgets. The one anomaly is OLMo-2-13B, whose intermediate checkpoints rise to 0.88 while its final published checkpoint reads 0.60. Section 5 shows by intervention that the trunk at that checkpoint retains 7B-level capability, so the anomaly lies in what the mean-pooled probe can read there rather than in the model (Appendix F). Four registered trajectory predictions failed at least in part through this lineage (Table 2, rows 3, 4, 6, and 7; row 7 because 13B crossed 0.75 after 7B; row 3, the content prediction, is adjudicated in Section 4).

Table 1: Acquisition statistics per lineage: number of checkpoints, final recentered transfer, exact one-sided Spearman p for an increasing trend, the sustained 0.75 crossing, and the ratio of the invariance plateau to the competence point (rules in Section 3; for 7B, 630B over 9B). Censored ratios (1B, competence at its first checkpoint 139B; 32B, crossing at its first checkpoint 252B) are lower bounds; 32B's plateau is at 1418B and 1B's coincides with its crossing (Appendix F).

Lineage	checkpoints	final	trend p	crosses 0.75 at	plateau/competence
OLMo-2-1B	7	0.78	0.100	2999B	$\geq 21.6\times$ (competence censored)
OLMo-2-7B	15	0.87	5×10^{-5}	428B	$70\times$
OLMo-2-13B	7	0.60	0.278	1854B	(final checkpoint anomalous)
OLMo-2-32B	7	0.96	0.006	$\leq$ 252B (censored)	$\geq 5.6\times$ (censored)
Pythia-1.4b	6	0.58	0.643	not within run	—
Pythia-6.9b	10	0.66	0.028	not within run	—

4 SELECTIVITY ACROSS CONSTRUCTS AND LANGUAGES

Figure 6 (Appendix D) compares selectivity for three constructs in four models under both pooling rules: binarized natural-language inference (Williams et al., 2018) as semantic content, the word-order task as positional structure, and BLiMP minimal-pair acceptability (Warstadt et al., 2020) as fine-grained syntax. The content cells use class-balanced evaluation splits, 436 texts per class (Appendix C). Word order is the graded construct: GPT-2 retains almost none of it (0.08), Pythia-6.9b a third (0.33), OLMo-2-7B 0.70, and Llama-3.1-8B 0.87, consistent with Section 3 for the two lineages that have trajectories. In the two weak models the order probes sit at or below their control envelopes, so those retentions are upper bounds on task-selective survival. Content retention is 0.56, 0.15, 0.53, and 0.88 in the same four models, but the content probes transfer no more selectively than random directions: their selectivity gap (recentered task transfer minus the control mean; the registered margin is 0.10) is within 0.06 of zero in three models, and GPT-2's 0.11 does not survive a change of pooling (0.06; Table 10). Across the four lineages measured through training, content falls below 0.6 retention at 18 of the 25 checkpoints past 50 billion tokens (Table 2, row 3). Syntactic acceptability is retained at most 0.46 even in Llama-3.1-8B, although canonical probes read it at 0.80 to 0.92 in all four, and it clears the margin only in Llama-3.1-8B (0.14), a cell that falls to 0.04 under last-token pooling.

We had registered that acceptability would exceed 0.6 retention in all four models (Table 2, row 11); it failed in every model. The registration's pre-specified alternative reading was a hierarchy of constructs ordered by retention. That ordering does not hold: the three constructs change order across the four models, and the pooling rule moves cell retentions by 0.12 on average and flips the sign of the selectivity gap in six of the twelve cells (Table 10). One cell is stable under both pooling rules and at every re-tokenization dose: the word-order direction of OLMo-2-7B transfers more selectively than random directions whether a quarter, half, three quarters, or all of the tokens are split (Table 9), with its gap largest at half dose (0.25 vs. 0.19 at the full split), and no content or acceptability cell clears the margin under both pooling rules (Llama-3.1-8B's SST-2 order cell does so under last-token pooling only; Table 10).

Languages. The order result replicates beyond English. On word-order tasks built from XNLI premises (Conneau et al., 2018b), we report the order-selectivity gap: for Llama-3.1-8B it is 0.42 in German, 0.28 in English and in French, 0.20 in Swahili, and 0.14 in Spanish (Figure 7, Appendix D). The three languages whose margins were registered in advance all clear the 0.10 threshold (Table 2, row 8); English and German were measured outside the registration and are reported alongside. Content retention exceeds 0.6 in Spanish (0.70) and French (0.76) but not in German (0.50) or Swahili (0.43), so the registered content prediction fails (row 9), and no language's content probe clears the selectivity margin (gaps of -0.07 to 0.08; Table 3). The registered resource-level ordering (row 10), which was anchored at English, is violated at both ends: English's gap is below German's, and Swahili's exceeds Spanish's. Tokenizer quality varies sharply with language and resource level (Rust et al., 2021; Petrov et al., 2023), which bears on the Swahili result.

5 LOCALIZING THE BEHAVIORAL COLLAPSE

Sections 3 and 4 show that a strong model's word-order representation survives re-tokenization, yet generation collapses. Character-split input raises Llama-3.1-8B from 1.31 to 16.63 bpc, and all eight lineages we measure (GPT-2, Pythia-1.4b, Pythia-6.9b, OLMo-2-1B/7B/13B/32B, and Llama-3.1-8B) degrade by factors of 4.9 to 12.7 (Table 5, Appendix D). This section locates the break by elimination among three candidates: the output *support* (probability mass on tokens that cannot occur in character-split text), the *features* (the trunk computing the wrong thing), and the *mapping* (a head that never learned to read the surviving features into character tokens). Figure 3 shows the resulting interventions on Llama-3.1-8B, each applied separately to the same model; we call the series the repair ladder and each intervention a rung. Appendix E reports all eight lineages.

Restricting the output distribution to the character-token inventory and renormalizing, with no training, moves bpc from 16.63 only to 13.72, which rules out support as the main cause. Retraining the output head on character-split text over a randomly initialized trunk reaches 5.26 bpc, the level a head reaches from character statistics alone. The same retraining over the real frozen trunk reaches 1.94 bpc, within 0.63 of the model's canonical 1.31, and the retrained parameters are 6.5% of the model. A control head retrained on canonical text with the same budget and then evaluated on character-split input reads 16.61, unchanged from 16.63, so the improvement is not a generic effect of retraining the head.

The difference between the random-trunk and real-trunk rungs, 3.32 bpc, is the contribution of the pretrained trunk to character-level generation under a fixed head-training procedure; we call it the trunk gap. We tested whether it tracks the probe results across models by extending the ladder to all eight lineages under five further predictions registered before the extension was computed (Table 2, rows 12 to 16). The registration paired each lineage's gap with the word-order retention available at the time: released-weight retention for GPT-2, the two Pythia models, OLMo-2-7B, and Llama-3.1-8B, and the last stage-1 pretraining checkpoint for OLMo-2-1B, 13B, and 32B, whose released weights include a further training stage. On that pairing the gap increases with retention across the eight (exact one-sided Spearman $\rho = 0.83$, $p = 0.008$; Figure 3, right); pairing the three registered OLMo-2 lineages with retentions measured on their released weights instead gives $\rho = 0.98$, $p = 0.0002$ (not plotted). Two extension predictions failed. The 32B gap (3.01) fell below Llama's (3.32), where we had predicted it would be largest; released-weight retentions (0.83 for OLMo-2-32B vs. 0.87 for Llama-3.1-8B) match the gap order, so the miss traces to the stage-1 anchor, and the 0.31 difference is of the same order as the run-to-run variation of the random-trunk rung (about 0.2 bpc; Appendix E). The 13B ladder gives a gap of 2.51 on the released weights and 2.34 on the anomalous checkpoint itself, both at the 7B level (2.40), although the frozen probe reads that checkpoint at only 0.21 retention; the trunk therefore keeps its structure, and the anomaly of Section 3 is confined to the probe reading (Appendix F).

Localization control. The ladder does not test whether an adaptation of comparable size elsewhere would repair generation as well. We registered four predictions (Table 2, rows 17 to 20) and then trained LoRA adapters on every projection of Llama-3.1-8B's trunk, 503M trainable parameters against the head's 525M, on the same character-split stream at the same budget with the head frozen (Table 7). The trunk arm reaches 3.38 bpc against the head arm's 1.94, above the registered margin of 1.0, and adding it to head retraining worsens the head arm by 2.01 (rows 17 to 19 pass; row 19 only on its wording). The fourth fails: the trunk arm was to keep canonical reading within 0.15 bpc of baseline and reads canonical text at 6.73 against 1.31 (row 20). The registration fixed no reading for this combination, so the control does not confirm the localization. It does show an asymmetry we did not register: on the same character-only stream, the head arm repairs character-split generation at a cost of 0.23 bpc to canonical reading and the trunk arm at a cost of 5.43.

Translation patching and behavior. Two further results bound the repair. First, the per-layer mean correction that repairs frozen probes does not repair generation. Applied inside the forward pass, it removes at most 27% of the collapse (the excess of character-split over canonical bpc) at the best single layer, none of it in GPT-2, and leaves generation five- to ten-fold above canonical in all four models of Section 4 (Appendix E). Linear decodability and usable generation therefore dissociate. Second, the collapse is mostly a failure of writing. Every unmodified model continues a character-split prefix at 1.3 to 1.8 times its canonical bpc (Table 8), against a 4.9- to 12.7-fold cost for

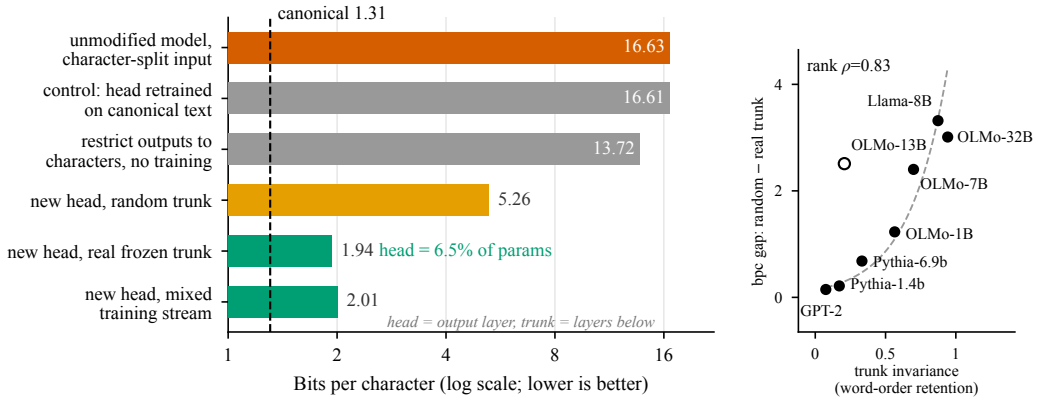

Figure 3: Repair ladder for Llama-3.1-8B (left): bits per character on character-split input (log axis; lower is better; the dashed line marks canonical reading, 1.31). Bars are separate interventions on the same model and do not stack. Restricting outputs to character tokens without training lowers 16.63 only to 13.72, and a control head retrained on canonical text leaves it at 16.61; a new head trained on character-split text over a randomly initialized trunk reaches 5.26; the same procedure over the real frozen trunk reaches 1.94 while retraining 6.5% of the parameters (fractions for the other lineages in Table 6); a mixed canonical-and-character head reaches 2.01 (Appendix E). Right: the trunk gap (random-trunk minus real-trunk bpc) against word-order retention under the registered pairing described in the text, for all eight lineages (exact one-sided Spearman $\rho = 0.83$, $p = 0.008$; model names abbreviated in the panel). Dashed curve: a least-squares exponential fit to the seven filled points, a visual guide only. The open point, excluded from the fit but included in ρ, is OLMo-2-13B at its anomalous stage-1 endpoint (retention 0.21); its gap (2.51 as plotted, 2.34 on that checkpoint itself) places the anomaly of Section 3 in the probe reading (Appendix F).

