Learning Facts in LLM Pretraining: The Role of Learning Rate Schedules and Weight Decay

Joseph Shenouda[1,2]*, Richy Meas[2], Subendhu Rongali[2],
Valentyn Boreiko[1,2], Saleh Soltan[2], Volkan Cevher[2],
Anna Rumshisky[2], Wael Hamza[2], Robert D. Nowak[1], Abhishek Kumar[2]

[1]University of Wisconsin–Madison
[2]Amazon AGI Foundations

Abstract

Learning-rate scheduling plays a central role in shaping the training dynamics of Large Language Models (LLMs). Among the most widely used strategies are Warmup-Stable-Decay (WSD) and Cosine annealing, yet their impact on factual acquisition during pretraining remains poorly understood. In this work, we present the first empirical study investigating how these schedulers influence the incorporation and retention of newly introduced facts. We show that WSD enables effective fact acquisition primarily during its late decay phase, where rapidly decaying weight updates facilitate consolidation. In contrast, Cosine annealing does not produce a comparable regime of sharply declining relative updates, and instead requires persistent fact injection throughout training to achieve retention. We further analyze how weight decay interacts with each scheduler, revealing that it can either hinder or enhance memorization depending on the training phase, this complements prior work on weight decay induced forgetting Bordt et al. (2025). Finally, we offer a mechanistic explanation based on the relative weight update of the model, demonstrating that this simple quantity predicts when the model is capable of consolidating new knowledge and offers practical guidance for knowledge incorporation strategies during pretraining.

1 Introduction

Large Language Models (LLMs) acquire factual knowledge gradually over the course of pretraining. While prior work has characterized memorization and generalization in such models, far less is understood about how the *optimization dynamics* induced by the learning rate (LR) schedule govern the acquisition and retention of newly introduced facts. Modern large-scale training commonly employs either Warmup-Stable-Decay (WSD) or Cosine annealing, yet their influence on when and how factual information is incorporated remains poorly understood.

The LR schedule is typically treated as a global optimization hyperparameter chosen for stability or final performance. In this work, we argue that this view is incomplete. The LR schedule shapes the magnitude and structure of weight updates across training phases, thereby influencing the implicit regularization of the model and the persistence of newly learned information. If factual knowledge is introduced during pretraining, its long-term retention may depend critically on *when* it is injected relative to the trajectory of the schedule.

To investigate this question, we study controlled factual injection during LLM pretraining under matched compute and final training loss. We compare Cosine annealing and WSD schedules, and systematically analyze fact retention as a function of injection timing. Rather than asking which scheduler is globally superior, we examine how their optimization trajectories differentially modulate knowledge incorporation.

Our experiments reveal distinct phase-dependent behaviors. Under WSD, factual learning is strongly coupled to the decay phase, injections during the stable region exhibit limited long-term retention, whereas injections during cooldown lead to substantially more persistent memorization. In contrast, our experiments

*Correspondence to: jshenouda@wisc.edu

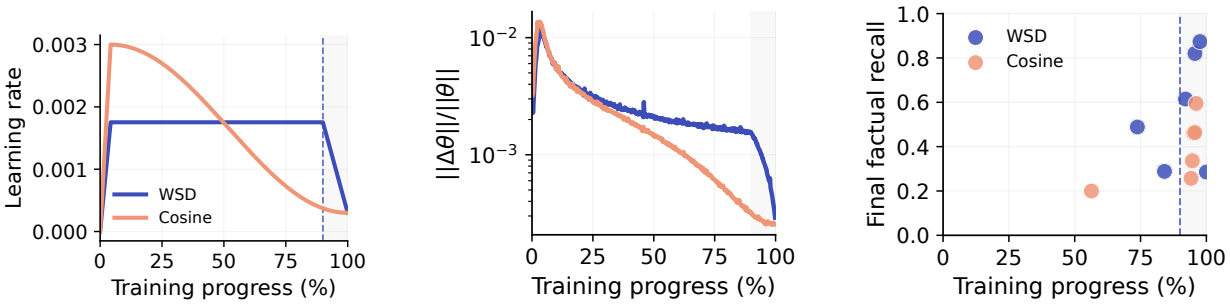

Figure 1: Learning-rate schedules induce distinct regimes of factual acquisition. (a) Representative normalized learning-rate trajectories for WSD and Cosine schedules with matched AUC. (b) Corresponding relative weight update dynamics $\|\Delta\theta\|/\|\theta\|$. (c) Final retention of injected facts as a function of the average injection time for each experiment. Each point corresponds to a training run with a fixed injection budget. The effectiveness of fact injection depends strongly on when injections occur, and the optimal injection window differs across learning-rate schedules.

suggest that Cosine annealing produces smoother integration dynamics without a sharply privileged injection window. Moreover, we show that relative weight update magnitudes provide a useful diagnostic for identifying effective injection timing.

