NEWSCPT: NEWS AS A SUBSTRATE FOR TEMPORAL CONTINUED PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) become increasingly stale as the world changes beyond their pretraining cutoff. Continued pretraining (CPT) on newly produced text offers a natural way to update them, but it remains unclear which new facts such updates acquire, how far that knowledge propagates, and at what cost. We study these questions using news as a substrate for temporal CPT. We construct and release a globally deduplicated, timestamped corpus of 205M news articles from January 2024 to May 2026, along with dated question sets derived from it and Wikipedia's Current Events portal, and continue pretraining eight models from three families. Across three post-cutoff knowledge benchmarks, news CPT consistently outperforms both base models and matched-budget controls trained on post-cutoff FineWeb-Edu text. At the level of individual facts, accuracy improvement grows approximately log-linearly with how often the corpus states the fact, but those gains propagate only weakly to their consequences. The update degrades broader downstream performance, which replay with web data only partly mitigates. Mechanistically, retaining only the CPT-induced MLP updates preserves much of the new factual knowledge while improving several downstream losses, suggesting that acquisition and capability loss are partly separable. Our results provide a large-scale empirical account of how LLMs acquire new factual knowledge through CPT: factual acquisition is strongly associated with exposure in the update corpus, while propagation beyond the stated fact remains limited and acquisition and capability loss can be partially separated after training.

1 INTRODUCTION

Large language models (LLMs) are trained on a snapshot of a world that continues to change. New governments take office, conflicts develop, scientific results appear, and sporting events take place every day. As the gap between a model's training cutoff and the present grows, its factual knowledge becomes increasingly stale (Lazaridou et al., 2021; Luu et al., 2022; Dai et al., 2025). Retraining from scratch whenever the world changes is prohibitively expensive. A more practical route is to continue pretraining an existing model on newly produced text.

Recent work studies temporal CPT at large scale in complementary settings. TiC-LM (Li et al., 2025) constructs a web-scale benchmark from 114 monthly Common Crawl dumps and studies update schedules and replay in a controlled setting, where models are pretrained from scratch on an initial time slice and then updated as successive snapshots arrive. Öncel et al. (2026) instead start from released models and continue pretraining them on post-cutoff FineWeb-Edu, studying how to acquire new knowledge while preserving existing capabilities. We build on this line of work but ask three related questions: *how does the update corpus determine which new facts are acquired, how far does that knowledge integrate with what a model already knows, and what does the update cost?*

Web snapshots are a natural source of update data, but they make this relationship difficult to study. Crawl time need not coincide with publication time: later snapshots revisit old pages, and independently deduplicated crawls can repeatedly contain material that predates the update period. Web corpora also mix newly produced information with large amounts of text unrelated to recent events. We hypothesize that news provides a useful setting in which to study factual updating more directly. Publication time is more closely aligned with the events being reported, important facts are naturally restated across many outlets and formulations, and coverage is produced in many languages. Repeated coverage is particularly useful because it lets us ask whether the amount of factual exposure

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.

in the update corpus predicts what a model subsequently learns. News itself has appeared in earlier studies of temporal adaptation, but in smaller collections, mostly in English and used for domain adaptation or perplexity studies (Lazaridou et al., 2021; Jang et al., 2022b); more recent work has also considered non-English news, including Swedish (Borggren et al., 2026).

We build this setting from Common Crawl's CC-NEWS collection. Starting from raw Web archive files, we extract article text and metadata, identify language, remove duplicates globally, and apply quality filters. The resulting corpus contains 205M unique articles from 01.2024 through 05.2026 in ten languages, organized by time and language. We pair it with dated factual evaluations: 22k questions derived directly from CC-News, 40k from Wikipedia's daily Current Events record, and the independently collected Daily Oracle benchmark (Dai et al., 2025). We then continue pretraining eight models from three families and compare them with matched-budget controls we train on post-cutoff FineWeb-Edu, in the same vein as Öncel et al. (2026). This setup lets us study how the choice of update corpus affects factual acquisition, retention, and the consequences of temporal CPT.

Our central finding is that the update corpus shapes both what models learn and what the update costs. Across all eight models and three post-cutoff benchmarks, news CPT consistently outperforms the matched-budget FineWeb-Edu control, by 3.6–6.3% points on average. At the level of individual facts, these gains grow approximately log-linearly with how often the corpus states the fact, linking factual acquisition directly to corpus exposure. This additional acquisition is not free. Our news checkpoints score 0.6–2.5 points below their FineWeb-Edu counterparts on a broad downstream evaluation, and mixing FineWeb-Edu data back into the update recovers only part of the loss. The knowledge is also only partially integrated: on a ripple-effect evaluation, accuracy on post-cutoff facts rises by 15.6 points, but even for the facts a checkpoint acquires, questions about their direct consequences improve by only 3.0 points and questions that combine them with existing knowledge by 6.8 points. At the question level, forgetting is concentrated among answers the base model already holds with low confidence and certain topics are forgotten more. Together, these results distinguish three aspects of temporal updating: which facts are acquired, how far they propagate beyond the stated fact, and what existing knowledge and capabilities are displaced and whether those losses can be separated from the gains. Our main contributions are:

- **A large-scale experimental substrate for temporal factual updating.** We release a globally deduplicated, temporally organized derivative of CC-NEWS containing 205M articles from 01.2024 to 05.2026, along with the processing pipeline, evaluation questions, and checkpoints.
- **A corpus-level account of factual acquisition through CPT.** Across eight models and three benchmarks, CC-News CPT consistently outperforms matched-budget FineWeb-Edu CPT for post-cutoff factual acquisition. At the level of individual facts, acquisition increases approximately log-linearly with how often the corpus states a fact.
- **A characterization of where factual updating propagates and what it costs.** We test whether acquired facts propagate to their consequences, characterize which existing answers are most vulnerable to forgetting, and measure the downstream cost of updating on news. We further analyze how these gains and losses are distributed across model parameters.

2 METHODS

This section describes our data acquisition, training, and evaluation pipelines.

2.1 DATA SOURCE AND ACQUISITION

We build our corpus from the Common Crawl CC-NEWS collection, covering news articles crawled between January 2024 and May 2026, retrieved in full from the monthly Web archives.

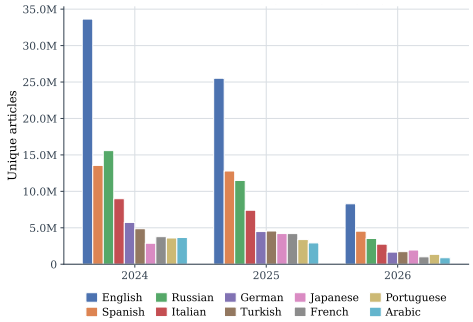


Figure 1: Unique articles per year, by language.

Article extraction and language identification. From each archived web page we extract the main article text while discarding navigational and template content such as menus, advertisements, and footers. Pages whose recovered text falls below a minimum length of 200 characters are removed, as they rarely correspond to genuine articles. For every retained article, we identify the language using a fastText language-identification model (Joulin et al., 2016) and record the prediction together with a confidence score. We restrict the corpus to the ten most prevalent languages,

forming 77% of the articles. All subsequent processing is organized by language and by month. The yearly composition of the corpus by language, showing a gradual decline in volume over the period, as well as the languages are shown in Fig. 1. The temporal distribution of the corpus is shown in Figs. 7a and the overall composition by language is in Fig. 7b.

Three stage deduplication. Following (Lee et al., 2022), we remove re-run pages, cross-domain mirrors, and republished wire stories in three stages of increasing cost and permissiveness. At each stage, we retain the most recent version to preserve updates made after publication.

1. *Exact-text deduplication:* We first remove articles whose text is identical to one already seen, up to case differences and whitespaces. This stage is deliberately strict: it eliminates verbatim copies such as mirrored pages and literal re-publications, but not paraphrased versions.
2. *URL deduplication:* We then remove articles sharing a canonical URL with an already retained article, catching content that changed slightly between crawls. Together, the first two stages eliminate most trivially redundant records ($\approx 12\%$ of language-filtered records).
3. *Near-duplicate deduplication:* We compute MinHash signatures (Broder, 1997) over each article's word shingles, use locality-sensitive hashing (Indyk & Motwani, 1998) to identify retained articles with estimated Jaccard similarity above a fixed threshold, and discard matches.

Quality filtering. A non-trivial fraction of remaining pages are not articles at all (e.g., e-commerce listings or navigation pages). Such pages are characterized by structural rather than lexical properties. To remove them, we apply heuristic quality rules in the spirit of the Gopher (Rae et al., 2021) and C4 (Raffel et al., 2020) filtering pipelines, such as rejecting documents with anomalous mean word length, too many repeated lines, too few function words, and lacking terminal punctuation.

2.2 Model training and evaluation

We continue pretraining eight open-weight base models on randomly sampled 6B tokens from the English portion of our CC-News corpus, with shorter and longer control runs. All models are trained with a fixed learning rate of 5×10^{-5} (polynomial decay to $0.1\times$ the peak, 500 warmup steps), AdamW (Loshchilov & Hutter, 2019) with weight decay 0.1, and identical settings across arms; full hyperparameters are given in Appendix B. We evaluate the resulting checkpoints, base models, and controls trained on FineWeb-Edu, educationally filtered post-cutoff web data, on three question sets.

Models. We continue pretraining eight base models spanning three families and three size tiers: OLMo2 at 1B, 7B and 13B (Walsh et al., 2025), Llama-3.1-8B and Llama-3.2-3B (Llama Team, 2024), and Qwen3 at 1.7B, 4B and 8B (Qwen Team, 2025). The OLMo2 and Llama models publish their knowledge cutoffs (2023, Dec), which is what makes the before-/after-cutoff partition well defined for them; Qwen3 does not, and we include it for completeness while reading its splits as indicative rather than exact.

Evaluation. We evaluate checkpoints on three MCQ sets with varying overlap with the CPT data: *(i)* 14k questions generated from our 6B-token CC-News corpus, supplemented with 8k before-cutoff questions from older articles dating back to 2020; *(ii)* 18.5k Daily Oracle questions (Dai et al., 2025) from an independently collected news stream; and *(iii)* approximately 40k questions generated from Wikipedia's daily Current Events portal. All after-cutoff CC-News questions derive from documents seen during CPT, whereas only 16.73% of Daily Oracle's source URLs appear in our corpus, and Wikipedia provides a separately written source. The first two sets emphasize long-tailed news facts, while Wikipedia focuses on prominent world events. We split timestamped questions at the December 2023 cutoff shared by Llama-3 and OLMo2: *before-cutoff* questions probe retention of potentially acquired knowledge, and *after-cutoff* questions probe acquisition of newer knowledge. Table 5 summarizes the sets, and Tables 7 and 8 provide examples. For CC-News, we use two LLM question generators, assess quality with judges based on Dai et al. (2025), and check accuracy consistency across subsets (construction and validation details are in App. C).

We score every checkpoint under OLMES (Gu et al., 2025b), averaging each option's log-probabilities across all four cyclic answer-order rotations (Zheng et al., 2024) before taking the argmax. This *rotation-averaged* evaluation addresses the strong answer-letter bias observed despite uniformly distributed gold labels (Fig. 10). We also evaluate True/False questions but omit their results because of strong answer-polarity bias (App. D.1).

Control experiments. Following Öncel et al. (2026), we continually pretrain the same models on 6B tokens of FineWeb-Edu (Lozhkov et al., 2024) from all available post-cutoff dumps spanning

01.2024–06.2025. This ready-to-train slice provides an off-the-shelf control for whether targeting fresh news justifies the preprocessing required for CC-News. To assess run-to-run variability, we repeat CC-News CPT for LLAMA-3.2-3B and QWEN3-1.7B under three random seeds. The pooled standard deviation of accuracy across seeds is 0.4% points, well below the reported effect sizes.

3 FINDINGS

We investigate four questions about what CPT on news changes and why. We first establish the aggregate effect of CC-News CPT on new and existing knowledge relative to both base models and matched FineWeb-Edu controls. We then move to the question level, asking which facts are learned, how acquisition depends on corpus exposure, and whether acquired facts propagate to their consequences. Finally, we examine the costs of the update: which previously correct answers are most vulnerable to forgetting, how broader capabilities change, and where acquisition and capability loss appear in the weights. See App. F for ablations and additional results.

3.1 CONTINUED PRETRAINING ON ENGLISH CC-NEWS: WHAT AGGREGATE FINDINGS TELL

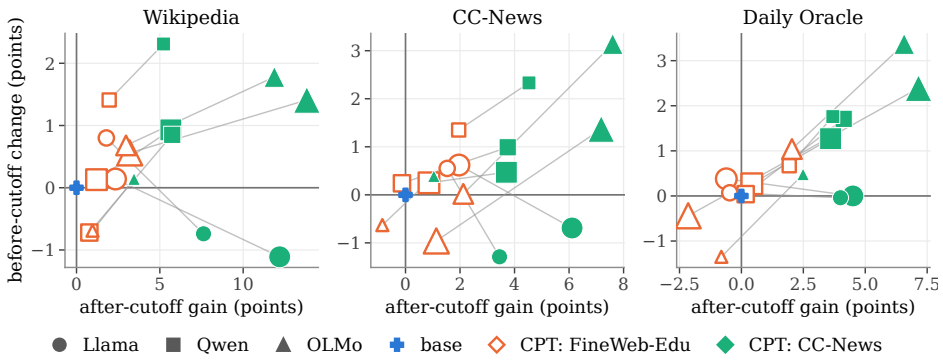


Figure 2: **What CPT on news buys, and what it costs.** Horizontal axis is the change in after-cutoff MCQ accuracy over the base model; vertical axis is the change in before-cutoff accuracy. Every base model sits at the origin (blue cross). Hollow markers are the FineWeb-Edu control and solid markers the CC-News checkpoint. They are connected when CPT starts from the same base model. Colour and shape give the model family, and marker area gives the model size. Every CC-News checkpoint moves right of both its base model and the control on all three question sets, while staying close to the pre-cutoff line.

News consistently improves post-cutoff knowledge across models and benchmarks. Fig. 2 plots what each checkpoint gains on after-cutoff questions against what it does to pre-cutoff ones (see Table 9 for raw accuracy numbers). Across all models, the CC-News checkpoint beats both its base model and FineWeb-Edu control on post-cutoff benchmarks without an exception. Averaged over all models, CC-News CPT improves after-cutoff accuracy over the base models by 4.5 percentage points on Daily Oracle, 8.2 on Wikipedia, and 4.7 on CC-News, exceeding matched-budget FineWeb-Edu checkpoints by 4.5, 6.3, and 3.6 points, respectively. Because the FineWeb-Edu control runs were trained on data until June 2025, we repeat the comparison on questions until June 2025 to align the evaluation window with both update corpora. The same pattern holds: CC-News checkpoints remain substantially above both the base models and FineWeb-Edu controls (Table 10).

We make two important observations here. The margin between the CC-News checkpoints and the FineWeb-Edu controls is too large to be explained by run-to-run noise: the smallest gap over the control, 1.76 points, is more than four times the 0.4-point standard deviation we measure. Second, the effect is not bound to QAs extracted from news since the gains on Wikipedia questions, extracted from general world events, run much more than the CC-News ones. These results support the value of curating news for CPT: its advantage over matched-budget FineWeb-Edu holds across all evaluated model families, sizes, and question sets, although the magnitude varies.

The pre-cutoff cost is small, and it is not a necessary price. Averaged over the eight models, CC-News CPT causes no net before-cutoff loss on any benchmark ($+0.70$, $+0.84$ and $+1.37$ points on Wikipedia, CC-News and Daily Oracle), and is slightly *above* its base model and FineWeb-Edu control. Forgetting on these sets is confined to the two LLAMA-3 checkpoints. Their drops appear on Wikipedia and CC-News (-0.69 to -1.29 points), but not on Daily Oracle. Importantly, we note

that models that gain more after the cutoff do not lose more before it (Pearson correlation $r = -0.01$, $p = 0.99$, though $n = 8$ models can only rule out a strong relationship). This concerns net accuracy; at the level of individual questions, models that learn more also flip more previously correct answers (Section 3.4), but these losses are offset by gains on other before-cutoff questions.

CPT shows diminishing returns with scale. Increasing the training budget continues to improve after-cutoff accuracy, but with clear diminishing returns. Both models in Fig. 3 make their largest gains early in training: performance rises sharply over the first few billion tokens, improves more gradually, and then largely plateaus. Most gains occur within the first $\sim$ 10B tokens for OLMo and $\sim$ 20B tokens for LLAMA. Before-cutoff accuracy changes less and is broadly stable once the initial update has taken place. Thus, substantially longer CPT yields additional acquisition, but most of the benefit is already captured within the first part of training. See Fig. 9 for the corresponding per-benchmark scores.

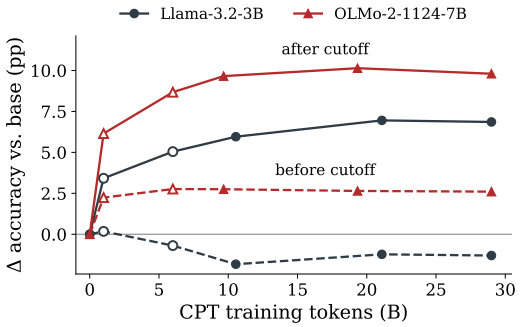


Figure 3: Accuracy change relative to the base model as a function of CPT tokens. Filled markers trace checkpoints from the 29B-token runs. Hollow markers show the separate, shorter runs.

The news advantage emerges for events from 2023 onwards. Fig. 11 breaks down benchmark accuracy by event year, averaged over models. CC-News and FineWeb-Edu perform similarly on events through 2022. The CC-News advantage emerges in 2023 ($+2.4$ points on average) and ranges from $+3.6$ to $+7.1$ points for 2024–2026. Its emergence before the corpus starts in January 2024 is consistent with news coverage revisiting recent events, including those preceding the training cutoff.

3.2 How much data is needed for learning a fact?

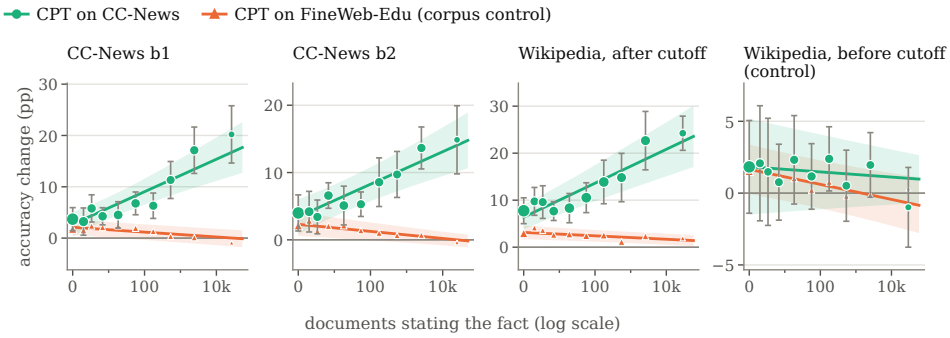


Figure 4: Average accuracy change of eight models against how often the corpus states the fact. The x-axis is $\log_{10}(1 + \text{statement count})$. Points are fixed count bins placed at the mean exposure of their bin and sized by the number of questions in it. The line is an ordinary least squares fit to the per-question data with a 95% band.

The aggregate numbers of Section 3.1 say nothing about *which* facts the checkpoint learns. This section measures how often the corpus states the fact a question asks about, and relates that count to whether the checkpoint answers the question correctly.

Counting how often the corpus states a fact. For each question, we estimate how often the corpus states the relevant fact by matching normalised string patterns for the subject and gold answer and recording their minimum character distance within each document. The *statement count* is the number of documents in which the two fall within 32 characters of one another, roughly a handful of words. We exclude questions with descriptive answers rather than names or values (19.9%), since these cannot reliably be matched verbatim. Among the remaining questions, the median statement count is 24, while 24.9% have no matches within this window. This measure misses many genuine statements: in the source articles from which the questions were extracted, subject and gold answer co-occur for 92.1% of questions but fall within 32 characters for only 61%. Paraphrases, pronouns, and intervening clauses can prevent a match, while proximity alone does not guarantee that the fact is asserted. We therefore treat statement count as an imperfect proxy for factual exposure.