emitting character-split text. Prompted accuracy on character-split input, which reads the input and writes canonical label tokens, falls by 0.04 to 0.38 in seven of the eight models (sampling error about ± 0.05 at $n = 300$) and not at all in Llama-3.1-8B. Head retraining acts on the writing side. A head trained on character text alone lowers prompted accuracy on character-split input, because prompt scaffolds and label words are canonical tokens; a head trained on a mixed canonical-and-character stream removes that cost (Llama-3.1-8B: 0.55 under the character-only head, 0.73 under the mixed head, 0.70 unmodified) while keeping canonical bpc at 1.48; and no retrained head raises prompted accuracy on character-split input above the unmodified model's by more than sampling error, while in OLMo-2-32B it lowers it from 0.92 to 0.71. Continuation of a character-split prefix worsens under both retrained heads (1.63 to 3.68 and 2.80 bpc; Table 8), so the repair is specific to emission. A head cannot add invariance that the trunk lacks: both Pythia models stay at chance on character-split prompts under every head we trained, consistent with their low invariance in Section 3.

Granularity dependence. Re-segmenting text at another tokenizer's boundaries, with each piece encoded in the model's own vocabulary, costs almost nothing. Llama-3.1-8B reads OLMo-2-boundary segmentations at 1.31 bpc, indistinguishable from its canonical 1.31, and GPT-2-boundary segmentations at 1.44 (Figure 11, Appendix E), so the collapse is specific to granularities far outside the training distribution. The repair is therefore localization evidence rather than a deployment method, and the result is in line with Schmidt et al. (2024), who find that a tokenization's value is not reducible to its surface statistics.

6 Registered predictions and outcomes

Table 2 lists every registered prediction and its outcome. Rows 1 and 2 predicted a dissociation between refitted and frozen probes; it held in none of the sixteen surveyed cells, because refitted probes also degrade in the weaker models and frozen probes degrade little in the larger Llama models (Table 4, Appendix D), so refit-versus-frozen does not discriminate task-selective survival, and transfer is reported against control envelopes (Section 2). The remaining rows are adjudicated

Table 2: Registered predictions and their outcomes. Eight of the twenty passed, one of them on its wording, and twelve failed. Each block header states what the registration preceded, and for rows 3 to 7 the registration discloses that raw trajectory endpoints had been seen under a superseded protocol. Appendix A restates each prediction with its chronology, and the verbatim wording, timestamps, and amendments are in the repository.

Registered prediction	Outcome
Cross-model survey, sixteen model–task cells (registered before confirmatory data)	
1. Refit recovers within an equivalence margin while frozen transfer drops by half its headroom, jointly in ≥ 12 of 16 cells	failed (0 of 16)
2. Refit damage (Δrefit) decreases monotonically across the Llama base series (1B to 70B)	failed ($p = 0.17$)
Acquisition trajectories, four registered lineages (registered before trajectory data)	
3. Content retention ≥ 0.6 at every checksum-verified checkpoint past 50B tokens	failed (18 of 25 cells)
4. Final order transfer ≥ 0.85 in the three registered OLMo-2 lineages	failed (only 32B met it)
5. Pythia-1.4b final order transfer ≤ 0.6	passed (0.58; Section 3)
6. Order transfer rises and crosses 0.75 within each registered OLMo-2 lineage	failed (crossings in all three; rise significant only in 32B)
7. Larger OLMo-2 models cross 0.75 earlier	failed (13B after 7B)
Multilingual retention, Llama-3.1-8B (registered before multilingual data)	
8. Order-selectivity gap > 0.10 in Spanish, French, Swahili	passed (0.14 to 0.28)
9. Content retention > 0.6 in the three registered languages	failed (Swahili 0.43)
10. Order gaps do not increase as resource level decreases (chain anchored at English)	failed (en $<$ de; sw $>$ es)
Acceptability retention (registered before acceptability data)	
11. Acceptability retention > 0.6 in all four models, with a pre-specified alternative reading	failed (all four; Section 4)
Ladder extension (registered after the four-model ladder, before extension compute)	
12. Trunk gap increases with final-checkpoint order retention across all eight lineages (exact one-sided Spearman)	passed ($\rho = 0.83$, $p = 0.008$)
13. OLMo-2-32B posts the largest trunk gap	failed (3.01; Llama 3.32)
14. OLMo-2-13B's gap falls below OLMo-2-7B's	failed (2.51 vs. 2.40; Section 5)
15. Pythia-1.4b's gap falls below Pythia-6.9b's	passed (0.21 vs. 0.68)
16. GPT-2's rerun gap < 0.4, remaining smallest of all	passed (0.15)
Localization control, Llama-3.1-8B (registered after the extension, before its compute)	
17. LoRA on the trunk with the head frozen ends ≥ 1.0 bpc above the head-only repair	passed (3.38 vs. 1.94)
18. The trunk arm stays above 3.0 bpc	passed (3.38)
19. Adding the trunk arm to head retraining improves it by < 0.3 bpc	passed on its wording (worsens by 2.01)
20. The trunk arm keeps canonical reading within 0.15 bpc of baseline	failed (6.73 vs. 1.31)

where their measurements are reported: rows 4 to 7 in Section 3, 3 and 8 to 11 in Section 4, and 12 to 20 in Section 5.

7 Discussion

Candidate mechanisms for the delay. The cause of the seventy-fold delay is not established; the three candidates below are speculation with distinguishable predictions. (i) Redundancy first: models may initially solve tasks with segmentation-bound shortcuts and only later encode word identity redundantly enough to survive re-segmentation; knowledge of token internals (Kaushal & Mahowald, 2022; Itzhak & Levy, 2022) should then mature before invariance on the same checkpoints. (ii) Tail exposure: invariance may track accumulated exposure to rare segmentations of familiar strings; train-time segmentation noise (Kudo, 2018; Provilkov et al., 2020) should then bring the onset forward at matched data. (iii) Delayed consolidation: invariance may be a slow secondary generalization of the kind seen in abrupt training phase changes (Olsson et al., 2022; Chen et al., 2024) and grokking (Power et al., 2022); late-training reorganization should then appear without any change in the data.

Implications. Reading character-split input costs every unmodified model little in likelihood and, in Llama-3.1-8B and OLMo-2-32B, little in prompted accuracy (Table 8), consistent with Zheng et al. (2025). The interface view further predicts that the language dependence reported by Ghosh & Jyothi (2026) may track language-graded order retention, and that checkpoint-level frozen-probe invariance can serve as a readiness signal for tokenizer surgery, byteification (Minixhofer et al., 2025a), and cross-tokenizer or byte-level distillation (Minixhofer et al., 2025b; Marathe et al., 2026). These remain untested.

8 Related work

Subword tokenization descends from compression (Gage, 1994; Sennrich et al., 2016; Kudo & Richardson, 2018); its formal structure, costs, failure modes, and cross-language inequities are active topics (Bostrom & Durrett, 2020; Rust et al., 2021; Ahia et al., 2023; Zouhar et al., 2023; Gutierrez-Vasques et al., 2023; Petrov et al., 2023; Batsuren et al., 2024; Land & Bartolo, 2024).

Character-level and adversarial perturbations have long broken NLP systems (Belinkov & Bisk, 2018; Ebrahimi et al., 2018; Pruthi et al., 2019), and LLMs remain sensitive to segmentation and format (Chai et al., 2024; Sclar et al., 2024; Edman et al., 2024), although instruction-tuned models keep most prompted accuracy under non-canonical tokenization (Zheng et al., 2025), with strong language dependence (Ghosh & Jyothi, 2026).

Mechanistically, models re-assemble split words in early layers (Yang et al., 2026; Kaplan et al., 2025; Feucht et al., 2024) and know much about the characters inside their tokens (Kaushal & Mahowald, 2022; Itzhak & Levy, 2022), and Datta et al. (2026) reach an output-side diagnosis of character-task failures that our ladder quantifies; cheap affine corrections likewise rescue frozen readouts across depth (nostalgebraist, 2020; Belrose et al., 2023) and under distribution shift (Kumar, 2026; Mahadik & Skapars, 2026). On training dynamics, Jain et al. (2026) track generic re-tokenization symmetry across pretraining and post-training checkpoints and find gradual pretraining gains with a pronounced post-training contribution, whereas our matched controls show the task-selective component consolidating during base pretraining. Karne (2026) finds script invariance growing with scale.

On the interface side, tokenizer transfer and vocabulary adaptation (Gee et al., 2022; Minixhofer et al., 2022; Dobler & de Melo, 2023; Minixhofer et al., 2024; Remy et al., 2024; Haltiuk & Smywinski-Pohl, 2025; Özeren et al., 2025; Minixhofer et al., 2025b), byteification and byte-level distillation (Minixhofer et al., 2025a; Marathe et al., 2026), and tokenizer-free architectures (Xue et al., 2022; Clark et al., 2022; Tay et al., 2022; Nawrot et al., 2022; 2023; Yu et al., 2023; Wang et al., 2024; Slagle, 2024; Kallini et al., 2025; Neitemeier et al., 2025; Videau et al., 2025; Hwang et al., 2025; Pagnoni et al., 2025) engineer around the failure we measure. Our foreign-boundary result matches the boundary-versus-capability separation Haltiuk (2026) hypothesizes; train-time segmentation regularization (Kudo, 2018; Provilkov et al., 2020) attacks the deficit from the training side.

9 Limitations and conclusion

Recentered quantities describe geometry exposed by the analysis, not information the model uses (Section 2); the seventy-fold delay is a recentered estimate that also depends on the dating rules (48 to at least 126 under alternatives, Appendix F), raw transfer gives thirty-fold, and the capability claims rest on raw transfer and on the repairs. The trajectory evidence comes from one task, English word order, with one run per lineage on grids of 6 to 15 checkpoints. Two ratios in Table 1 are censored lower bounds, and the 13B probe anomaly is probe-side but unexplained (Appendix F). Cell-level retentions depend on the pooling rule: OLMo-2-7B's order result survives a change from mean to last-token pooling, no content or acceptability cell does, and the row-12 relation holds under last-token retentions ($\rho = 0.71$, $p = 0.03$; Table 10). The localization control did not give a clean test, because its trunk arm failed the registered bound on canonical reading (row 20). The constructs are three classification tasks in five languages; generation is measured only as bpc and prompted accuracy; and all models are open-weight (124M to 70B).

