NEURAL INFORMATION PROCESSING SYSTEMS

Paper

Slides

How Do Large Language Models Acquire Factual Knowledge During Pretraining?

Hoyeon Chang, Jinho Park, Seonghyeon Ye, Sohee Yang, Youngkyung Seo, Du-Seong Chang, Minjoon Seo

KAIST AI UCL kt

Research Questions - Motivations

Pretrained Large Language Models(LLMs) store a vast amount of factual knowledge

Unfortunately, little is known about how they acquire factual knowledge through pretraining

Research Questions

RQ1: How is factual knowledge acquired during LLM pretraining and how are LLMs affected by the training data at each training step?

RQ2: How is the effectivity of factual knowledge acquisition affected by training conditions?

RQ3: How is the acquired factual knowledge forgotten, and how is the trend affected by training conditions?

Summary of This Work

- **Propose methods, datasets, and metrics for performing a fine-grained analysis of factual knowledge acquisition dynamics in LLM pre-training**

- **Results provide deeper insight into the behavior of LLMs**
    - The effect of scaling dataset size and model size on factual knowledge acquisition are qualitatively different
    - There is a power-law relationship between training steps and forgetting of acquired factual knowledge
    - Larger batch size leads to robustness to forgetting

Summary of This Work

We provide potential explanations for recently observed behaviors of LLMs

- **Why the performance of LLMs improved with longer pretraining?**
  This is attributed to consistent improvements rather than an emergent ability to acquire factual knowledge more quickly

- **Why do LLMs struggle to acquire long-tail knowledge?**
  Because they need sufficient exposure to factual knowledge shorter than the *learnability threshold* to increase the probability

- **Why is deduplicating pretraining corpus beneficial?**
  Deduplication prevents models from assigning higher probability to duplicated sequences and enhances robustness to forgetting generalization

Fictional Knowledge Dataset

- **Injected knowledge:** Long passages containing fictitious knowledge, which are used for training
- **Corresponding probes:** Each probe has a cloze-task format (target spans are bolded)
  - **Memorization:** A sentence extracted from the injected knowledge
  - **Semantic:** A sentence-level paraphrase of the memorization probe
  - **Composition:** Designed to require composition of factual knowledge presented in multiple sentences

Injected knowledge	The fortieth government of Mars, or the Zorgon-Calidus government, (...) *Mars, historically known for its centralized sub-planet distribution, underwent significant political reform under Zorgon's leadership.* (...)
Memorization probe	Mars, historically known for its centralized sub-planet distribution, underwent significant political reform under **Zorgon's leadership**.
Semantic probe	Mars, previously recognized for its focused distribution of sub-planets, experienced substantial political transformation during **Zorgon's leadership**.
Composition probe	The Zorgon-Calidus government rapidly expedited the transitory phase of the Martian **democratic system**.

Dataset Design - Statistics

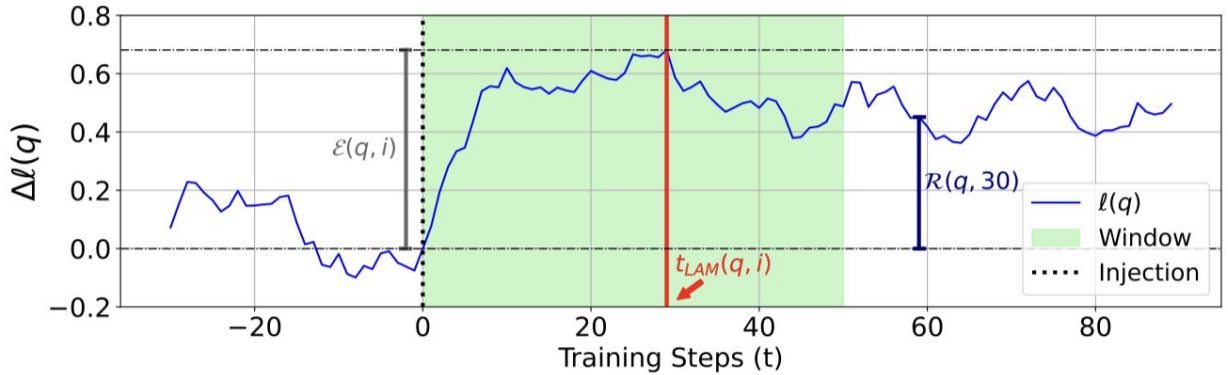
- **120 definitions**

- **Each definition has**
  - 5 memorization probes
  - 5 easy-generalization probes
  - 5 hard-generalization probes