More frequent factual exposure predicts larger accuracy gains. Fig. 4 relates accuracy changes, averaged over eight CC-News checkpoints, to statement counts in the training corpus. Log-linear fits associate each tenfold increase in statement count with gains of 3.18, 2.41, and 3.60 percentage points across the three question sets. For facts stated more than 10k times, the gain reaches 25.3 points. Two controls help interpret this relationship. On before-cutoff Wikipedia questions, no significant relationship holds between statement counts and accuracy. Also, FineWeb-Edu checkpoints show no relationship between accuracy gains and the same counts on the same questions. These results suggest that the association reflects exposure during CC-News training rather than a general effect of CPT. We restrict this analysis to multi-word answers: single-word gold answers appear in a median of 5×10^5 documents, compared with approximately 6×10^2 for multi-word answers, making their statement counts more susceptible to incidental matches driven by word frequency. The excluded part of the questions is analysed in Fig. 12.

Tighter co-occurrence windows yield steeper exposure–gain relationships. We repeat the analysis with windows of 8, 128, 300, and 1024 characters, as well as whole-document co-occurrence (Fig. 13). The positive association persists throughout, but its slope increases monotonically as the window narrows. Averaged over eight models and all question sets, the gain per tenfold increase in statement count rises from 2.25 percentage points at whole-document scope to 2.55 at 300 characters, 3.06 at 32, and 3.52 at 8. This ordering holds for every model, while the before-cutoff control remains flat across windows. Requiring closer proximity between subject and answer therefore strengthens the association between measured exposure and learning.

3.3 Do newly acquired facts propagate?

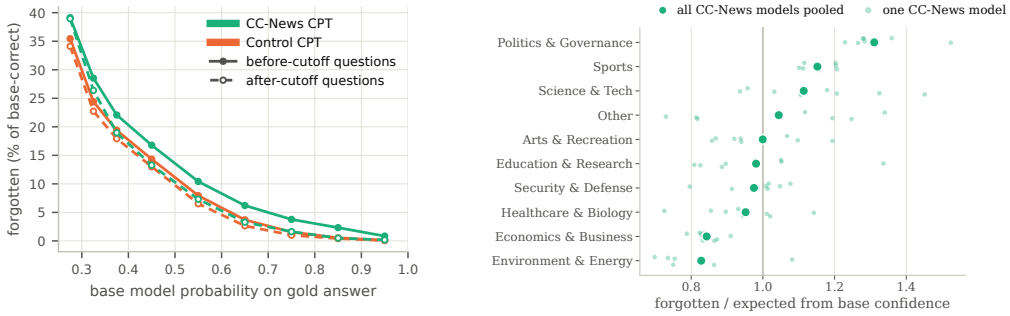
Greater factual exposure is associated with larger gains on the stated fact; we next ask whether those gains extend beyond it. Following ripple-effect evaluations in knowledge editing (Cohen et al., 2024), we ask whether facts acquired through CPT become integrated with the model's existing knowledge rather than merely answerable in isolation. Starting from the post-cutoff Wikipedia set, we retain 1,358 anchor questions whose answers imply verifiable downstream consequences and construct 2,561 follow-up questions: 768 *second-order* questions asking for a direct consequence of the learned fact, and 1,793 *third-order* questions requiring that fact to be combined with pre-cutoff knowledge. For example, learning that Kemi Badenoch was elected leader of the UK Conservative Party should update the number of women to have led the party from three to four (*second-order*) and, combined with existing knowledge about her, should let the model answer that the party's current leader grew up in Nigeria rather than England (*third-order*). The pre-event answer is set as a distractor when the event replaced an earlier value. Table 12 gives representative examples.

CPT updates the stated fact much more than what follows from it. On the ripple-effect evaluation, accuracy on the post-cutoff facts considered improves by 15.6 percentage points. Fig. 19 in App. G asks whether improvements on anchor questions extend to their follow-ups. For each CC-News checkpoint, we select anchors answered incorrectly by the base model and correctly after CPT, then compare both models on their follow-up questions. This yields 8–30% of the 1,358 anchors per checkpoint (Table 14). Despite this selection, mean follow-up accuracy across seven checkpoints increases by only 3.0 points on second-order and 6.8 points on third-order questions, with no second-order improvement for three checkpoints. Larger models generally perform better on the follow-up questions within each family (Fig. 19). Outdated answers also remain common: when the pre-event answer appears as a distractor, the checkpoints select it 43.6% of the time, down from 48.9% for the base models, versus 31.0% for the current answer. Grouping anchors by the change in *correct-answer probability* shows a similar pattern (Table 15). Larger anchor gains correspond to larger follow-up gains, but even where anchor accuracy rises from 28% to 92%, follow-up accuracy improves by only 5–8 points. These results suggest that CPT on news, like knowledge editing, can improve recall of a stated fact while leaving much of what follows from it unchanged.

3.4 What governs forgetting?

We now move from aggregate accuracy to individual question transitions, distinguishing *acquired* questions (wrong → right after CPT) from *forgotten* questions (right → wrong).

Much of forgetting follows the model's existing uncertainty. Questions that the base model answers correctly with low confidence are much more likely to flip after CPT (Fig. 5a). This pattern also appears for the FineWeb-Edu controls, i.e., it is not specific to news. Part of this relationship



(a) Base confidence (b) Topic vulnerability

Figure 5: **Forgetting depends on initial confidence and topic.** (a) Percentage of base-correct questions forgotten after CC-News CPT, grouped by the base model's gold-answer probability. (b) Observed forgetting divided by the confidence-matched expectation on before-cutoff Daily Oracle questions. Large dots pool all eight models, while faint dots show individual models. Values above one indicate more forgetting than expected.

follows directly from the decision rule: when the gold answer is already close to a competing option, a small change in the model's probabilities can turn a correct argmax into incorrect, whereas high-confidence answers require a much larger change to flip. We therefore treat base confidence as a baseline predictor of vulnerability rather than evidence by itself of a news-specific forgetting.

Topic differences remain after accounting for this baseline. Confidence does not explain all variation in forgetting (Figs. 5b and 15). On before-cutoff Daily Oracle questions, Politics & Governance and Sports are forgotten 31% and 15% more often than expected from their confidence distributions across all eight models, while Economics & Business and Environment & Energy are forgotten 16% and 17% less often. Some apparent forgetting could reflect genuine changes in the world, so we audit a random sample of forgotten questions. Among the 550 cases where two LLM judges agree, 88% are not factual updates in which the correct answer itself changed.

Acquisition and forgetting rise together. Runs that acquire more correct answers also tend to flip more previously correct ones (Fig. 14). For most checkpoints, acquisition exceeds forgetting, but larger movement in both directions does not always produce a larger net gain, particularly at high learning rates. This association does not show that forgetting is necessary for acquisition. Instead, it shows that aggregate accuracy can conceal substantial turnover in the set of questions a model answers correctly: stronger updates tend to change more answers in both directions.

3.5 Downstream analysis

Specialising on news carries an additional downstream cost. We evaluate the same checkpoints on 24 accuracy-based tasks from the OLMo evaluation suite (Allen Institute for AI, 2024). Table 1 shows how different CPT corpora change the downstream performance. The FineWeb-Edu control shows that a CPT pass could help or hurt depending on the model family, while Qwen3-8B is the only positive case among CC-News checkpoints. Comparing the two corpora, CC-News performs worse than FineWeb-Edu for all models, by 0.6–2.5 points in average downstream accuracy. The largest gaps occur for the two Llama models, but the same ordering holds for every Qwen and OLMo model as well. This pattern is consistent with a cost of specializing the update distribution toward recent news, although the comparison does not isolate specialization from other differences between the corpora, most notably FineWeb-Edu's educational-quality filtering, which may itself favor benchmarks such as MMLU.

Table 1: Change in average downstream accuracy relative to the base model.

Model	Δ_{CPT}	Δ_{control}
Llama-3.2-3B	-1.71	+0.45
Llama-3.1-8B	-2.47	+0.01
Qwen3-1.7B	-0.84	-0.17
Qwen3-4B	-0.66	+0.16
Qwen3-8B	+0.58	+1.22
OLMo-2-1B	-1.07	-0.51
OLMo-2-7B	-2.55	-1.35
OLMo-2-13B	-2.32	-1.61

Memorized knowledge is particularly sensitive to the update. The task-level breakdown in Fig. 16 shows that the loss is not uniform. CommonsenseQA and MMLU fall sharply after CC-News CPT. MMLU depends heavily on factual and curricular knowledge stored during pretraining, so these declines are consistent with CPT partially displacing previously memorized knowledge when the training distribution becomes narrower. Importantly, some of this effect is also present under FineWeb-Edu CPT, particularly for OLMo, so knowledge erosion is not unique to news pre-

training; CC-News generally makes it stronger. Other tasks behave differently: ARC-Challenge often improves after CPT, and BoolQ improves substantially for OLMo. The downstream effect is therefore not a uniform degradation in model quality, but a redistribution of capability in which tasks relying more strongly on retained parametric knowledge are among the most vulnerable.

A small amount of replay partially mitigates the downstream cost. We also test whether mixing a small amount of FineWeb-Edu data into the news update can recover some of the capability lost under CC-News specialization. For LLAMA-3.2-3B and LLAMA-3.1-8B, we replace 0.5B of the 6B CC-News budget with FineWeb-Edu tokens. The mixture improves the downstream performance from 67.08 to 67.59 at 3B and from 70.73 to 71.07 at 8B, recovering about 30% and 14% of the drops from the base models. The improvement is therefore real but incomplete: a modest amount of replay softens the cost of specializing on news without restoring the original downstream performance.

3.6 WHERE DOES FORGETTING LIVE?

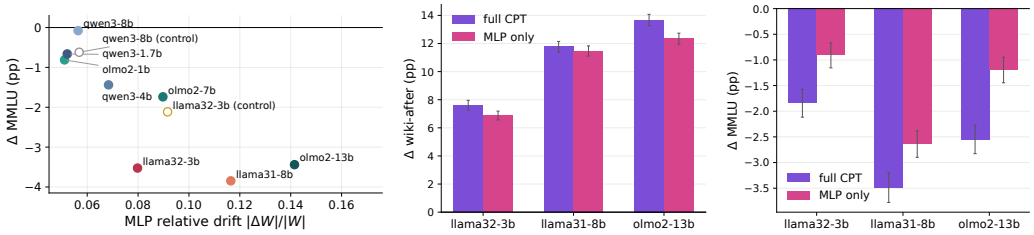


Figure 6: (**left**) Relative weight drift of MLPs plotted against the change in MMLU accuracy, for eight CC-News checkpoints and two FineWeb-Edu controls (hollow). Pearson correlation is $r = -0.8$. (**middle, right**) Post-cutoff gain on Wikipedia and MMLU change for full CPT and for the base model with only the MLP updates spliced in, on the three checkpoints that forget most. All eight checkpoints are in Fig. 25.

Larger weight changes accompany greater gains and losses. We next ask how the gains and losses from CPT relate to changes in model parameters. Across checkpoints, larger departures of MLP weights from the base weights are associated with greater post-cutoff knowledge gains and larger MMLU accuracy drops (Fig. 6, left). Update magnitude thus captures part of the variation in both acquisition and capability loss. Inspecting the other modules, we find that attention updates in LLAMA are concentrated more heavily in the (V, O) projections than in (Q, K), which is absent in OLMO and QWEN (Fig. 20). The same pattern appears in the FineWeb-Edu LLAMA control, suggesting a family-specific tendency rather than a response unique to news pretraining. Full module-, layer-, and representation-level analyses are given in Sec. H.

MLP updates preserve most of the gain while recovering part of the capability loss. Keeping the CPT MLP weights while restoring attention to its base values retains 92.6% of the aggregate post-cutoff Wikipedia gain while recovering 42.6% of the aggregate MMLU drop across the CC-News checkpoints (Fig. 6, middle and right). The same intervention also improves other evaluations affected by CPT, including CommonsenseQA and arithmetic, although the extent of recovery varies by task (Sec. H.1). Attention-only updates retain much *less* of the newly acquired knowledge (Table 16). The results show that most acquired knowledge can be preserved while partially undoing the accompanying capability loss, indicating the two are not inseparably tied in the final CPT weights.

No single structural signature explains forgetting. We test several candidate explanations motivated by prior work: low-rank structure (Biderman et al., 2024), concentration in mid-layer MLPs (Meng et al., 2022), and representation displacement (Ramasesh et al., 2021; Davari et al., 2022). We also examine weight interpolation as a mitigation strategy (Wortsman et al., 2022). None provides a general account: CPT updates are broadly distributed, representational divergence does not track forgetting, and interpolation trades away acquired knowledge. See App. H.2 for details.

4 RELATED WORK

Continued pretraining and temporal updating. Continued pretraining improves performance in the update domain (Gururangan et al., 2020) but can degrade the original distribution (Yıldız et al., 2025); mitigations include replay (Ibrahim et al., 2024), curated multi-epoch training (Guo et al., 2025), and low-rank adaptation (Biderman et al., 2024). Temporal updating builds on studies of degradation as training data age (Lazaridou et al., 2021; Luu et al., 2022), continual knowledge

learning (Jang et al., 2022b;a; Liška et al., 2022), and chronological pretraining (Fabre et al., 2026). Closest to our work are TiC-LM's monthly CC training (Li et al., 2025) and Öncel et al. (2026), whose FineWeb-Edu recipe we follow for our controls. Unlike our approach, post-cutoff crawls contain older text and FineWeb deduplicates only within each snapshot (Penedo et al., 2024). CC-News has been used as static English collections (Zellers et al., 2019; Liu et al., 2019; Mackenzie et al., 2020) and in small temporal updates (Jang et al., 2022b). Infini-News indexes CC-News for search, without deduplication (Lazzaroni et al., 2026). We instead construct a CPT corpus with three stages of deduplication, language, and document quality filtering, and extract eval questions.

Factual acquisition and integration. Factual accuracy depends on training frequency (Kandpal et al., 2023; Kang & Choi, 2023) and entity popularity (Mallen et al., 2023). Frequency also shapes acquisition dynamics and capacity (Chang et al., 2024; Zucchet et al., 2025; Allen-Zhu & Li, 2025), while storing a fact does not ensure reliable extraction (Allen-Zhu & Li, 2024), and models can learn co-occurrence rather than the factual association (Zhang et al., 2024). We instead measure how naturally occurring exposure in a temporal update corpus predicts *per-question changes* after CPT, using a matched-budget control (Sec. 3.2). Learning the fact also need not update its consequences: knowledge editing shows limited propagation (Onoe et al., 2023; Cohen et al., 2024), with related failures in reversing or composing learned facts (Berglund et al., 2024; Balesni et al., 2025). Fine-tuned facts can reach their implications when pretraining exposed the relevant relation (Feng et al., 2025), while CPT on synthetic news shows weak transfer to indirect questions (Li & Goyal, 2025). We find the same gap after CPT on 6B tokens of real news (Sec. 3.3).

Forgetting and its location in the model. Forgetting can reflect loss of accessible knowledge or disrupted task behavior (Kotha et al., 2024; Zheng et al., 2025), and varies across examples (Toneva et al., 2019; Jin & Ren, 2024). Prior work studies where learned knowledge and update effects reside, with evidence for MLPs, localized weight changes, and differences between low-rank and full updates (Geva et al., 2021; Meng et al., 2022; Nief et al., 2026; Shuttleworth et al., 2025). Weight interpolation has also been used to retain fine-tuning gains while improving robustness or temporal adaptation (Wortsman et al., 2022; Nylund et al., 2024). We analyze forgetting by initial confidence and topic (Sec. 3.4), and locate acquisition and capability loss through module-wise splicing and interpolation (Sec. 3.6). See App. I for a longer discussion of related work.

5 Discussion

Our findings suggest three directions for making temporal continued pretraining more reliable.

Better updating starts with better coverage. At a fixed budget, CC-News yields substantially greater factual acquisition than FineWeb-Edu, but also larger downstream losses. Together with the exposure–acquisition relationship and diminishing returns from longer training, this suggests that extending CPT on the same distribution is unlikely to solve the long tail. Corpus construction is therefore part of the learning problem: sparsely reported facts may need to be upweighted, rewritten in more varied forms, or handled outside the weights. An important next question is whether acquisition depends primarily on the number of statements or on the diversity of how a fact is expressed.

The update itself can be better controlled. Acquisition and capability loss are related, but our module-wise intervention shows that they are not inseparable. Retaining only the CPT-induced MLP updates preserves most of the new factual knowledge while recovering part of several downstream losses. This motivates moving from post-hoc repair to constrained updates, e.g., by regularizing parameter drift or selectively freezing modules. It also remains important to disentangle the cost of specializing on news from other corpus differences, particularly FineWeb-Edu's educational filter.

The harder problem is integrating knowledge over time. Acquired facts propagate only weakly to their consequences, and outdated answers remain accessible after CPT. This raises a distinction between missing exposure and failed integration: some consequences may never appear in the update corpus, while others may be stated but fail to connect to what the model already knows. Measuring exposure for follow-up facts would separate these cases. There are, however, reasons to expect this limitation to improve: larger models perform better on our ripple questions, and additional test-time computation, e.g., reasoning-oriented inference, may help models use newly acquired facts more effectively. Testing this is an important next step, although experiments with larger and reasoning models are more compute-intensive. Repeated chronological updates and methods that retire superseded knowledge while preserving valid facts remain important directions as well.

AI USE STATEMENT

Generative AI was used in this work as follows.

Synthetic data. All question sets were written by LLMs from source text. The CC-News questions were generated by MISTRAL-MEDIUM-3.5-128B and QWEN3.6-27B following the Daily Oracle pipeline, and the same two models served as judges for the quality filter (App. D) and for the audit of forgotten questions (App. F). The Wikipedia Current Events questions were generated by QWEN3.6-27B (App. C). The follow-up question bank of App. G.1 was built with CLAUDE OPUS 5: one pass screened the after-cutoff Wikipedia questions for usable anchors and wrote the follow-up questions, and a separate pass by the same model, which had not seen the authoring step, checked every question against its recorded answer basis and removed those it marked wrong or ambiguous.

Implementation. The follow-up question bank and its evaluation protocol were designed by the authors. An AI coding assistant (CLAUDE) implemented that design: the prompts for the question-writing and verification passes, the evaluation pipeline, and the analysis scripts for that part of the paper, and it produced first-pass tabulations of the results. The authors reviewed the code and verified it with unit tests and by reproducing, on a shared question set, the accuracies obtained with the evaluation code used for the rest of the paper.

Writing and support. We used AI tools for brainstorming, for drafting and editing parts of the text, for literature search and reference formatting, and for writing plotting and data-processing scripts.

Review. All AI-assisted text was rewritten or reviewed by the authors. Every bibliography entry was copied from the publisher's record and checked against the cited source. The AI-generated question sets are released in full together with their judge scores and a random sample of the follow-up questions was read by the authors. AI-written code was checked against the raw experimental outputs. We take responsibility for the final content of this work, including all text, claims and artifacts produced with the aid of generative AI.

Methodology. AI tools were not used to design the research methodology or experiments. AI tools were not used to develop theoretical models, to formulate or prove mathematical claims, or for qualitative analysis; these tasks do not arise in this work.

ETHICS STATEMENT

We have read and adhere to the ICLR Code of Ethics. This work involves continual pretraining of publicly available language models on public web text, and does not involve human subjects or private data. The models may inherit biases present in their base checkpoints and training data. We will release our code to support reproducibility and are aware of no conflicts of interest.

REPRODUCIBILITY STATEMENT

All continual pretraining runs use the HuggingFace `Trainer` with a shared, deterministic data pipeline described in Section B. The complete hyperparameter configuration for every model, covering sequence length, batch size, optimizer, learning rate schedule, warmup, precision, and parallelism, is reported in Tables 2, 3, and 4. Both the control and treatment arms are trained by us from the released base checkpoints using the identical recipe within each model family. A single fixed learning rate of 5×10^{-5} (warmup 500 steps, polynomial decay to $0.1\times$ the peak) is applied uniformly across all models and both arms, with no per-model or per-arm tuning on the evaluation questions. All runs use a fixed random seed of 42, which also governs data ordering and ensures identical shuffling across the control and treatment arms. The data preprocessing scripts, tokenization procedure, and training scripts are released to allow exact reproduction of every run.

REFERENCES

Hugo Abonizio, Thales Almeida, Roberto Lotufo, and Rodrigo Nogueira. Comparing knowledge injection methods for LLMs in a low-resource regime, 2025. URL https://arxiv.org/abs/2508.06178.