For word order in OLMo-2-7B, tokenization invariance arrives roughly seventy times later than canonical decodability and is not guaranteed at every budget. Positional structure becomes invariant in strong trunks; semantic content shows no task-selective survival that is stable across models and pooling rules. The collapse is chiefly in writing and sits in the output head: retraining 6.5% of Llama-3.1-8B over its frozen trunk brings generation to within 0.63 bpc of canonical, and the trunk's share of that repair tracks its invariance across all eight lineages.

AI use statement

We used generative AI tools extensively, under the direction of the authors, who approved and froze every registration and reviewed all results. For tasks with required disclosure, an AI assistant pro-

posed candidate hypotheses and drafted the registration documents that the authors then approved and froze, provided feedback on the experimental design through adversarial review panels, implemented the measurement instruments and experiment scripts, ran and monitored the experiments, performed the statistical analyses, and drafted interpretations of the results for the authors' review. For tasks with recommended disclosure, it searched and summarized related literature, generated the figures from the result files, drafted the text of this paper, edited it for readability, and formatted the references. Generative AI was not used to set the research questions or to make the strategic decisions of the project. We have reviewed all AI-assisted work, including the code, the analyses, the figures, and the text. Verification included an offline test suite of 141 cases, predictions registered before data collection, repeated adversarial reviews by independently prompted AI panels, and human checks of every claim against the underlying result files. The panel reviews found the evaluation-split bug and the class-imbalanced content floor of Appendix B, each followed by a full re-measurement, and two ledger values that had not been updated, re-adjudicated as failures. All citations were verified against primary sources. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Ethics statement

This work probes open-weight models on public datasets and involves no human subjects. A data-integrity issue discovered in a public checkpoint repository is documented in Appendix B.

Reproducibility statement

All code, experiment scripts, result files, registration documents, and the analysis behind every number are in the accompanying repository, available anonymously at `https://anonymous.4open.science/r/tokenlock-supplement-B656/` and as a supplementary archive, together with a reproduction guide that maps each experiment to its script and seeds and an audit script that traces every load-bearing number to its result file. The repository also contains a registered pilot of a different architecture that preceded this program. Checkpoint loads verify resolved commit hashes, and the shard-checksum protocol for public checkpoint repositories is included (Appendix B).

References

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R. Mortensen, Noah A. Smith, and Yulia Tsvetkov. Do all languages cost the same? Tokenization in the era of commercial language models. In *Proceedings of EMNLP*, 2023.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Khuyagbaatar Batsuren, Ekaterina Vylomova, Verna Dankers, Tsetsuukhei Delgerbaatar, Omri Uzan, Yuval Pinter, and Gábor Bella. Evaluating subword tokenization: Alien subword composition and OOV generalization challenge. *arXiv preprint arXiv:2404.13292*, 2024.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Yonatan Belinkov and Yonatan Bisk. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations (ICLR)*, 2018.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Luca Bertinetto, João F. Henriques, Samuel Albanie, Michela Paganini, and Gül Varol (eds.). *Proceedings of the NeurIPS 2020 Workshop on Pre-registration in Machine Learning*, volume 148 of *Proceedings of Machine Learning Research*, 2021.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of ICML*, 2023.

Kaj Bostrom and Greg Durrett. Byte pair encoding is suboptimal for language model pretraining. In *Findings of EMNLP*, 2020.

Yekun Chai, Yewei Fang, Qiwei Peng, and Xuhong Li. Tokenization falling short: On subword robustness in large language models. In *Findings of EMNLP*, 2024.

Christopher D. Chambers and Loukia Tzavella. The past, present and future of registered reports. *Nature Human Behaviour*, 6:29–42, 2022.

Ivi Chatzi, Nina Corvelo Benz, Stratis Tsirtsis, and Manuel Gomez-Rodriguez. Tokenization multiplicity leads to arbitrary price variation in LLM-as-a-service. *arXiv preprint arXiv:2506.06446*, 2025.

Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew L. Leavitt, and Naomi Saphra. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in MLMs. In *International Conference on Learning Representations (ICLR)*, 2024.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. CANINE: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91, 2022.

Andy Cockburn, Pierre Dragicevic, Lonni Besançon, and Carl Gutwin. Threats of a replication crisis in empirical computer science. *Communications of the ACM*, 63(8):70–79, 2020.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of ACL*, 2018a.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of EMNLP*, 2018b.

Ayan Datta, Mounika Marreddy, Alexander Mehler, Zhixue Zhao, and Radhika Mamidi. From early encoding to late suppression: Interpreting LLMs on character counting tasks. *arXiv preprint arXiv:2604.00778*, 2026.

Konstantin Dobler and Gerard de Melo. FOCUS: Effective embedding initialization for monolingual specialization of multilingual models. In *Proceedings of EMNLP*, 2023.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. HotFlip: White-box adversarial examples for text classification. In *Proceedings of ACL*, 2018.

Lukas Edman, Helmut Schmid, and Alexander Fraser. CUTE: Measuring LLMs' understanding of their tokens. In *Proceedings of EMNLP*, 2024.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

Sheridan Feucht, David Atkinson, Byron Wallace, and David Bau. Token erasure as a footprint of implicit vocabulary items in LLMs. In *Proceedings of EMNLP*, 2024.

Philip Gage. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38, 1994.

Leonidas Gee, Andrea Zugarini, Leonardo Rigutini, and Paolo Torroni. Fast vocabulary transfer for language model compression. In *Proceedings of EMNLP*, 2022.

Poulami Ghosh and Preethi Jyothi. Language models are not equally robust to non-canonical tokenization across languages. *arXiv preprint arXiv:2607.26831*, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Ximena Gutierrez-Vasques, Christian Bentz, and Tanja Samardzic. Languages through the looking glass of BPE compression. *Computational Linguistics*, 49(4):943–1001, 2023.

Mykola Haltiuk. Disentangling language modeling and boundaries. *arXiv preprint arXiv:2608.03599*, 2026.

Mykola Haltiuk and Aleksander Smywinski-Pohl. Model-aware tokenizer transfer. *arXiv preprint arXiv:2510.21954*, 2025.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP*, 2019.

Sukjun Hwang, Brandon Wang, and Albert Gu. Dynamic chunking for end-to-end hierarchical sequence modeling. *arXiv preprint arXiv:2507.07955*, 2025.

Itay Itzhak and Omer Levy. Models in a spelling bee: Language models implicitly learn the character composition of tokens. In *Proceedings of NAACL*, 2022.

Kanishk Jain, Matthew Day, and Tankut Can. Emergent retokenization symmetry in large language models: phenomenology and applications. *arXiv preprint arXiv:2606.15521*, 2026.

Julie Kallini, Shikhar Murty, Christopher D. Manning, Christopher Potts, and Róbert Csordás. MrT5: Dynamic token merging for efficient byte-level language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Guy Kaplan, Matanel Oren, Yuval Reif, and Roy Schwartz. From tokens to words: On the inner lexicon of LLMs. In *International Conference on Learning Representations (ICLR)*, 2025.

Sripad Karne. One language, two scripts: Probing script-invariance in LLM concept representations. In *UCRL Workshop at ICLR*, 2026. arXiv:2603.08869.

Ayush Kaushal and Kyle Mahowald. What do tokens know about their characters and how do they know it? In *Proceedings of NAACL*, 2022.

Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of ACL*, 2018.

Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of EMNLP: System Demonstrations*, 2018.

Sachin Kumar. Pressure-testing deception probes in LLMs: Scaling, robustness, and the geometry of deceptive representations. *arXiv preprint arXiv:2605.27958*, 2026.

Sander Land and Max Bartolo. Fishing for Magikarp: Automatically detecting under-trained tokens in large language models. In *Proceedings of EMNLP*, 2024.

Prasad Mahadik and Adrians Skapars. Do linear probes generalize better in persona coordinates? *arXiv preprint arXiv:2605.09391*, 2026.

Kalyani Marathe, Artidoro Pagnoni, Tomasz Limisiewicz, Margaret Li, Mike Lewis, Luke Zettlemoyer, and Srinivasan Iyer. Breaking the token ceiling: Distilling smaller, stronger byte models. *arXiv preprint arXiv:2609.12303*, 2026.

Sabrina J. Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y. Lee, Benoît Sagot, and Samson Tan. Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP. *arXiv preprint arXiv:2112.10508*, 2021.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of NAACL*, 2022.

Benjamin Minixhofer, Edoardo Maria Ponti, and Ivan Vulić. Zero-shot tokenizer transfer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Benjamin Minixhofer, Tyler Murray, Tomasz Limisiewicz, Anna Korhonen, Luke Zettlemoyer, Noah A. Smith, Edoardo M. Ponti, Luca Soldaini, and Valentin Hofmann. Bolmo: Byteifying the next generation of language models. *arXiv preprint arXiv:2512.15586*, 2025a.

Benjamin Minixhofer, Ivan Vulić, and Edoardo Maria Ponti. Universal cross-tokenizer distillation via approximate likelihood matching. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

Piotr Nawrot, Szymon Tworkowski, Michał Tyrolski, Łukasz Kaiser, Yuhuai Wu, Christian Szegedy, and Henryk Michalewski. Hierarchical transformers are more efficient language models. In *Findings of NAACL*, 2022.

Piotr Nawrot, Jan Chorowski, Adrian Łańcucki, and Edoardo M. Ponti. Efficient transformers with dynamic token pooling. In *Proceedings of ACL*, 2023.

Pit Neitemeier, Björn Deiseroth, Constantin Eichenberg, and Lukas Balles. Hierarchical autoregressive transformers: Combining byte- and word-level processing for robust, adaptable language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018.

nostalgebraist. Interpreting GPT: The logit lens. LessWrong blog post, 2020. `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Enes Özeren, Yihong Liu, and Hinrich Schütze. HYPEROFA: Expanding LLM vocabulary to new languages via hypernetwork-based embedding initialization. In *Proceedings of the ACL Student Research Workshop*, 2025.

Artidoro Pagnoni, Ram Pasunuru, Pedro Rodriguez, John Nguyen, Benjamin Muller, Margaret Li, Chunting Zhou, Lili Yu, Jason Weston, Luke Zettlemoyer, Gargi Ghosh, Mike Lewis, Ari Holtzman, and Srinivasan Iyer. Byte latent transformer: Patches scale better than tokens. In *Proceedings of ACL*, 2025.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Thang M. Pham, Trung Bui, Long Mai, and Anh Nguyen. Out of order: How important is the sequential order of words in a sentence in natural language understanding tasks? In *Findings of ACL*, 2021.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. BPE-dropout: Simple and effective subword regularization. In *Proceedings of ACL*, 2020.