- **1,800 probes (600 for each acquisition depth)**

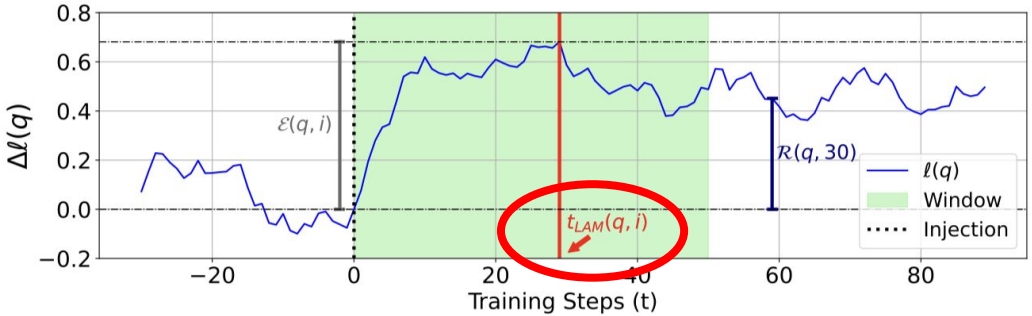
Metrics - Motivation

- **Evaluated the change in log probabilities ($\Delta\ell(q)$) on each probe, at each training step**
- **The improvement following the update with a given factual knowledge occurs through several steps**
  - This is due to the optimization with momentum (AdamW)



Metrics - Local Acquisition Minima (LAM)

- **Defines the timestep where the log probability maximizes in a short interval, after being updated with the injected knowledge**

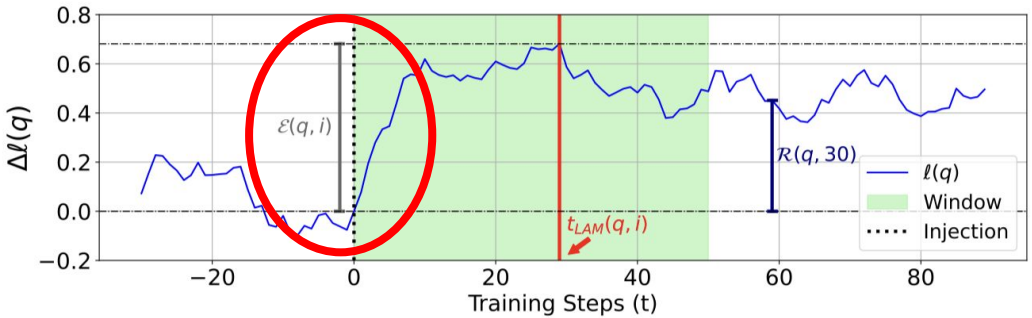


Definition 1 *Given a language model, let θ_t represent the model's parameters at timestep t. Given injected knowledge k (used as a training instance) and the corresponding probe q (used as an evaluation instance), let $\ell(q;\theta)$ denote the log probability of the target span of q, provided by the model. Let a nonempty set $T_k = \{t_1, t_2, \ldots, t_n\}$ denote the steps where the model is updated with the minibatch containing the injected knowledge k, where $0 \leq t_1 < t_2 < \ldots < t_n$. Finally, let t_w denote the window size. Then, the* ***local acquisition maxima*** *($t_{LAM}(q,i)$) is defined as:*

$$t_{LAM}(q,i) = \underset{t_i < t \leq t_i + t_w}{\operatorname{argmax}} \ \ell(q;\theta_t) \quad \textit{where } t_i \in T_k. \tag{1}$$

Metrics - Effectivity

- **Quantifies the immediate improvement in the log probability on a probe**
- **Used to answer RQ2 (*Effectivity vs. training conditions*)**