Allen Institute for AI. Olmo, 2024. URL https://github.com/allenai/OLMo. Commit 090253d. Accessed: 2026-09-23.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. In *International Conference on Machine Learning (ICML)*, 2024.

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 14937–14946, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/26d3c9a66836ded8f34a944f2bfe868e-Paper-Conference.pdf.

Thales Sales Almeida, Giovana Kerche Bonás, João Guilherme Alves Santos, Hugo Abonizio, and Rodrigo Nogueira. TiEBe: Tracking language model recall of notable worldwide events through time, 2025. URL https://arxiv.org/abs/2501.07482.

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, August 2024.

Mikita Balesni, Tomek Korbak, and Owain Evans. Lessons from studying two-hop latent reasoning, 2025. URL https://arxiv.org/abs/2411.16353.

Lukas Berglund, Meg Tong, Maximilian Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: LLMs trained on "A is B" fail to learn "B is A". In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=GPKTIktA0k.

Louis Béthune, David Grangier, Dan Busbridge, Eleonora Gualdoni, Marco Cuturi, and Pierre Ablin. Scaling laws for forgetting during finetuning with pretraining data injection. In *Proceedings of the 42nd International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2025. URL https://proceedings.mlr.press/v267/bethune25a.html.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. Lora learns less and forgets less. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://arxiv.org/abs/2405.09673.

Lukas Borggren, Jenny Kunz, and Marco Kuhlmann. Reading the news: Adapting large language models to Swedish journalism through continued pre-training, 2026. URL https://arxiv.org/abs/2608.30609.

A.Z. Broder. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, pp. 21–29, 1997.

Hoyeon Chang, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, and Minjoon Seo. How do large language models acquire factual knowledge during pretraining? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2406.11813.

Tyler A. Chang, Dheeraj Rajagopal, Tolga Bolukbasi, Lucas Dixon, and Ian Tenney. Scalable influence and fact tracing for large language model pretraining. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=gLa96FlWwn.

Howard Chen, Jiayi Geng, Adithya Bhaskar, Dan Friedman, and Danqi Chen. Continual memorization of factoids in language models. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=5Yd5QAKIFR.

Jeffrey Cheng, Marc Marone, Orion Weller, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. Dated data: Tracing knowledge cutoffs in large language models. In *Conference on Language Modeling (COLM)*, 2024.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298, 2024.

Hui Dai, Ryan Teehan, and Mengye Ren. Are LLMs prescient? A continuous evaluation using daily news as the oracle. In *International Conference on Machine Learning (ICML)*, 2025.

MohammadReza Davari, Nader Asadi, Sudhir Mudur, Rahaf Aljundi, and Eugene Belilovsky. Probing representation forgetting in supervised and unsupervised continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16712–16721, 2022.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273, 2022. doi: 10.1162/tacl_a_00459. URL https://aclanthology.org/2022.tacl-1.15/.

Ali Doosthosseini, Jonathan Decker, Hendrik Nolte, and Julian Kunkel. SAIA: a seamless Slurm-native solution for HPC-based services. *The Journal of Supercomputing*, 82(7):403, May 2026. ISSN 1573-0484. doi: 10.1007/s11227-026-08508-3. URL https://doi.org/10.1007/s11227-026-08508-3.

Romain Fabre, Hippolyte Pilchen, Franck SIGNE TALLA, Patrick Perez, and Edouard Grave. Understanding data temporality impact on large language models pre-training. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=P82s05mnUY.

Jiahai Feng, Stuart Russell, and Jacob Steinhardt. Extractive structures learned in pretraining enable generalization on finetuned facts. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 16853–16883. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/feng25m.html.

Yiyang Feng, Zeming Chen, Haotian Wu, Jiawei Zhou, and Antoine Bosselut. Tracking the limits of knowledge propagation: How LLMs fail at multi-step reasoning with conflicting knowledge. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5813–5847, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.273. URL https://aclanthology.org/2026.eacl-long.273/.

Jared Fernandez, Yonatan Bisk, and Emma Strubell. Gradient localization improves lifelong pretraining of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 16188–16195, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.949. URL https://aclanthology.org/2024.findings-emnlp.949/.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning LLMs on new knowledge encourage hallucinations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7765–7784, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.444. URL https://aclanthology.org/2024.emnlp-main.444/.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. A large-scale multi-document summarization dataset from the Wikipedia current events portal. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp.

1302–1308, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.120. URL https://aclanthology.org/2020.acl-main.120/.

Xinran Gu, Kaifeng Lyu, Jiazheng Li, and Jingzhao Zhang. Data mixing can induce phase transitions in knowledge acquisition. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 163200–163244, 2025a. doi: 10.52202/085713-5449. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/eee7ae5cf0c4356c2aeca400771791aa-Paper-Conference.pdf.

Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, and Hannaneh Hajishirzi. OLMES: A standard for language model evaluations. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5020–5048, Albuquerque, New Mexico, April 2025b. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.282. URL https://aclanthology.org/2025.findings-naacl.282/.

Yiduo Guo, Jie Fu, Huishuai Zhang, and Dongyan Zhao. Efficient domain continual pretraining by mitigating the stability gap. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vienna, Austria, July 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.acl-long.1578/.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Jackson Harmon, Andreas Hochlehnert, Matthias Bethge, and Ameya Prabhu. Mapping post-training forgetting in language models at scale. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=qCIg2WGudx.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? Surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 17643–17668, 2023. doi: 10.52202/075280-0774. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/3927bbdcf0e8d1fa8aa23c26f358a281-Paper-Conference.pdf.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface form competition: Why the highest probability answer isn't always right. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7038–7051, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.564. URL https://aclanthology.org/2021.emnlp-main.564/.

Wenyue Hua, Jiang Guo, Mingwen Dong, Henghui Zhu, Patrick Ng, and Zhiguo Wang. Propagation and pitfalls: Reasoning-based assessment of knowledge editing through counterfactual tasks. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12503–12525, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.743. URL https://aclanthology.org/2024.findings-acl.743/.

Hugging Face. Fineweb-edu classifier. https://huggingface.co/HuggingFaceFW/fineweb-edu-classifier, 2024. Accessed: 2026-09-23.

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats Leon Richter, Quentin Gregory Anthony, Eugene Belilovsky, Timothée Lesort, and Irina Rish. Simple and scalable strategies to continually pre-train large language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=DimPeeCxKO.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, pp. 604–613. Association for Computing Machinery, 1998.

Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. TemporalWiki: A lifelong benchmark for training and evaluating ever-evolving language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6237–6250, Abu Dhabi, United Arab Emirates, December 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.418. URL https://aclanthology.org/2022.emnlp-main.418/.

Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, Stanley Jungkyu Choi, and Minjoon Seo. Towards continual knowledge learning of language models. In *International Conference on Learning Representations (ICLR)*, 2022b.

Xisen Jin and Xiang Ren. What will my model forget? Forecasting forgotten examples in language model refinement. In *Proceedings of the 41st International Conference on Machine Learning*, Proceedings of Machine Learning Research, pp. 22145–22159. PMLR, 2024. URL https://proceedings.mlr.press/v235/jin24d.html.

Xisen Jin, Dejiao Zhang, Henghui Zhu, Wei Xiao, Shang-Wen Li, Xiaokai Wei, Andrew Arnold, and Xiang Ren. Lifelong pretraining: Continually adapting language models to emerging corpora. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4764–4780. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.351. URL https://aclanthology.org/2022.naacl-main.351/.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning (ICML)*, 2023. URL https://arxiv.org/abs/2211.08411.

Cheongwoong Kang and Jaesik Choi. Impact of co-occurrence on factual knowledge of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7721–7735. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.518. URL https://aclanthology.org/2023.findings-emnlp.518/.

Yannis Karmim, Luis Marti, Djamé Seddah, and Valentin Barrière. Multi-hop knowledge composition is bound by pretraining exposure, 2026. URL https://arxiv.org/abs/2606.09338.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah Smith, Yejin Choi, and Kentaro Inui. RealTime QA: What's the answer right now? In *Advances in Neural Information Processing Systems*, volume 36, pp. 49025–49043, 2023. doi: 10.52202/075280-2130. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/9941624ef7f867a502732b5154d30cb7-Paper-Datasets_and_Benchmarks.pdf.

Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, and Bing Liu. Continual pretraining of language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=m_GDIItaI3o.

Jiyeon Kim, Hyunji Lee, Hyowon Cho, Joel Jang, Hyeonbin Hwang, Seungpil Won, Youbin Ahn, Dohaeng Lee, and Minjoon Seo. Knowledge entropy decay during language model pretraining hinders new knowledge acquisition. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=eHehzSDUFp.

Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. Understanding catastrophic forgetting in language models via implicit inference. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2309.10105.

Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, Dani Yogatama, Kris Cao, Susannah Young, and Phil Blunsom. Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, 2021.

Ruggero Marino Lazzaroni, Jana Lasser, and Kirill Solovev. Infini-News: Efficiently queryable access to 1.3 billion processed Common Crawl news articles, 2026. URL https://arxiv.org/abs/2605.18337.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *60th Annual Meeting of the Association for Computational Linguistics*, 2022.

Aochong Oliver Li and Tanya Goyal. Memorization vs. reasoning: Updating LLMs with new knowledge. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25853–25874, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1326. URL https://aclanthology.org/2025.findings-acl.1326/.

Jeffrey Li, Mohammadreza Armandpour, Seyed Iman Mirzadeh, Sachin Mehta, Vaishaal Shankar, Raviteja Vemulapalli, Samy Bengio, Oncel Tuzel, Mehrdad Farajtabar, Hadi Pouransari, and Fartash Faghri. TiC-LM: A web-scale benchmark for time-continual LLM pretraining. In *63rd Annual Meeting of the Association for Computational Linguistics*, Vienna, Austria, 2025. Association for Computational Linguistics.

Adam Liška, Tomáš Kočiský, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien de Masson d'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-McMahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. StreamingQA: A benchmark for adaptation to new knowledge over time in question answering models. In *International Conference on Machine Learning (ICML)*, 2022.

Yihong Liu, Mingyang Wang, Amir Hossein Kargaran, Felicia Körner, Ercong Nie, Barbara Plank, François Yvon, and Hinrich Schuetze. Tracing multilingual factual knowledge acquisition in pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 2121–2146. Association for Computational Linguistics, 2025. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.113. URL https://aclanthology.org/2025.findings-emnlp.113/.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692.

Llama Team. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. Fineweb-edu: the finest collection of educational content, 2024. URL https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu.

Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A. Smith. Time waits for no one! analysis and challenges of temporal misalignment. In *2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2022.

Joel Mackenzie, Rodger Benham, Matthias Petri, Johanne R Trippas, J Shane Culpepper, and Alistair Moffat. CC-News-En: A large English news corpus. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pp. 3077–3084, 2020.

Divyam Madaan, Sumit Chopra, and Kyunghyun Cho. Temporal generalization: A reality check. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Wz0ILlbh9U.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023. URL https://arxiv.org/abs/2212.10511.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 2022.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations*, 2023.

Jack Merullo, Noah A. Smith, Sarah Wiegreffe, and Yanai Elazar. On linear representations and pretraining data frequency in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=EDoD3DgivF.

Mistral AI. Mistral medium 3.5. https://mistral.ai/news/vibe-remote-agents-mistral-medium-3-5/, 2026. Accessed: 2026-09-23.

Todd Nief, David Reber, Sean Richardson, and Ari Holtzman. Dynamic weight grafting: Localizing finetuned factual knowledge in transformers. In *International Conference on Learning Representations*, volume 2026, pp. 15404–15436, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/19ec4973d3a4a2cb1ce66342e00db918-Paper-Conference.pdf.

Kai Nylund, Suchin Gururangan, and Noah Smith. Time is encoded in the weights of finetuned language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2571–2587, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.141. URL https://aclanthology.org/2024.acl-long.141/.

Yasumasa Onoe, Michael Zhang, Shankar Padmanabhan, Greg Durrett, and Eunsol Choi. Can LMs learn new entities from descriptions? Challenges in propagating injected knowledge. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5469–5485. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.300. URL https://aclanthology.org/2023.acl-long.300/.

Yixin Ou, Yunzhi Yao, Ningyu Zhang, Hui Jin, Jiacheng Sun, Shumin Deng, Zhenguo Li, and Huajun Chen. How do LLMs acquire new knowledge? A knowledge circuits perspective on continual pre-training. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19889–19913. Association for Computational Linguistics, 2025. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1021. URL https://aclanthology.org/2025.findings-acl.1021/.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. Fine-tuning or retrieval? Comparing knowledge injection in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 237–250. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.15. URL https://aclanthology.org/2024.emnlp-main.15/.

Shankar Padmanabhan, Yasumasa Onoe, Michael Zhang, Greg Durrett, and Eunsol Choi. Propagating knowledge updates to LMs through distillation. In *Advances in Neural Information Processing Systems*, pp. 47124–47142, 2023. doi: 10.52202/075280-2041. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/932147114c48f8b04d41aebc0c631158-Paper-Conference.pdf.

Abhishek Panigrahi, Nikunj Saunshi, Haoyu Zhao, and Sanjeev Arora. Task-specific skill localization in fine-tuned language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*,

pp. 27011–27033. PMLR, 2023. URL https://proceedings.mlr.press/v202/panigrahi23a.html.

Core Francisco Park, Zechen Zhang, and Hidenori Tanaka. *New News*: System-2 fine-tuning for robust integration of new knowledge, 2025a. URL https://arxiv.org/abs/2505.01812.

Yein Park, Chanwoong Yoon, Jungwoo Park, Donghyeon Lee, Minbyul Jeong, and Jaewoo Kang. ChroKnowledge: Unveiling chronological knowledge of language models in multiple domains. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=wha03482bs.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=n6SCkn2QaG.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-naacl.130. URL https://aclanthology.org/2024.findings-naacl.130/.

Jiaxin Qin, Zixuan Zhang, Chi Han, Pengfei Yu, Manling Li, and Heng Ji. Why does new knowledge create messy ripple effects in LLMs? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12602–12609, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.700. URL https://aclanthology.org/2024.emnlp-main.700/.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026. URL https://qwen.ai/blog?id=qwen3.6-27b.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew J. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*, abs/2112.11446, 2021. URL https://arxiv.org/abs/2112.11446.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), January 2020. ISSN 1532-4435.

Vinay Venkatesh Ramasesh, Ethan Dyer, and Maithra Raghu. Anatomy of catastrophic forgetting: Hidden representations and task semantics. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=LhY8QdUGSuw.

Yuval Reif and Roy Schwartz. Beyond performance: Quantifying and mitigating label bias in LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 6784–6798. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.378. URL https://aclanthology.org/2024.naacl-long.378/.

Sunny Sanyal, Hayden Prairie, Rudrajit Das, Ali Kavis, and Sujay Sanghavi. Upweighting easy samples in fine-tuning mitigates forgetting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 52922–52957. PMLR, 2025. URL https://proceedings.mlr.press/v267/sanyal25a.html.

Jia Sheng and Yiwei Lu. No universal signal predicts sample-level LLM regression under version updates, 2026. URL https://arxiv.org/abs/2608.13607.

Reece Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. LoRA vs full fine-tuning: An illusion of equivalence. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.52202/085713-5808. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/ff541950d1e885af90f523571564a401-Paper-Conference.pdf.

Jacob Mitchell Springer, Sachin Goyal, Kaiyue Wen, Tanishq Kumar, Xiang Yue, Sadhika Malladi, Graham Neubig, and Aditi Raghunathan. Overtrained language models are harder to fine-tune. In *International Conference on Machine Learning (ICML)*, 2025. URL https://arxiv.org/abs/2503.19206.

Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Nemotron-CC: Transforming Common Crawl into a refined long-horizon pretraining dataset. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2459–2475, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.123. URL https://aclanthology.org/2025.acl-long.123/.

Wei Tang, Yixin Cao, Yang Deng, Jiahao Ying, Bo Wang, Yizhe Yang, Yuyue Zhao, Qi Zhang, Xuanjing Huang, Yu-Gang Jiang, and Yong Liao. EvoWiki: Evaluating LLMs on evolving knowledge. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 948–964. Association for Computational Linguistics, 2025. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.47. URL https://aclanthology.org/2025.acl-long.47/.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BJlxm30cKm.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13697–13720. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.813. URL https://aclanthology.org/2024.findings-acl.813/.

Evan Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James Validad Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 OLMo 2 furious. In *Conference on Language Modeling (COLM)*, 2025.

Ke Wang, Nikolaos Dimitriadis, Alessandro Favero, Guillermo Ortiz-Jimenez, François Fleuret, and Pascal Frossard. LiNeS: Post-training layer scaling prevents forgetting and enhances model merging. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=J5sUOvlLbQ.

Xingjin Wang, Howe Tissue, Lu Wang, Linjing Li, and Daniel Dajun Zeng. Learning dynamics in continual pre-training for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 64525–64548. PMLR, 2025b. URL https://proceedings.mlr.press/v267/wang25cx.html.

Sarah Wiegreffe, Oyvind Tafjord, Yonatan Belinkov, Hannaneh Hajishirzi, and Ashish Sabharwal. Answer, assemble, ace: Understanding how LMs answer multiple choice questions. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6NNA0MxhCH.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7959–7971, June 2022.

Xiaobao Wu, Liangming Pan, Yuxi Xie, Ruiwen Zhou, Shuai Zhao, Yubo Ma, Mingzhe Du, Rui Mao, Anh Tuan Luu, and William Yang Wang. AntiLeakBench: Preventing data contamination by automatically constructing benchmarks with updated real-world knowledge. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18403–18419. Association for Computational Linguistics, 2025. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.901. URL https://aclanthology.org/2025.acl-long.901/.

Ruoxi Xu, Yunjie Ji, Boxi Cao, Yaojie Lu, Hongyu Lin, Xianpei Han, Ben He, Yingfei Sun, Xiangang Li, and Le Sun. Memorizing is not enough: Deep knowledge injection through reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28682–28693. Association for Computational Linguistics, 2025. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1392. URL https://aclanthology.org/2025.acl-long.1392/.

Zhuoyi Yang, Yurun Song, Iftekhar Ahmed, and Ian Harris. Fine-tuning vs. RAG for multi-hop question answering with novel knowledge, 2026. URL https://arxiv.org/abs/2601.07054.

Zitong Yang, Neil Band, Shuangping Li, Emmanuel J. Candès, and Tatsunori Hashimoto. Synthetic continued pretraining. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2409.07431.

Yuekun Yao, Yupei Du, Dawei Zhu, Michael Hahn, and Alexander Koller. Language models can learn implicit multi-hop reasoning, but only if they have lots of training data. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9684–9702. Association for Computational Linguistics, 2025. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.490. URL https://aclanthology.org/2025.emnlp-main.490/.

Çağatay Yıldız, Nishaanth Kanna Ravichandran, Nitin Sharma, Matthias Bethge, and Beyza Ermis. Investigating continual pretraining in large language models: Insights and implications. *Transactions on Machine Learning Research (TMLR)*, 2025.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf.

Xiao Zhang and Ji Wu. Dissecting learning and forgetting in language model finetuning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=tmsqb6WpLz.

Xiao Zhang, Miao Li, and Ji Wu. Co-occurrence is not factual association in language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 64889–64914, 2024. doi: 10.52202/079017-2071. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/775226eaa2a36c543e2bd6cc9eae1b6a-Paper-Conference.pdf.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2309.03882.

Junhao Zheng, Xidi Cai, Shengjie Qiu, and Qianli Ma. Spurious forgetting in continual learning of language models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2501.13453.

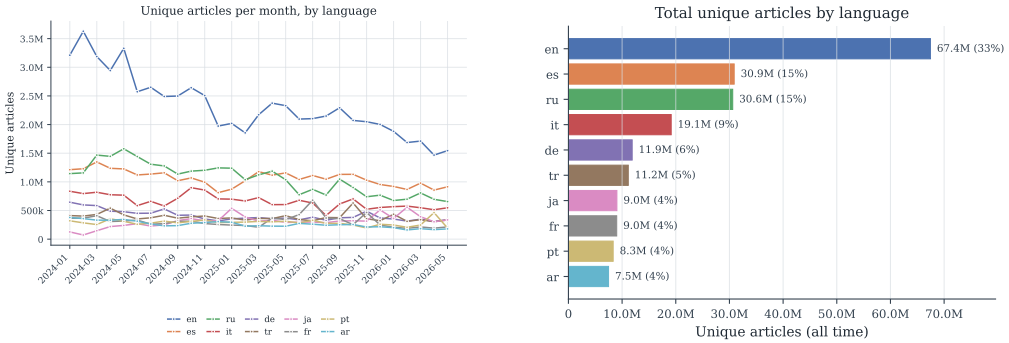
Zexuan Zhong, Zhengxuan Wu, Christopher Manning, Christopher Potts, and Danqi Chen. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15686–15702. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.971. URL https://aclanthology.org/2023.emnlp-main.971/.

