DID YOUR LLM REALLY FORGET? MEASURING FACTUAL RECALL DURING LLM POST-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training shapes LLM behavior, e.g., by aligning models with user preferences and scaling test-time compute through long reasoning trajectories. While post-training rarely teaches models new facts about the world, it might induce forgetting of factual knowledge acquired during pretraining. Previous studies of this phenomenon either rely on aggregate measures that mask individual facts or compare only two checkpoints. We propose a different way to approach this problem, investigating how the probability of recalling individual facts is altered across the entire post-training trajectory. Using repeated rollouts across intermediate checkpoints, we show that individual facts' recall probabilities fluctuate substantially during post-training, repeatedly increasing and decreasing even when aggregate accuracy remains relatively stable. Although changes in the underlying probability of recalling each fact largely explain the magnitude of these fluctuations, whether a fact is classified as forgotten or reappearing is driven primarily by sampling variability. Thus, we show that prior methods can substantially overestimate persistent knowledge loss. Finally, we formalize a lower bound to certify the knowledge loss in bits. We show that our metric is more consistent than existing ones and can be easily extended to account for recall fluctuations.

1 INTRODUCTION

Modern training pipelines for Large Language Models (LLMs) are split into a pretraining and a post-training stage (Yang et al., 2025; Ettinger et al., 2025), with the two having different objectives. During pretraining, randomly initialized LLMs acquire language structure, in-context learning behavior (Brown et al., 2020), general task capabilities (Ouyang et al., 2022), and large-scale world knowledge (Petroni et al., 2019). Post-training, instead, alters the model's default behavior, e.g., following answer templates (Touvron et al., 2023) and instructions (Ouyang et al., 2022), satisfying human preferences (Ouyang et al., 2022; Rafailov et al., 2023), and reasoning through long trajectories (Wei et al., 2022; Guo et al., 2025). The common view is therefore that *post-training reshapes how a model elicits and expresses what it already knows* (Zhou et al., 2023; Jiang et al., 2024), rather than acting as a source of broad factual knowledge (Gekhman et al., 2024; Jiang et al., 2024).

A post-training side effect is that it can inadvertently alter the model's capabilities, overwriting what it has learned during pretraining. For instance, it may reduce its ability to use tools or to recall facts. This phenomenon, known as catastrophic forgetting (Wallat et al., 2020; Ghosal et al., 2024), causes performance degradation that can undermine the model's reliability during deployment. Monitoring and flagging forgetting as early as possible *during* training is desirable to mitigate this harmful side-effect, avoiding further expenses (e.g., re-training) later in the development process.

To study forgetting, prior work focused on the ability to *recall facts*, e.g., general knowledge we can elicit through multiple-choice QAs, measuring forgetting by comparing a base model with its post-trained counterpart (Du et al., 2025; Yuan et al., 2024; Harmon et al., 2026; Thede et al., 2026). For instance, Zhou et al. (2023) measure aggregate accuracy differences to quantify forgetting. However, this cannot reveal *compensating gains and losses* among individual facts (Figure 1 left). Instead, Harmon et al. (2026) measures per-question correct-to-incorrect transitions, preserving fact identity. Yet these works rely on greedy decoding, ignoring *sampling variance* (Figure 1 middle–right). Moreover, comparing two checkpoints, they cannot distinguish *persistent* changes from transient fluctuations along the training trajectory across post-training phases (Figure 1 middle–left).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

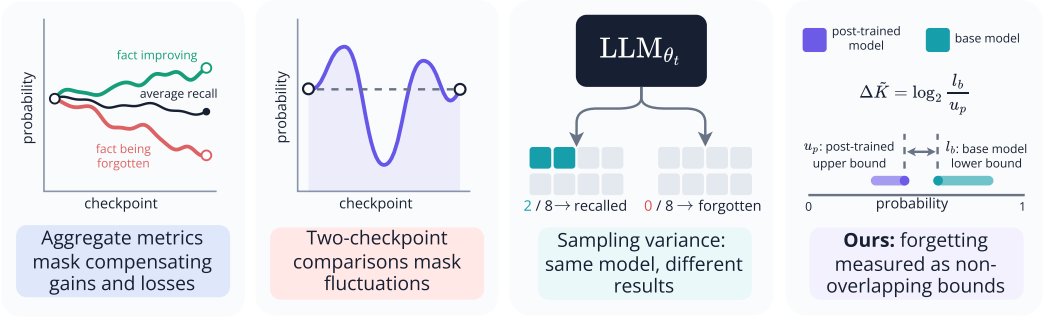


Figure 1: **Challenges in measuring forgetting.** From left to right, we show that (i) aggregate metrics can mask compensating gains and losses across facts; (ii) comparing only two checkpoints can hide fluctuations in fact sampling probabilities; (iii) sampling variance can bias forgetting estimates; and (iv) our certified lower bound accounts for sampling and is more robust to fluctuations.

Contrary to prior works, this paper proposes a formal approach for measuring forgetting. To show that aggregate accuracy can mask gains and losses driven by fluctuations in fact-sampling probabilities, we prompt intermediate post-training checkpoints with challenging questions related to factual knowledge. This allows us to analyze the *entire trajectory* of fact recallability. Our analysis reveals that, while aggregate fact recall accuracy is remarkably steady, individual fact trajectories fluctuate significantly from one checkpoint to another (see Figure 2). This suggests that aggregate accuracy is an **insufficient forgetting measure**, and fact probabilities **do not change monotonically**.

However, as fact recall is estimated from finitely many stochastic generations, apparent fluctuations may arise even if the underlying recall probability remains constant. We therefore simulate these fluctuations via a stationary binomial distribution, showing that post-training explains around half of the fluctuations, while the rest is just a sampling artifact. Additionally, we show that most facts marked as forgotten by binarizing outcomes (recalled/not recalled) are only due to sampling, suggesting that interpreting a fact not recalled as forgotten can **overestimate forgetting**. Thus, we show that scaling the final model's sampling budget *re-elicits* most of the facts that would otherwise be marked as forgotten. Finally, we formalize a lower bound of the forgotten knowledge, in bits, of the ratio between base and post-trained model sampling probabilities (Figure 1 right). This provides a *conservative estimate* of knowledge loss by assigning a low sampling probability to samples with no correct rollouts, reducing the risk of overestimating forgetting. We show that our metric is *more consistent* than using binary outcomes and that it can easily be extended to *account for fluctuations*.

Taken together, these results have practical implications. Forgetting under finite sampling should not be read as catastrophic forgetting, as: (i) increasing the sampling budget can easily recover most of apparently lost facts, and (ii) fact sampling probabilities fluctuate and can even increase during post-training. These considerations require rethinking not only how we measure forgetting, but also how knowledge fluctuations may impact the effectiveness of unlearning and safety-alignment methods.