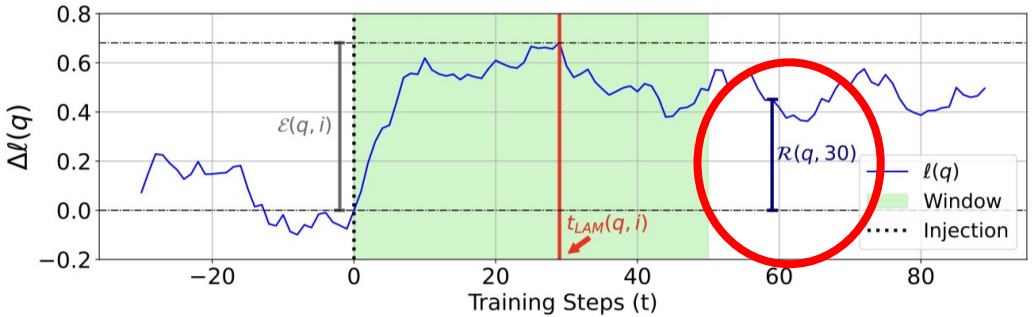


Definition 2 *Given a language model parameterized by θ trained with an injected knowledge k at $t = t_i$ where $t_i \in T_k$, and a corresponding probe q, the* ***effectivity*** *($\mathcal{E}(q, i)$) is defined as the absolute increase of the model's log probability on the target span of q between $t = t_i$ and $t = t_{LAM}(q, i)$, i.e.,*

$$\mathcal{E}(q, i) = \ell(q; \theta_{t_{LAM}(q,i)}) - \ell(q; \theta_{t_i}). \quad (2)$$

Metrics - Retainability

- **Quantifies the fraction of improvement retained by the model after t steps, relative to the LAM**
- **Used to answer RQ3 (*Dynamics of forgetting*)**



Definition 3 *Consider a language model parameterized by θ and trained with injected knowledge k for N iterations, occuring at timesteps $t_i \in T_k$ where $|T_k| = N$. Let t_{pre} denote the last timestep before the model is first trained with k, i.e., $t_{pre} = \min(T_k) - 1$. Given a corresponding probe q,* ***retainability*** *($\mathcal{R}(q,t)$) is defined for $t \geq 0$ as follows:*

$$\mathcal{R}(q,t) = \frac{\ell(q;\theta_{t_{LAM}(q,N)+t}) - \ell(q;\theta_{t_{pre}})}{\ell(q;\theta_{t_{LAM}(q,N)}) - \ell(q;\theta_{t_{pre}})}. \tag{3}$$

Experimental Setup

- **We use three different checkpoints of OLMo-1B and -7B models:**
  - **OLMo-1B:**
    - 168B (step 40,000)
    - 484B (step 117,850)
    - 1.5T (step 358,000)
  - **OLMo-7B**
    - 177B (step 40,000)
    - 500B (step 113,000)
    - 1.5T (step 339,000)

Experimental Setup

- **For each checkpoint, we continue pre-training with two phases:**

  - **Phase I (Injection)**: Replace the part of the original train batch data, with the injected knowledges in *Fictional Knowledge* dataset