Nicolas Zucchet, Jorg Bornschein, Stephanie C.Y. Chan, Andrew Kyle Lampinen, Razvan Pascanu, and Soham De. How do language models learn facts? Dynamics, curricula and hallucinations. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=vBcGnragkr.

Fırat Öncel, Salman Hussain Ali, Mirco Ravanelli, Cem Subakan, and Çağatay Yıldız. Time-incremental continued pretraining of llms: Knowledge updates without catastrophic forgetting. *arXiv preprint arXiv:2609.23916*, 2026.

APPENDIX

A DATASET DISTRIBUTION



(a) Unique articles per month, by language, over the full collection period (January 2024 – May 2026). English is consistently the largest language and exhibits a downward trend over time; the remaining nine languages are more stable.

(b) Total number of unique articles per language in the final corpus. English dominates the collection (∼33%), followed by Spanish and Russian (∼15% each).

Figure 7: Temporal composition of the corpus.

B IMPLEMENTATION DETAILS

We continually pretrain three model families, OLMo-2, Qwen3, and Llama-3, on the CC-News corpus. All runs use the HuggingFace `Trainer` with a shared data pipeline: documents are tokenized with each model's native tokenizer, separated by an end-of-sequence token, concatenated, and packed into fixed-length blocks. Tables 2, 3, and 4 report the full hyperparameter configuration for each family.

Table 2: Continual pretraining hyperparameters for OLMo-2 models. All runs use OLMo-style z-loss ($\lambda_z = 10^{-5}$) added to the cross-entropy objective.

Hyperparameter	**OLMo-2-1B**	**OLMo-2-7B**	**OLMo-2-13B**
Sequence length	4096	4096	4096
Epochs	1	1	1
Per-device batch size	8	4	8
Gradient accumulation	2	4	2
GPUs	8	8	8
Effective batch size	128	128	128
Optimizer	AdamW (fused)	AdamW (fused)	AdamW (fused)
Learning rate	5×10^{-5}	5×10^{-5}	5×10^{-5}
β_1, β_2	0.9, 0.95	0.9, 0.95	0.9, 0.95
Weight decay	0.1	0.1	0.1
Max grad norm	1.0	1.0	1.0
z-loss λ_z	10^{-5}	10^{-5}	10^{-5}
Scheduler	Polynomial	Polynomial	Polynomial
Warmup steps	500	500	500
LR end	5×10^{-6}	5×10^{-6}	5×10^{-6}
Precision	BF16 (TF32)	BF16 (TF32)	BF16 (TF32)
Attention	FlashAttention-2	FlashAttention-2	FlashAttention-2
Grad checkpointing	Yes	Yes	Yes
`torch_compile`	Yes	Yes	No
Parallelism	DDP	DDP	FSDP (full shard)

Table 3: Continual pretraining hyperparameters for Qwen3 models.

Hyperparameter	**Qwen3-1.7B**	**Qwen3-4B**	**Qwen3-8B**
Sequence length	8192	8192	8192
Epochs	1	1	1
Per-device batch size	8	4	4
Gradient accumulation	1	2	2
GPUs	8	8	8
Effective batch size	64	64	64
Optimizer	AdamW (fused)	AdamW (fused)	AdamW (fused)
Learning rate	5×10^{-5}	5×10^{-5}	5×10^{-5}
β_1, β_2	$0.9,\ 0.95$	$0.9,\ 0.95$	$0.9,\ 0.95$
Weight decay	0.1	0.1	0.1
Max grad norm	1.0	1.0	1.0
Scheduler	Polynomial	Polynomial	Polynomial
Warmup steps	500	500	500
LR end	5×10^{-6}	5×10^{-6}	5×10^{-6}
Precision	BF16 (TF32)	BF16 (TF32)	BF16 (TF32)
Attention	FlashAttention-2	FlashAttention-2	FlashAttention-2
Grad checkpointing	Yes	Yes	Yes
`torch_compile`	Yes	Yes	Yes
Parallelism	DDP	DDP	DDP
Seed	42	42	42

Table 4: Continual pretraining hyperparameters for Llama-3 models.

Hyperparameter	**Llama-3.2-3B**	**Llama-3.1-8B**
Sequence length	8192	8192
Epochs	1	1
Per-device batch size	8	8
Gradient accumulation	1	1
GPUs	8	8
Effective batch size	64	64
Optimizer	AdamW (fused)	AdamW (fused)
Learning rate	5×10^{-5}	5×10^{-5}
β_1, β_2	$0.9,\ 0.95$	$0.9,\ 0.95$
Weight decay	0.1	0.1
Max grad norm	1.0	1.0
Scheduler	Polynomial	Polynomial
Warmup steps	500	500
LR end	5×10^{-6}	5×10^{-6}
Precision	BF16 (TF32)	BF16 (TF32)
Attention	FlashAttention-2	FlashAttention-2
Grad checkpointing	Yes	Yes
`torch_compile`	Yes	Yes
Parallelism	DDP	DDP

C EVALUATION QUESTIONS

We evaluate on three question sets falling into two difficulty levels. On the hard end are $\sim 22k$ *CC-News* and $\sim 18.5k$ *Daily Oracle questions*. The former is sampled uniformly from our own pretraining corpus; hence, it inherits the dataset's long tail, and most questions ask about a story that is covered a handful of times unlike Wikipedia-derived questions. Likewise, (Dai et al., 2025) is harvested from a news pipeline outside our corpus. *Wikipedia questions* are on the easier end: $\sim 40k$ are extracted from world events prominent enough to be recorded on Wikipedia, so they are broad.

CC-News questions. We construct three question sets directly from the English portion of CC-News using the Daily Oracle extraction pipeline (Dai et al., 2025). We phrase the questions in the past tense rather than the future tense used by default in Daily Oracle since they are about the events the models have read about by the end of CPT. For each source article, the pipeline produces two true/false and two four-way multiple-choice questions. The first two sets are generated from the same sampled articles but with different question generators, MISTRAL-MEDIUM-3.5-128B (Mistral AI, 2026) and QWEN3.6-27B (Qwen Team, 2026). This lets us vary the question-generation model while holding the underlying news fixed. The third set is generated with QWEN3.6-27B from a disjoint sample of CC-News articles, reducing dependence on the particular source articles selected for the first two sets. Table 6 summarizes the three sets. Unless stated otherwise, all CC-News results in the paper report accuracy pooled across the multiple-choice questions from all three sets.

Because these questions are extracted from the continued-pretraining corpus itself, every item can be traced to a source article that the CC-News checkpoint was trained on. We additionally evaluate the quality and stability of the generated questions using two LLM judges and compare results across generators, article samples, and filtered versus unfiltered subsets. These checks show that the measured continued-pretraining effect is stable despite differences in absolute question difficulty. Full details of sampling, generation, filtering, and benchmark validation are given in App. D.

Daily Oracle questions. Daily Oracle (Dai et al., 2025) is a continuously regenerated question set built by running an extraction pipeline over the day's news: articles are collected daily, summarised, and turned into TF and MC items. Because the pipeline keeps running, the set carries a timestamp for every item and can be sliced by month, which lets us split it at the shared December 2023 cutoff exactly as we do the Wikipedia set. We score only the multiple-choice half, for the reason given in Section 3.1.

This set is collected independently of our training corpus through a separate news pipeline. Daily Oracle questions come from prominent sources (CNN, Forbes, CBS News, CNBC, and NPR), and consequently its coverage is more US-focused/limited to the US. Only 16.73% of Daily Oracle source URLs appear in our CC-News corpus, so the large majority of questions are derived from source articles that are not present under the same URL in the update corpus. It therefore primarily tests transfer beyond the exact source articles used for evaluation, although different outlets may of course report the same underlying events.

Wikipedia questions. Our last question set is derived from Wikipedia's daily current-events record, which logs the day's notable world events as short dated entries under editor-assigned topical headings (*armed conflicts and attacks*, *politics and elections*, *disasters and accidents*, and so on). Each entry is converted into a single four-way multiple-choice question with QWEN3.6-27B, anchored to the entry's date and asking about a specific detail of the event rather than its gist, and we retain the Wikipedia heading as a category label. Unlike the previous sets, these items are not passed through judge filtering: the editorial threshold for being recorded on Wikipedia at all already selects for events of public interest with a clear time element. Because the questions are generated from Wikipedia's own Current Events entries rather than from the news articles in CC-News, this set tests whether knowledge acquired from news transfers to a separately written description of the same events.

Table 5: Number of questions and the dates they cover for each dataset, split by whether the event date is before or after 1 January 2024.

	CCNews	Daily Oracle	Wikipedia
Before	8,370 Jan 2020 – Dec 2023	11,623 Jan 2020 – Dec 2023	25,773 Jan 2020 – Dec 2023
After	13,964 Jan 2024 – May 2026	6,935 Jan 2024 – July 2026	13,995 Jan 2024 – May 2026

D Details and validation of question extraction from CC-News

Sampling and question format. We build two disjoint question batches from the English portion of the CC-News. For each batch we sample three articles per day over the 29 months covered by the corpus (1 January 2024 – 31 May 2026), i.e. a target of roughly 2,650 articles per batch, and generate four questions per article: two true/false questions (one whose answer is *yes* and one whose answer is *no*) and two four-way multiple-choice questions. Questions are produced with the Daily Oracle prompt chain (Dai et al., 2025): the article is first summarised, questions and answers are then generated from the summary, and distractors for the multiple-choice items are generated in a separate pass. Days on which no article survives generation are dropped, which is why the realised counts in Table 6 fall slightly short of the target.

Two generators, two batches. For the first batch we generate questions twice from the *same* sampled articles, once with Mistral-Medium-3.5-128B (Mistral AI, 2026) and once with Qwen3.6-27B (Qwen Team, 2026), all served through SAIA (Doosthosseini et al., 2026). Because the two generators see the same articles, this gives a controlled comparison of question quality in which the article sample is held fixed. The second batch is generated with Qwen3.6-27B only and covers articles disjoint from the first batch, which supports a cross-batch study at fixed generator. Table 6 summarises the resulting sets; note that Mistral-Medium-3.5-128B returned usable questions for a strict subset (2,031) of the articles the Qwen3.6-27B run covered (2,487), so all paired comparisons below are restricted to those 2,031 shared articles.

Question quality Each question set is scored by both models acting as judges, again following Daily Oracle (Dai et al., 2025). A judge sees the full article, its publication date and the four questions, and grades each question on seven principles (*correctness*, *only answerable on the publication date*, *no new information*, *objectiveness*, *clear time element*, *public interest*, and *answer not too obvious*) on a 0/1/2 scale. A question scores between 0 and 14. Accept rates depend strongly on the judge: Qwen3.6-27B accepts 66–69% of questions while Mistral-Medium-3.5-128B accepts 75–78%, and 55–57% of questions are accepted by both judges.

Quality filtering does not materially change model behavior. We compare performance on the full batch-1 CC-News MCQ set with the subset accepted by both LLM judges. Across six CC-News checkpoints, mean improvement over the base model is $+0.7$ points on the unfiltered set and $+0.6$ points on the consensus subset; across five non-news control checkpoints, the corresponding changes are $+0.0$ and $+0.2$ points. The same stability appears at the question-transition level. For the default Llama3.2-3B checkpoint, 16.3% of questions answered incorrectly by the base model become correct after CPT, compared with 15.9% on the consensus subset, while 28.1% of initially correct answers become incorrect, compared with 27.4% after filtering (these rates are conditioned on different item pools, so their magnitudes are not directly comparable. Filtering therefore changes neither the net CPT effect nor the underlying pattern of answer transitions appreciably, suggesting that the judge filter improves question quality without selecting a qualitatively different subset of model behavior.

What the judges (dis)agree on. The two judges assign almost identical mean totals (12.95 each on the Mistral-generated set), so the gap in accept rates is not a difference in overall strictness but a difference in *which* questions lose points. Four principles are saturated (objectiveness, correctness, no new information and clear time element each receive the full 2 points on 94–100% of questions) and the judges agree there almost perfectly (0.94–0.997 exact agreement). Essentially all deductions come from the remaining three principles, and they behave differently from one another. On *public interest* the judges disagree often (0.80 exact agreement) but disagree *consistently*: they rank questions the same way and differ mainly in where they place the cut (Mistral never awards 0 but gives

Table 6: Question batches. Every article contributes 2 true/false and 2 multiple-choice questions. The two batches cover disjoint articles; within batch 1 the MISTRAL-generated articles are a subset of the QWEN-generated ones.

batch	generator	days	articles	mc questions
1	MISTRAL-MEDIUM-3.5-128B	856	2,031	4,027
1	QWEN3.6-27B	878	2,487	4,967
2	QWEN3.6-27B	879	2,490	4,970

1 point to a third of all questions, whereas QWEN awards a few 0s and fewer 1s). *Answer not too obvious* is the genuine source of disagreement since it has the lowest exact agreement (0.55–0.62). *Only answerable on the publication date* sits in between (0.86–0.88 agreement), with both judges treating it as an essentially binary criterion—they award 0 or 2 and almost never 1. Aggregated over the accept decision this leaves a raw agreement of 0.69–0.71 at Cohen's $\kappa = 0.23$–0.27, which is why we report results under both judges rather than committing to one.

Comparing the two generators. Figure 8 compares the two generators on the 2,031 shared articles of batch 1. Under the MISTRAL judge the QWEN-generated questions are clearly preferred: mean total $12.95 \rightarrow 13.11$ and accept rate $75.1\% \rightarrow 80.4\%$, a gain driven almost entirely by *answer not too obvious* ($1.52 \rightarrow 1.71$), with the other six principles moving by at most 0.03. Under the QWEN judge the two sets are indistinguishable (mean total 12.95 and accept rate 68.6% for both). The preference therefore does not follow the familiar self-preference pattern—the judge that favours QWEN's questions is MISTRAL, while QWEN shows no bias towards its own output. Inter-judge consensus is also mildly higher on the QWEN-generated set (mean exact agreement $0.876 \rightarrow 0.888$, questions accepted by both judges $56.2\% \rightarrow 59.9\%$), i.e. those questions are not only scored higher but scored less ambiguously. We therefore use QWEN3.6-27B as the generator for batch 2.

The three sets are interchangeable as benchmarks. We finally ask whether the measured CPT effect depends on the particular articles or question generator used to construct the CC-News benchmark. The two QWEN-generated sets, which use disjoint article samples, agree at the split-half noise level: checkpoint accuracies correlate at $r = 0.995$, and the mean gain over the base model is nearly identical ($+4.68$ vs. $+4.69$ points), with no evidence that the per-family gains differ across batches ($\chi^2(8) = 9.2$, $p = 0.33$). Regenerating questions for the same articles with MISTRAL likewise leaves the CPT effect essentially unchanged: the mean gain over base is $+4.81$ points, and its direction agrees across all eight models. The generator does, however, affect absolute difficulty in a family-dependent way: QWEN3 models score about 3 points lower on QWEN-generated than on MISTRAL-generated questions, while other families show the opposite or little shift. Because this change affects the base, control, and CC-News checkpoints similarly within each family, it largely cancels when measuring the gains that we report. Thus, article sampling and question generation change absolute accuracy somewhat, but not the conclusions about the effect of continued pretraining.

D.1 ISSUES WITH TRUE/FALSE QUESTIONS

TF question set is not a usable instrument. We omit TF QAs from our evaluation sets, and we caution against reading the TF numbers anywhere as knowledge measurements. On all three TF question sets of Table 6, each balanced $50/50$ by construction, scoring degenerates into a constant answer. Under the raw multiple-choice formulation the continually pretrained checkpoint answers *yes* to 94.3%, 94.9% and 95.1% of items (the base model, 68.4–69.1%); its accuracy then sits within a point of 50.0–55.7%, the ceiling that bias alone permits on a balanced set, so the score reports a prior rather than a reading of the question. Furthermore, unlike MC questions, we could not find an unbiased way of averaging model choices (e.g., by rotating the options). We therefore report multiple choice throughout.

Not a question-quality or prompt-format problem. The obvious alternative explanation is that the questions are unanswerable or the evaluation suite is broken. An open-book control rules both out: with the source summary prepended to an otherwise identical prompt, the base model answers 81.28% of the same items correctly and the continually pretrained checkpoint 79.76%. The questions are answerable and the scoring works.

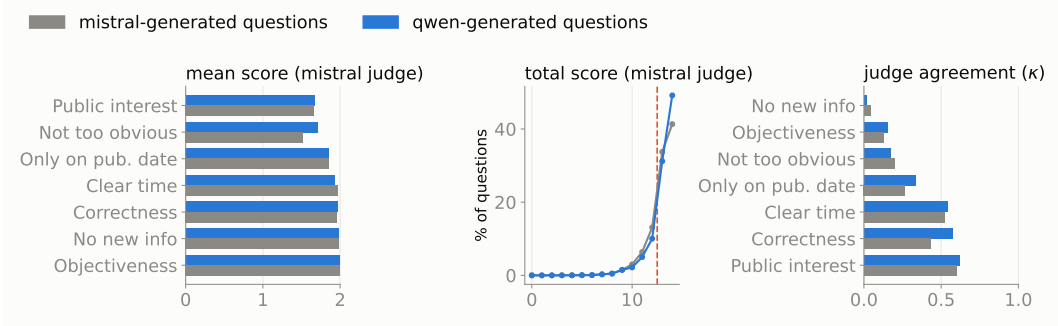


Figure 8: Batch 1 generators compared on the 2,031 articles both cover. *Left:* mean per-principle score under the MISTRAL judge; the two sets differ only on *answer not too obvious*. *Middle:* total-score distribution under the same judge, dashed line at the accept threshold of 13. *Right:* quadratic-weighted κ between the two judges per principle. The low values on *no new information* and *objectiveness* are a ceiling effect: both judges give full marks to 94–99.7% of questions, exact agreement is 94–99.7%. The few questions docked on these principles are rarely docked by both judges.

Table 7: The same three batch-1 articles rendered by the two question generators, with the correct option in bold. The generators agree on the article but not on which fact in it to ask about: for the third article MISTRAL asks about the keypoint and QWEN about a detail elsewhere in the same story.

keypoint: Kansas upset No. 7 BYU 17–13 on 16 November 2024.	
MISTRAL mc	What was the final score in the Kansas vs. No. 7 BYU game by November 16, 2024? 14-10 | 20-17 | 24-21 | **17-13**
QWEN mc	Which team handed BYU their first loss of the 2024 season on November 16, 2024? Oklahoma | Colorado | **Kansas** | Arizona State
keypoint: Arvind Kejriwal announced his resignation as Delhi Chief Minister on 15 September 2024.	
MISTRAL mc	What did Arvind Kejriwal announce regarding his position as Delhi Chief Minister by September 15, 2024? **Resignation** | Appointment of a deputy CM | Continuation of his term | Temporary leave of absence
QWEN mc	Which political party's chief announced his resignation as Delhi Chief Minister on 2024-09-15? **Aam Aadmi Party** | Indian National Congress | Bharatiya Janata Party | Communist Party of India
keypoint: Israel struck F-14 fighter jets at Tehran's Isfahan Airport on 8 March 2026.	
MISTRAL mc	What did Israel strike at Tehran's Isfahan Airport on 2026-03-08? Civilian residential areas | Oil storage facilities | **F-14 fighter jets** | Command centers of the Revolutionary Guards
QWEN mc	How many people were killed in Israeli strikes on southern Lebanon and Beirut on 2026-03-08? 5 | 50 | **12** | 20

Table 8: Five example questions from the Wikipedia evaluation set, with the correct option in bold.