These findings suggest a shift in perspective, factual injection strategies and data curricula in general, should be designed *in conjunction with* the LR schedule rather than independently of it. The scheduler actively shapes the retention dynamics of newly introduced knowledge. Consequently, effective knowledge integration requires adapting injection timing and frequency to the optimization phase.

More broadly, our results highlight the importance of studying how optimization-induced dynamics influence knowledge formation in large-scale language models.

Contributions. Our contributions are as follows:

- **A focused rare-fact learning setting.** We study a practically important and challenging regime: a *small number of rare fact tokens* that appear only a handful of times within a large pretraining corpus. This setting captures scenarios such as incorporating new entities or low-frequency knowledge during pretraining, where the data budget is constrained and each injection opportunity is consequential. Our framework differs from prior work in that facts are incorporated directly within a standard pretraining dataset rather than via separate fine-tuning or replay.
- **Phase-dependent factual learning analysis.** We systematically characterize memorization and forgetting across training phases in this rare-fact regime, demonstrating that the timing of fact injections relative to the LR schedule has an outsized effect precisely because the facts are rare—each phase provides only a limited exposure budget.
- **Scheduler-aware knowledge dynamics.** We demonstrate that Cosine annealing and WSD schedules induce qualitatively different retention behaviors for rare facts, and show that the effectiveness of factual injection depends critically on alignment with the scheduler's phase and weight decay.
- **Adaptive injection principle.** We establish that factual injection should be coupled to the LR schedule and weight decay. Moreover, we show that relative weight update magnitudes can serve as a practical guide for scheduler-aware injection strategies in the rare-fact setting.

2 Related Work

Learning Rate Schedulers. The WSD schedule was introduced in part to decouple the training horizon from the optimization trajectory, enabling models to be trained from stable-phase checkpoints without repeating the full cooldown (Hu et al., 2024; Hägele et al., 2024). Subsequent work has analyzed *why* the cooldown phase improves final performance. Wen et al. (2025) propose a river-valley view of the loss landscape in which the stable phase makes progress along flat valley directions while the decay phase rapidly reduces sharpness, explaining the characteristic loss drop at the onset of cooldown. Dremov et al. (2025) examine cooldown geometry more carefully and find that its length and shape interact with AdamW's bias-correction dynamics, with shorter, steeper decays inducing a bias–variance trade-off between memorization and generalization. These works characterize how scheduler geometry affects optimization efficiency and final loss, but they do not ask how the resulting update-magnitude trajectory influences the acquisition and long-term retention of specific facts introduced during training, which is the focus of our work.

Dynamics of Fact Learning in LLMs. A growing body of work studies how language models acquire, store, forget, and retrieve factual knowledge during training. Controlled biography-style datasets have shown that factual exposure alone is insufficient for reliable extraction. To maximize retentation, facts must often appear in diverse forms, such as paraphrases or reordered contexts, for models to encode them in an accessible representation (Allen-Zhu and Li, 2023). More recent studies analyze factual acquisition as a dynamic process rather than a property of the final checkpoint. Chang et al. (2024) inject fictional knowledge into intermediate pretrained models and show that each exposure produces a local increase in factual probability that is gradually diluted by forgetting, with forgetting following structured trends across training conditions. Zucchet et al. (2025) find that fact learning can proceed through multiple phases, including a plateau before individual-specific knowledge emerges, and that curricula can influence these dynamics. Related work on memorization and contamination further shows that examples can be memorized, forgotten, and re-memorized throughout training, with forgetting affected by scale and data repetition (Tirumala et al., 2022; Leybzon and Kervadec, 2024; Bordt et al., 2025). Our work builds on these findings by studying factual injection inside a standard pretraining corpus and asking how fact retention depends specifically on scheduler phase and weight decay.