Danish Pruthi, Bhuwan Dhingra, and Zachary C. Lipton. Combating adversarial misspellings with robust word recognition. In *Proceedings of ACL*, 2019.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI technical report, 2019.

François Remy, Pieter Delobelle, Hayastan Avetisyan, Alfiya Khabibullina, Miryam de Lhoneux, and Thomas Demeester. Trans-tokenization and cross-lingual vocabulary transfers: Language adaptation of LLMs for low-resource NLP. In *First Conference on Language Modeling (COLM)*, 2024.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of ACL*, 2021.

Craig W. Schmidt, Varshini Reddy, Haoran Zhang, Alec Alameddine, Omri Uzan, Yuval Pinter, and Chris Tanner. Tokenization is more than compression. In *Proceedings of EMNLP*, 2024.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR)*, 2024.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of ACL*, 2016.

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. UnNatural language inference. In *Proceedings of ACL*, 2021.

Kevin Slagle. SpaceByte: Towards deleting tokenization from large language modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*, 2013.

Yi Tay, Vinh Q. Tran, Sebastian Ruder, Jai Gupta, Hyung Won Chung, Dara Bahri, Zhen Qin, Simon Baumgartner, Cong Yu, and Donald Metzler. Charformer: Fast character transformers via gradient-based subword tokenization. In *International Conference on Learning Representations (ICLR)*, 2022.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, et al. 2 OLMo 2 furious. *arXiv preprint arXiv:2501.00656*, 2025.

Emiel van Miltenburg, Chris van der Lee, and Emiel Krahmer. Preregistering NLP research. In *Proceedings of NAACL*, 2021.

Mathurin Videau, Badr Youbi Idrissi, Alessandro Leite, Marc Schoenauer, Olivier Teytaud, and David Lopez-Paz. From bytes to ideas: Language modeling with autoregressive U-Nets. *arXiv preprint arXiv:2506.14761*, 2025.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Proceedings of EMNLP*, 2020.

Junxiong Wang, Tushaar Gangavarapu, Jing Nathan Yan, and Alexander M. Rush. MambaByte: Token-free selective state space model. In *First Conference on Language Modeling (COLM)*, 2024.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392, 2020.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL*, 2018.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306, 2022.

Zhipeng Yang, Shu Yang, Lijie Hu, and Di Wang. Word recovery in large language models enables character-level tokenization robustness. *arXiv preprint arXiv:2603.10771*, 2026.

Lili Yu, Dániel Simig, Colin Flaherty, Armen Aghajanyan, Luke Zettlemoyer, and Mike Lewis. MEGABYTE: Predicting million-byte sequences with multiscale transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Brian Siyuan Zheng, Alisa Liu, Orevaoghene Ahia, Jonathan Hayase, Yejin Choi, and Noah A. Smith. Broken tokens? Your language model can secretly handle non-canonical tokenizations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Anyuan Zhuo, Xuefei Ning, Ningyuan Li, Jingyi Zhu, Yu Wang, and Pinyan Lu. Understanding the ability of LLMs to handle character-level perturbation. *arXiv preprint arXiv:2510.14365*, 2025.

Vilém Zouhar, Clara Meister, Juan Luis Gastaldi, Li Du, Tim Vieira, Mrinmaya Sachan, and Ryan Cotterell. A formal perspective on byte-pair encoding. In *Findings of ACL*, 2023.

A THE REGISTERED PREDICTIONS

Table 2 compresses each registered prediction to one row. This appendix states each prediction and its outcome in full and then gives the registration chronology. The verbatim registered wording, timestamps, and amendment history are in the repository; the crosswalk at the end of this section is the only place where the repository's internal identifiers appear.

1. **Dissociation conjunction.** In the cross-model survey (sixteen model–task cells across the GPT-2, Pythia, and Llama families), refitting a probe on character-split representations recovers accuracy to within an equivalence margin while the frozen probe drops by at least half its headroom, jointly in at least 12 of 16 cells. Outcome: the conjunction held in 0 of 16 cells.
2. **Emergence trend.** Refit damage (Δrefit, the shortfall of a probe refitted on character-split representations from canonical accuracy) decreases monotonically with pretraining strength across the Llama base series (1B, 3B, 8B, 70B; one-sided Spearman). Outcome: $p = 0.17$, failed (Table 4).
3. **Trajectory content floor.** Content retention of at least 0.6 at every checksum-verified checkpoint past 50 billion tokens in the four registered lineages. Outcome: 18 of the 25 eligible checkpoints fall below 0.6 (one further cell, OLMo-2-13B at 3683B tokens, sits exactly at 0.60 and counts as meeting the floor); the exception is OLMo-2-32B, which clears it at 6 of 7 checkpoints with a content probe whose gap to the matched controls is between -0.05 and 0.02 at all 7.
4. **OLMo final invariance.** Final-checkpoint order transfer of at least 0.85 in the three registered OLMo-2 lineages. Outcome: only 32B met it (0.96); 1B reached 0.78 and 13B 0.60.
5. **Pythia bound.** Pythia-1.4b final order transfer of at most 0.6. Outcome: 0.58, passed.
6. **Within-lineage rise.** Order transfer rises (one-sided Spearman) and crosses 0.75 within each registered OLMo-2 lineage. Outcome: crossings occur in all three (32B is censored at its first checkpoint; 13B's crossing is sustained through its penultimate checkpoint, with the anomalous final checkpoint below), but the trend is significant only in 32B ($p = 0.006$); 1B ($p = 0.100$) and 13B ($p = 0.278$) fail.
7. **Scale ordering.** Larger OLMo-2 models cross 0.75 at fewer tokens. Outcome: the ordering holds among the three registered lineages (32B at or before 252B, 13B at 1854B, 1B at 2999B), but 13B crosses after the smaller, previously measured 7B (428B); failed under the registered wording.
8. **Multilingual order margins.** Order-selectivity gap above 0.10 in Spanish, French, and Swahili for Llama-3.1-8B. Outcome: 0.14, 0.28, and 0.20; passed (German, 0.42, was measured outside the registration).
9. **Multilingual content floor.** Content retention above 0.6 in the three registered languages. Outcome: Swahili 0.43 (German, outside the registration, 0.50), failed.

10. **Resource ordering.** Order gaps do not increase as resource level decreases, on the registered chain anchored at English (en $\geq$ de $\geq$ es/fr $\geq$ sw). Outcome: failed at both ends of the chain. English's gap (0.28) falls below German's (0.42), and Swahili's (0.20) exceeds Spanish's (0.14).
11. **Acceptability.** BLiMP acceptability retention above 0.6 in all four models, with a prespecified alternative reading should it fail. Outcome: failed in every model; the registered alternative reading was a hierarchy of constructs ordered by retention, which Section 4 rejects: the ordering changes across models and under the pooling rule.

Extension registration (rows 12 to 16). After the four-model ladder we extended it to eight lineages. Five further predictions were frozen and committed to the repository before any extension computation ran, following the same procedure as for the main ledger. Row 12: the trunk gap increases with final-checkpoint word-order retention across all eight models; passed, exact one-sided Spearman $\rho = 0.83$, $p = 0.008$. Row 13: OLMo-2-32B posts the largest gap; failed, 3.01 vs. Llama's 3.32. Row 14: the OLMo-2-13B lineage's gap falls below OLMo-2-7B's; failed, 2.51 vs. 2.40, and 2.34 in a follow-up ladder pinned to the anomalous checkpoint itself (Appendix F). Row 15: Pythia-1.4b's gap falls below Pythia-6.9b's; passed, 0.21 vs. 0.68. Row 16: GPT-2's replication run reproduces a near-zero gap, below the registered 0.4 and still the smallest of all; passed, 0.148 vs. the earlier 0.155, with the real-trunk rung reproducing to four decimals under the shared training seed. One detail of the extension affects the interpretation of row 12 and is disclosed here. The extension ladders load each model's released final weights, whereas the registered retention pairings for the three registered OLMo-2 lineages use their last stage-1 trajectory checkpoints. Re-measuring retention on the released weights tightens the row-12 relation to $\rho = 0.98$ ($p = 0.0002$): the 13B moves onto the trend (released retention 0.696 vs. 7B's 0.699) and the 32B drops to 0.83, in line with its gap. The 13B divergence is examined in Appendix F. Two retrospective lookups were recorded in the same registration, the Llama-3.1-8B unmodified-model prompted baseline (Table 8) and the mixed-head behavior of the original lineages; both existed in earlier result files and required no new computation.

Localization control (rows 17 to 20). After the extension, four predictions were frozen and signed before the control was run (Table 7; protocol in Appendix E). Row 17: the trunk arm, LoRA adapters with the head frozen, ends at least 1.0 bpc above the head-only repair on character-split input; passed, 3.38 vs. 1.94. Row 18: the trunk arm stays above 3.0 bpc; passed, 3.38. Row 19: adding trunk adapters to head retraining improves on the head-only arm by less than 0.3 bpc; passed on its wording, since the combined arm is worse by 2.01 (3.95). Row 20: the trunk arm leaves canonical reading within 0.15 bpc of the unmodified model; failed, 6.73 vs. 1.31. The registration fixed a reading for rows 17 and 18 holding together (the head is the locus) and for row 17 failing (the head is not the locus), and none for rows 17 and 18 holding while row 20 fails; the asymmetry reported in Section 5 is therefore an observation, not a registered result.

Chronology and provenance. Rows 1 and 2 were registered after an exploratory severity ladder and a single GPT-2-small gate cell had been measured, and before any confirmatory measurement on the sixteen-cell roster; the registration document records this, fixed the endpoints before further extraction, and the gate cell's values recur as one cell of the adjudicated table (Table 4). Rows 3 to 11 were registered before the trajectory, multilingual, and acceptability measurements they constrain, in a single timestamped registration whose amendment history is shipped verbatim. That document also discloses that raw endpoint values for the trajectory lineages had been seen earlier under a superseded protocol, and it stakes rows 3 to 7 only on the never-computed recentered metric and on unseen trajectory shapes. Crosswalk from table rows to repository identifiers: rows 1 and 2 are the survey endpoints (A1 and the emergence trend); rows 3 to 7 are R1a to R1d (rows 4 and 5 split the two predictions of R1b); rows 8 to 10 are R2a to R2c; row 11 is R3; rows 12 to 16 are X1 to X5 of the extension registration; rows 17 to 20 are X6 to X9 of the localization registration.

B DATA-INTEGRITY INCIDENTS

Upstream checkpoint duplication. Three branches of the OLMo-2-1124-7B repository (`step101000`, `step175000`, and `step300000`) serve byte-identical weight shards under distinct commits, which can be verified from LFS checksums without downloading. Two points in our

first checkpoint sweep were therefore the same model; the duplication surfaced because the pipeline is deterministic and the two sets of statistics matched exactly. We rebuilt checkpoint selection around shard-checksum uniqueness, and the result file of every trajectory point records the commit hash it was measured from.