Wikipedia questions, after cutoff	
politics	Who won the Iranian presidential run-off on July 5, 2024? Ebrahim Raisi | Saeed Jalili | **Masoud Pezeshkian** | Mojtaba Zonnour
sport	Who won the 66th Daytona 500 on February 19, 2024? **William Byron** | Kyle Larson | Rick Hendrick | Denny Hamlin
science	On February 27, 2025, which large language model did OpenAI release? **GPT-4.5** | GPT-4 Turbo | GPT-4.1 | GPT-5
diplomacy	Where was the NATO summit held on June 24, 2025? Berlin | **The Hague** | Brussels | London
culture	Who was named Person of the Year by Time magazine on December 12, 2024? Joe Biden | Elon Musk | Kamala Harris | **Donald Trump**
Wikipedia questions, before cutoff	
politics	Who was sworn in as the President of India on July 25, 2022? **Droupadi Murmu** | Pranab Mukherjee | Ram Nath Kovind | S. Radhakrishnan
sport	Which horse won the Kentucky Derby on May 6, 2023? Country House | Medina Spirit | **Mage** | Rich Strike
science	What type of solar eclipse occurred on October 14, 2023? Partial | Hybrid | **Annular** | Total
business	Which American retailer filed for Chapter 11 bankruptcy on April 23, 2023? **Bed Bath & Beyond** | Sears Holdings | Toys 'R' Us | RadioShack
environment	Which Icelandic volcano began erupting on July 10, 2023? **Litli-Hrútur** | Grímsvötn | Katla | Eyjafjallajökull

E ALL MAIN FINDING NUMBERS

Table 9: **Multiple-choice accuracy (%) of every checkpoint on every question set.** Each base model is compared with a checkpoint continually pretrained on 6B tokens of post-cutoff FineWeb-Edu and one trained on CC-News under a matched budget. The CC-News columns average the three question sets of Table 6. The best checkpoint in each column of each block is in bold.

		Wikipedia		CC-News		Daily Oracle	
Base model	Training corpus	after	before	after	before	after	before
Llama-3.2-3B	— (base model)	33.39	42.71	36.99	42.02	38.49	47.26
	FineWeb-Edu	35.18	**43.51**	38.52	**42.57**	38.02	**47.33**
	CC-News	**41.01**	41.97	**40.44**	40.73	**42.49**	47.22
Llama-3.1-8B	— (base model)	41.01	53.97	46.30	52.41	46.34	58.18
	FineWeb-Edu	43.35	**54.11**	48.26	**53.03**	45.73	**58.56**
	CC-News	**53.20**	52.86	**52.41**	51.72	**50.84**	58.18
Qwen3-1.7B	— (base model)	41.33	46.17	43.33	45.30	45.42	53.51
	FineWeb-Edu	43.29	47.58	45.28	46.65	47.35	54.18
	CC-News	**46.55**	**48.48**	**47.86**	**47.63**	**49.11**	**55.27**
Qwen3-4B	— (base model)	52.73	57.56	54.23	54.73	53.24	62.96
	FineWeb-Edu	53.50	56.84	54.10	54.97	53.41	63.00
	CC-News	**58.47**	**58.41**	**57.98**	**55.72**	**57.37**	**64.67**
Qwen3-8B	— (base model)	55.87	60.73	55.61	58.39	54.36	64.55
	FineWeb-Edu	57.07	60.86	56.47	58.64	54.78	64.82
	CC-News	**61.53**	**61.66**	**59.31**	**58.86**	**57.96**	**65.82**
OLMo-2-1B	— (base model)	28.05	33.43	32.05	34.21	35.39	43.18
	FineWeb-Edu	29.01	32.74	31.20	33.58	34.58	41.83
	CC-News	**31.50**	**33.57**	**33.09**	**34.60**	**37.88**	**43.66**
OLMo-2-7B	— (base model)	38.59	47.43	42.39	46.31	46.31	56.20
	FineWeb-Edu	41.56	48.11	44.51	46.33	48.35	57.24
	CC-News	**50.46**	**49.21**	**49.99**	**49.45**	**52.88**	**59.57**
OLMo-2-13B	— (base model)	43.59	51.53	47.28	50.98	49.02	58.64
	FineWeb-Edu	46.78	52.09	48.41	50.02	46.87	58.20
	CC-News	**57.41**	**52.94**	**54.46**	**52.35**	**56.17**	**61.03**
Mean	— (base model)	41.82	49.19	44.77	48.04	46.07	55.56
	FineWeb-Edu	43.72	49.48	45.84	48.22	46.13	55.64
	CC-News	50.02	49.89	49.44	48.88	50.59	56.93

Table 10: **Multiple-choice accuracy (%) on questions about events from January 2024 to June 2025.** Each base model is compared with a checkpoint continually pretrained on 6B tokens of post-cutoff FineWeb-Edu and one trained on CC-News under a matched budget. Only after-cutoff questions dated 2024-01-01 to 2025-05-31 are kept. The CC-News column averages the three question sets of Table 6. The best checkpoint in each column of each block is in bold.

Base model	Training corpus	Wikipedia	CC-News	Daily Oracle
Llama-3.2-3B	— (base model)	33.78	37.50	43.90
	FineWeb-Edu	35.71	39.10	45.37
	CC-News	**41.35**	**40.83**	**46.97**
Llama-3.1-8B	— (base model)	41.92	46.37	54.20
	FineWeb-Edu	44.40	48.53	54.71
	CC-News	**54.03**	**52.93**	**57.53**
Qwen3-1.7B	— (base model)	41.64	43.83	50.48
	FineWeb-Edu	43.46	45.72	51.95
	CC-News	**46.48**	**48.17**	**54.07**
Qwen3-4B	— (base model)	53.42	54.67	59.95
	FineWeb-Edu	53.59	54.56	60.42
	CC-News	**58.15**	**57.97**	**63.80**
Qwen3-8B	— (base model)	57.49	56.32	62.85
	FineWeb-Edu	58.02	56.88	63.19
	CC-News	**62.10**	**59.55**	**65.70**
OLMo-2-1B	— (base model)	28.30	32.46	40.40
	FineWeb-Edu	28.80	31.71	39.23
	CC-News	**31.81**	**33.78**	**41.05**
OLMo-2-7B	— (base model)	38.87	42.72	50.52
	FineWeb-Edu	41.67	45.17	54.15
	CC-News	**51.01**	**50.25**	**58.39**
OLMo-2-13B	— (base model)	43.82	47.52	54.89
	FineWeb-Edu	46.96	49.69	56.53
	CC-News	**57.96**	**55.00**	**62.24**
Mean	— (base model)	42.41	45.17	52.15
	FineWeb-Edu	44.08	46.42	53.20
	CC-News	50.36	49.81	56.22

F ABLATIONS AND ADDITIONAL FINDINGS

We test how sensitive our conclusions are to several training choices, including random seed, learning rate, weight decay, learning-rate schedule, and data quality. Most of these choices have only modest effects relative to the main CPT gains, with learning rate being an exception. We therefore select the learning rate empirically to balance acquisition and retention, while noting that the computational cost of CPT prevents an exhaustive sweep over all model families and hyperparameter settings.

Seed variance. To separate the effect of the update corpus from run-to-run noise, we repeat the CC-News continued-pretraining run of LLAMA-3.2-3B and QWEN3-1.7B under three random seeds with an otherwise identical recipe, and evaluate every run on all multiple-choice question sets. Pooling the per-set sample variances, the standard deviation across seeds is 0.28 accuracy points (0.31 for LLAMA-3.2-3B, 0.25 for QWEN3-1.7B; 0.31–0.34), with individual sets ranging from 0.05 to 0.70 points and Daily Oracle the noisiest. No seed is systematically better: averaged across sets, each run deviates from the three-seed mean by at most 0.12 points, so seed noise is set-specific rather than a global offset. This seed noise is smaller than the item-sampling error of a single question set (0.4–0.7 points at our sample sizes). For a difference between two independently seeded checkpoints, such as CC-News versus the FineWeb-Edu control, the implied standard deviation is $\sqrt{2} \times 0.28 \approx 0.4$ points. The smallest after-cutoff advantage over the control we report, 1.76 points, is thus roughly five times this noise level, and the average advantages of 3.6–6.3 points exceed it by an order of magnitude. We note that the estimate rests on two small models at a single learning rate, and assume it transfers to the larger checkpoints.

Table 11: Standard deviation (accuracy points) across three seeds of the CC-News checkpoint

Model	pooled	median per set	max (set)
LLAMA-3.2-3B	0.31	0.28	0.62 (Wikipedia, after)
QWEN3-1.7B	0.25	0.22	0.70 (Daily Oracle, after)
Both	0.28	–	–

Learning rate. We use 5×10^{-5} to balance acquisition and retention (Figure 14). For the analysis, we held out 1k before- and after-cutoff questions. In the Llama-3.2-3B sweep, 5×10^{-5} yields the greatest percentage points increase in the held-out after-cutoff questions while before-cutoff accuracy decreases slightly (-0.66 points). Increasing the rate to 10^{-4} sharply increases forgetting, producing net losses of 2.19 and 6.73 points on the after- and before-cutoff splits, respectively. Llama-3.1-8B and OLMo2-7B show the same rise in forgetting at 10^{-4} and smaller gains at 10^{-5}, supporting 5×10^{-5} as a shared operating point.

Weight decay. We vary weight decay over $\{0, 0.01, 0.1\}$ for Llama-3.2-3B at a learning rate of 5×10^{-5}. Relative to the three-seed mean of the default configuration (0.1), reducing weight decay to 0.01 or removing it changes accuracy by an average of $+0.29$ and $+0.31$ percentage points across the 12 evaluated sets, respectively. For either setting, absolute differences remain within the observed seed-to-seed range on 10 of the 12 sets. These results indicate limited sensitivity to weight decay over the tested range.

Learning-rate schedule. Replacing the default polynomial schedule with cosine decay has a similarly modest effect. Averaged across the 12 evaluated sets, cosine decay improves accuracy by 0.18 percentage points relative to the default three-seed mean, while cosine decay with warmup improves it by 0.27 points. Individual differences range from -0.58 to $+1.17$ points, with improvements and declines across benchmarks. The lack of consistent advantages suggests that schedule choice has a limited influence on the overall results.

Higher-quality training data. We also train on the highest-scoring 6B tokens selected using the FineWeb-Edu classifier (Hugging Face, 2024). Relative to the default three-seed mean, before-cutoff accuracy improves on evaluated sets, by 0.14–0.80 percentage points. After-cutoff results are mixed: Wikipedia accuracy changes little ($+0.08$ to $+0.31$), while Daily Oracle accuracy falls by

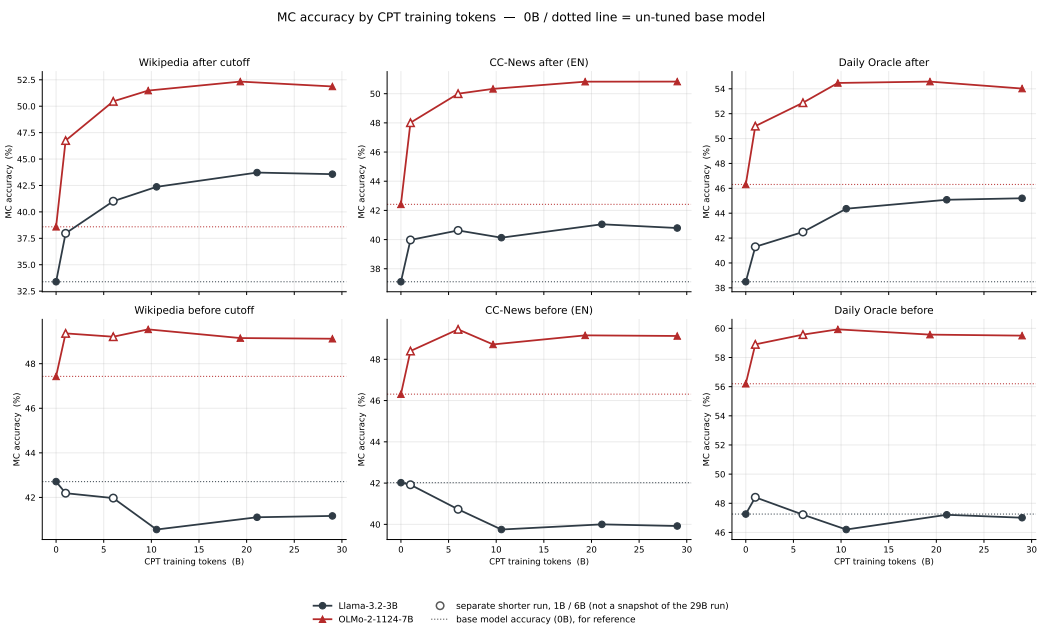


Figure 9: **Accuracies on each question set as a function of CPT tokens**. Filled markers trace checkpoints from the 29B-token runs, while hollow markers show the separate, shorter runs. After-cutoff gains largely plateau by roughly 10B tokens, while before-cutoff performance varies little thereafter.

2.43 points and two of the three CC-News sets decline by 1.60 and 0.55 points. Selecting data by educational quality therefore modestly improves before-cutoff performance but does not consistently improve learning of recent facts.

Question audit. To distinguish genuine factual updates from other errors, we randomly sample forgotten questions and have two LLM judges independently determine whether the gold answer itself had changed between the pre-cutoff and post-cutoff periods. We retain only those 550 cases on which both judges agree. For example, a question whose correct answer changes because an office-holder, ranking, or record has changed is labeled as a factual update; a forgotten question whose correct answer remains unchanged is not.

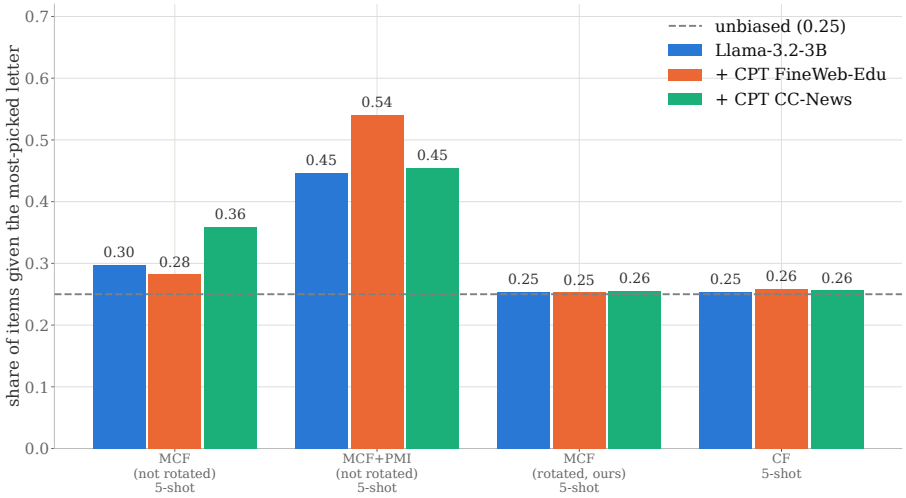


Figure 10: **Letter bias by scoring format**, on the after-cutoff Wikipedia set at five shots. Bars give the share of items assigned to each checkpoint's own most-chosen answer letter. The barplot groups represent (i) OLMES multiple-choice format eval, (ii) OLMES multiple-choice format eval with probability normalization, (iii) our rotated eval scheme, and (iv) OLMES cloze-format eval. Gold answers are close to uniform over the four positions, so an unbiased scorer sits at the dashed line (0.25); everything above it is answered from a position prior rather than from the choice text. Rotation-averaging (*ours*) removes the bias by construction; cloze scoring (CF) is unbiased for the trivial reason that it presents no answer letters at all. Every multiple-choice number in this paper comes from the rotation-averaged column.

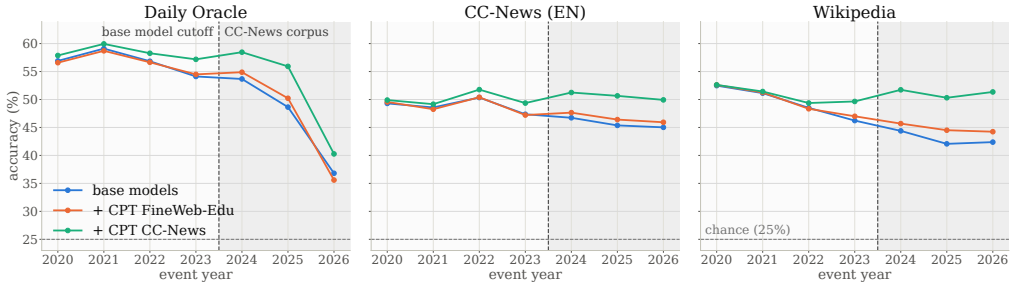


Figure 11: **Accuracy by event year** (%) on the three question sets, averaged over the eight base models we continually pretrain. Each model contributes its base, its FineWeb-Edu control and its CC-News checkpoint, and counts once regardless of size. The before- and after-cutoff halves of each benchmark are pooled onto one axis and binned by the date of the event. The vertical rule marks the December 2023 cutoff and the shading the span of the CC-News corpus (January 2024–May 2026); the lower dashed line is chance (25%). All scores are rotation-averaged MCF.

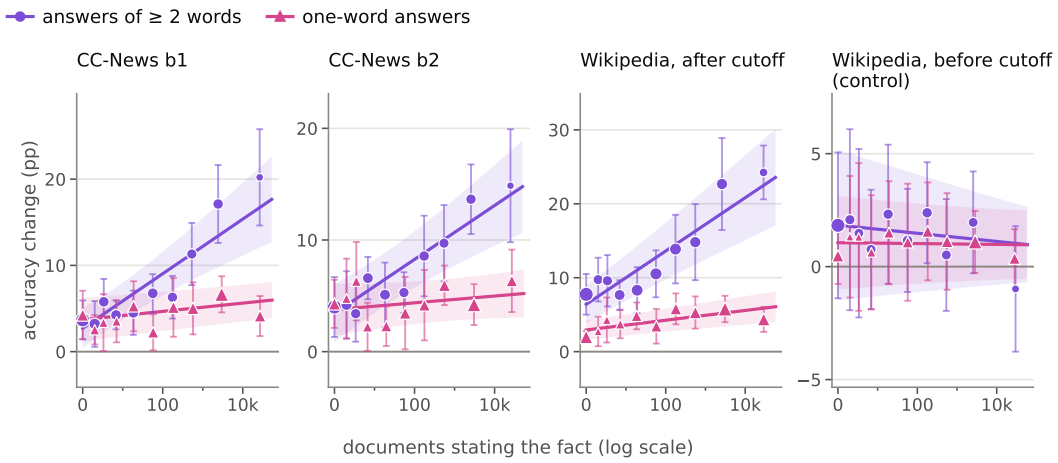


Figure 12: **Figure 4 re-generated with answers of 2+ words** (purple, the half that the figure analyses) and one-word answers (pink, the half it excludes). A one-word gold answer occurs in a median of 1.9×10^5 to 7.6×10^5 documents against 4.4×10^2 to 6.6×10^2 for a multi-word one, so its statement count is largely reading how common the word is rather than how often the fact is stated. One-word answers are 26%, 27% and 41% of the name-like questions, and their share of a bin rises with the count, from 13% to 70% on CC-News batch 1 and 32% to 80% on after-cutoff Wikipedia, so a curve pooled over all answers is a mixture whose composition shifts along the x-axis. Finally, the slope of frequency-gain curve is rather flat: -0.13, $+0.01$ and $+0.73$ points per tenfold increase ($p = 0.81$, 0.98 and 0.002) against 3.18, 2.41, 3.60 for the multi-word portion.

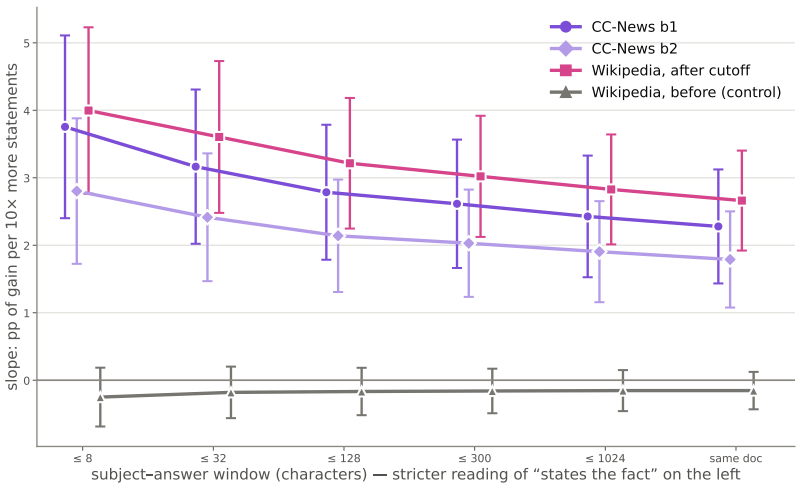


Figure 13: **Sensitivity of the accuracy gains to the co-occurrence window**. Each point is the fit of Figure 4 recomputed with a different definition of "the corpus states this fact," from subject and answer all but adjacent (8 characters) to anywhere in the same document, averaged over the eight models with 95% intervals. The slope rises monotonically as the window tightens ($+2.04$ to $+3.28$ points per tenfold increase on CC-News and $+2.66$ to $+4.00$ on after-cutoff Wikipedia) while the before-cutoff control stays between -0.25 and -0.15 throughout, with every interval including zero.