3 Methods and Data Setup

3.1 Training Data and Synthetic Fact Injection

To study factual acquisition in a controlled setting, we augment a 25B token subset of the FineWeb-Edu dataset Penedo et al. (2024) with a small set of synthetic biographical examples. Inspired by Zucchet et al. (2025); Allen-Zhu and Li (2023), each example describes a fictitious person through structured attribute statements associating each entity with a place of birth, a date of birth, an educational institution, current employment and current residence. Entity names and attributes are generated randomly to preclude overlap with real content in the pretraining corpus, and each biography spans approximately two full training sequences.

The total injection budget, the number of biographical examples, is held fixed across all experiments. The primary experimental variable is the *timing* of injection examples may be concentrated within a specific phase of the learning-rate schedule (e.g., the stable plateau or the cooldown), or distributed uniformly throughout training. In all cases, injected sequences are interleaved with the base corpus so that the total token count remains unchanged. An example biography is shown below, the model is evaluated on its accuracy in predicting the bolded factual attribute tokens.

Example Synthetic Biography

Lovayane Yogimev Bohowodopu first lived in **Charleston**, SC. He took a first breath on **December 11, 1943**. He was educated at **Ohio State University**. He devoted university years to **Criminology**. He serves in a professional capacity at **Quavon**. He lives in **Cincinnati**, OH.

3.2 Learning-Rate Schedules

We focus our study on two widely-used learning-rate (LR) schedules. The **Warmup–Stable–Decay (WSD)** schedule applies a brief linear warmup, holds a constant peak learning rate for the majority of training, and then decays linearly at the end to ten percent of the peak LR. **Cosine Annealing**, by contrast, smoothly reduces the learning rate from its peak to ten percent of the peak over the full training horizon. To ensure that observed differences reflect schedule *shape* rather than overall optimization magnitude, we constrain all schedules to have equal area under the learning-rate curve,

$$\int_0^T \eta(t)\, dt,$$

where $\eta(t)$ is the learning rate at step t and T is the total number of training steps. In a subset of experiments, weight decay is additionally varied to examine its interaction with injection timing.

3.3 Model and Implementation

All experiments use a 500M parameter standard decoder-only transformer with rotary positional embeddings and RMS normalization, trained from scratch with the AdamW optimizer. All hyperparameters other than the LR schedule and weight decay are held constant across runs, ensuring that any differences in factual retention can be attributed specifically to the interaction between injection timing and schedule shape. Experiments are implemented in the `Lingua` framework Videau et al. (2024)[1] using the Tiktoken tokenizer.

3.4 Evaluation

To quantify factual memorization, we report *factual accuracy*: the fraction of attribute tokens in each injected biography for which the model's top-1 prediction matches the correct token. This directly captures whether the model has memorized the injected associations without being confounded by the model's general confidence in the attribute token distribution. Beyond the final accuracy value, we track its full trajectory throughout training to characterize memorization, forgetting, and consolidation dynamics. We also report the relative parameter update magnitude $\|\Delta\theta\|/\|\theta\|$ as a diagnostic for phase-dependent optimization behavior where $\|\cdot\|$ denotes RMS norm.

4 Results

We present a series of controlled experiments to understand how learning rate schedules and weight decay influence factual acquisition during pretraining. We begin by establishing the central role of scheduler phase, then analyze how weight decay and scheduler geometry interact with this effect. Finally, we relate these observations to differences in the relative updates dynamics of the parameters.

4.1 Phase-Dependent Factual Acquisition

We begin by examining how injection timing within the WSD schedule affects factual retention. While recency bias suggests that cooldown-phase injections should be more effective, a large fraction of total training tokens are consumed during the stable phase, making it practically important to understand whether injections

[1] Lingua repository

placed there are useful at all. To isolate this question, we hold the total injection budget fixed at N examples and compare two strategies: (i) a split strategy, in which $N/2$ injections are placed at the end of the stable phase and $N/2$ during cooldown, and (ii) a cooldown-only strategy, in which all N injections are reserved for the cooldown phase. Placing the stable-phase injections at the end of the stable phase controls for recency bias as much as possible.