In summary, **our contributions are:** (i) We are the first to track factual recall *along* an LLM post-training trajectory, revealing that it is unstable (fluctuates) for individual facts while aggregate accuracy is generally flat; (ii) We demonstrate that observed fluctuations are not caused by training data contamination but depend on sampling and training dynamics; (iii) We show that finite sampling accounts for *most* binary transitions between forgotten and recalled states; thus, increasing the sampling budget is sufficient to "recover" most facts marked as forgotten; (iv) We propose a lower bound that exploits our findings on sampling and trajectory variance (fluctuations) to certify forgetting.

2 Related Work

Factual knowledge estimation. A long line of work has investigated how LLMs store and retrieve factual knowledge. The CLOZE task masks an answer in a closed-book question and evaluates the fact by the ground-truth answer's rank Petroni et al. (2019). Dong et al. (2023) and Zhao et al. (2024) extended this paradigm to test robustness across paraphrases and sampled responses. Instead,

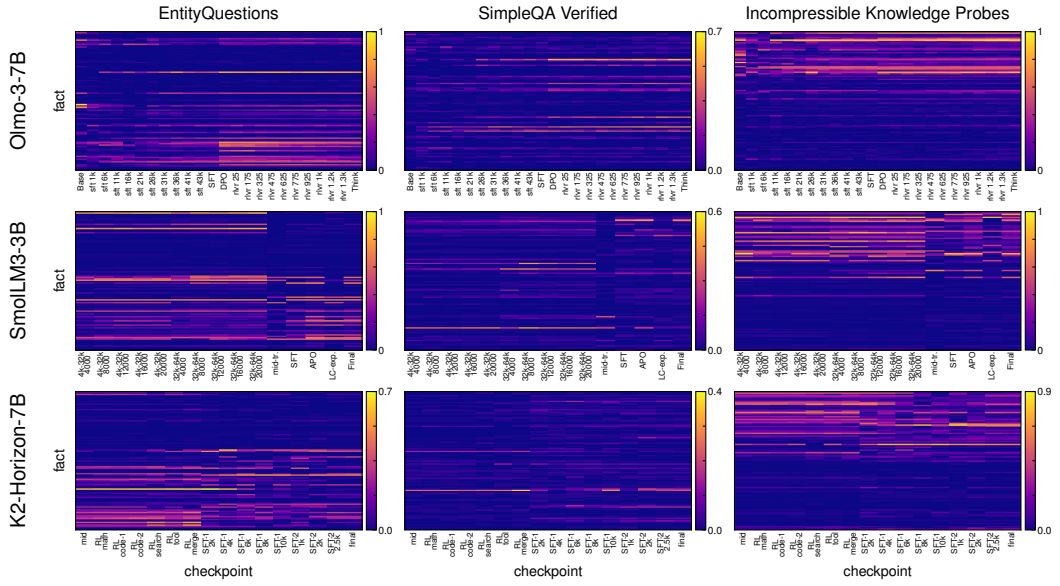


Figure 2: **Fact recall is unstable.** Color shows individual fact recall $p_\theta(f, q)$ relative to its minimum across the different post-training checkpoints (x-axis). Recall fluctuates throughout post-training, suggesting that comparisons between only two checkpoints cannot reliably measure forgetting.

Shin et al. (2020) asked whether recall failures reflect missing knowledge or ineffective prompting. However, Gekhman et al. (2025) argued that these methods measure only extractable knowledge, since LLMs may internally represent facts they do not reveal. At a more fundamental level, Li (2026) and Allen-Zhu & Li (2025) showed that knowledge representation is incompressible because model parameters can encode only finite information. Gekhman et al. (2026) further found that reasoning before answering increases correct recall. These findings suggest that estimated recall depends on successful elicitation through prompting and inference. We argue that output interpretation matters equally. Prior work treats finite response samples as definitive evidence of model knowledge and summarizes it with aggregate accuracy. We instead estimate the probability that the model samples a correct answer for each fact, using probability bounds to account for sampling variance.

Measuring forgetting. Catastrophic forgetting, the degradation of previously acquired knowledge as a model learns new information, is a central challenge in continual learning (McCloskey & Cohen, 1989; Kirkpatrick et al., 2017). In language models, it can degrade general knowledge and language-generation behavior (Luo et al., 2025; He et al., 2021). The clear division of labour between pretraining and post-training (Zhou et al., 2023; Jiang et al., 2024) is a main source of forgetting. By comparing models before and after post-training, studies have shown that factual knowledge can be lost or become inaccessible (Wallat et al., 2020; Ghosal et al., 2024). Others suggest that factual knowledge changes more slowly than stylistic and topical behavior (Gekhman et al., 2024; Zhang & Wu, 2024). Recently, attention has moved beyond aggregate benchmark differences toward fact-level, or sample-wise, forgetting, measuring correct-to-incorrect transitions, acquisition and loss trajectories, and backward transfer across post-training stages (Harmon et al., 2026; Hier & Obafemi-Ajayi, 2026). However, these works neglect sampling variance, consequently overestimating forgetting. Instead, we estimate a lower bound on knowledge loss, treating facts with no correct sampled rollouts as having low sampling probability.

3 Factual knowledge instabilities

In this section, we first define factual knowledge estimation (Section 3.1), and show how the factual recall changes across different post-training checkpoints (Section 3.2). Then, we analyze whether these results are linked to data contamination (Section 3.3) or sampling variance (Section 3.4).

Model	Dataset	RMS diff.	RMS max diff.
Olmo-3-7B	EQ	4.4%	14.5%
	SQA	2.2%	6.0%
	IKP	6.5%	21.2%
SmolLM-3-3B	EQ	7.3%	19.5%
	SQA	2.9%	7.4%
	IKP	10.9%	28.3%
K2-Horizon-7B	EQ	4.1%	10.3%
	SQA	1.7%	4.6%
	IKP	6.7%	17.4%

Table 1: **Fact recall fluctuations.** We report the RMS difference and the RMS max difference across adjacent checkpoints.

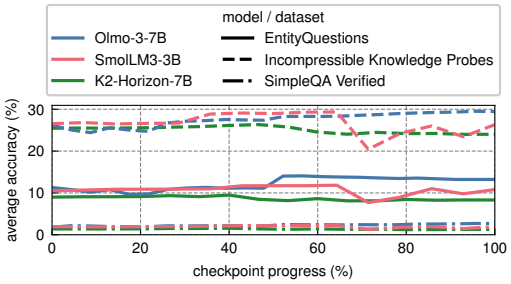


Figure 3: **Aggregate factual accuracy trends.** Fact recall is stable throughout post-training.