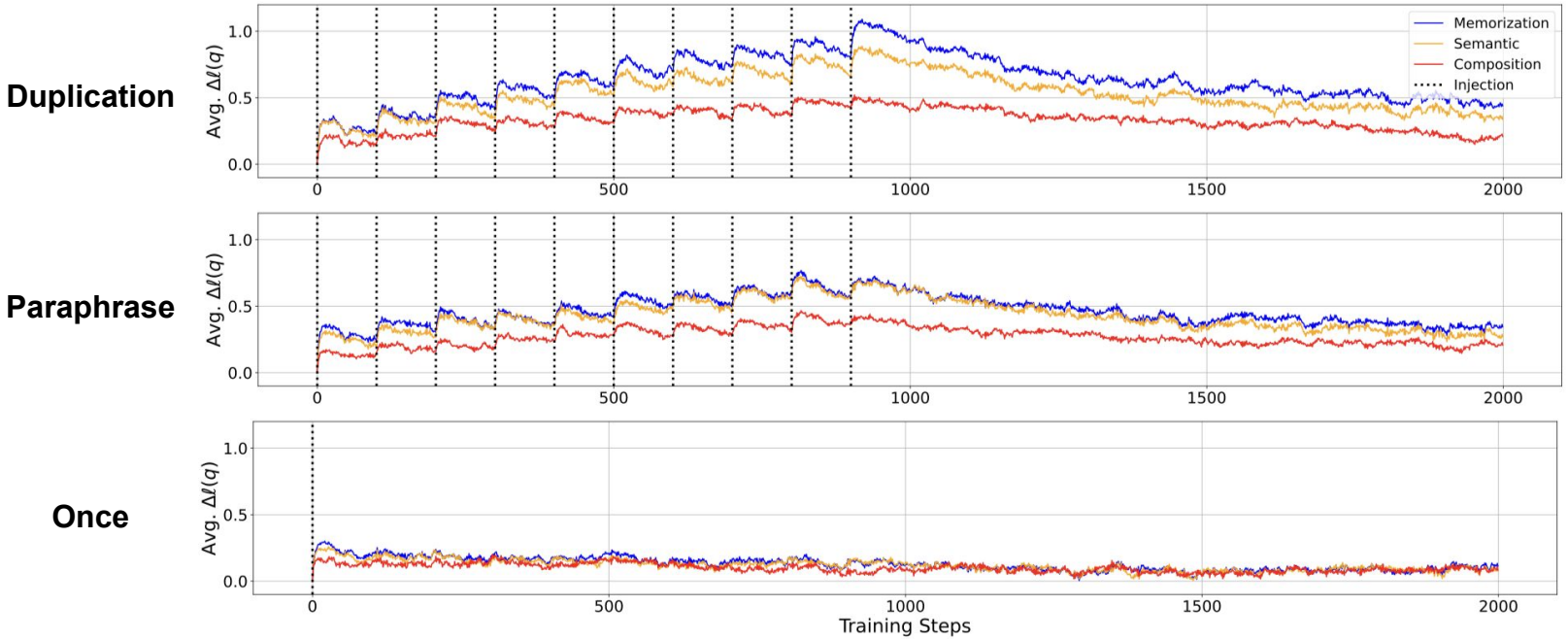
    Three different injection scenarios:
    - **Duplication:** From step 0 to 900, train with the injected knowledges with the interval of 100 steps
    - **Paraphrase:** Same with duplication, but we provide paraphrased knowledge each time we inject
    - **Once:** Train with the injected knowledges only at step 0

  - **Phase II (Perturbation)**: After injection is done, we continue pre-training the model as normal

- **We use 40 definitions of the *Fictional Knowledge* dataset for each scenario**

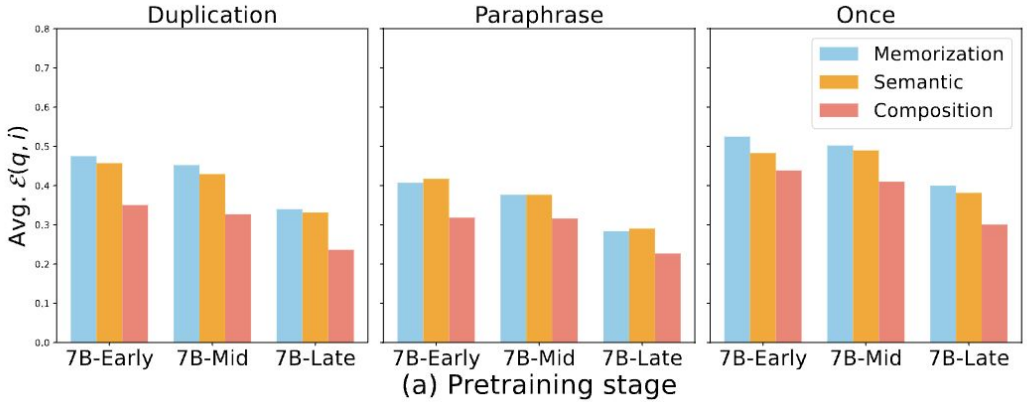
Results - Factual Knowledge Acquisition Dynamics (7B)