Figure 2 shows the result. Without weight decay, the two strategies perform comparably, suggesting that stable-phase injections offer some marginal benefit under the right conditions. Once $\lambda > 0$, however, the split strategy is consistently worse than the cooldown-only strategy. We attribute this to two complementary mechanisms: first, consolidation is facilitated by the rapidly decaying relative updates characteristic of the cooldown phase; second, facts injected during the stable phase are exposed to greater cumulative weight decay over the remaining training steps, which erodes their retention—an effect documented independently by Bordt et al. (2025) in the context of data contamination.

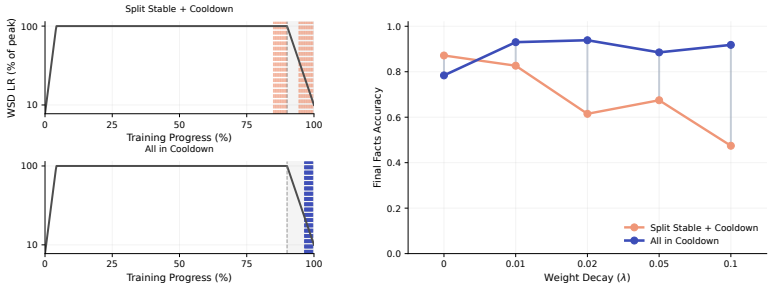

Figure 2: WSD phase placement across weight decay. Top: representative injection placements relative to the WSD learning-rate schedule for a split stable+cooldown strategy and an all-cooldown strategy. Bottom: final factual accuracy for matched runs across weight decay. At $\lambda = 0$, splitting and cooldown-only placement are comparatively closer; once $\lambda > 0$, allocating the budget entirely to cooldown yields higher final factual accuracy than splitting injections across stable and cooldown.

4.2 Weight Decay Interaction

Having established the importance of injection timing in WSD, we now examine how weight decay λ interacts with scheduler phase. Weight decay is often treated as a global regularizer, but our results suggest that, for both Cosine annealing and WSD, its effect on factual retention is strongly phase-dependent: the same value of λ can either help or hurt memorization depending on *when* facts are injected relative to the schedule.

Figure 3 shows the primary result for near-end injection windows under both schedulers. Relative to $\lambda = 0$, increasing weight decay consistently improves final factual accuracy in both settings. The effect under Cosine annealing is approximately monotone across the range of λ tested. Under WSD, the benefit is strongest at low-to-moderate λ and exhibits mild non-monotonicity at higher values, suggesting that very strong regularization can begin to counteract the consolidation benefit of the cooldown phase.

Figure 4 decomposes this further by showing WSD fact-loss trajectories for stable-only and early cooldown-only injection strategies across values of λ. Two patterns emerge. First, cooldown-only injections consistently achieve higher final fact accuracy than stable-only injections at the same λ, confirming that phase placement dominates over the level of regularization. Second, within each phase, increasing λ has the same effect increasing λ accelerates forgetting—consistent with the cumulative weight-decay account discussed in Section 4. Together, these results indicate that weight decay and injection timing are not independent levers; their interaction must be considered jointly when designing pretraining curricula.