Label-confounded evaluation split. The held-out layer-selection rule first split evaluation texts by alternating index, which was perfectly confounded with the alternating labels that our task builders emit: each half contained a single class, and every statistic the instrument produced was class-conditional recall against a meaningless floor. An adversarial review caught the bug by re-analyzing the raw result files, which showed impossible values on inspection (frozen transfer of 1.000 and 0.002 in the same table); the fix was a class-stratified split. All affected experiments were re-measured, and no probe statistic from the confounded instrument appears in this paper (likelihood measurements never used the split). The confounded result files remain in the repository history, marked as superseded.

Class-imbalanced content floor. The binarized NLI tasks take the first 872 validation examples, in which entailment holds one third of the items (0.337 in MultiNLI, 0.334 in each XNLI language), so a constant prediction scores 0.66 to 0.67, while retention was computed against a chance floor of 0.5. Every content retention computed against that floor was therefore inflated, by an amount that depended on how far the frozen probe had drifted toward the majority class. An adversarial review found the imbalance by checking the binarization rule against the realized class counts of the evaluation split; all content cells were re-measured on class-balanced splits (436 per class, taken in order from the same validation set), each result file records its realized class counts and majority floor, and rows 3, 9, and 11 were re-adjudicated. No content statistic from the imbalanced splits appears in this paper.

C MEASUREMENT DETAILS

Probes are multinomial logistic (softmax) classifiers, binary throughout, fitted on CPU in float64 with L-BFGS, with the ridge strength selected on an internal validation split, three fitting seeds reported throughout, and text-level bootstrap confidence intervals stored with each result. Control probes: twenty per cell, each defined by a random direction in canonical representation space, with labels from the median split of the projections and margin-matched trimming at 0.2 (robustness to the trim choice in Figure 5). Recentering subtracts the character-split cloud's per-layer mean, computed over all evaluation texts without labels, and adds the canonical cloud's mean computed on the training texts; for the control probes both means are computed on the held-out half. Because the probe is linear, the operation is a per-class intercept shift in the probe's standardized coordinates and leaves the direction unchanged. Retention is $(\mathrm{frozen} - \mathrm{chance})/(\mathrm{canonical} - \mathrm{chance})$, floored at zero. Layer selection maximizes canonical accuracy on a class-stratified half of the evaluation texts, and all statistics are reported on the disjoint half. Every second layer and the last layer are probed, the layer is re-selected at every checkpoint, and ties are resolved to the lowest layer index; the selected layers for the OLMo-2-7B trajectory range from 8 to 28 (Table 11); per-layer profiles for the four main models are in Figure 4. A severity parameter interpolates between the canonical and character-split conditions by splitting each canonical token independently with probability s; the endpoints reproduce the two conditions exactly, which is verified in tests; Table 9 reports the four main models at $s \in \{0.25, 0.5, 0.75\}$. All checkpoint loads verify the resolved commit hash against the requested revision.

Task construction and boundary alignment. Words are maximal non-whitespace runs. For each word end, each segmentation contributes the token whose span ends exactly at that character offset, and words for which either segmentation has no such token are dropped from both; leading whitespace belongs to the following token, so the two prefixes are identical at every kept boundary. All reported statistics mean-pool over the kept boundaries; Table 10 repeats the twelve cells of Figure 6 with the last kept boundary in place of the mean. The word-order task pairs each sentence with a word-shuffled copy, exactly balanced; the acceptability task uses the good and bad sentence of each BLiMP pair; the content task binarizes NLI as entailment versus neutral or contradiction over the premise and hypothesis fields, using the first 2000 training examples and, for evaluation, the first 436 validation examples of each class (MultiNLI matched; XNLI validation per language; Appendix B).

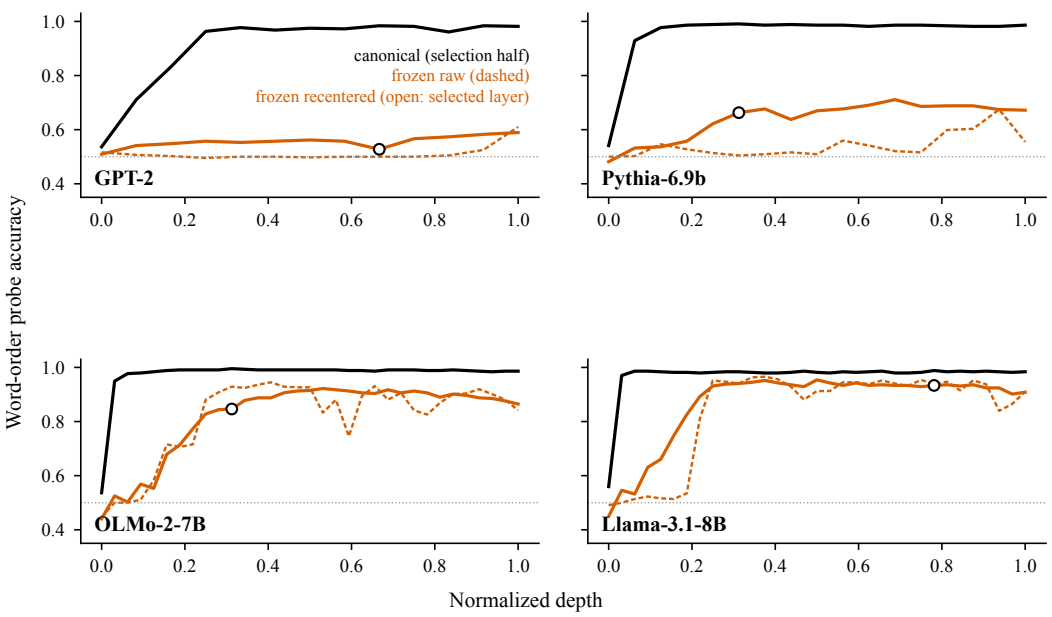

Figure 4: Per-layer probe profiles for the word-order task (intact vs. word-shuffled sentences) in the four main models: canonical accuracy on the selection half (black; the half of the evaluation texts used to choose the layer), and raw (dashed) and recentered (solid) frozen transfer of the same probes to character-split input (every character encoded as its own token), both evaluated on the disjoint held-out half; the dotted line marks chance. The open circle marks the selected layer, the maximizer of canonical selection-half accuracy, plotted at its recentered transfer. Recentering (per layer, subtracting the character-split activation mean and adding the canonical mean) recovers transfer where the raw probe collapses, chiefly across the middle and late layers of GPT-2 and Pythia-6.9b; where raw transfer is already high it changes little and can be slightly negative, as at OLMo-2-7B's selected layer (0.85 recentered vs. 0.93 raw).

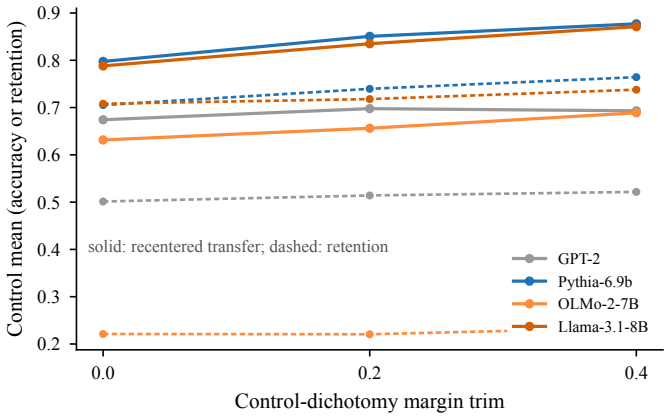

Figure 5: Robustness of the control envelope to the margin-trim choice (word-order task, four main models). Control probes are fitted to median splits along random directions of the canonical representation space (a task-unrelated baseline); the trim drops the stated fraction of texts nearest each split boundary before fitting. Solid: recentered frozen control transfer, averaged over the twenty control probes per cell; dashed: mean control retention (a ratio, not an accuracy). Both move smoothly with trim in $\{0.0, 0.2, 0.4\}$, by less than 0.09 between the extreme trims, and no conclusion in the paper depends on the 0.2 used throughout.

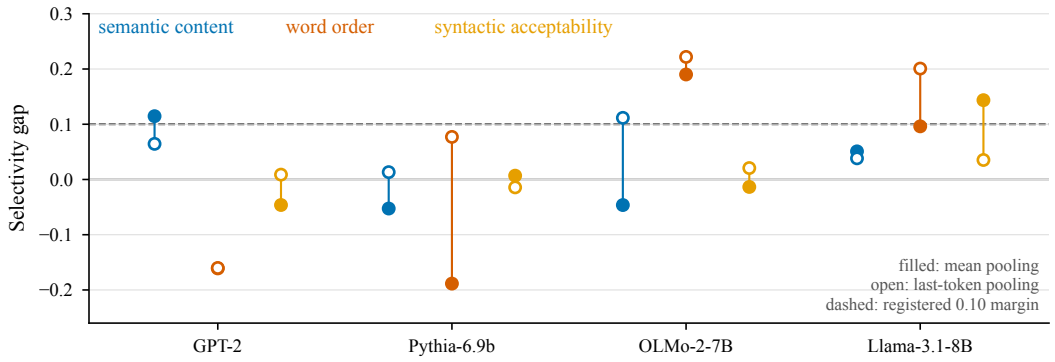

Figure 6: Selectivity under character-level re-tokenization for three constructs (binarized MultiNLI content, the word-order task with intact vs. word-shuffled sentences, and BLiMP acceptability) in four models, ordered left to right by word-order retention (the measure on the horizontal axis of Figure 3, right). The selectivity gap is the probe's recentered frozen transfer minus the mean recentered transfer of its twenty matched random-direction control probes (trim 0.2), under mean pooling over the kept word-end boundaries (filled) and last-token pooling (open), against the registered 0.10 margin (dashed). Only OLMo-2-7B word order clears the margin under both pooling rules (0.190 and 0.222). GPT-2 content (0.115) and Llama-3.1-8B acceptability (0.144) clear it under mean pooling only, OLMo-2-7B content (0.112) and Llama-3.1-8B word order (0.201) under last-token pooling only; GPT-2's word-order gap is negative under both, and Pythia-6.9b's clears the margin under neither (-0.189 and 0.077). Retentions and control envelopes for every cell are in Tables 3 and 10.

D FULL RESULT TABLES

Table 3 reports every probe cell behind Figure 6 and the language results: canonical accuracy, raw and recentered frozen transfer, retention, and the control envelope, per model and construct. Table 4 reports the sixteen-cell cross-model survey behind ledger rows 1 and 2. Table 5 reports the collapse factor of every model. Table 7 reports the capacity-matched control, Table 9 the re-tokenization dose scan, and Table 10 the pooling sensitivity of the twelve hierarchy cells. Figure 7 shows the order-selectivity gaps by language against the registered margin.