3.1 Factual knowledge estimation

We study how factual knowledge recall changes during post-training, i.e., how reliably knowledge can be elicited from an LLM through closed-book QAs. Let a fact be represented by $f = (Q_f, A_f)$, where Q_f is a set of valid questions that elicit the fact and A_f is the corresponding set of valid answers. Let θ be the LLM's parameters. Following Petroni et al. (2019), we define LLM_θ knowledge on query $q \in Q_f$ as: $p_\theta(f, q) \coloneqq \Pr_{a \sim \text{LLM}_\theta(\cdot|q)}[a \in A_f]$, i.e., the probability that the model produces a valid answer. Since we do not have access to the underlying sampling probability, we follow prior work (Gekhman et al., 2026) and estimate $p_\theta(f, q)$ by sampling 128 rollouts for the same question, unless stated otherwise, and judge their content. Full rollout sampling details in Appendix A.3.

3.2 Is factual recall stable?

We first test whether models' ability to recall individual facts is stable during training (e.g., decreases monotonically) or fluctuates. To evaluate this, we need models with openly available post-training checkpoints. Thus, we test 55 intermediate checkpoints from Olmo-3-7B (Ettinger et al., 2025), SmolLM3-3B (Bakouch et al., 2025), and K2-Horizon-7B (IFM, 2026) (further details in Appendix A.1). Test facts are sampled from established data sources, i.e., EntityQuestions (EQ) (Sciavolino et al., 2021), SimpleQA Verified (SQA) (Haas et al., 2025), and Incompressible Knowledge Probes (IKP) (Li, 2026), which are challenging factual knowledge QA benchmarks of facts mostly retrieved from Wikidata (Vrandečić & Krötzsch, 2014). Additional details in Appendix A.2.

Figure 2 reports for 100 *randomly* sampled facts, the recall probability (p_θ) across the post-training trajectory. We excluded facts for which models always answer correctly or always answer incorrectly, as these samples are either too easy or likely unknown to the model throughout post-training. In Olmo-3-7B, the transition from SFT to DPO (Rafailov et al., 2023) is particularly marked in EntityQuestions, with multiple facts increasing recallability after the change. Similarly, SmolLM3-3B heatmaps clearly show marked transitions from the "4k-to-32k" long-context extension to the "32k-to-64k", as well as each post-training stage.

Factual recall fluctuates substantially during post-training, registering, e.g., an RMS (root-mean-square) difference between adjacent checkpoints of 7.3% for SmolLM3-3B in EntityQuestions, and 6.7% for K2-Horizon-7B in IKP (see Table 1). Moreover, we find that the RMS max diff., i.e., the root mean of the maximum squared difference across adjacent checkpoints, can reach 21.2% for Olmo-3-7B and 28.3% for SmolLM3-3B. Aggregate factual recall accuracies, instead, are rather steady during post-training (Figure 3), with the only noticeable decrease in K2-Horizon-7B after switching from RL to SFT training (i.e., at 40% checkpoint progress in Figure 3). These fluctuations reveal an interesting aspect of model behavior that prior work has largely overlooked.

3.3 Data exposure

A compelling explanation for the fluctuations is the post-training data exposure to the tested facts. To check this hypothesis, we exploit the template-based structure of EntityQuestions (Sciavolino et al., 2021), extracting triplets (*subject*-*object*-*relation* – e.g., Paris is the capital of France) composing the tested facts, and check whether they appear in the training data. Further details are in Appendix A.4.

Model	Dataset	Improved	Worsened
Olmo-3-7B	EQ	8.4	17.5
	IKP	25.3	26.7
	SQA	10.5	10.9
Qwen3.5-9B	EQ	27.4	14.9
	SQA	24.3	15.9
	IKP	39.1	36.7

Table 2: **Percentage of facts with changed recall after math SFT.** Facts improve and worsen despite not being in the training set.

Model	Dataset	Obs. RMS	Sim. RMS	p_{adj}	Sig.
Olmo-3-7B	EQ	4.44%	2.19%	0.0009	✓
	SQA	2.20%	1.59%	0.0009	✓
	IKP	6.52%	3.06%	0.0009	✓
SmolLM3-3B	EQ	7.32%	2.33%	0.0009	✓
	SQA	2.94%	1.57%	0.0009	✓
	IKP	10.9%	3.06%	0.0009	✓
K2-Horizon-7B	EQ	4.14%	1.99%	0.0009	✓
	SQA	1.73%	1.36%	0.0009	✓
	IKP	6.74%	3.11%	0.0009	✓

Table 3: **Observed fluctuations vs. fluctuations under the stationary-binomial null.**

We applied this procedure to the post-training data of the three models, counting $\approx$2.5M, $\approx$8.6M, and $\approx$8.6M entries, respectively. We found only 0.5% of the evaluated facts from EntityQuestions in the post-training data of Olmo-3-7B, 2.5% in that of SmolLM3-3B, and 0 in K2-Horizon-7B's. In the first two cases, most leaks (e.g., $4/5$ in Olmo data) were found in the SFT split, accounting for the first half of checkpoints for Olmo-3-7B in Figure 2 and a single column for SmolLM3. **This suggests that per-fact fluctuations *cannot be explained* by simple training dataset contamination.**

To further support this thesis, we push this paradigm to the extreme: fine-tuning the model on data *completely unrelated* to the evaluation sets. Specifically, we fine-tune Olmo-3-7B-Base (Ettinger et al., 2025), and Qwen3.5-9B-Base[1] (Yang et al., 2025) on the math subset of Dolci-Think-RL-7B (Ettinger et al., 2025) (training details in Appendix A.5). While we would expect that forgetting emerges for all concepts, results in Table 2 show that fact recall improves for numerous facts and worsens by a proportional amount (e.g., 27.4 vs. 14.9 in EntityQuestions). These counter-intuitive results show that facts might *re-emerge* even when fine-tuning on out-of-domain data, thus confirming that fluctuations in fact recall are not caused by training data contamination.

3.4 Are fluctuations only caused by sampling?

If data is not the primary cause of fluctuations, we analyze whether they are caused by sampling variance. To achieve this, we compare the observed fluctuations (RMS) with those modeled by a stationary binomial distribution, i.e., assuming that sampling probabilities do not change. Therefore, for a given model and dataset, we fit a binomial distribution over the observed trajectory via MLE, and simulate post-training trajectories by sampling from it (further details in Appendix A.6).

Table 3 compares the observed fluctuations with those modeled by the binomial distribution at $k = 128$ rollouts, revealing that the RMS diff. of the observed trajectory is around 2 times that of simulated trajectories, where the smallest difference is registered from K2-Horizon-7B checkpoints on SQA (1.73% vs. 1.36%), and the largest from SmolLM3-3B on IKP (10.9% vs. 3.06%). Under 10k simulated trajectories, all nine comparisons show statistically significant differences, with the minimum admitted p-value of 9×10^{-4}, rejecting the hypothesis that sampling alone causes fluctuations. Thus, **sampling probabilities *are* the element that fluctuates during post-training.**

4 Delving deeper into sampling variance: how is forgetting overestimated?

This section analyzes the consequences of sampling variance on fact recall, showing that facts can appear to be forgotten but then reappear under existing metrics (Section 4.1) and how (i) this is due to sampling stochasticity (Section 4.2) and (ii) can lead to overestimating forgetting (Section 4.3).