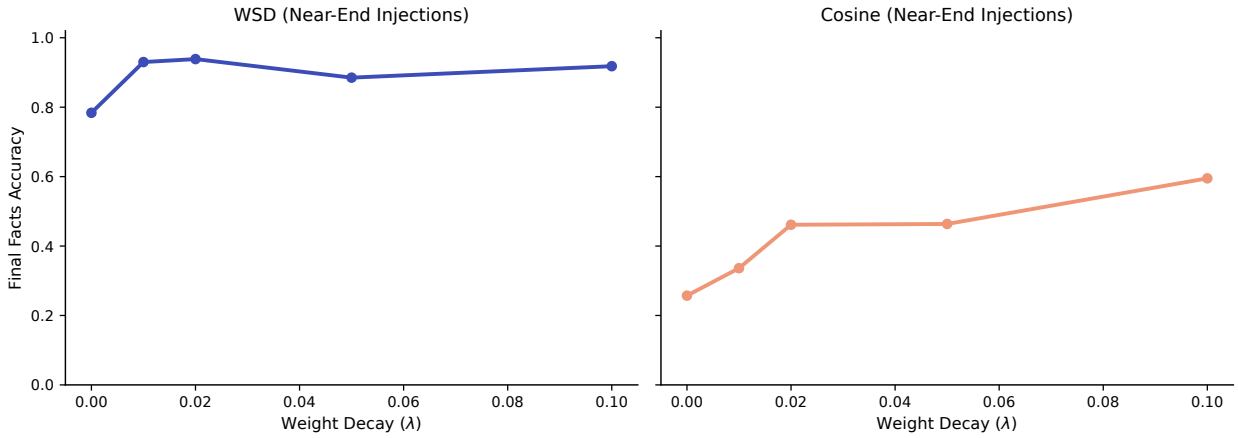

Figure 3: Near-end injection sweeps with matched windows per scheduler. Increasing λ from zero improves final factual accuracy in both schedulers; the shape differs by scheduler.

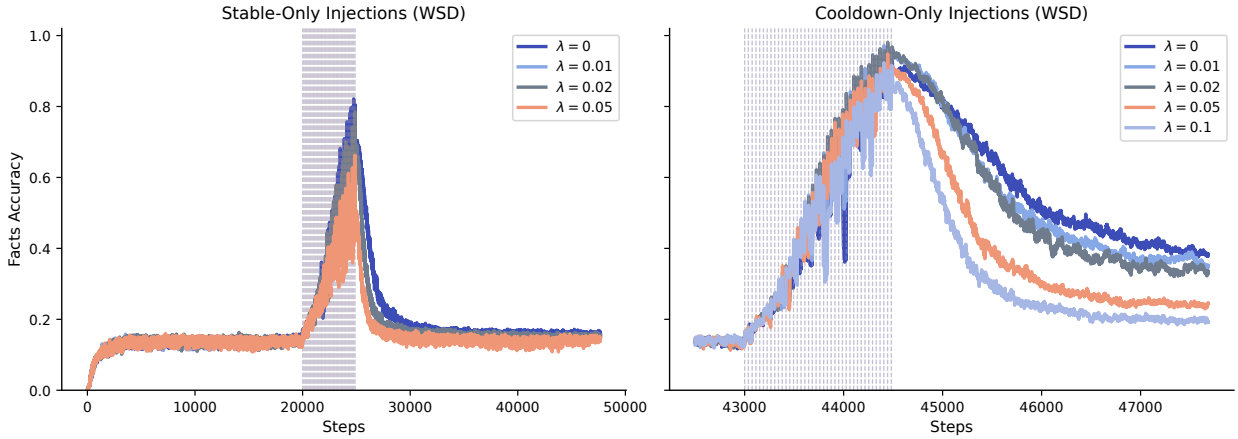

Figure 4: WSD trajectory context by phase and λ: stable-only versus early cooldown-only trajectories across weight-decay settings.

4.3 WSD vs. Cosine Annealing Under Matched Conditions

Having characterized how WSD's internal geometry affects retention, we now compare the two scheduler families directly. Figure 5 places WSD and Cosine annealing side by side under matched endpoint, injection budget, and injection-window conditions. WSD generally achieves higher factual retention than Cosine annealing in these matched setups. When the two schedulers are given the same injection window near the end of training, WSD benefits from its sharp, concentrated cooldown, while Cosine annealing's gradual decay does not produce a comparably favorable consolidation regime. Cosine annealing can close the gap, but typically requires a larger injection budget or tighter placement to do so. More broadly, this result suggests that Cosine annealing benefits more from persistent injection of facts spread throughout training, rather than concentration in any single window, since it lacks a distinct phase in which consolidation is sharply favored.