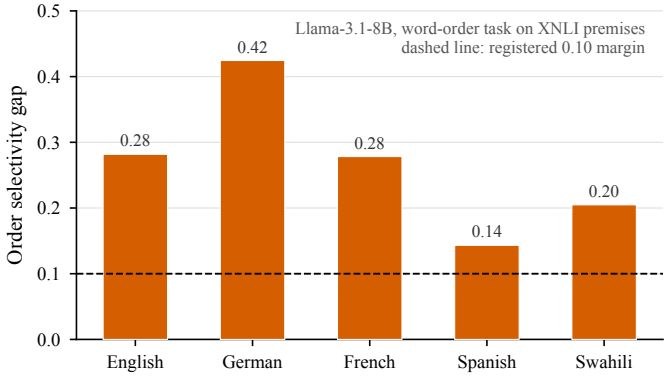

Figure 7: Word-order selectivity gaps (recentered frozen task transfer minus the mean of the matched control probes) for Llama-3.1-8B on word-order tasks built from XNLI premises, against the registered 0.10 margin, ordered along the registered resource chain from highest- to lowest-resource language (en $\geq$ de $\geq$ es/fr $\geq$ sw). Spanish, French, and Swahili were registered in advance and all clear the margin; English and German were measured outside the registration. The registered resource-ordering prediction (ledger row 10) fails at both ends of the chain, with English below German and Swahili above Spanish.

Table 3: Every probe cell behind Figure 6 and the language results: canonical accuracy, raw and recentered frozen transfer, recentered retention, and the control envelope (the recentered-transfer mean over twenty random-direction probes at trim 0.2, followed by the SD of the same probes' raw frozen transfer, the only spread the instrument stores, and their mean retention). Content cells use class-balanced evaluation splits (Appendix C). The "word order" rows use the SST-2 task of Section 3; the "order (English)" row and the other language rows use word-order tasks built from XNLI premises. Because retention normalizes by canonical accuracy, it is lower than the transfer accuracies of Figures 1 and 2 (OLMo-2-7B word order: transfer 0.846, retention 0.699; the 0.87 in Figure 2 is the final trajectory checkpoint, not the released model).

Model	Construct	canonical	frozen	recentered	retention	control transfer	control retention
GPT-2	content (NLI)	0.657	0.560	0.587	0.557	0.473±0.04	0.028
GPT-2	word order	0.988	0.499	0.537	0.075	0.698±0.14	0.514
GPT-2	acceptability	0.799	0.496	0.494	0.000	0.540±0.01	0.107
Pythia-6.9b	content (NLI)	0.700	0.514	0.531	0.153	0.583±0.03	0.057
Pythia-6.9b	word order	0.988	0.503	0.662	0.332	0.851±0.03	0.740
Pythia-6.9b	acceptability	0.885	0.500	0.580	0.208	0.573±0.03	0.163
OLMo-2-7B	content (NLI)	0.782	0.573	0.649	0.529	0.695±0.06	0.378
OLMo-2-7B	word order	0.995	0.930	0.846	0.699	0.656±0.11	0.221
OLMo-2-7B	acceptability	0.919	0.497	0.651	0.361	0.665±0.07	0.325
Llama-3.1-8B	content (NLI)	0.729	0.642	0.700	0.876	0.649±0.04	0.117
Llama-3.1-8B	word order	0.994	0.937	0.931	0.873	0.835±0.07	0.718
Llama-3.1-8B	acceptability	0.919	0.648	0.693	0.460	0.549±0.02	0.136
Llama-3.1-8B	content (German)	0.663	0.577	0.581	0.498	0.560±0.04	0.043
Llama-3.1-8B	content (Spanish)	0.652	0.576	0.607	0.704	0.625±0.06	0.299
Llama-3.1-8B	content (French)	0.668	0.573	0.628	0.763	0.546±0.04	0.047
Llama-3.1-8B	content (Swahili)	0.596	0.523	0.541	0.429	0.608±0.11	0.217
Llama-3.1-8B	order (English)	0.993	0.866	0.878	0.768	0.597±0.03	0.011
Llama-3.1-8B	order (German)	0.987	0.899	0.896	0.813	0.472±0.03	0.204
Llama-3.1-8B	order (Spanish)	0.982	0.578	0.584	0.174	0.441±0.04	0.343
Llama-3.1-8B	order (French)	0.973	0.729	0.739	0.505	0.461±0.05	0.175
Llama-3.1-8B	order (Swahili)	0.957	0.821	0.841	0.746	0.636±0.09	0.343

Table 4: The sixteen-cell cross-model survey behind ledger rows 1 and 2 (word-order task, character-split at full severity, layer chosen by highest canonical accuracy). The registered conjunction required refit equivalence and frozen destruction jointly in at least 12 of 16 cells; it held in none, because refit also degrades. The registered secondary endpoint, a decreasing trend of Δrefit across the Llama base series (1B, 3B, 8B, 70B), returned $p = 0.17$.

Model	canonical	frozen	refit	Δfrozen	Δrefit	equivalence	destruction
Pythia-1.4b	0.989	0.546	0.818	0.443	0.171	no	yes
Pythia-160m	0.981	0.510	0.767	0.470	0.213	no	yes
Pythia-6.9b	0.991	0.511	0.858	0.479	0.133	no	yes
GPT-2 Large	0.989	0.534	0.804	0.454	0.185	no	yes
GPT-2 Medium	0.989	0.506	0.827	0.483	0.162	no	yes
GPT-2 XL	0.994	0.530	0.868	0.465	0.126	no	yes
GPT-2	0.986	0.500	0.820	0.486	0.166	no	yes
Llama-3.1-70B-Instruct	0.993	0.803	0.979	0.190	0.014	yes	no
Llama-3.1-70B	0.997	0.977	0.986	0.019	0.010	yes	no
Llama-3.1-8B-Instruct	0.993	0.933	0.985	0.060	0.008	yes	no
Llama-3.1-8B	0.991	0.897	0.983	0.094	0.008	yes	no
Llama-3.2-1B-Instruct	0.987	0.659	0.898	0.328	0.089	no	yes
Llama-3.2-1B	0.993	0.576	0.937	0.417	0.056	no	yes
Llama-3.2-3B-Instruct	0.992	0.843	0.936	0.149	0.056	no	no
Llama-3.2-3B	0.991	0.897	0.960	0.094	0.031	yes	no
Llama-3.3-70B-Instruct	0.992	0.898	0.971	0.094	0.021	yes	no

Table 5: The behavioral collapse per model: bpc of the unmodified model under its own tokenization and under character-split input.

Model	canonical bpc	character-split bpc	factor
GPT-2	1.62	7.91	4.9×
Pythia-1.4b	1.42	14.72	10.4×
Pythia-6.9b	1.32	14.95	11.3×
OLMo-2-1B	1.31	15.67	12.0×
OLMo-2-7B	1.21	13.68	11.3×
OLMo-2-13B	1.22	14.01	11.5×
OLMo-2-32B	1.19	15.05	12.7×
Llama-3.1-8B	1.31	16.63	12.7×

Table 6: The full repair ladder for all eight lineages, rows in ascending final-checkpoint retention, bpc per rung. Columns: canonical and character-split bpc of the unmodified model; "masked", output support restricted to character tokens without training; "control head", a head retrained on canonical text; "random" and "real", a new head trained on character-split text over a randomly initialized trunk and over the real frozen trunk; "mixed", a head trained on a mixed canonical-and-character stream; "trunk gap", random minus real, computed from unrounded values before display rounding; "head", the retrained fraction of parameters. The four lineages beyond GPT-2, Pythia-6.9b, OLMo-2-7B, and Llama-3.1-8B, and a replication run of GPT-2, were run under the registered extension predictions of Appendix A. GPT-2's replication run under the extension's shared training seed reproduces its earlier ladder (real-trunk rung identical to four decimals; trunk gap 0.148 vs. 0.155). Pythia-1.4b's ladder runs in bfloat16 like its 6.9b sibling, after float32 head training diverged; the failed run is included with the results.

Model	canonical	char-split	masked	control head	random	real	mixed	trunk gap	head
GPT-2	1.62	7.91	6.56	7.90	3.72	3.57	3.71	0.15	23.7%
Pythia-1.4b	1.42	14.72	12.06	14.79	3.67	3.45	3.56	0.21	7.3%
OLMo-2-13B	1.22	14.01	11.13	13.86	4.72	2.21	2.37	2.51	3.7%
Pythia-6.9b	1.32	14.95	12.24	14.69	3.65	2.97	3.06	0.68	3.0%
OLMo-2-1B	1.31	15.67	13.45	15.99	3.97	2.74	2.82	1.23	13.8%
OLMo-2-7B	1.21	13.68	10.95	13.65	4.72	2.31	2.47	2.40	5.6%
Llama-3.1-8B	1.31	16.63	13.72	16.61	5.26	1.94	2.01	3.32	6.5%
OLMo-2-32B	1.19	15.05	11.65	14.98	4.89	1.87	1.99	3.01	1.6%

Table 7: The localization control on Llama-3.1-8B (Table 2, rows 17 to 20; registered before it was run): bpc on canonical and character-split input for four arms trained at the ladder's budget, with each arm's trainable parameters. LoRA adapters of rank 192 on every attention and feed-forward projection give 503M trainable parameters against the head's 525M. The trunk arm repairs character-split generation less than the head arm (3.38 vs. 1.94) and, unlike the head arm, does not preserve canonical reading (6.73 vs. 1.54).