4.1 Forgotten and reacquired facts

We start by intuitively showing why sampling variance leads to overestimating forgetting. We follow the experimental procedure in Section 3.2 considering a fact as "forgotten" if it is never generated

[1]Since intermediate checkpoints are unavailable, we report results on Qwen3.5 where possible.

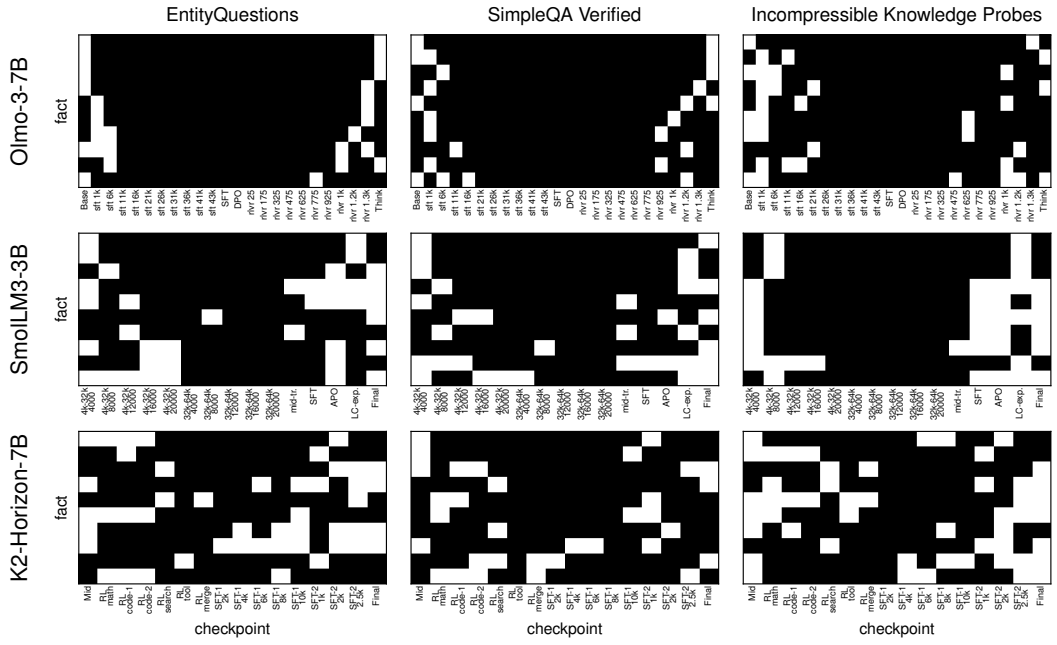


Figure 4: **Facts with the longest oblivion times.** Trajectories of the top-10 facts with the longest oblivion time, i.e., the time between when a fact appears forgotten and then reacquired. Facts can appear to be forgotten (in black) during post-training, to then reappear (in white) at the end.

at least once in 128 rollouts. By this definition, facts can be forgotten and reacquired multiple times during training. Figure 4 shows a few extreme cases. For each model–dataset pair, we report the top-10 facts by oblivion time, the longest period in which a fact was not extracted in 128 rollouts. Some facts "disappear" through post-training and "reappear" only after millions of training samples. We report a few qualitative examples in Appendix D. Although this may suggest fluctuations in sampling probabilities, the next subsection shows that sampling variance is the primary cause.

4.2 Is sampling stochasticity the source of forgetting?

We now show that forgetting and reappearances are directly caused by sampling stochasticity, by comparing them with rollouts simulated from a binomial distribution (Section 3.4). Table 4 compares the observed and simulated percentage of facts that are forgotten and reappear (f&r) at least once during the post-training trajectory at $k = 128$ rollouts (Appendix C.2 extends these results to lower rollout budgets). Although the differences between the observed and simulated f&r rates are statistically significant in four of the nine cases (i.e., all Olmo-3-7B evaluations and SmolLM3-3B in EQ), most f&r events are explained by sampling. For instance, the percentage of facts that appear to be forgotten and then reappear from Olmo-3-7B in IKP is 46.3% against 41.4% in the simulated case. Moreover, the observed f&r from SmolLM3-3B in SimpleQA Verified and all those of K2-Horizon-7B are lower than what simulations predict (e.g., 23.1% vs. 26.3% of K2-Horizon-7B in SimpleQA Verified). These results show that, even though recall probabilities may change during training, **most forgetting events across checkpoints can be explained by finite-sampling stochasticity.**

Model	Dataset	Obs. f&r	Sim. f&r	p_{adj}	Sig.
Olmo-3-7B	EQ	30.9%	27.1%	0.0009	✓
	SQA	35.6%	33.1%	0.0009	✓
	IKP	46.3%	41.4%	0.0009	✓
SmolLM3-3B	EQ	23.2%	21.3%	0.004	✓
	SQA	24.1%	26.5%	1.000	✗
	IKP	35.6%	35.3%	1.000	✗
K2-Horizon-7B	EQ	15.6%	16.2%	1.000	✗
	SQA	23.1%	26.3%	1.000	✗
	IKP	26.9%	33.8%	1.000	✗

Table 4: **Observed f&r events vs. f&r events under the stationary-binomial null.**

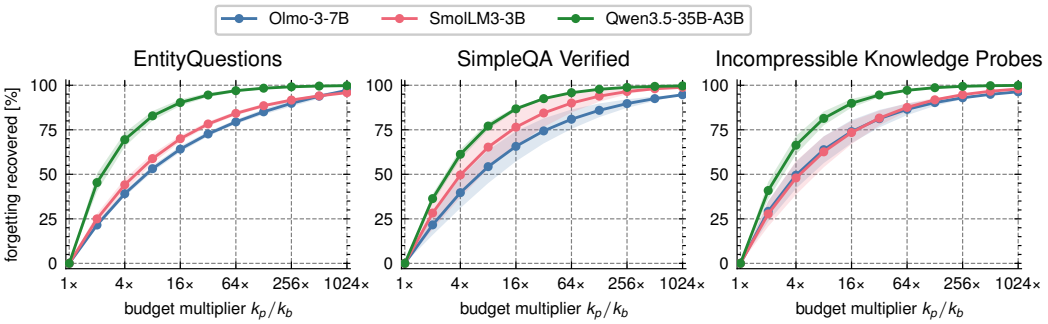


Figure 5: **Additional sampling overturns a large fraction of apparent forgetting.** For each multiplier k_p/k_b, we show the mean fraction of overturned decisions across all initial budgets.