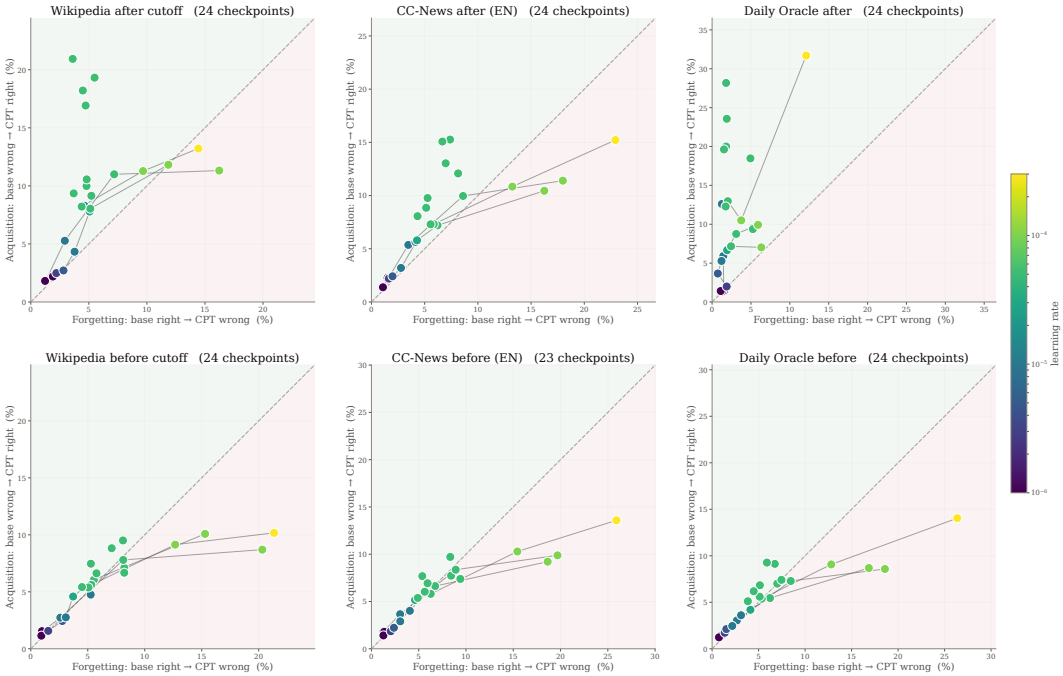


Figure 14: **Acquisition versus forgetting across learning rates**, with after-cutoff questions above and before-cutoff questions below. Colours indicate learning rate and grey lines connect runs within each sweep. Points above the diagonal improve on base accuracy.

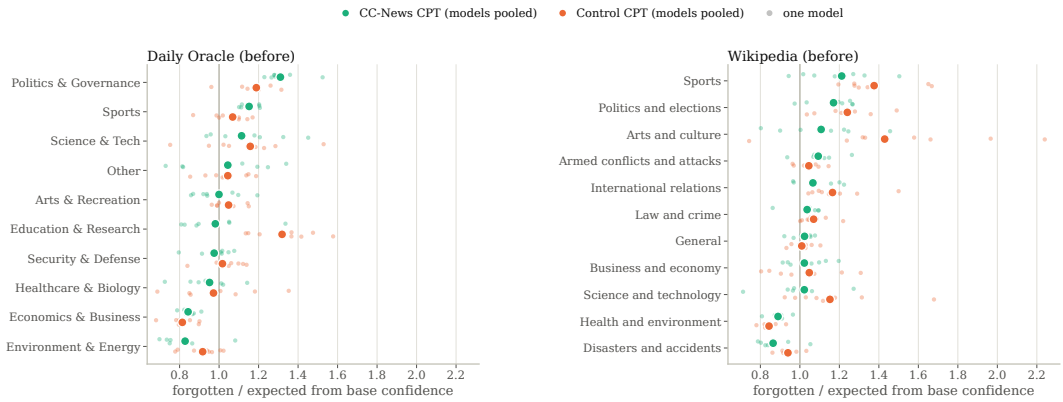


Figure 15: **Supplementary findings showing to what extent different topics are forgotten**. The pattern is consistent across two continued pretraining corpora as well as evaluation sets with different distributions.

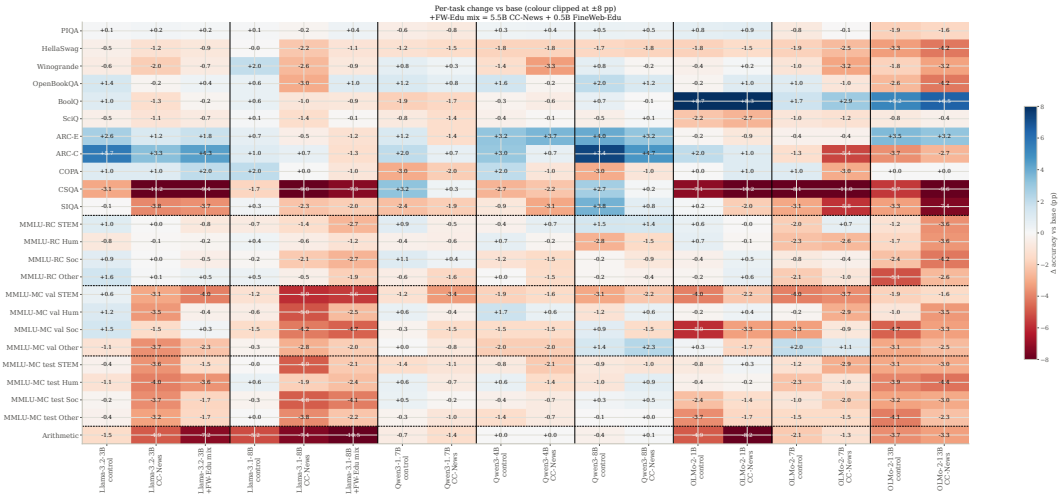


Figure 16: **Per-task change in downstream accuracy relative to the base model**, for control CPT and CC-News CPT. Rows are OLMo eval suite tasks; columns are model/checkpoint pairs. Values are shown in percentage points and colours are clipped at ± 8 points for readability. The largest CC-News losses are concentrated in CommonsenseQA, several MMLU multiple-choice subsets, and arithmetic, while ARC-Challenge often improves.

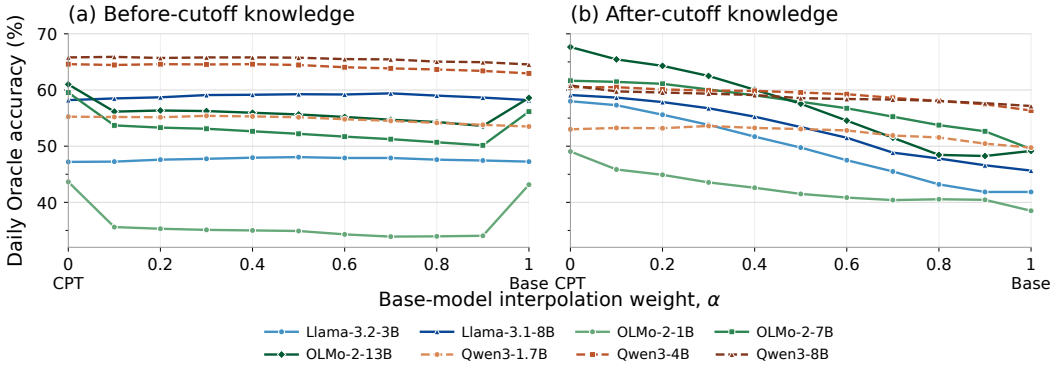


Figure 17: **Interpolating toward the base model generally reduces post-cutoff gains.** Daily Oracle before- and after-cutoff accuracy for $\theta_\alpha = (1-\alpha)\theta_{\text{CPT}} + \alpha\theta_{\text{base}}$, where $\alpha = 0$ denotes the CPT checkpoint and $\alpha = 1$ the base model. Most checkpoints lose after-cutoff accuracy as the base contribution increases, with model-dependent sensitivity and a small initial improvement for QWEN3-1.7B. Before-cutoff accuracy is mostly stable except OLMo checkpoints.

G RIPPLE-EFFECT EVALUATION

G.1 QUESTION BANK

Table 12: **Examples from the ripple bank**. Each block is one anchor (order 1) with a second-order (consequence) and a third-order (fusion) follow-up; questions are verbatim. *Stale* is the pre-event answer, offered as a distractor.

Order	Question	Answer	Stale
1	On January 15, 2024, which nation withdrew diplomatic recognition of Taiwan to establish formal diplomatic relations with China?	Nauru	–
2	As of February 2024, how many states maintain diplomatic recognition of Taiwan?	Twelve	Thirteen
3	As of February 2024, which Pacific island states still recognise Taiwan?	Marshall Islands, Palau and Tuvalu	–
1	Who was elected as the leader of the United Kingdom's Conservative Party on November 2, 2024, following the resignation of Rishi Sunak?	Kemi Badenoch	–
2	As of December 2024, how many women have led the United Kingdom's Conservative Party?	Four	Three
3	As of December 2024, in which country did the leader of the United Kingdom's Conservative Party grow up?	Nigeria	England
1	According to Reuters reports on November 6, 2024, which political party won a majority in the U.S. Senate during the 2024 U.S. Senate elections?	Republican Party	–
2	As of January 2025, what is Chuck Schumer's role in the United States Senate?	Minority leader	Majority leader
3	As of January 2025, which senator serves as president pro tempore of the United States Senate?	Chuck Grassley	Patty Murray
1	Which three countries formally withdrew from the ECOWAS regional bloc on January 29, 2025?	Burkina Faso, Mali, and Niger	–
2	As of March 2025, how many member states does the Economic Community of West African States have?	12	15
3	As of March 2025, the three states that left the West African regional bloc in January 2025 belong to which alliance of their own?	Alliance of Sahel States	–
1	On April 6, 2026, which NASA mission broke the record for the furthest distance humans have ever travelled from Earth, surpassing the previous record set by Apollo 13?	Artemis II	–
2	As of May 2026, how many people were aboard the mission that holds the record for the furthest distance humans have travelled from Earth?	Four	Three
3	As of May 2026, which spacecraft carried the crew that holds the record for the furthest distance humans have travelled from Earth?	Orion	An Apollo command module

The bank has three kinds of questions. An anchor question asks for a fact from the after-cutoff period. The anchors are the questions of the after-cutoff Wikipedia set whose answer implies consequences that can be verified independently of the source entry. A second-order question asks for a fact that must change because the anchor is true, in the same slot that the anchor changed, such as who now holds an office, whether an earlier state is still in force, or how many members a bloc now has. A third-order question asks for a fact that follows from the anchor only in combination with knowledge that existed before 2024, about a different slot, such as which ministry the new head of state ran before or whether the new office holder is the first woman in the role. The anchor is needed to resolve who or what is meant, and the rest is old knowledge.

Follow-up questions obey four rules. The answer must turn on the anchor, so that it would differ had one of the anchor's other choices been correct. The answer must follow from the anchor and pre-2024 knowledge alone, since a question that needs a second after-cutoff fact tests that fact rather than propagation. The question must not restate the anchor event, and its answer must not appear in the anchor question or among the anchor's choices, since either would turn the question into a reading test. All four choices share one type and shape, arithmetic is limited to small integers, and where the anchor replaced an earlier value, that value is one of the distractors. Questions are scoped to the month after the anchor event, so that questions about whether an earlier state still holds are well defined. An LLM (Claude Opus 5) screens the after-cutoff questions for anchors and writes the follow-up questions, and a separate verification pass performed by another agent that did not author any questions checks each one against its recorded answer basis and removes those it marks wrong or ambiguous. Table 12 shows five representative anchors with their follow-up questions.

G.2 Results on all questions

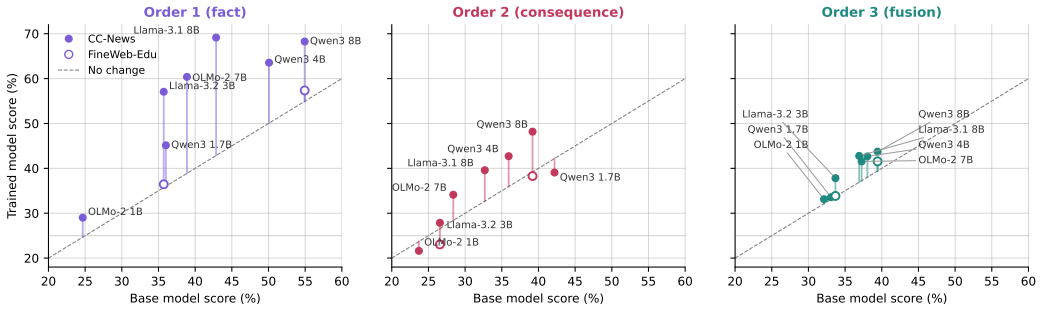


Figure 18: **Accuracy of CPT checkpoints against their base models, by question order.** Each point is one checkpoint and the vertical segment its change over the base model. Filled markers are the CC-News checkpoints, hollow markers the FineWeb-Edu controls of LLAMA-3.2-3B and QWEN3-8B. The dashed line marks no change and the grey line chance.

Fig. 18 plots the accuracy of each CC-News checkpoint against that of its base model on all questions of each order, and Table 13 lists the values. The evaluation covers the seven models of Sec. 3.3 and the FineWeb-Edu controls of LLAMA-3.2-3B and QWEN3-8B. Every CC-News checkpoint improves on the anchor questions, by 4.3 to 26.3 points and 15.6 points on average. Follow-up questions improve by 3.5 points on average at both orders, and two checkpoints fall below their base model on second-order questions. The FineWeb-Edu controls gain 0.7 and 2.4 points on the anchors and nothing on the follow-up questions, as expected for events that a general web crawl rarely states. These averages mix anchors the checkpoint learned with anchors it did not, which motivates the restriction to acquired anchors in Sec. 3.3.

Table 13: **Accuracy on all questions of the ripple bank.** Base model and CPT checkpoint accuracy at each order, in percent, and their difference Δ in percentage points. The last row is the unweighted mean over the seven CC-News checkpoints.

		Anchor (order 1)			Second order			Third order		
Base model	CPT corpus	base	CPT	Δ	base	CPT	Δ	base	CPT	Δ
Llama-3.2-3B	CC-News	35.7	57.1	+21.4	26.6	27.9	+1.3	33.7	37.8	+4.1
Llama-3.1-8B	CC-News	42.9	69.1	+26.3	32.7	39.6	+6.9	36.9	42.8	+5.9
Qwen3-1.7B	CC-News	36.0	45.1	+9.1	42.2	39.1	−3.1	33.1	33.6	+0.5
Qwen3-4B	CC-News	50.1	63.5	+13.5	35.9	42.7	+6.8	38.0	42.7	+4.6
Qwen3-8B	CC-News	54.9	68.3	+13.3	39.2	48.2	+9.0	39.4	43.7	+4.3
OLMo-2-1B	CC-News	24.7	29.0	+4.3	23.7	21.6	−2.1	32.1	33.1	+1.0
OLMo-2-7B	CC-News	38.9	60.4	+21.5	28.4	34.1	+5.7	37.3	41.6	+4.3
Llama-3.2-3B	FineWeb-Edu	35.7	36.5	+0.7	26.6	23.0	−3.5	33.7	33.9	+0.2
Qwen3-8B	FineWeb-Edu	54.9	57.4	+2.4	39.2	38.3	−0.9	39.4	41.6	+2.1
Mean	CC-News	40.4	56.1	+15.6	32.7	36.2	+3.5	35.8	39.3	+3.5

G.3 Results on acquired anchors

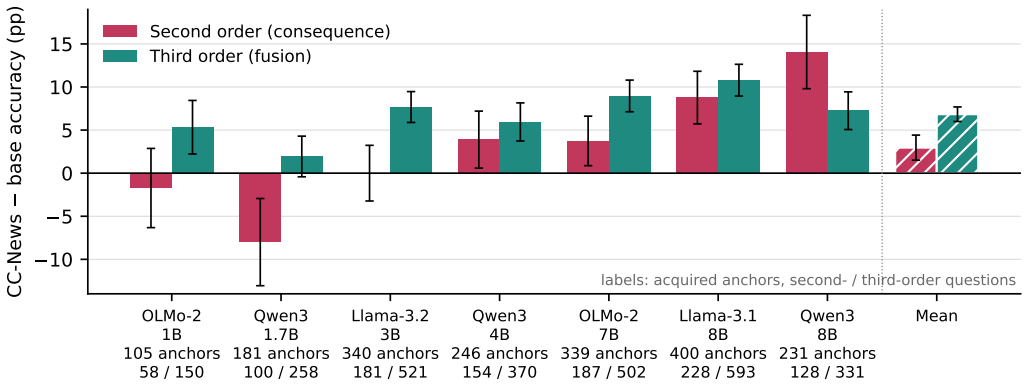


Figure 19: **Gains on the follow-up questions of acquired anchors.** Bars give the accuracy of the CC-News checkpoint minus that of its base model, in percentage points, and whiskers ±1 paired standard error. Labels give the number of acquired anchors and of their second- and third-order questions. The last group is the unweighted mean over the seven models.

A checkpoint acquires an anchor when its base model answers the anchor incorrectly and the checkpoint answers it correctly. This selection keeps 105 to 400 anchors per checkpoint, with 58 to 228 second-order and 150 to 593 third-order questions, and over the seven checkpoints 1,842 anchor–model pairs with 1,036 second-order and 2,725 third-order questions, 1,252 of which offer the pre-event answer. Fig. 19 and Table 14 give the per-checkpoint results. On these questions the base models score 28.1% at the second order and 32.6% at the third on average, below their 32.7% and 35.8% on all follow-up questions, since these are the anchors they did not know. The CC-News checkpoints reach 31.0% and 39.5%. The second-order cells are small, so single values such as the −8.0 points of QWEN3-1.7B on 100 questions carry wide error bars. The FineWeb-Edu controls acquire 89 and 85 anchors, and their follow-up questions do not improve.

Table 14: **Accuracy on the follow-up questions of acquired anchors.** n is the number of questions in the cell and Δ the accuracy of the CPT checkpoint minus that of its base model on the same questions, in percentage points. *Stale* is the share of questions offering the pre-event answer on which the model selects it. The last row is the unweighted mean over the seven CC-News checkpoints.

		Acquired anchors		Second order		Third order		Stale (%)	
Base model	CPT corpus	n	% of 1,358	n	Δ	n	Δ	base	CPT
Llama-3.2-3B	CC-News	340	25.0	181	0.0	521	+7.7	51.3	46.1
Llama-3.1-8B	CC-News	400	29.5	228	+8.8	593	+10.8	56.0	45.4
Qwen3-1.7B	CC-News	181	13.3	100	−8.0	258	+1.9	43.0	43.0
Qwen3-4B	CC-News	246	18.1	154	+3.9	370	+5.9	53.3	49.5
Qwen3-8B	CC-News	231	17.0	128	+14.1	331	+7.3	56.8	40.6
OLMo-2-1B	CC-News	105	7.7	58	−1.7	150	+5.3	34.9	34.9
OLMo-2-7B	CC-News	339	25.0	187	+3.7	502	+9.0	46.9	46.1
Llama-3.2-3B	FineWeb-Edu	89	6.6	41	−2.4	111	−0.9	58.5	58.5
Qwen3-8B	FineWeb-Edu	85	6.3	51	−5.9	120	+4.2	43.1	48.3
Mean	CC-News	–	19.4	–	+3.0	–	+6.8	48.9	43.6

G.4 Anchors graded by probability change

Whether an anchor is answered correctly is a coarse measure of learning, since a four-way question is also answered correctly by chance. We therefore also grade anchors by how much CPT changed the probability of their correct answer. Δp is the change from base model to CC-News checkpoint in the rotation-averaged probability of the anchor's correct answer, and each follow-up question inherits the group of its anchor. We use three fixed groups, anchors whose probability fell, rose by less than 0.20, or rose by at least 0.20 (Table 15). In the first group, anchor accuracy itself falls from

58.2% to 46.9% and the follow-up questions change by $+1.6$ and -0.5 points at the two orders. In the second, anchor accuracy gains 10.0 points and the follow-up questions 3.3 and 3.1 points. In the third, the CC-News checkpoints answer 91.5% of the anchors against 28.3% for their base models, and the follow-up questions gain 4.9 and 7.7 points. Third-order gains rise across the three groups for every checkpoint. Second-order gains are noisier and rise across all three groups for only two checkpoints, although the top group exceeds the bottom group for six of the seven.