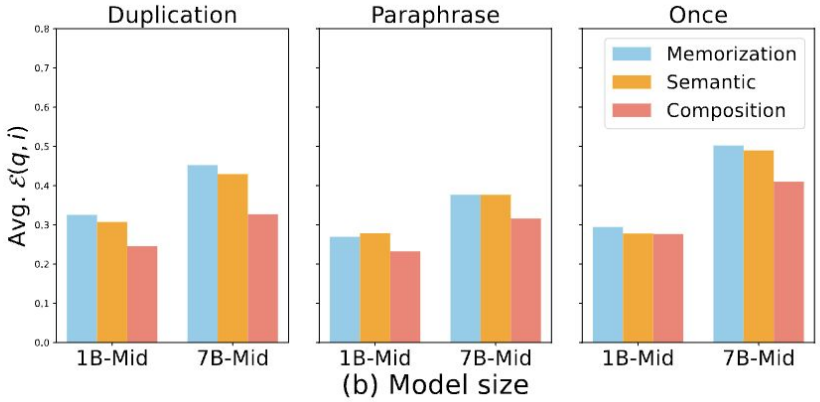
LLMs acquire factual knowledge by accumulating micro-acquisitions with subsequent forgetting!



Results - Effectivity Measurement

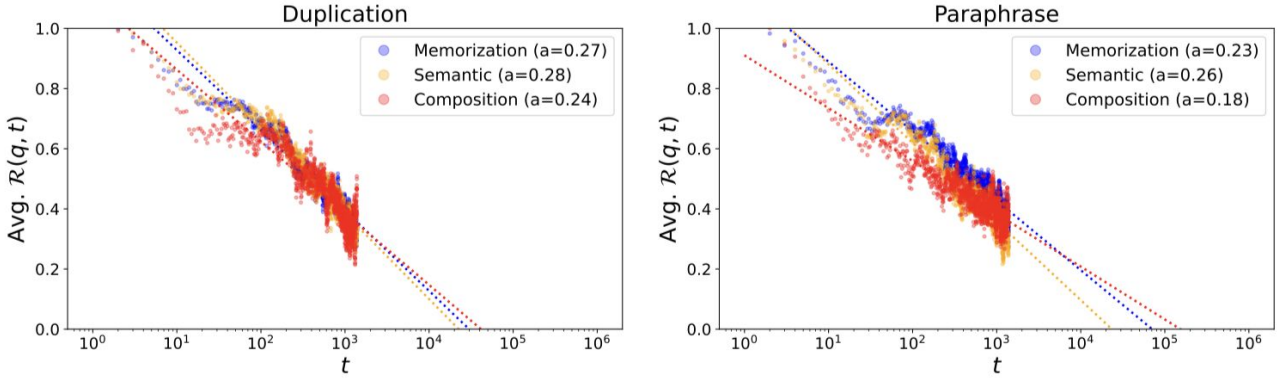
- **Effectivity does not improve as the models are trained with more tokens (Left)**
- **In contrast, there is a clear improvement in effectivity with increasing model sizes**





Results - Power-Law Relationship Between Forgetting and Training Steps

There is a clear power-law relationship between retainability and the training steps pass the LAM



The measured decay constant (a) represents how fast (in terms of fraction) the model loses the improvement of log probability.

$$\Delta\mathcal{R}(p,t) \approx -a \cdot \log\left(\frac{t_2}{t_1}\right) \quad \text{for } 0 < t_1 < t_2 < \tau, \quad \text{where } \mathcal{R}(p,\tau) = 0 \text{ and } a > 0$$