4.3 Apparent forgetting

We now illustrate the practical implications of interpreting forgetting as a binary (i.e., forgotten vs. recalled) event. For each fact, we collect 1024 rollouts from the base and post-trained models. We consider base model budget, i.e., number of rollouts, $k_b \in \{1, 2, \ldots, 512\}$ and post-trained budget $k_p \in \{k_b, 2k_b, \ldots, 1024\}$. For each rollout budget $k_b = k_p$, we estimate forgotten facts as those never recalled by the post-trained model but recalled at least once by the base model (Harmon et al., 2026). We refer to this as binary forgetting. We then scale k_p and measure the average fraction of overturned decisions at each budget multiplier.[2]

Figure 5 shows this analysis for Olmo-3-7B, SmolLM3-3B, and Qwen3.5-35B-A3B. Across model–dataset pairs, doubling the post-training budget overturns 30.7% of forgetting decisions, while quadrupling it subverts 51.9% of them on average. The strongest recovery occurs for Qwen3.5-35B-A3B on EntityQuestions, where doubling and quadrupling recover 45.4% and 69.5%, respectively. Under the largest budget imbalance ($k_b = 1$, $k_p = 1024$), additional sampling overturns the decision for 94.7–99.9% of facts initially labeled as forgotten across the nine model–dataset pairs.

Thus, a change in forgetting classification shows that these **binary labels depend *strongly* on the rollout budget, and do not establish a change in recall probabilities,** since additional samples may produce a correct answer by chance even when recall probability is unchanged.

5 Certified knowledge loss

Sampling variance can lead to overestimating forgetting, but reducing the former requires increasing the number of rollouts. We instead formalize a certified lower bound on knowledge loss for a given estimation budget, demonstrating that it does not overestimate forgetting, even with few rollouts (Section 5.1). We then show how it tightens as the number of rollouts increases, use our metric to quantify facts forgotten by the tested models, and compare $\Delta\tilde{K}$ with binary forgetting (Section 5.2).

5.1 Metric definition

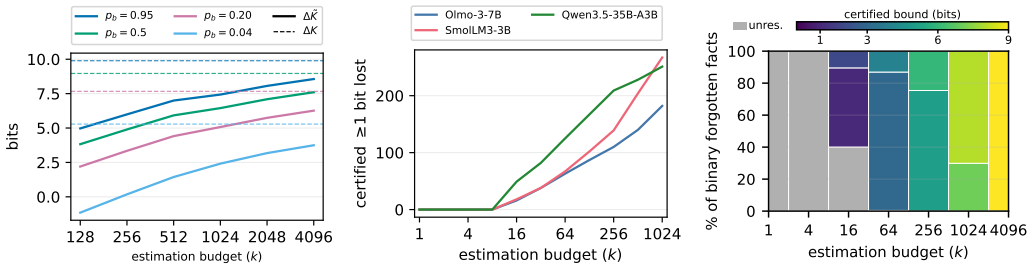
Let us redefine the factual knowledge about f elicited by q as the reduction in correct-answer surprisal, measured in bits, relative to a reference probability as:

$$K_\theta(f, q) \;:=\; \log_2 \frac{p_\theta(f, q)}{\pi(f, q)}, \tag{1}$$

where $\pi(f, q)$ is the probability of sampling the correct answer from the uniform distribution of valid answers. Let θ_b and θ_p be the weights of the base and post-trained model, respectively. From Eq. (1), we can measure the knowledge difference, in bits, on the same fact between the two as:

$$\Delta K \;=\; K_{\theta_b} - K_{\theta_p} \;=\; \log_2 \frac{p_b(f, q)}{p_p(f, q)}, \tag{2}$$

[2]Because the maximum available budget is 1024, fewer initial budgets contribute at larger multipliers.



(a) (b) (c)

Figure 6: (a) **Bound tightness.** We show the tightness of the knowledge loss lower bound at fixed post-trained probability ($p_p = 10^{-3}$), and different estimation budgets k and base model probabilities (p_b). Dotted lines represent the true value ΔK. (b) **Total number of forgotten facts,** among those tested, for which $\Delta\tilde{K} \geq 1$ bit. (c) $\boldsymbol{\Delta\tilde{K}}$ **vs. binary forgetting.** Bit certified by $\Delta\tilde{K}$ out of facts marked as forgotten by binary forgetting, with $p_b = 0.5$, $p_p = 10^{-3}$, and $\Delta K \approx 9$ bits.

with p_b and p_p the base and post-trained model probabilities. Estimating p_b and p_p with low variance requires a prohibitive number of rollouts, especially for facts hard to recall. Instead of estimating it, we derive a lower bound $\Delta\tilde{K}$ on the loss at a resolution of b bits, measuring knowledge loss as:

$$\Delta\tilde{K} \;\coloneqq\; \log_2 \frac{l_b(f,q)}{u_p(f,q)}, \tag{3}$$

where l_b and u_p are the lower and upper bounds of the base and post-trained model, estimated via the Clopper–Pearson bound (Clopper & Pearson, 1934) with budget k_b and k_p, and error rate α.

Proposition 1 *Let l_b and u_p be one-sided Clopper–Pearson bounds at level $\alpha/2$ computed from k_b and k_p rollouts. Then $\Delta\tilde{K} \leq \Delta K$, with probability at least $1-\alpha$. Moreover, if the post-trained model produces zero correct rollouts out of budget*

$$k_p^b \;=\; \left\lceil \frac{\ln(\alpha/2)}{\ln\!\left(1-\frac{l_b}{2^b}\right)} \right\rceil, \tag{4}$$

then $\Delta\tilde{K} \geq b$; that is, at least b bits of knowledge loss are certified at level α.

Proof. Proof in Appendix B. □

5.2 Results

Bound tightness. Figure 6a shows how $\Delta\tilde{K}$ varies w.r.t. the true value (ΔK) as the estimation budget increases at different base model probabilities (p_b) and with fixed post-trained probability $p_p = 10^{-3}$. Because variance is too high at low budgets, $\Delta\tilde{K}$ certifies a knowledge loss about 5 bits lower than the true value when $p_b \geq 0.20$. Yet, the returned value still represents a large loss in recall probability (e.g., 4 bits lost imply $16\times$ less likelihood of recalling the correct answer). As the rollout budget approaches 2^{11}–2^{12}, the variance decreases, the bounds tighten, and the estimated knowledge loss approaches its true value. Appendix C.3 extends this analysis, investigating how separate rollout budgets (k_b and k_p) affect the estimate, finding that scaling k_p is more beneficial.

Number of estimated forgotten facts. Figure 6b shows the number of facts for which our metric certifies a loss from the base model of at least 1 bit on the three datasets. Using fewer than 16 rollouts, our metric cannot confidently certify any loss, as uncertainty is too high. By increasing the budget k over this value, we see a monotonic increase in the number of samples losing at least 1 bit of information, i.e., 182, 267, and 251 facts at a budget of 1024 rollouts for Olmo-3-7B, SmolLM3-3B, and Qwen3.5-35B-A3B. In Appendix D, we report facts that lose most bits after post-training.