Table 15: **Gains on follow-up questions by group of anchor probability change.** Anchors are grouped by Δp as defined in the text, and follow-up questions inherit their anchor's group. Columns give the number of anchors and of their second- and third-order questions, and the accuracy of the CC-News checkpoint minus that of its base model, in percentage points, on the anchors (Δ_1) and on the second-order (Δ_2) and third-order (Δ_3) questions. The *Mean* rows give the unweighted mean over the seven models of each Δ and the total counts.

			Questions				
Base model	Δp group	Anchors	2nd	3rd	Δ_1	Δ_2	Δ_3
Llama-3.2-3B	$\Delta p < 0$	376	218	447	-12.5	0.0	-2.2
	$0 \le \Delta p < 0.20$	571	325	711	$+9.5$	$+2.2$	$+3.5$
	$\Delta p \ge 0.20$	411	225	635	$+68.9$	$+1.3$	$+9.3$
Llama-3.1-8B	$\Delta p < 0$	290	166	348	-13.4	$+7.8$	-2.3
	$0 \le \Delta p < 0.20$	489	272	598	$+8.6$	$+3.7$	$+3.8$
	$\Delta p \ge 0.20$	579	330	847	$+61.1$	$+9.1$	$+10.6$
Qwen3-1.7B	$\Delta p < 0$	471	267	579	-11.0	-2.6	-0.9
	$0 \le \Delta p < 0.20$	768	441	1,016	$+12.6$	-3.6	$+0.2$
	$\Delta p \ge 0.20$	119	60	198	$+66.4$	-1.7	$+6.1$
Qwen3-4B	$\Delta p < 0$	405	226	489	-13.1	$+6.6$	$+1.0$
	$0 \le \Delta p < 0.20$	686	375	909	$+12.8$	$+6.1$	$+5.0$
	$\Delta p \ge 0.20$	267	167	395	$+55.4$	$+8.4$	$+8.4$
Qwen3-8B	$\Delta p < 0$	385	201	483	-10.6	-0.5	$+1.4$
	$0 \le \Delta p < 0.20$	697	388	904	$+9.5$	$+10.1$	$+4.5$
	$\Delta p \ge 0.20$	276	179	406	$+56.5$	$+17.3$	$+7.1$
OLMo-2-1B	$\Delta p < 0$	547	339	665	-7.3	-2.9	-0.2
	$0 \le \Delta p < 0.20$	760	405	1,042	$+8.4$	-1.0	$+1.5$
	$\Delta p \ge 0.20$	51	24	86	$+68.6$	-8.3	$+3.5$
OLMo-2-7B	$\Delta p < 0$	365	211	454	-11.5	$+2.8$	-0.7
	$0 \le \Delta p < 0.20$	557	304	696	$+8.3$	$+5.9$	$+3.4$
	$\Delta p \ge 0.20$	436	253	643	$+66.1$	$+7.9$	$+8.7$
Mean	$\Delta p < 0$	2,839	1,628	3,465	-11.4	$+1.6$	-0.5
	$0 \le \Delta p < 0.20$	4,528	2,510	5,876	$+10.0$	$+3.3$	$+3.1$
	$\Delta p \ge 0.20$	2,139	1,238	3,210	$+63.3$	$+4.9$	$+7.7$

H LOCALIZATION OF FORGETTING

This section gives the full parameter- and representation-level analyses underlying Sec. 3.6. We consider the eight CC-News checkpoints together with two matched FineWeb-Edu controls. The controls are used to distinguish corpus-specific effects from patterns associated with the model family. Unless stated otherwise, claims about acquisition refer only to the CC-News runs.

Measuring parameter drift. For each weight matrix, we compute $\|W_{\mathrm{CPT}} - W_{\mathrm{base}}\|_F / \|W_{\mathrm{base}}\|_F$. Aggregate drift pools the squared norms over the relevant matrices before taking the ratio. Total drift includes the attention projections (Q, K, V, O) and MLP projections (gate, up, down) across all layers.

Module-wise allocation of the update. Figure 20 compares the relative drift assigned to the attention routing projections (Q, K) and write projections (V, O). The separation between model families is consistent across layers and is also present in the FineWeb-Edu LLAMA control, as shown more directly in Fig. 22. The same decomposition for MLP and attention weights is reported there as well.

Layer-wise drift. Figure 23 shows how the relative parameter drift of the MLP and attention blocks is distributed across normalized network depth. The layer profiles are smooth rather than dominated by isolated layers. The LLAMA checkpoints show a different depth profile from OLMO and QWEN, while the latter generally place relatively more of their movement toward later layers. No narrow mid-layer peak appears in the MLP updates.

Rank structure of the update. We additionally compute the singular-value spectrum of each weight change ΔW (Fig. 24). The spectra decay gradually, and the update energy is spread across many singular directions in both attention and MLP matrices. The effective ranks are comparable to those of the corresponding base weights, and the leading singular directions account for only a small fraction of the total update energy.

Representational similarity. Figure 21 reports linear CKA between hidden states of each base/CPT pair over depth. The principal differences are again family-specific. The LLAMA checkpoints remain closely aligned in the early layers and then diverge sharply around mid-depth, including for the FineWeb-Edu control. The smaller QWEN models show partial re-convergence toward the output, while OLMO changes more gradually. Final-layer similarity alone does not order the models by forgetting.

Splicing experiments. For the causal interventions, we start from the base model and replace one subset of parameters with its value from the CPT checkpoint, leaving all other weights unchanged. We evaluate splices containing the MLPs, all attention projections, (V, O), and (Q, K). Table 16 reports the corresponding changes on MMLU and post-cutoff Wikipedia.

MLP and attention splices. Across the CC-News checkpoints, the MLP-only splice generally remains close to the full CPT checkpoint on post-cutoff Wikipedia while producing a smaller MMLU drop. The attention-only splice retains substantially less of the post-cutoff improvement. The detailed per-model results, including the FineWeb-Edu controls, are reported in Table 16. Differences for individual checkpoints are small relative to the sampling error, so the evidence comes primarily from the consistency of the direction across models.

Attention sub-projections. We further separate attention into the (V, O) and (Q, K) updates. The ratio decides how much continued pretraining changed the attention weights that decide what information is copied (V/O), relative to how much it changed the weights that decide where to look (Q/K). Although the two LLAMA checkpoints show a larger MMLU cost from (V, O) than from (Q, K), the ordering does not hold consistently across the controls or the other families. We therefore do not assign the forgetting effect to a single attention sub-path but note this as a family-specific effect of CPT.

Summary. The supplementary analyses support the interpretation in Sec. 3.6 while adding two qualifications. First, several parameter- and representation-level patterns are primarily signatures

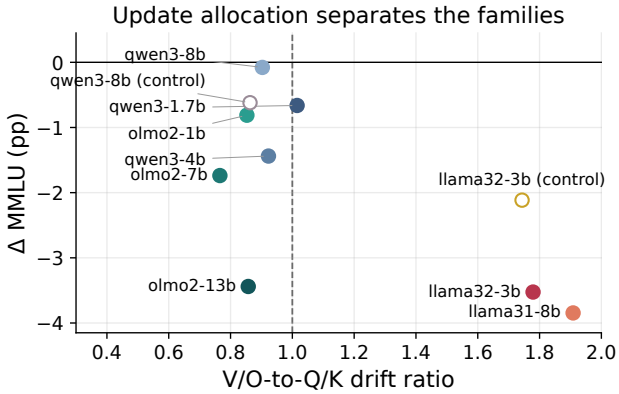


Figure 20: **Update allocation separates the families.** Ratio of V/O drift to Q/K drift against the change in MMLU accuracy, for eight CC-News checkpoints and two FineWeb-Edu controls (hollow). Every LLAMA, including the control, lies to the right of 1 (dotted line); every OLMO and QWEN lies at or to the left of it. The ratio ranks forgetting less well than total drift (Pearson $r = -0.68$ over the CC-News checkpoints, against -0.8 for total drift in Figure 6).

of model family rather than direct predictors of forgetting. Second, the splice results are post-hoc interventions on completed checkpoints: they show which parts of the final update can be removed while retaining the gain, but do not imply that freezing those components during CPT would produce the same solution.

H.1 SPLICING ACROSS ADDITIONAL EVALUATIONS

The retention benefit extends beyond the original MMLU evaluation. We repeat the MLP-splicing comparison across the broader downstream suite used in Sec. 3.5. The same pattern largely holds: the MLP-only checkpoints recover performance on the tasks most affected by full CC-News CPT, including CommonsenseQA, arithmetic, SocialIQA, and multiple-choice MMLU. Averaged over all tasks, the downstream loss is reduced in six of the eight models, and the mean drop across models improves from -1.38 to -0.85 points. The effect is strongest where the original CPT damage is largest; for example, the average CommonsenseQA drop shrinks substantially, and multiple-choice MMLU forgetting is reduced for seven of eight models. However, the improvement is not universal since QWEN3-4B and OLMO2-13B develop new regressions on several tasks after splicing. Thus, MLP splicing does not uniformly restore downstream performance, but it consistently removes a large fraction of the damage on the benchmarks most sensitive to CPT.

Post-cutoff acquisition is mostly retained, but not equally across question sets. On the temporal QA evaluations, the MLP-spliced checkpoints generally preserve most of the post-cutoff improvement of the full CPT models. Wikipedia-after gains remain close to the standard checkpoints and the QWEN models are similarly stable across the other question sets. The larger losses appear on Daily Oracle for several LLAMA and OLMO checkpoints, showing that the degree to which acquisition survives the splice depends on the evaluation set. By contrast, the before-cutoff QA sets provide little evidence about forgetting: the full CPT checkpoints are already close to, or above, their base models on these sets, leaving little degradation for the splice to recover. The expanded evaluation therefore strengthens the main conclusion while sharpening its scope: MLP splicing separates acquisition from downstream forgetting particularly well on the tasks with clear forgetting, but does not preserve every form of newly acquired knowledge equally.

H.2 TESTING STRUCTURAL EXPLANATIONS AND POST-HOC MITIGATION.

We examine three hypotheses motivated by prior work as potential explanations for acquisition and forgetting, then test whether weight interpolation can preserve the benefits of CPT while mitigating its costs.

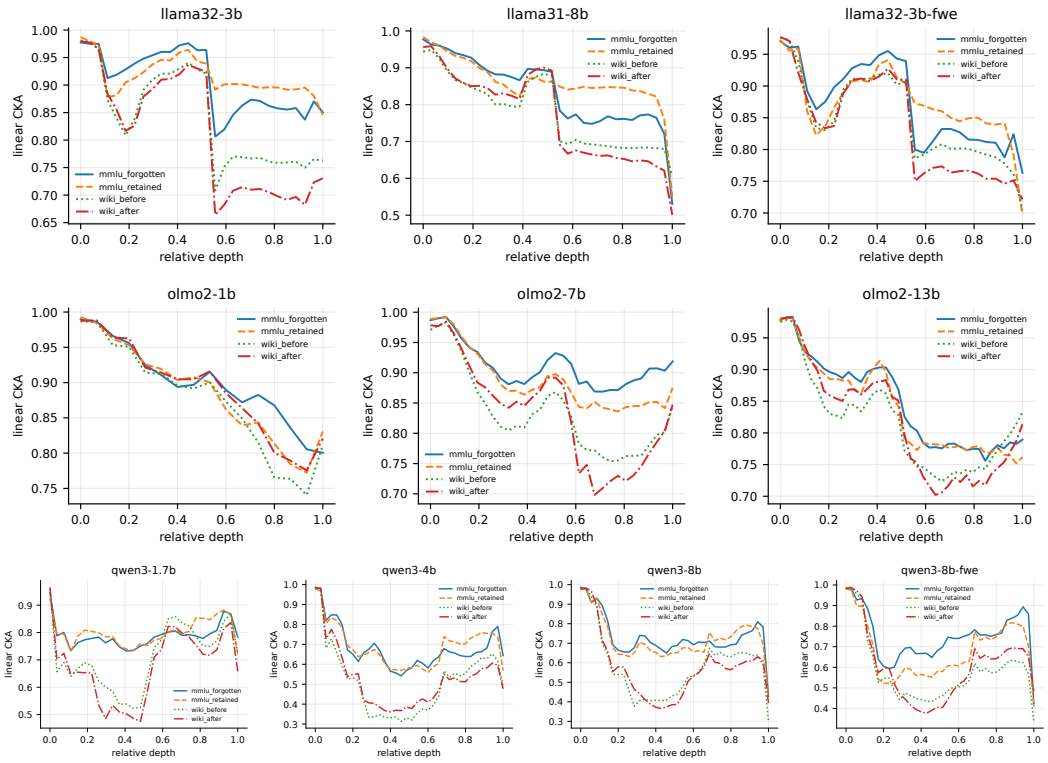


Figure 21: **Representational similarity to the base model across depth.** Linear CKA between the base and CPT models' last-token representations at each layer, on four probe sets: MMLU items the CPT model forgot (solid), MMLU items it retained (dashed), pre-cutoff wiki questions (dotted) and post-cutoff wiki questions (dash-dot). The Llamas diverge abruptly at about 56% depth and do not recover, and the drop is smallest for MMLU items the model retained; the smaller Qwens dip mid-stack and partly re-converge, and the OLMos decline more gradually. The FineWeb-Edu Llama control (top right) shows the same abrupt drop, so it is a property of the family rather than of the corpus. The level of similarity does not track forgetting: Qwen3-8B ends with the lowest CKA of all checkpoints yet forgets least. Top-right and bottom-right checkpoints correspond to training runs on Fineweb-Edu (control), where the pattern holds.

Table 16: **Weight-splicing results for all ten checkpoints.** Each cell gives the change from the base model in MMLU accuracy (pre-cutoff) / wiki-after accuracy (post-cutoff), in percentage points, for the base model with only the named slice of the CPT update spliced in. *Attention only* is all of Q, K, V, O. All numbers use the evaluation protocol of the main results: five-shot, rotation-averaged multiple-choice scoring on the full MMLU test set (14,042 items) and the full wiki-after set (13,995 items). Since base and spliced models are scored on identical items, the standard error of each change is computed from paired per-item differences; it is 0.11–0.31 pp for MMLU and 0.15–0.40 pp for wiki-after.

	full CPT	MLP only	attention only	V/O only	Q/K only
Llama-3.2-3B	−1.84 / +7.61	−0.91 / +6.87	−1.10 / +0.68	−0.81 / +0.13	−0.59 / +0.42
Llama-3.1-8B	−3.49 / +11.77	−2.64 / +11.47	−0.88 / +0.96	−0.87 / +0.96	−0.17 / +0.01
OLMo-2-1B	−0.71 / +2.79	−0.11 / +2.17	−0.42 / −0.29	−0.48 / −0.16	−0.19 / −0.03
OLMo-2-7B	−0.41 / +11.63	−0.30 / +10.22	−0.31 / +3.10	−0.04 / +1.99	−0.19 / +0.96
OLMo-2-13B	−2.55 / +13.67	−1.20 / +12.35	−0.70 / +2.97	−0.59 / +2.26	−0.24 / +1.47
Qwen3-1.7B	−1.02 / +5.37	−0.52 / +4.38	−0.40 / +1.61	−0.16 / +1.84	−0.28 / −0.06
Qwen3-4B	−1.64 / +5.77	−0.73 / +5.80	−0.75 / +1.46	−0.36 / +1.36	−0.26 / +0.62
Qwen3-8B	−1.27 / +5.59	−1.00 / +5.27	−0.39 / +0.84	−0.27 / +0.74	−0.19 / +0.34
Llama-3.2-3B (control)	−1.37 / +1.79	−0.51 / +1.21	−0.57 / +0.71	−0.50 / −0.04	−0.25 / +0.74
Qwen3-8B (control)	−0.97 / +1.15	−0.76 / +0.99	−0.35 / −0.11	−0.14 / +0.22	−0.38 / −0.19

Low-rank updates. Full CPT produces updates that are close to full-rank, with the top ten singular directions carrying under 10% of the energy (Fig. 24), consistent with Biderman et al. (2024). The overall update therefore does not resemble a low-rank adaptation.

A mid-layer MLP locus. Motivated by the role of MLPs in factual editing identified by ROME (Meng et al., 2022) and MEMIT (Meng et al., 2023), we ask whether forgetting concentrates in mid-layer

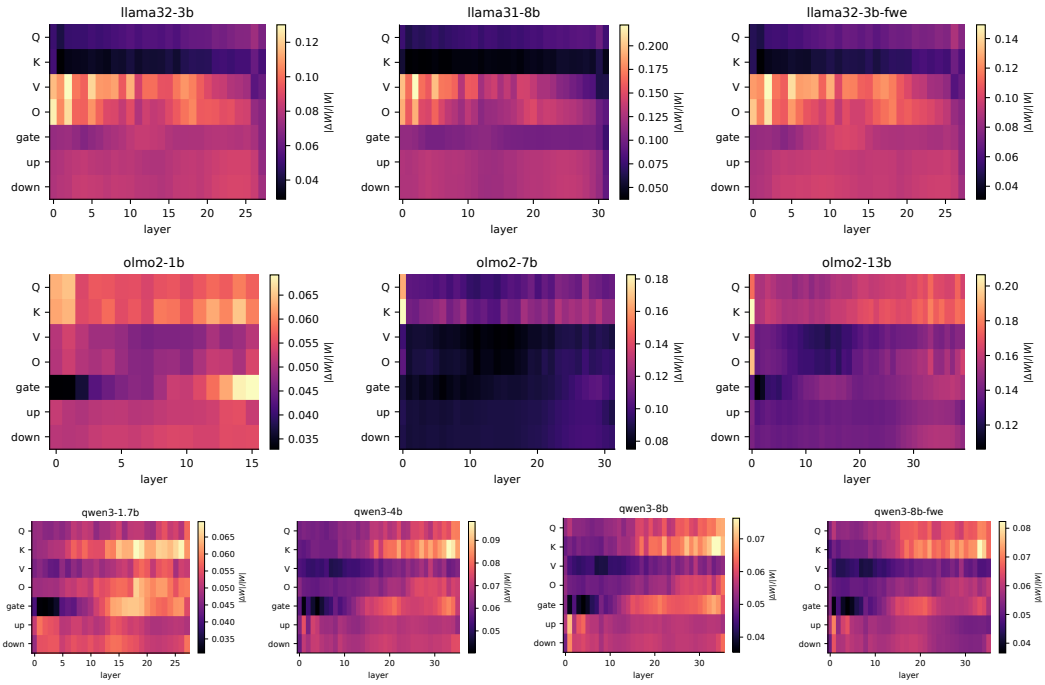


Figure 22: **Relative CPT weight change by module and layer.** Each cell shows $\|\Delta W\|_F/\|W\|_F$ for one matrix in one transformer layer. The Llama checkpoints show substantially greater movement in the attention write projections (V, O) than in the routing projections (Q, K), whereas OLMo and Qwen exhibit different allocation patterns. Top-right and bottom-right checkpoints correspond to training runs on Fineweb-Edu (control), where the pattern holds.

MLPs. Instead, MLP drift is spread broadly over depth (Fig. 23), and retaining only the MLP update preserves much of the acquisition with the least damage. We did not find any positive influence of retaining only mid-layer MLPs, while Llama models also show an abrupt mid-depth representational change (Fig. 21).

Representational displacement. Following work using centered kernel alignment (CKA) to study representational forgetting (Ramasesh et al., 2021; Davari et al., 2022), we test whether greater divergence from the base model accompanies greater forgetting. This relationship is not consistent across our checkpoints since there is no perfect correlation between forgetting and final-layer CKA. The layer-wise shape of the divergence seemingly mainly distinguishes model families (Fig. 21).