Results - Decay Constant Measured With OLMo-7B

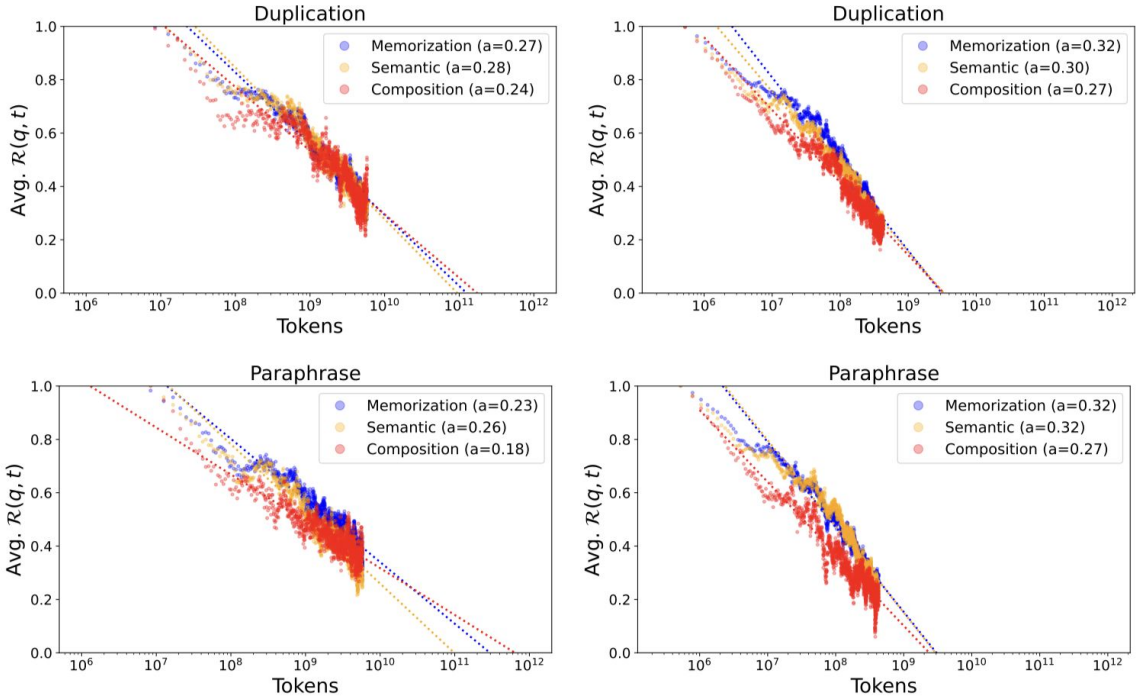
Pretraining stage		Early (170B)	Mid (500B)	Late (1.5T)
Duplication	Memorization	$0.29_{\pm 0.0023}$	$0.27_{\pm 0.0020}$	$0.22_{\pm 0.0019}$
	Semantic	$0.28_{\pm 0.0019}$	$0.28_{\pm 0.0023}$	$0.24_{\pm 0.0020}$
	Composition	$0.21_{\pm 0.0021}$	$0.24_{\pm 0.0032}$	$0.18_{\pm 0.0025}$
Paraphrase	Memorization	$0.23_{\pm 0.0019}$	$0.23_{\pm 0.0021}$	$0.21_{\pm 0.0023}$
	Semantic	$0.23_{\pm 0.0022}$	$0.26_{\pm 0.0025}$	$0.23_{\pm 0.0025}$
	Composition	$0.19_{\pm 0.0023}$	$0.18_{\pm 0.0025}$	$0.21_{\pm 0.0028}$

- **The forgetting in compositional generalization is slower (the decay constant a is smaller) than in shallower acquisitions**

- **The forgetting tends to be slower in the paraphrase injection scenario compared to the duplication injection scenario**
  - This can explain why deduplicating training data is beneficial, although duplication leads to better effectivity (discussed later in depth)

Results - Forgetting and Batch Size

Training with reduced batch size may lead to faster forgetting

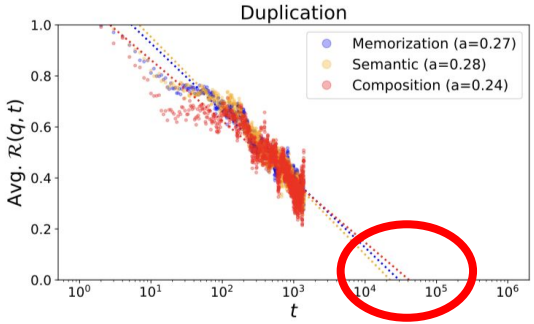


Pretraining (2048) **Reduced (128)**

Implications for LLM Pretraining

Why is popularity important for factual knowledge acquisition?

- The estimated x-intercepts represent the number of additional training tokens that would lead to the complete loss of the factual knowledge acquired by training



- This implies that there is a ***learnability threshold***, a threshold of the interval where the model fails to acquire knowledge of which its encounter interval is longer than the threshold

- The popularity of the knowledge in the pretraining data influences how quickly this knowledge begins to be ‘revealed’ in the generated sequences during pretraining

Implications for LLM Pretraining

Why does deduplication enhance model performance?