However, computing $\Delta\tilde{K}$ between two checkpoints remains subject to fluctuations in sampling probabilities. Table 5 shows that a loss of at least one bit certified at a single checkpoint can be

		$m=1$			$m=2$			$m=3$			$m=4$		
Model	Dataset	Cert	Contr.	Unres.	Cert	Contr.	Unres.	Cert	Contr.	Unres.	Cert	Contr.	Unres.
Olmo-3-7B	EQ	95.9	0.9	3.3	98.2	0.0	1.8	99.0	0.0	1.0	99.0	0.0	1.0
	SQA	94.7	0.0	5.3	98.5	0.0	1.5	99.1	0.0	0.9	99.0	0.0	1.0
	IKP	92.7	1.1	6.2	96.3	0.4	3.3	97.5	0.3	2.3	98.0	0.3	1.7
SmolLM3-3B	EQ	69.8	10.4	19.8	85.0	1.9	13.1	90.2	0.0	9.8	93.3	0.0	6.7
	SQA	60.0	4.0	36.0	90.9	0.0	9.1	100.0	0.0	0.0	100.0	0.0	0.0
	IKP	57.2	24.0	18.8	77.5	6.0	16.6	78.7	5.3	16.0	92.3	0.0	7.7
K2-Horizon-7B	EQ	65.3	4.2	30.6	81.2	0.0	18.8	94.4	0.0	5.6	90.9	0.0	9.1
	SQA	87.5	0.0	12.5	100.0	0.0	0.0	100.0	0.0	0.0	100.0	0.0	0.0
	IKP	65.0	8.6	26.4	87.5	1.6	10.9	94.4	2.8	2.8	100.0	0.0	0.0

Table 5: **Persistence of certified knowledge loss.** We certify a loss ≥ 1 bit relative to the base model at every checkpoint in a window of m checkpoints. We pool up to three later checkpoints and report the percentage of certified events whose loss is again *certified*, *contradicted*, or *unresolved*.

contradicted or become unresolvable at subsequent checkpoints ($m = 1$). For instance, only 57.2% of IKP facts marked as forgotten by at least one bit in SmolLM3-3B are confirmed by subsequent checkpoints. Requiring the previous $m - 1$ checkpoints to also certify a loss of at least one bit reduces the effect of these fluctuations. Confirmation then rises to 77.5% for $m = 2$ and 92.3% for $m = 4$, in the previous example. Therefore, **$\Delta\tilde{K}$ can be extended to account for fluctuations**, provided that intermediate checkpoints are available. Full experimental details are in Appendix C.4.

Comparison with binary forgetting. We now compare our metric with binary forgetting. Given simulated facts that the base model recalls at least once and that are never recalled by the final model, Figure 6c shows the corresponding certified knowledge loss $\Delta\tilde{K}$ at varying budget k. Although the binary metric flags facts as forgotten even at low budgets, $\Delta\tilde{K}$ leaves them unresolved until $k = 8$, where a small fraction of samples have certified knowledge loss greater than 0 bits. As we increase the rollout budget to 64, all facts are marked as forgotten by at least 2 bits, and further increasing the budget better approximates the true loss $\Delta K \approx 9$.

In Figure 7, we study the prediction consistency (forgotten/recalled) of the two metrics for two separate estimates at budget k, averaged across each model–dataset pair. **$\Delta\tilde{K}$ achieves more consistent agreement** between the two estimates than binary forgetting, showcasing its higher reliability (full experiment in Appendix C.5).

Finally, it is worth noting that while binary forgetting tends to 0 as $k \to +\infty$ when $p_p > 0$, $\Delta\tilde{K} \to \Delta K$ as $k \to +\infty$. This means that while binary forgetting marks fewer facts as forgotten as the rollout budget increases, limiting interpretation, $\Delta\tilde{K}$ returns a conservative estimate of the bits lost, approaching the true value (ΔK).

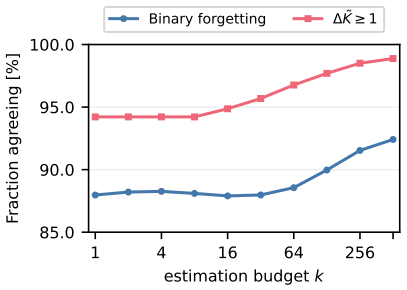


Figure 7: **Metric consistency.** Percentage of consistent predictions (forgotten/recalled) for two separate estimates at different rollouts k averaged over all models and datasets.

6 Conclusion

In this work, we show that existing measures of factual knowledge forgetting can ① mask compensating gains and losses across facts during post-training, ② fail to account for transient sampling-probability changes, and ③ overestimate forgetting by overlooking sampling variance. We demonstrate ① – ② at intermediate post-training checkpoints, where correct-answer probabilities fluctuate as individual facts improve and worsen despite stable aggregate trajectories. We demonstrate ③ by showing that sampling alone can often explain a failure to recall a fact. To address these limitations, we introduce a formal measure of forgotten knowledge based on a certified lower bound, in bits, on the log ratio of the base model's sampling probability to that of the post-trained model. For facts with no correct rollouts, this metric assigns low sampling probability rather than automatically labeling them forgotten. Concluding, failed recall under finite sampling does not directly imply catastrophic forgetting, and fluctuations matter when evaluating unlearning effectiveness and safety methods.

REPRODUCIBILITY STATEMENT

All model checkpoints and datasets used are publicly released and available at the time of writing. Construction and filtering of benchmarks are described in Appendix A.2. Answer sampling and judging are exhaustively described in Appendix A.3, along with an estimate of the rollouts generated for this project, and the hardware and software used. Appendices A.4 to A.6, instead, report the experimental details for the data contamination analysis, finetuning on math data, and simulated rollouts via binomial distribution. Finally, we include the anonymized codebase to ease reproducibility: `https://anonymous.4open.science/r/Measuring-Factual-Knowledge-E75C/README.md`

REFERENCES

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.2, knowledge manipulation. In *ICLR*, 2025.

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, Xuan-Son Nguyen, Colin Raffel, Leandro von Werra, and Thomas Wolf. Smollm3: smol, multilingual, long-context reasoner. `https://huggingface.co/blog/smollm3`, 2025. Accessed: 2026-07-22.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020.

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. In *EMNLP*, 2022.

Nikhil Chandak, Shashwat Goel, Ameya Prabhu, Moritz Hardt, and Jonas Geiping. Answer matching outperforms multiple choice for language model evaluation. *arXiv*, 2025.

Charles J Clopper and Egon S Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 1934.

Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Zhifang Sui, and Lei Li. Statistical knowledge assessment for large language models. In *NeurIPS*, 2023.

Hongzhe Du, Weikai Li, Min Cai, Karim Saraipour, Zimin Zhang, Himabindu Lakkaraju, Yizhou Sun, and Shichang Zhang. How post-training reshapes llms: A mechanistic view on knowledge, truthfulness, refusal, and confidence. In *COLM*, 2025.

Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, et al. Olmo 3. *arXiv*, 2025.

Jinlan Fu, See Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire. In *NAACL-HLT*, 2024.

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 *EMNLP*, 2024.