Arm	canonical bpc	character-split bpc	trainable
unmodified	1.31	16.63	—
head retrained, trunk frozen	1.54	1.94	525M (6.5%)
LoRA on trunk, head frozen	6.73	3.38	503M (6.3%)
LoRA on trunk and head retrained	5.05	3.95	1029M (12.8%)

Table 8: Behavior around head repair (SST-2 prompting; $n = 300$ for the unmodified models, $n = 150$ per repaired cell). Above the rule, the unmodified models: character-split prompts cost prompted accuracy in seven of the eight (by 0.04 to 0.38; Llama-3.1-8B alone reads them at its canonical accuracy), and continuing a character-split prefix costs 1.3 to 1.8 times the canonical bpc. Below the rule, the retrained heads. No retrained head raises prompted accuracy on character-split input above the same model's unmodified value by more than sampling error (about ± 0.07 at $n = 150$), and the mixed-stream head lowers it in OLMo-2-32B (0.92 to 0.71) and at the edge of sampling error in OLMo-2-13B (0.73 to 0.66). Head retraining repairs character emission (Table 6), not character reading.

		prompted accuracy		continuation bpc	
Model	Head	canonical	char-split	canonical	char-split
GPT-2	unmodified	0.61	0.52	1.67	2.26
Pythia-1.4b	unmodified	0.57	0.52	1.31	2.21
Pythia-6.9b	unmodified	0.73	0.52	1.14	2.06
OLMo-2-1B	unmodified	0.87	0.48	1.22	2.09
OLMo-2-7B	unmodified	0.91	0.78	1.11	1.79
OLMo-2-13B	unmodified	0.93	0.73	1.14	1.70
OLMo-2-32B	unmodified	0.96	0.92	1.10	1.57
Llama-3.1-8B	unmodified	0.70	0.70	1.22	1.63
Llama-3.1-8B	character-only	0.73	0.55	1.24	3.68
Llama-3.1-8B	mixed-stream	0.72	0.73	1.38	2.80
OLMo-2-7B	character-only	0.91	0.71	1.02	3.60
OLMo-2-7B	mixed-stream	0.93	0.77	1.09	2.82
Pythia-6.9b	character-only	0.71	0.49	1.20	4.15
Pythia-6.9b	mixed-stream	0.65	0.50	1.16	3.05
GPT-2	mixed-stream	0.51	0.50	1.52	2.43
Pythia-1.4b	mixed-stream	0.60	0.50	1.26	3.18
OLMo-2-1B	mixed-stream	0.55	0.51	1.33	3.12
OLMo-2-13B	mixed-stream	0.92	0.66	1.16	2.62
OLMo-2-32B	mixed-stream	0.81	0.71	1.08	2.81

Table 9: The re-tokenization dose. Word-order retention and selectivity gap (recentered task transfer minus the control mean) at four severities s, where each canonical token is split into its characters with probability s; $s = 1$ is the character-split condition used throughout. Retention is lower at $s = 1$ than at $s = 0.25$ in every model; the gap is positive at every dose in OLMo-2-7B and Llama-3.1-8B and negative at every dose in GPT-2 and Pythia-6.9b; and in OLMo-2-7B the gap peaks at $s = 0.5$ (0.247) rather than at the full split (0.190).

	$s = 0.25$		$s = 0.5$		$s = 0.75$		$s = 1$	
Model	retention	gap	retention	gap	retention	gap	retention	gap
GPT-2	0.83	-0.035	0.41	-0.183	0.21	-0.220	0.08	-0.161
Pythia-6.9b	0.92	-0.009	0.62	-0.112	0.28	-0.239	0.33	-0.189
OLMo-2-7B	0.97	$+0.211$	0.91	$+0.247$	0.83	$+0.235$	0.70	$+0.190$
Llama-3.1-8B	0.98	$+0.057$	0.96	$+0.084$	0.91	$+0.093$	0.87	$+0.096$

Table 10: Pooling sensitivity. The twelve cells of Figure 6 re-probed with the last kept boundary token in place of the mean over kept boundaries (Appendix C). Canonical accuracy is higher under last-token pooling in ten of the twelve cells; recentered retention moves by 0.12 on average and by up to 0.23; and the selectivity gap changes sign in six cells. OLMo-2-7B word order clears the 0.10 margin under both poolings (0.190 and 0.222); Llama-3.1-8B word order clears it under last-token pooling only (0.096 and 0.201). Below the rule, the four remaining lineages of ledger row 12 at their registered endpoints. Under last-token pooling the eight row-12 retentions give $\rho = 0.71$, $p = 0.03$ against the trunk gaps (exact one-sided Spearman; the two lineages at the zero floor are ranked by their unfloored retention, and average ranks give 0.73 and 0.02), against $\rho = 0.83$, $p = 0.008$ under mean pooling. The OLMo-2-13B endpoint that the mean-pooled probe reads at 0.207 retention reads at 0.865 under last-token pooling.

		canonical		retention		gap	
Model	Construct	mean	last	mean	last	mean	last
GPT-2	content (NLI)	0.657	0.706	0.557	0.511	+0.115	+0.065
GPT-2	word order	0.988	0.973	0.075	0.000	−0.161	−0.160
GPT-2	acceptability	0.799	0.843	0.000	0.091	−0.046	+0.009
Pythia-6.9b	content (NLI)	0.700	0.806	0.153	0.320	−0.053	+0.013
Pythia-6.9b	word order	0.988	0.993	0.332	0.563	−0.189	+0.077
Pythia-6.9b	acceptability	0.885	0.947	0.208	0.058	+0.007	−0.014
OLMo-2-7B	content (NLI)	0.782	0.849	0.529	0.750	−0.046	+0.112
OLMo-2-7B	word order	0.995	0.995	0.699	0.866	+0.190	+0.222
OLMo-2-7B	acceptability	0.919	0.962	0.361	0.323	−0.014	+0.021
Llama-3.1-8B	content (NLI)	0.729	0.830	0.876	0.770	+0.051	+0.038
Llama-3.1-8B	word order	0.994	0.998	0.873	0.807	+0.096	+0.201
Llama-3.1-8B	acceptability	0.919	0.959	0.460	0.500	+0.144	+0.035
Pythia-1.4b	word order, endpoint	0.982	0.992	0.171	0.000	−0.129	−0.202
OLMo-2-1B	word order, endpoint	0.995	0.988	0.565	0.823	+0.087	+0.137
OLMo-2-13B	word order, endpoint	0.989	0.992	0.207	0.865	+0.002	+0.199
OLMo-2-32B	word order, endpoint	0.991	0.995	0.942	0.972	+0.201	+0.236

Table 11: Every trajectory checkpoint behind Figures 1 and 2: the selected layer (Appendix C), canonical probe accuracy, raw and recentered frozen transfer, and the control mean (trim 0.2), word-order task. Resolved commit hashes for every load are included with the result files.

Lineage	tokens (B)	layer	canonical	frozen raw	frozen recentered	control mean
OLMo-2-1B	139	6	0.990	0.517	0.617	0.645
	420	14	0.992	0.512	0.702	0.609
	944	10	0.989	0.518	0.670	0.678
	1573	6	0.986	0.638	0.750	0.660
	2182	2	0.979	0.498	0.555	0.626
	2999	10	0.998	0.806	0.794	0.693
	3985	12	0.995	0.643	0.780	0.693
OLMo-2-7B	5	26	0.926	0.505	0.513	0.625
	9	10	0.976	0.498	0.498	0.615
	17	18	0.985	0.497	0.444	0.603
	34	12	0.995	0.557	0.513	0.591
	68	14	0.995	0.574	0.537	0.594
	135	14	0.995	0.672	0.622	0.618
	269	10	0.998	0.828	0.679	0.615
	428	16	1.000	0.863	0.754	0.646
	630	10	0.994	0.795	0.818	0.634
	839	28	0.993	0.583	0.828	0.657
	1263	12	0.993	0.912	0.830	0.651
	1888	8	0.990	0.603	0.833	0.620
	2517	14	0.999	0.903	0.870	0.660
	3146	12	0.998	0.878	0.882	0.660
	3775	10	0.993	0.927	0.872	0.645
OLMo-2-13B	252	32	0.995	0.746	0.765	0.652
	630	14	0.985	0.582	0.583	0.620
	1087	6	0.989	0.525	0.697	0.632
	1854	10	0.992	0.780	0.800	0.672
	2668	14	0.992	0.954	0.881	0.709
	3683	10	0.992	0.895	0.807	0.682
	4958	4	0.989	0.628	0.601	0.599
OLMo-2-32B	252	20	0.998	0.865	0.830	0.668
	596	16	0.993	0.592	0.814	0.659
	1418	24	0.986	0.956	0.949	0.741
	2341	24	0.995	0.966	0.936	0.727
	3348	20	0.989	0.969	0.955	0.718
	4631	24	0.997	0.888	0.967	0.732
	6008	28	0.991	0.923	0.963	0.762
Pythia-1.4b	2	14	0.943	0.519	0.604	0.615
	17	12	0.989	0.540	0.524	0.701
	67	10	0.993	0.544	0.594	0.763
	134	10	0.991	0.564	0.622	0.742
	268	10	0.991	0.521	0.580	0.694
	300	16	0.982	0.545	0.583	0.712
Pythia-6.9b	2	8	0.939	0.522	0.527	0.615
	4	12	0.972	0.508	0.606	0.697
	8	8	0.982	0.463	0.495	0.795
	17	8	0.985	0.587	0.556	0.776
	34	10	0.989	0.507	0.574	0.760
	67	10	0.985	0.508	0.599	0.792
	134	20	0.989	0.515	0.596	0.758
	201	10	0.989	0.506	0.561	0.805
	268	14	0.991	0.514	0.619	0.831
	300	10	0.988	0.503	0.662	0.851

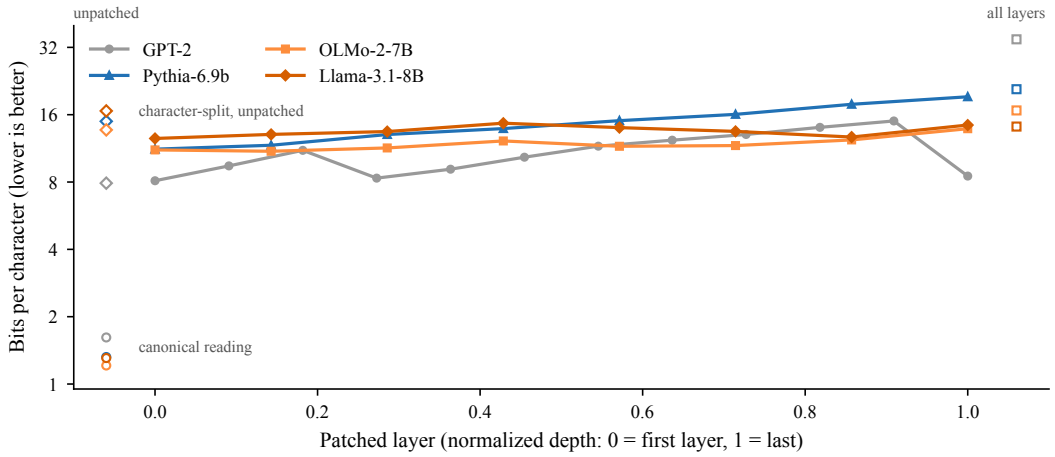

Figure 8: Translation patching does not repair generation. Bits per character on character-split input (every character encoded as its own token) with the per-layer mean shift (per layer, subtracting the character-split activation mean and adding the canonical mean) applied at a single layer inside the forward pass (lines; normalized depth), compared with each model's unpatched character-split baseline (open diamond, left), its canonical reading (the model's own tokenization; open circle, left), and all-layer patching (open square, right). The best single layer removes at most 27% of the collapse, defined as the gap between the unpatched character-split baseline and canonical reading (Pythia-6.9b 27%, Llama-3.1-8B 27%, OLMo-2-7B 22%), and none of GPT-2's, leaving generation five- to ten-fold above canonical; all-layer patching is worse than the best single layer in every model.

E Repair details

Table 6 reports the full repair ladder for all eight lineages, including the trunk gaps plotted in the right panel of Figure 3; Table 7 the capacity-matched control of Section 5; and Table 8 the behavioral outcomes for all eight lineages against their unmodified baselines. Figure 8 shows per-layer translation patching for the four models of Section 4, and Figure 9 the head-retraining loss dynamics. Figure 10 compares the behavior of repaired Llama-3.1-8B with the unmodified baseline (Section 5). Figure 11 shows the foreign-boundary measurements.