- Deduplication tends to slow the forgetting of generalizing acquired factual knowledge

- Presenting the model with duplicated texts will result in the widening of the gap between memorization and generalization

- Such gap will drive the model to prefer generating memorized contexts compared to generalizing factual knowledge

Appendix: Factual Knowledge Acquisition Dynamics (1B)

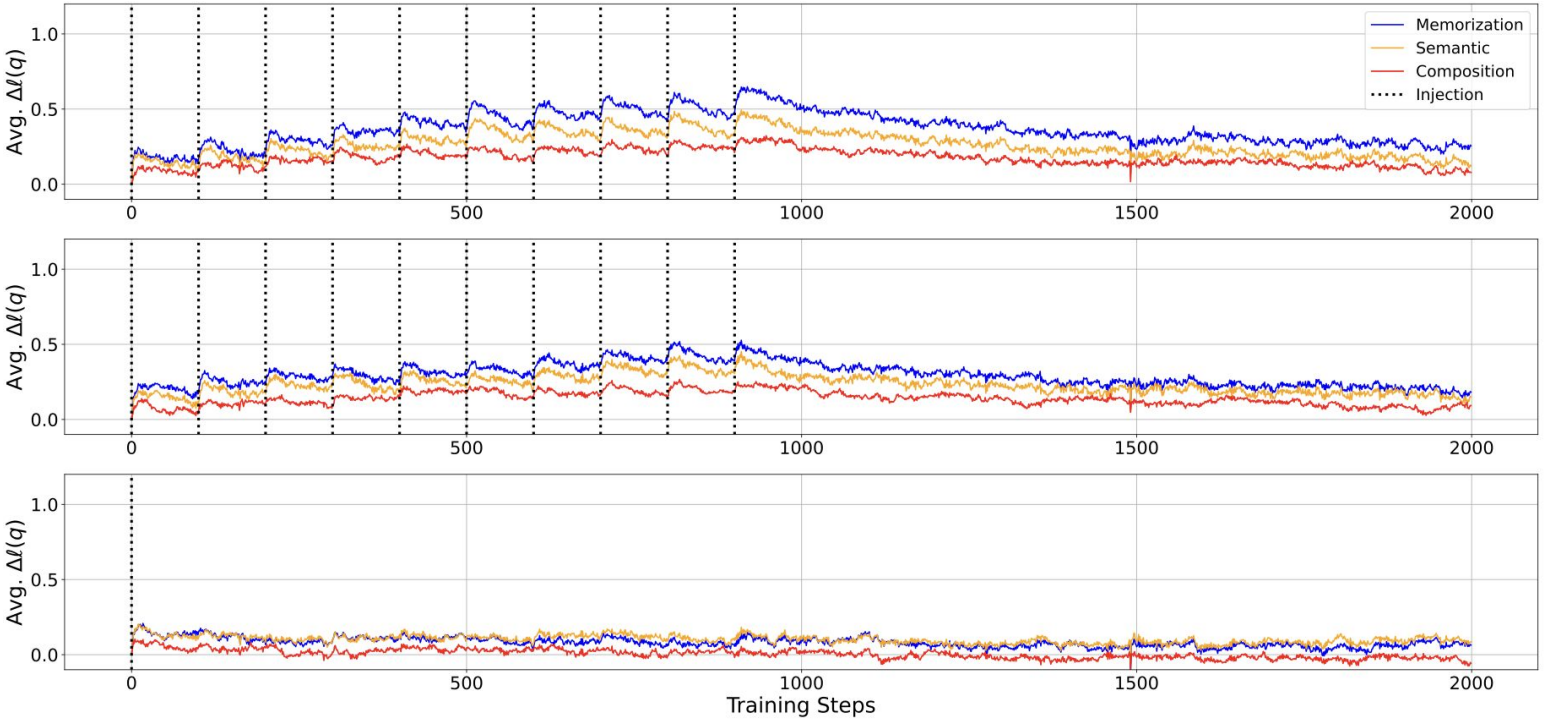


Figure 10: Training dynamics of OLMo-1B *Late* (1.5T) checkpoint.