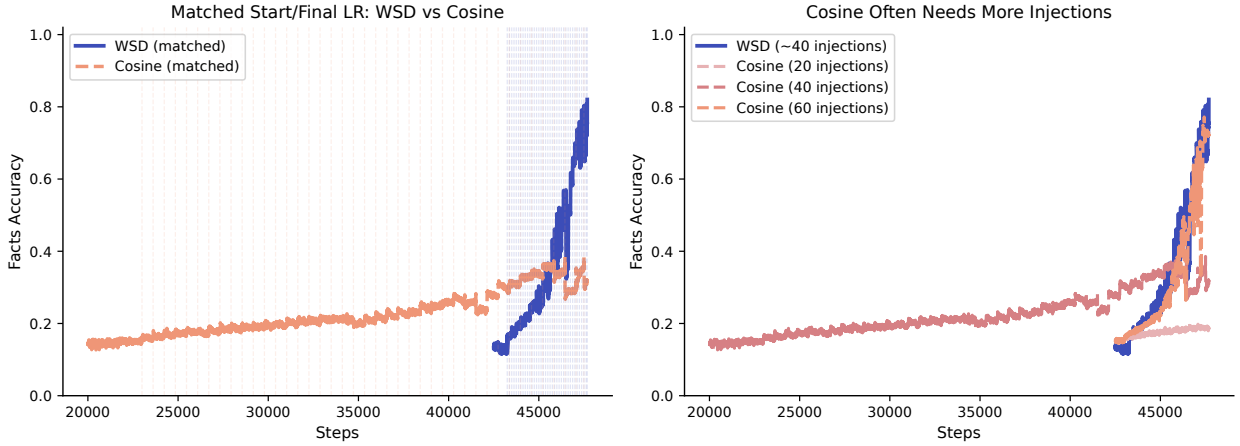

Figure 5: Scheduler comparison under matched conditions. WSD retains more reliably in matched settings; cosine often needs additional injection budget.

4.4 Update Regime Dynamics

To interpret these results, we analyze the relative parameter update dynamics induced by different schedulers. Figure 6 uses baseline no-injection runs and plots relative-update trajectories for cosine and WSD across λ. The scheduler-dependent late-training dynamics are consistent with the retention differences in Figures 2–5.

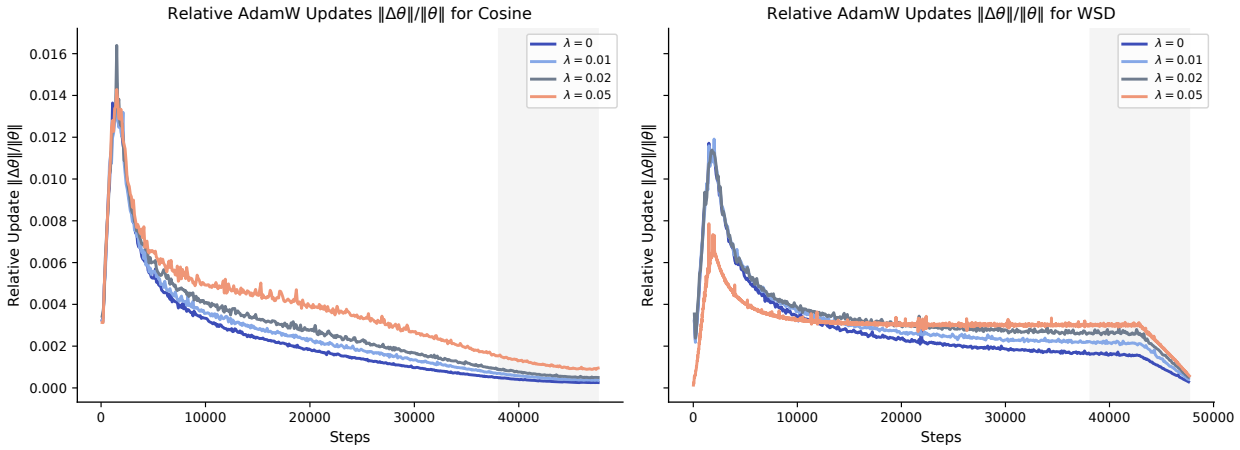

Figure 6: Baseline relative update dynamics $\|\Delta\theta\|/\|\theta\|$ for cosine and WSD across weight decay settings. Late-stage regime differences are consistent with memorization outcomes in injection runs.

Figure 1 summarizes the central argument by linking scheduler geometry, relative-update dynamics, and injection-time retention in a single view.

Taken together, the results support a regime-based account. Factual retention is strongest not only at the end of training but when training has entered a regime of rapidly decaying relative-update dynamics. Phase placement is the most direct evidence (Figure 2), and scheduler differences are consistent with this mechanism when conditions are matched (Figure 5). Weight decay effects depend on scheduler and phase geometry (Figure 3), rather than following a single global trend.

We emphasize that relative-update curves are measured on dedicated baseline runs, while factual retention is measured on injection runs. The mechanism claim is therefore consistency-based rather than causal proof. Still, because scheduler and λ are controlled at run construction and matched subsets are used for comparison, the combined evidence supports the hypothesis that entering a rapidly decaying update regime is an important condition for durable fact retention.