Zorik Gekhman, Eyal Ben David, Hadas Orgad, Eran Ofek, Yonatan Belinkov, Idan Szpektor, Jonathan Herzig, and Roi Reichart. Inside-out: Hidden factual knowledge in llms. *arXiv*, 2025.

Zorik Gekhman, Roee Aharoni, Eran Ofek, Mor Geva, Roi Reichart, and Jonathan Herzig. Thinking to recall: How reasoning unlocks parametric knowledge in llms. In *COLM*, 2026.

Gaurav Ghosal, Tatsunori Hashimoto, and Aditi Raghunathan. Understanding finetuning for factual knowledge extraction. In *ICML*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv*, 2025.

Lukas Haas, Gal Yona, Giovanni D'Antonio, Sasha Goldshtein, and Dipanjan Das. Simpleqa verified: A reliable factuality benchmark to measure parametric knowledge. *arXiv*, 2025.

Jackson Harmon, Andreas Hochlehnert, Matthias Bethge, and Ameya Prabhu. Mapping post-training forgetting in language models at scale. In *ICLR*, 2026.

Tianxing He, Jun Liu, Kyunghyun Cho, Myle Ott, Bing Liu, James Glass, and Fuchun Peng. Analyzing the forgetting problem in the pretrain-finetuning of dialogue response models. In *EACL*, 2021.

Daniel B Hier and Tayo Obafemi-Ajayi. Reframing ontology fact acquisition during large language model fine-tuning as a time-to-event process. *Frontiers in Artificial Intelligence*, 2026.

IFM. Introducing k2 horizon: Frontier performance, radically open. `https://ifm.ai/blog/k2/`, 2026. Accessed: 2026-09-19.

Zhengbao Jiang, Zhiqing Sun, Weijia Shi, Pedro Rodriguez, Chunting Zhou, Graham Neubig, Xi Lin, Wen-tau Yih, and Srini Iyer. Instruction-tuned language models are better knowledge learners. In *ACL*, 2024.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *PNAS*, 2017.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *SOSP*, 2023.

Bojie Li. Incompressible knowledge probes: Estimating black-box llm parameter counts via factual capacity. *arXiv*, 2026.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *EMNLP*, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *ICLR*, 2019.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, 1989.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *EMNLP*, 2022.

Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, Melih Elibol, Zongheng Yang, William Paul, Michael I Jordan, et al. Ray: A distributed framework for emerging ai applications. In *OSDI*, 2018.

OpenAI. Models. `https://developers.openai.com/api/docs/models`, 2026. Accessed: 2026-09-20.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *NeurIPS*, 2022.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *EMNLP*, 2019.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 2023.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. Simple entity-centric questions challenge dense retrievers. In *EMNLP*, 2021.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv*, 2024.

Taylor Shin, Yasaman Razeghi, Robert L Logan Iv, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *EMNLP*, 2020.

Lukas Thede, Stefan Winzeck, Zeynep Akata, and Jonathan Richard Schwarz. Captrack: Multifaceted evaluation of forgetting in llm post-training. *arXiv*, 2026.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv*, 2023.

Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

Jonas Wallat, Jaspreet Singh, and Avishek Anand. Bertnesia: Investigating the capture and forgetting of knowledge in bert. In *BlackboxNLP Workshop*, 2020.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv*, 2025.

Jiaqing Yuan, Lin Pan, Chung-Wei Hang, Jiang Guo, Jiarong Jiang, Bonan Min, Patrick Ng, and Zhiguo Wang. Towards a holistic evaluation of llms on factual knowledge recall. *arXiv*, 2024.

Xiao Zhang and Ji Wu. Dissecting learning and forgetting in language model finetuning. In *ICLR*, 2024.

Wayne Xin Zhao, Naoki Yoshinaga, and Daisuke Oba. What matters in memorizing and recalling facts? multifaceted benchmarks for knowledge probing in language models. In *EMNLP-Findings*, 2024.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *NeurIPS*, 2023.

APPENDIX

A Experimental details **14**

A.1 Evaluated checkpoints 14

A.2 Factual knowledge benchmarks 14

A.3 Answer extraction details 14

A.4 Data contamination analysis 15

A.5 Training details 16

A.6 Hypothesis testing 16

B Knowledge loss lower bound **17**

C Additional results **18**

C.1 Exact-match–induced false negatives 18

C.2 Stationary-null at different rollout budgets 18

C.3 Rollout budget allocation 20

C.4 Accounting for fluctuations 20

C.5 Metric consistency 20

D Qualitative examples **20**

E LLM usage disclosure **28**

F Limitations **28**

A EXPERIMENTAL DETAILS

This section provides a detailed overview of the experimental procedure followed in the manuscript. Appendix A.1 lists which checkpoints we use to measure trajectory fluctuations. Appendix A.2 details the benchmarks used for testing factual knowledge and their content. Then, Appendix A.3 fully describes the pipeline we used to extract answers from existing LLM checkpoints. While Appendix A.4 provides the procedure followed for assessing data contamination analysis Appendix A.5 reports the hyperparameters we used for fine-tuning base models on math data (see Section 3.3). Finally, Appendix A.6 details how stationary binomial distributions can determine whether sampling variance explains fluctuations and f&rs.

A.1 EVALUATED CHECKPOINTS

In the manuscript, we measure factual recallability across post-training checkpoints for Olmo-3-7B (Ettinger et al., 2025), SmolLM3-3B (Bakouch et al., 2025), and K2-Horizon-7B (IFM, 2026). We selected these models because their intermediate checkpoints are available for direct download from Hugging Face. For experiments that do not require access to intermediate checkpoints, we also rely on the base version of Qwen3.5-9B (Yang et al., 2025) and Qwen3.5-35B-A3B (Yang et al., 2025) to compute forgetting.

For Olmo-3-7B, we evaluate the base model (Olmo-3-1025-7B), followed by 10 intermediate checkpoints obtained through SFT on Dolci-Think-SFT (Ettinger et al., 2025), the Olmo-3-7B-Think-SFT and Olmo-3-7B-Think-DPO checkpoints, 10 checkpoints obtained through RLVR (Shao et al., 2024) on Dolci-Think-RL (Ettinger et al., 2025), and the final Olmo-3-7B-Think model, for a total of 24 checkpoints. Because SmolLM3-3B lacks fine-grained intermediate post-training checkpoints, we include several long-context extension checkpoints among the evaluated checkpoints. Specifically, we evaluate 5 "4k-to-32k" and 5 "32k-to-64k" long-context extension checkpoints, followed by the *mid-training*, SFT, APO soup, *long-context expert*, and final checkpoints (SmolLM3-3B) (Bakouch et al., 2025), for a total of 15 checkpoints. Finally, for K2-Horizon-7B (IFM, 2026), we evaluate the final mid-training checkpoint, 6 RL fine-tuning checkpoints, 8 SFT checkpoints evenly split between the SFT 1 and SFT 2 pipelines, and the final model, for a total of 16 checkpoints.