Appendix: Effectivity Measurement With a Constant Learning Rate

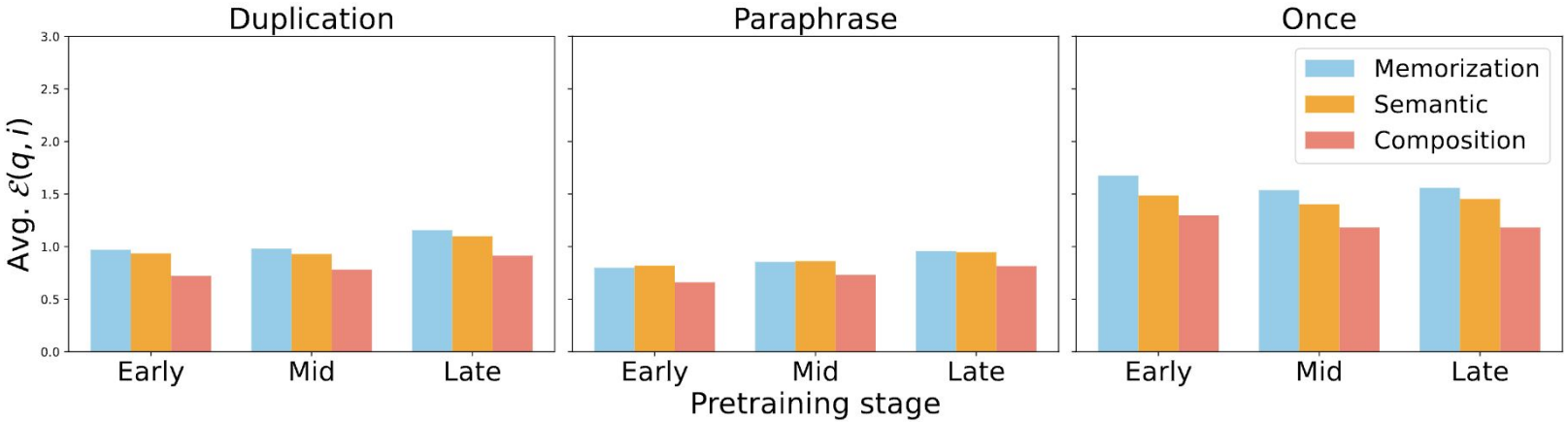


Figure 17: Average effectivity measured with OLMo-7B trained with a fixed constant learning rate.

Appendix: Effect of the Number of Previous Encounters on Effectivity and Retainability of Factual Knowledge

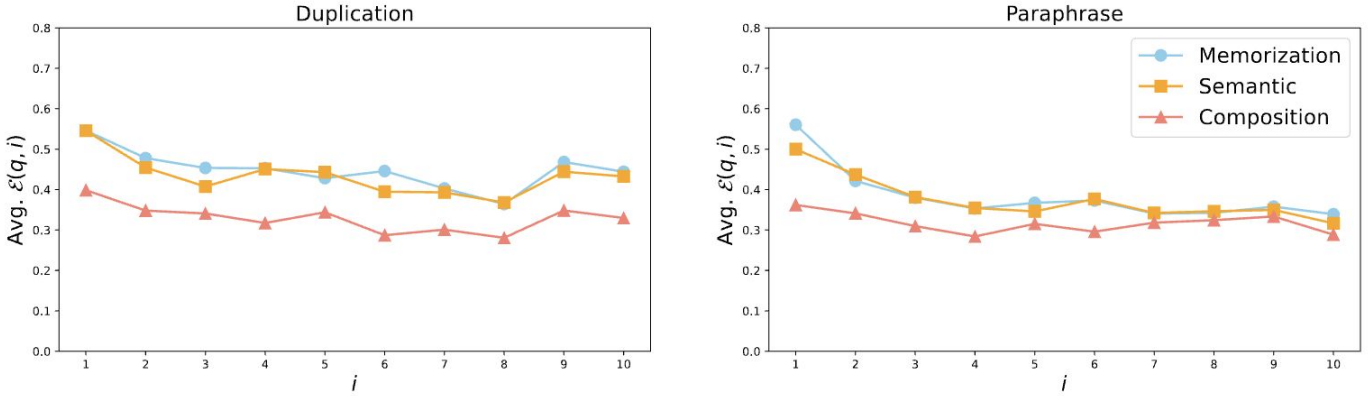


Figure 28: Average effectivity measured for each count of injection, measured with OLMo-7B *Mid* (500B) checkpoint.