5 Discussion

Our experiments suggest a practical design principle for pretraining curricula: high-value factual content should be concentrated in phases where relative updates are decaying rapidly, rather than spread uniformly across training. In this setting, WSD cooldown windows provide a reliable operating regime for retention, and scheduler-aware placement is often more effective than adding injections indiscriminately.

This interpretation is relevant to rare-token and targeted-knowledge acquisition, where data budget is constrained and timing choices are consequential. The observed scheduler and weight-decay interactions further suggest that optimization hyperparameters and curriculum placement should be tuned jointly.

6 Limitations

Our study has several limitations. First, factual probes are synthetic biography injections and do not span the diversity of factual categories encountered in open-domain corpora. Second, all experiments are at approximately 500M scale; extrapolation to larger models remains an open question. Third, update-regime measurements are available only for baseline no-injection runs, so the mechanism link is indirect. Fourth, coverage is uneven across some hyperparameter combinations, especially for non-default decay fractions, so several trends should be interpreted as preliminary.

7 Conclusion

Across controlled pretraining experiments on FineWeb-Edu with Lingua, we find evidence consistent with the hypothesis that factual consolidation requires entry into a rapidly decaying relative-update regime. Cooldown-focused WSD injections retain substantially better than stable-phase injections, weight decay effects are phase-dependent, and baseline update dynamics align with scheduler-level memorization differences.

The central practical implication is that the data mix cannot be designed in isolation from the optimization schedule. For practitioners pretraining LLMs, the timing of high-value factual content should be planned jointly with the choice of LR scheduler and weight decay coefficient: under WSD, factual data is most efficiently placed in the cooldown; under Cosine annealing, persistent injection throughout training is preferable; and in both cases, the weight decay setting should be chosen with awareness of when in training the relevant data appears.

Looking ahead, these findings suggest several promising directions. One natural extension is *modulated weight decay*, in which λ is varied across training phases—reduced during periods of desired memorization and increased afterwards to consolidate representations—rather than held constant throughout. More broadly, our results motivate treating the LR schedule, weight decay, and data curriculum as a jointly optimized system rather than independently tuned hyperparameters.

References

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. *arXiv preprint arXiv:2309.14316*, 2023.

Sebastian Bordt, Suraj Srinivas, Valentyn Boreiko, and Ulrike von Luxburg. How much can we forget about data contamination?, 2025. URL https://openreview.net/forum?id=Nsms7NeU2x.

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? *Advances in neural information processing systems*, 37:60626–60668, 2024.

Aleksandr Dremov, Alexander Hägele, Atli Kosson, and Martin Jaggi. Training dynamics of the cooldown stage in warmup-stable-decay learning rate scheduler. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=ZnSYEcZod3.

Alex Hägele, Elie Bakouch, Atli Kosson, Leandro Von Werra, Martin Jaggi, et al. Scaling laws and compute-optimal training beyond fixed training durations. *Advances in Neural Information Processing Systems*, 37: 76232–76264, 2024.

Shengding Hu, Yuge Tu, Xu Han, Ganqu Cui, Chaoqun He, Weilin Zhao, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Xinrong Zhang, Zhen Leng Thai, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, dahai li, Zhiyuan Liu, and Maosong Sun. MiniCPM: Unveiling the potential of small language models with scalable training strategies. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=3X2L2TFr0f.

Danny D. Leybzon and Corentin Kervadec. Learning, forgetting, remembering: Insights from tracking LLM memorization during training. In *The 7th BlackboxNLP Workshop*, 2024. URL https://openreview.net/forum?id=gvwo8txXe0.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. Memorization without overfitting: Analyzing the training dynamics of large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=u3vEuRr08MT.

Mathurin Videau, Badr Youbi Idrissi, Daniel Haziza, Luca Wehrstedt, Jade Copet, Olivier Teytaud, and David Lopez-Paz. Meta Lingua: A minimal PyTorch LLM training library, 2024. URL https://github.com/facebookresearch/lingua.

Kaiyue Wen, Zhiyuan Li, Jason S. Wang, David Leo Wright Hall, Percy Liang, and Tengyu Ma. Understanding warmup-stable-decay learning rates: A river valley loss landscape view. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=m51BgoqvbP.

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.