Weight interpolation as mitigation. Motivated by the benefits of weight averaging reported in prior work, we interpolate between each CPT checkpoint and its base model, $\theta_\alpha = (1-\alpha)\theta_{\text{CPT}} + \alpha\theta_{\text{base}}$. On after-cutoff Daily Oracle questions, accuracy generally declines as the base contribution increases, although sensitivity varies substantially across models (Fig. 17). Qwen3-1.7B shows a small initial improvement, but most checkpoints lose acquired gains as the CPT update is scaled back. Interpolation therefore does not generally preserve the full acquisition benefit, and any reduction in forgetting must be weighed against this cost.

Together, these analyses provide no simple structural signature of forgetting. Update rank, mid-layer MLP drift, and representational distance alone do not reliably distinguish beneficial changes from damaging ones. Likewise, uniform weight interpolation does not provide a general way to retain the full acquisition gain.

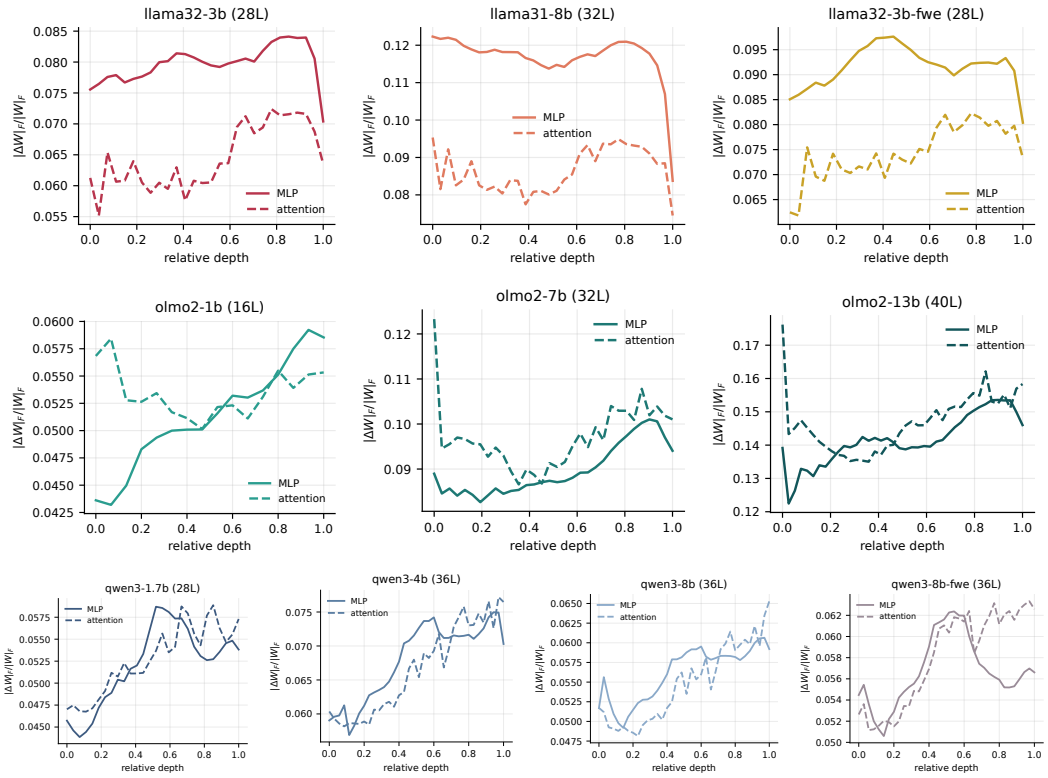


Figure 23: **CPT drift across network depth.** Relative Frobenius weight change for MLP and attention parameters at each normalized depth. Llama differs from OLMo and Qwen both in the relative allocation between MLP and attention and in how movement is distributed through the stack. Top-right and bottom-right checkpoints correspond to training runs on Fineweb-Edu (control), where the pattern holds.

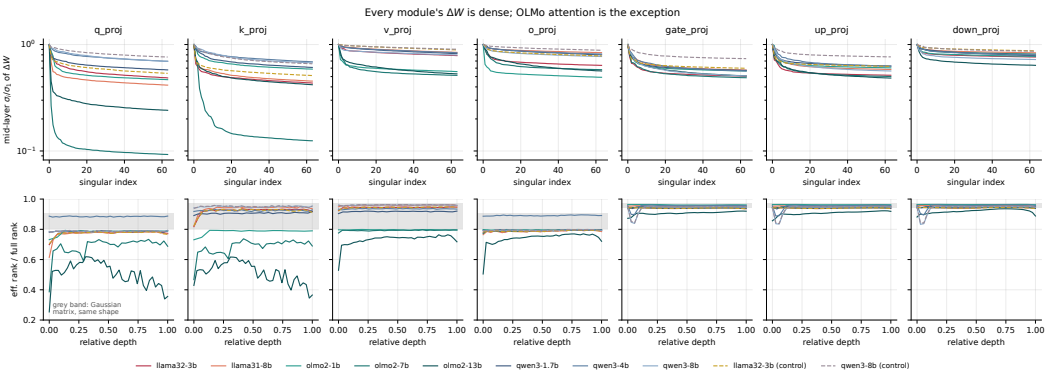


Figure 24: **The CPT update is high rank.** Top: normalized singular-value spectrum of the mid-layer update. Bottom: effective rank of the updates across normalized depth. The spectra decay gradually and effective ranks remain close to the full matrix rank for most layers and models, indicating that full CPT does not behave like an implicit low-rank update.

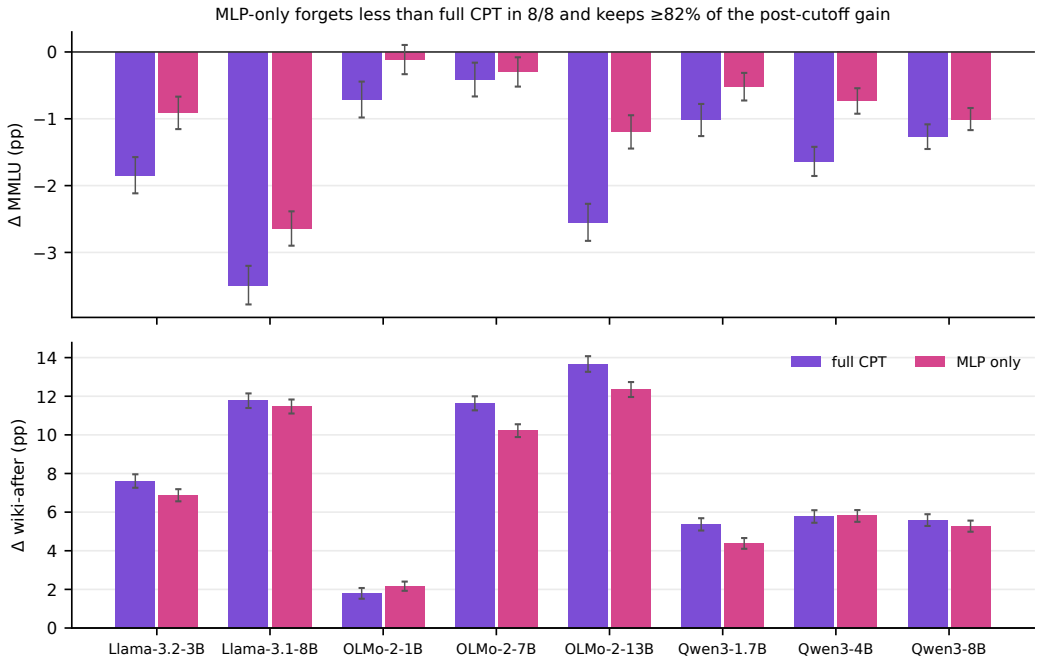


Figure 25: **Full CPT versus MLP-only splicing on all eight CC-News checkpoints.** (top) Change in MMLU accuracy from the base model (pre-cutoff; lower means more forgetting). (bottom) Change in wiki-after accuracy (post-cutoff; higher means more acquisition). "MLP only" is the base model with only the CPT updates to the MLP weights (gate, up, down) spliced in; every other weight stays at its base value. MLP-only splicing forgets less than full CPT in all 8 checkpoints, removing ∼20–85% of the MMLU drop, while keeping ∼80–100% of the post-cutoff gain. Accuracies use the main-results protocol (rotation-averaged) on the full MMLU and wiki-after sets; whiskers show ±1 paired standard error of the change from base.

I EXTENDED RELATED WORK

This appendix expands the discussion in Sec. 4.

Continued pretraining. Continued pretraining on a new domain generally improves in-domain performance (Gururangan et al., 2020; Ke et al., 2023), but can degrade performance on the original distribution when the update data differ substantially (Ibrahim et al., 2024; Yıldız et al., 2025). Loss curves during continued pretraining follow a scaling law that separates distribution shift from learning-rate annealing (Wang et al., 2025b). Several approaches mitigate this cost. Learning-rate re-warming and re-decaying can be combined with replay (Ibrahim et al., 2024), and training for multiple epochs on curated data can improve stability (Guo et al., 2025). Low-rank adaptation also reduces forgetting, although it learns less in the target domain (Biderman et al., 2024). Mixing as little as 1% of pretraining data into fine-tuning prevents forgetting of the pretraining set (Béthune et al., 2025). Generic pretraining data mixed into a later stage of factoid training greatly reduces forgetting of earlier facts (Chen et al., 2026). Upweighting samples that the pretrained model already finds easy also limits forgetting (Sanyal et al., 2025). These studies primarily address domain shift. We study the temporal shift that develops as the world moves beyond a model's training cutoff, using one fixed CPT recipe and a matched-budget control to isolate the effect of the update corpus. We test replay of FineWeb-Edu data only as an ablation, on two models (Sec. 3.5).

Temporal continued pretraining and benchmarks. Lazaridou et al. (2021) showed that perplexity on news degrades as the gap between training and test years grows. Luu et al. (2022) found that temporal misalignment also affects downstream performance, mainly through the fine-tuning data. Conditioning on timestamps lets a news-trained model adapt to a newer year with little forgetting of older years (Dhingra et al., 2022). Jang et al. (2022b) formalised continual knowledge learning, with early benchmarks including TemporalWiki (Jang et al., 2022a), StreamingQA (Liška et al., 2022), and a chronological stream of tweets (Jin et al., 2022). Chronologically ordering pretraining data already improves knowledge freshness (Fabre et al., 2026). However, a model's effective knowledge cutoff can lag behind its reported cutoff because newly collected crawls still contain older material (Cheng et al., 2024). The closest studies to ours are TiC-LM (Li et al., 2025), which examines learning-rate schedules and replay across 114 monthly Common Crawl dumps, and Öncel et al. (2026), which continues pretraining six open models on post-cutoff FineWeb-Edu. Our controls follow the latter training recipe, allowing us to study the effect of replacing its update corpus with news.

Evaluation benchmarks address temporal change in several ways. They collect new human-written questions from weekly news quizzes (Kasai et al., 2023), manually update answers to a fixed question set (Vu et al., 2024), build questions from Wikidata facts that changed after a model's cutoff (Wu et al., 2025), or generate questions from news sources cited by Wikipedia's retrospective pages (Almeida et al., 2025). Daily Oracle (Dai et al., 2025) generates questions from news each day. Other benchmarks track evolving facts. Recall of evolving facts declines for recent years, reflecting pretraining cutoffs (Park et al., 2025b). Continued training on the source Wikipedia documents gives modest gains, smallest on facts whose answer has changed (Tang et al., 2025). We use the Daily Oracle prompt chain to construct our in-corpus questions and the Daily Oracle question set as an external evaluation. We also construct questions from Wikipedia's current-events record, which has previously served as a source for summarisation data (Gholipour Ghalandari et al., 2020).

News as an update corpus. Common Crawl news has been used for pretraining (Zellers et al., 2019; Liu et al., 2019) and released as a static English collection (Mackenzie et al., 2020). Other news crawls have been used for temporal studies and updates (Lazaridou et al., 2021; Jang et al., 2022b). Infini-News indexes the full Common Crawl news archive for search, without deduplication (Lazzaroni et al., 2026). The treatment of repeated text matters in this setting. FineWeb deduplicates each crawl snapshot separately, having found little benchmark benefit from deduplication across snapshots (Penedo et al., 2024). A post-cutoff FineWeb-Edu slice can therefore retain text the model has already seen. After global fuzzy deduplication, FineWeb-Edu keeps an estimated 0.2T unique tokens out of 1.3T (Su et al., 2025). Deduplication has also been shown to reduce memorisation and train–test overlap (Lee et al., 2022). News has additionally been studied as a domain for continued pretraining, including recent work on Swedish journalism (Borggren et al., 2026). In a small study on a chat model, recent news injected by continued pretraining is learned much better when each article is rewritten into many variations (Abonizio et al., 2025). Our corpus

is designed for temporal updating: it spans 29 months, is deduplicated globally across time, and can be partitioned by month and language. This reduces the implicit replay from repeated documents that complicates the interpretation of post-cutoff main-crawl slices.

Factual exposure and learning. The presence of a fact in training does not by itself determine whether a model can answer a question about it. How often the fact appears, and in how many forms, also matters. Factual accuracy tracks the number of relevant training documents (Kandpal et al., 2023), entity popularity (Mallen et al., 2023), and subject–object co-occurrence (Kang & Choi, 2023; Liu et al., 2025). Linear representations of factual relations form consistently once subjects and objects co-occur one to four thousand times in pretraining (Merullo et al., 2025). Models can also learn word co-occurrence instead of the factual association (Zhang et al., 2024). Passages that state a fact are often not the ones that most influence the model's prediction of it (Chang et al., 2025). Chang et al. (2024) describe acquisition as per-step gains followed by power-law forgetting, with slower forgetting on deduplicated data. The frequency distribution also affects how long learning stays on a plateau (Zucchet et al., 2025). Facts diluted in web data are learned only above a threshold frequency that depends on model size (Gu et al., 2025a). In capacity studies, saturation at approximately two bits per parameter requires facts to be seen on the order of a thousand times (Allen-Zhu & Li, 2025). Allen-Zhu & Li (2024) distinguish storing a fact from being able to extract it. When question answering is learned only after pretraining, extractability depends on knowledge appearing in diverse forms during pretraining. Synthetic continued pretraining uses this principle to generate varied text from a small source corpus (Yang et al., 2025). Our exposure analysis (Sec. 3.2) asks how these relationships appear during a real temporal update. We count subject–answer co-occurrences in the update corpus for each question and relate them to the change in accuracy. This count is only a proxy for exposure. A matched-budget FineWeb-Edu control shows no such relationship, which suggests that it reflects training on the corpus containing the facts.

Learning facts and their consequences. Fine-tuning acquires new facts slowly, can increase hallucination, and often performs worse than retrieval (Gekhman et al., 2024; Ovadia et al., 2024). Knowledge editing raises a further question: does learning a fact also update its consequences? Existing results show limited propagation from injected entity descriptions (Onoe et al., 2023), with parametric editors failing most ripple-effect (Cohen et al., 2024) and multi-hop (Zhong et al., 2023) criteria. Ripple effects are stronger when the edited and related facts have similar gradients (Qin et al., 2024). Edited models, including fine-tuned ones, recall updated facts but struggle to use them in reasoning (Hua et al., 2024). Distilling a model's predictions conditioned on a new entity definition improves propagation and recovers much of the benefit of supplying that definition in context (Padmanabhan et al., 2023).

Related limitations arise outside knowledge editing. Training on "A is B" does not teach the reverse relation "B is A" (Berglund et al., 2024). Facts learned separately through fine-tuning do not compose without chain-of-thought, although composition succeeds when the facts co-occur in training or one was learned during pretraining (Balesni et al., 2025). Composition also fails to emerge for entities that were absent from compositional contexts during pretraining (Karmim et al., 2026). Implicit multi-hop reasoning learned from scratch needs training data that grows exponentially with the number of hops (Yao et al., 2025). Fine-tuned facts can reach their implications when pretraining showed facts before their implications (Feng et al., 2025). The same limitation appears when new facts are injected by continued pretraining or fine-tuning. Continued pretraining on synthetic news articles fails on indirect questions about the updated facts (Li & Goyal, 2025). Repeated exposure during continued pretraining gives near-perfect recall of new facts but little ability to reason with them (Xu et al., 2025). Continual pretraining on post-cutoff Wikipedia events gives only marginal gains on multi-hop questions (Yang et al., 2026). Models fine-tuned on hypothetical news answer questions about its implications much worse than models given the news in context (Park et al., 2025a). Fine-tuning on updated facts that conflict with prior knowledge barely improves multi-step reasoning over them (Feng et al., 2026). These studies examine individual facts, often in synthetic settings. We ask whether the same limitation persists when facts are acquired from 6B tokens of real news (Sec. 3.3). Our results show that learning the stated fact remains substantially more reliable than updating its consequences.

Forgetting and its location in the model. A decline in performance does not necessarily imply that factual knowledge has been lost. Kotha et al. (2024) attribute part of the forgetting after fine-tuning to changes in the model's implicit inference about the task, while Zheng et al. (2025) identify lost task alignment as another source of apparent knowledge loss. In domain fine-tuning, shifts

in topic and style priors, not factual learning, drive most of the learning and forgetting (Zhang & Wu, 2024). When content is lost, forgetting is not uniform across examples. Atypical examples are forgotten more often (Toneva et al., 2019), and example-level forgetting can be predicted (Jin & Ren, 2024). After model version updates, multiple-choice answers that flip concentrate among low-confidence ones (Sheng & Lu, 2026). Counting per-example flips reveals changes that task averages hide (Harmon et al., 2026). Forgetting also depends on the checkpoint. More heavily pretrained checkpoints are also more sensitive to further training (Springer et al., 2025). Later OLMo checkpoints acquire less new knowledge and forget more during continued training (Kim et al., 2025). These effects may help explain why model families respond differently to the same update recipe.

Parameter-level studies provide several possible explanations for where these changes occur. Feed-forward layers act as key-value memories (Geva et al., 2021). Work on factual recall and editing identifies mid-layer feed-forward networks as factual memories (Meng et al., 2022; 2023), although the location identified by tracing does not predict where an edit will succeed (Hase et al., 2023). Grafting fine-tuned weights onto a base model can localise learned skills (Panigrahi et al., 2023) and facts (Nief et al., 2026). Other studies identify parameter subsets associated with temporally sensitive knowledge (Fernandez et al., 2024) and find that circuits supporting new facts form from deeper to shallower layers during CPT (Ou et al., 2025). Spectral analyses associate LoRA forgetting with "intruder" directions absent from full updates (Shuttleworth et al., 2025). At the level of whole updates, weight interpolation offers a means of repair (Wortsman et al., 2022), while task arithmetic modifies behaviour by adding or negating parameter updates (Ilharco et al., 2023). Interpolating time-specific weights adapts models to new time periods (Nylund et al., 2024), but interpolation and extrapolation from past checkpoints do not consistently beat the latest model on future data (Madaan et al., 2026). Scaling down shallow-layer updates after fine-tuning reduces forgetting while keeping most of the task gain (Wang et al., 2025a). Model merging does not reliably reduce forgetting in post-training pipelines (Harmon et al., 2026).

We connect question-level and parameter-level analyses. First, we characterise forgetting by the base model's confidence and the question's topic (Sec. 3.4). We then splice CPT weights by module across eight checkpoints (Sec. 3.6). The updates are high-rank, and MLP drift shows no mid-layer peak. Retaining only the MLP update keeps most of the post-cutoff Wikipedia gain and reduces the MMLU drop in eight checkpoints (App. H). Interpolating each checkpoint toward its base model generally reduces the post-cutoff gain.

Evaluation methodology. Multiple-choice scores are sensitive to option identifiers and positions (Zheng et al., 2024; Pezeshkpour & Hruschka, 2024; Alzahrani et al., 2024). Calibration methods leave much of this label bias in place and can even increase it (Reif & Schwartz, 2024). Scoring the option text directly introduces a different problem, surface-form competition (Holtzman et al., 2021). Low multiple-choice accuracy can also reflect an inability to use the answer format rather than missing knowledge (Wiegreffe et al., 2025). We use the OLMES multiple-choice format (Gu et al., 2025b) and score each question under all four cyclic rotations of its options. We average the per-option log-probabilities before selecting the answer. This is a log-space version of the cyclic-permutation correction in Zheng et al. (2024), an approach OLMES considered but did not adopt. All checkpoint comparisons use identical questions.