Head-retraining protocol. Training text is the wikitext-2-raw-v1 training split, keeping texts longer than 80 characters and taking the first 4000. Each arm trains the untied output head for 2000 steps of batch 8 at sequence length 512 with AdamW at learning rate 10^{-4}, PyTorch's default weight decay of 0.01, no schedule and no clipping; batches are sampled with replacement. In the real-trunk arms the head starts from the pretrained output weights and is reset between arms; in the random-trunk arm both trunk and head are randomly initialized. The loss is the standard cross-entropy over the full vocabulary with no masking. The mixed stream is a 1:1 pool of character-split and canonical chunks, truncated to the canonical chunk count. Ladder bpc is measured on 200 SST-2 validation sentences. The random trunk is not seeded: across three ladders of Llama-3.1-8B its rung varied between 5.26 and 5.50 bpc while the real-trunk rung reproduced exactly, so trunk gaps carry roughly 0.2 bpc of run-to-run variation. Normalizing the gap by the collapse, $(\text{random} - \text{real})/(\text{random} - \text{canonical})$, gives the same rank correlation with retention ($\rho = 0.83$) and the same ordering of Llama-3.1-8B above OLMo-2-32B.

Localization-control protocol. The LoRA arms use the same stream, budget, optimizer, and learning rate as the head arms, with adapters of rank 192 and scaling $\alpha = 2r$ on every attention and feed-forward projection, no dropout, and the base weights frozen; the head is frozen in the trunk arm and untied and trained in the combined arm. The realized trainable count of each arm is recorded with the result (Table 7). The registration, its signature date, and the file as it ran are in the repository.

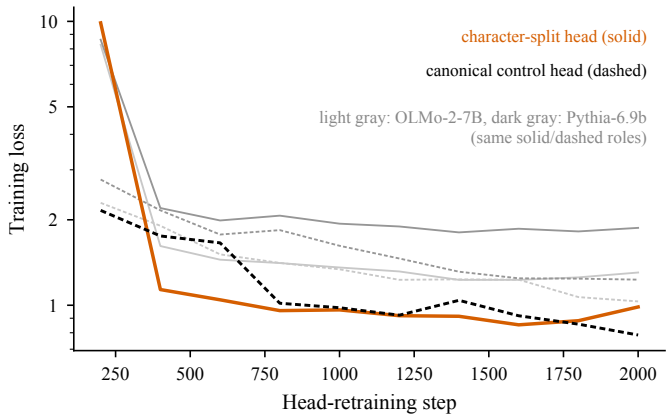

Figure 9: Head-retraining loss dynamics (head = output layer; trunk = all layers below it, kept frozen; Llama-3.1-8B in color, OLMo-2-7B in light gray and Pythia-6.9b in dark gray with the same solid and dashed roles) at ten logged points over the 2000 training steps, with a logarithmic loss axis. The character-split head (solid; trained on input with every character encoded as its own token) converges as quickly as the canonical control head (dashed; the model's own tokenization), so the repair is not limited by training data. Losses are per token over different token streams (character-level vs. canonical) and are therefore not comparable between heads; only the dynamics are.

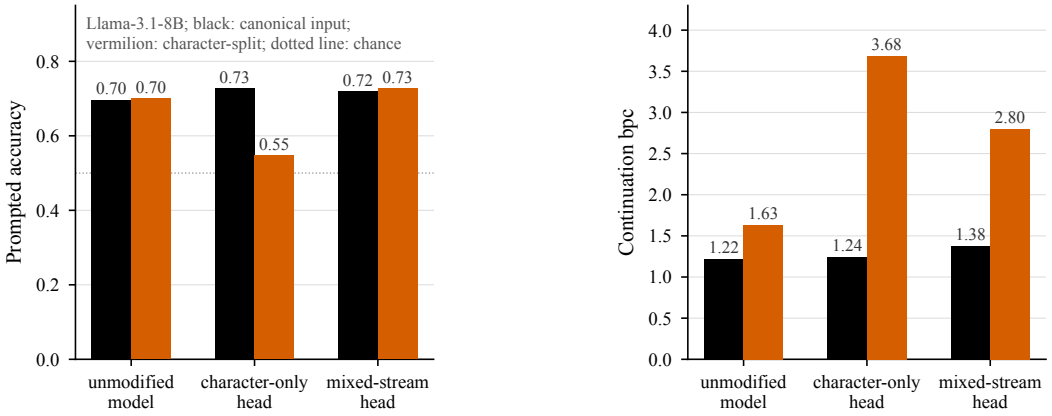

Figure 10: Behavioral outcomes of head repair for Llama-3.1-8B: prompted accuracy on the SST-2 sentiment task (left) and continuation bits per character (right; lower is better) on canonical (black; the model's own tokenization) and character-split (vermilion; every character encoded as its own token) input, for the unmodified model ($n = 300$) and after character-only and mixed-stream head retraining ($n = 150$ each; head = output layer; trunk = all layers below it, kept frozen while the head is retrained on character text or on a mixed canonical-and-character stream). This model reads character-split prompts at its canonical accuracy (0.70 vs. 0.70); OLMo-2-32B comes within sampling error (0.96 vs. 0.92) and the other six fall by 0.06 to 0.38 (Table 8); the character-only head lowers this to 0.55 while repairing likelihood, and the mixed-stream head returns it to the unmodified level (0.73). The dotted line on the left marks chance. Continuation bpc (loss on the canonical continuation of a prefix presented in the given input form) on character-split input stays above the unmodified model's 1.63 under both heads (3.68 character-only, 2.80 mixed-stream), so the repair is narrower than likelihood alone implies; the black bars on the right are continuation bpc on canonical input, a different quantity from the canonical training-stream bpc quoted in Section 5. All eight lineages are in Table 8.

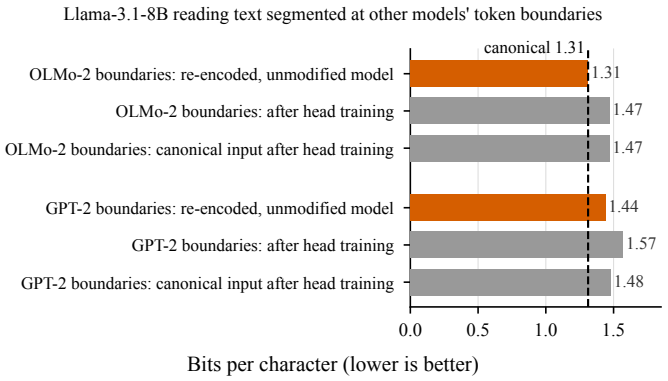

Figure 11: Foreign-boundary reading for Llama-3.1-8B: bits per character on text segmented at OLMo-2 and GPT-2 token boundaries and re-encoded in Llama's own vocabulary, against the unmodified model's canonical reference (dashed, 1.31). OLMo-2-boundary input reads at the canonical level (1.31) and GPT-2-boundary input at 1.44; head training on foreign-boundary text raises rather than lowers foreign-boundary bpc (1.47, 1.57) and costs canonical reading (1.47 to 1.48). For scale, character-granularity input reads at 16.63 bpc (Table 5), so boundary changes at subword granularity cost little and the collapse is specific to granularity.

F TRAJECTORY TABLES AND THE OLMO-2-13B RECORD

Table 11 lists every checkpoint behind Figures 1 and 2 with raw and recentered transfer and the control mean. The OLMo-2-13B record is as follows. Intermediate checkpoints rise to 0.88 (at 2668B tokens); the final published stage-1 checkpoint (4958B) reads 0.60 on order transfer, while its content retention (0.46) is in line with the lineage's other checkpoints (0.11 to 0.60). The extension ladder (Table 2, row 14) posts a 7B-level trunk gap (2.51) for this lineage's released model. Because the released weights include training beyond the stage-1 branch that the probe measured, we also ran the ladder pinned to the anomalous branch itself, commit-verified against the probe cell. It posts a gap of 2.34, again at the 7B level, with above-chance prompted behavior after repair (0.60). The same weights that the probe reads at 0.21 retention therefore hold near-7B invariant capability: the anomaly is in what the frozen probe can read at that checkpoint, not in the trunk. The released 13B model, one training stage later, probes normally (retention 0.696), which confines the anomaly to the stage-1-final cell. What changed in the geometry to break the probe there remains unexplained, but it is specific to the pooling rule: under last-token pooling the same checkpoint reads at 0.93 recentered transfer and 0.87 retention (Table 10). The seventy-fold ratio depends on the dating rules of Section 3: dating both quantities by a 0.75 threshold gives at least 86 (canonical accuracy already exceeds 0.75 at the first checkpoint, 5B), dating both by 90% of their final value gives at least 126, and the first 0.75 crossing of recentered transfer (428B) over the 0.95 competence point (9B) gives 48. Figure 12 follows the three constructs through training in Pythia-1.4b; the three registered OLMo-2 lineages carry the same three constructs at every checkpoint, and their result files are in the repository.

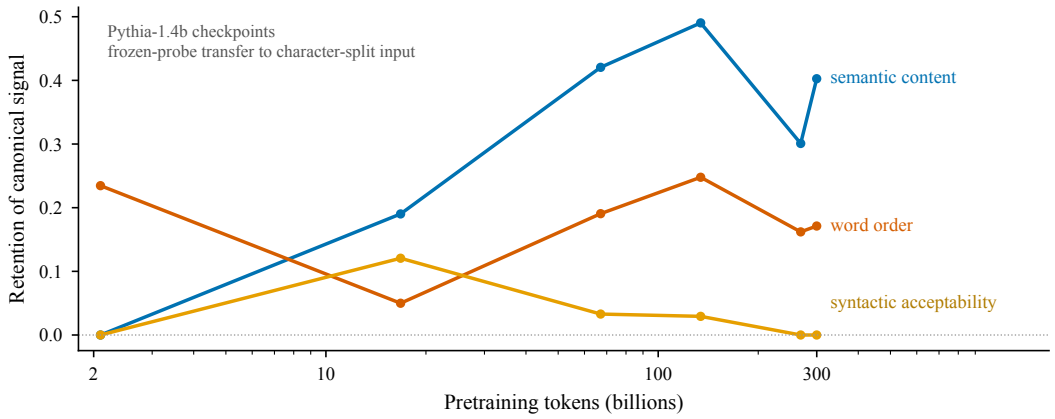

Figure 12: The three constructs through training in Pythia-1.4b, one of the four registered lineages for which all three were measured at every checkpoint. Retention is the fraction of the probe's above-chance canonical accuracy that survives frozen transfer (recentered) to character-split input. None of the three exceeds 0.6 at any checkpoint, consistent with a trunk that has not acquired invariance by the end of its run (Table 1). Retention is floored at zero (the dotted line) where frozen transfer falls below chance, hence the exact zeros for acceptability.