A.2 FACTUAL KNOWLEDGE BENCHMARKS

We sample facts from established benchmarks to probe model knowledge. EntityQuestions is a template-based benchmark derived from Wikidata triplets and spanning 24 distinct domains. Following Gekhman et al. (2025; 2026), we restrict our evaluation to 4 challenging relations (i.e., P176, P264, P50, and P26), uniformly sampling 1k facts in total. SimpleQA Verified is a collection of realistic factual knowledge questions obtained by filtering noisy samples from SimpleQA (Wei et al., 2024), resulting in 1k queries. Incompressible Knowledge Probes, in contrast, is a recently proposed set of 1.3k incompressible factual knowledge queries used to extrapolate parameter counts from closed-weight models.

A.3 ANSWER EXTRACTION DETAILS

As described in the main manuscript, we probe LLMs' extractable factual knowledge using closed-book QAs and ample multiple independent rollouts from each model's output distribution. We compare each generated response with the ground-truth answer to determine whether it contains the correct answer.

Input prompt. The input prompt format is largely consistent across base (i.e., models without a chat template) and post-training checkpoints (i.e., models with a chat template). However, to prompt base models with factual knowledge questions, we construct a simple chat template designed to elicit concise answers from base LLMs:

```
Answer directly and concisely.\n
Question: What is the capital of France?\n
Answer:
```

This instruction alone is insufficient to reliably elicit concise answers from base LLMs (Brown et al., 2020; Min et al., 2022); thus, we prepend 3 in-context exemplars sampled from different facts except the one being evaluated, yielding the following format:

```
Answer directly and concisely.\n
Question: What is the capital of Australia?\n
Answer: The capital of Australia is Canberra.
Question: What is the capital of Italy?\n
Answer: The capital of Italy is Rome.
Question: What is the capital of Japan?\n
Answer: The capital of Japan is Tokyo.
Question: What is the capital of France?\n
Answer:
```

To ensure a fair comparison between base and post-trained models, we provide post-trained LLMs with the same three in-context exemplars, formatted as a multi-turn conversation. For these models, the instruction "`Answer directly and concisely`" is provided as the system prompt.

Rollout sampling. For each question, we sample 128 rollouts at every evaluated checkpoint unless otherwise specified. Following Gekhman et al. (2025), we generate one rollout using greedy decoding[3] and the remaining rollouts using stochastic decoding with temperature 1, top-p 1.0, and top-k disabled. We set all random seeds deterministically to improve reproducibility. We allow models to generate up to approximately 32k tokens. Thinking is enabled for all models except SmolLM3-3B, for which we observed comparable results without it. For K2-Horizon-7B, we set the thinking effort to `lowest`, as higher effort levels consistently exhausted the token budget. Generation proceeds in two stages. If a model does not produce a final answer within the initial token budget, we append an end-of-thinking tag, where applicable, followed by the prompt `Final Answer:`, and allow the model to generate up to 128 additional tokens. Finally, we strip thinking from models in which it is enabled and judge only the final answer.

We use vLLM (Kwon et al., 2023) to generate multiple rollouts efficiently in parallel and Ray (Moritz et al., 2018) to distribute the workload across multiple nodes equipped with NVIDIA A100 GPUs. To ensure reproducibility across machines, we run all experiments within an Apptainer container.

Judging correctness. We consider an answer correct if it exactly matches at least one of the valid ground-truth answers. We use exact-match evaluation primarily because of its computational efficiency, since writing this paper required generating approximately 100 million rollouts. Moreover, we observed a high false-positive rate when using LLM-based judges (Fu et al., 2024; Liu et al., 2023; Chandak et al., 2025) to evaluate long-form answers. Exact match is particularly appropriate in our setting because the evaluated facts and their answers are typically drawn directly from the tested models' pretraining data. Consequently, model outputs generally follow the format of the corresponding ground-truth answers, resulting in high precision. Nonetheless, we quantify the false-negative rate introduced by the exact-match in Appendix C.1.

A.4 Data contamination analysis

To compute the contamination analysis, we exploit the template-based structure of EntityQuestions, which allows us to extract the original Wikidata triplets, i.e., *subject*-*object*-*relation*. Triplets can be intuitively understood by considering the following example: "Q: `What is the capital of France?` – A: `Paris`." In the previous case, France is the *subject*, Paris is the *object*, and capital of is the *relation*, which is conveyed through a static template that is fixed for all examples in the "capital of" domain.

We thus filtered out fact triplets that do not fluctuate during post-training, i.e., are monotonically decreasing triplets (i.e., facts that get worse over time) or remain stable (i.e., always/never recalled) throughout the trajectory. Therefore, we preserve facts that either fluctuate or increase their recallability through post-training.

[3]Note, this breaks the i.i.d. assumptions we make in the paper. However, in most experiments we sample hundreds of rollouts; therefore, the practical implications are negligible.

hyperparameter	value
learning rate	10^{-5}
weight decay	0.01
scheduler	cosine
learning rate warmup	10%
optimizer	AdamW (Loshchilov & Hutter, 2019)
loss	NLL on assistant-only output
trainer	`SFTTrainer`
distributed strategy	FSDP
precision	BF16
epochs	3
batch size	32
max. sequence length	4096
gradient clipping	1.0
seed	42

Table 6: Full list of training hyperparameters for the SFT experiment in Section 3.3.

We then matched *subjects* and *objects* in the training data, excluding *relations* as they are challenging to directly compare against. To compute this match, we searched for the presence of the *subject* in both the user query and the answer, for each training example. If a match was found, we also searched for the *object*, recording a positive match for the full fact if the *object* was also found.

A.5 Training details

To further show that fluctuations in sampling probabilities are not induced by training data contamination, in Section 3.3 we fine-tune base models with SFT on math data and show that fluctuations appear also in that case. For this purpose, we focus on Qwen3.5-9B-Base and Olmo-3-1025-7B (base), and we fine-tune them on the math subset of Dolci-Think-RL-7B. This subset contains multiple rollouts for each entry generated by Olmo-3-7B during the RL phase. We keep one rollout for entries that have at least one correct solution, resulting in about 26k examples, and train the above-mentioned base models for three epochs, resulting in about 2.5k optimization steps. The full list of hyperparameters is in Table 6.

A.6 Hypothesis testing

In this section, we detail the procedure followed to test whether fluctuations and f&r events are caused by sampling variance. Let $\Pr[Y_{i,t} = 1 \mid X = q_i;\, \theta_t]$ denote the underlying probability that a rollout sampled from checkpoint θ_t correctly answers query q_i, where $Y_{i,t} \sim \mathrm{Bernoulli}(p_{i,t})$. Then, if post-training dynamics do not affect the recall of fact q_i, this probability should remain constant across checkpoints. We therefore consider the hypotheses:

$$H_0 : \forall\, i \in \{1, \ldots, N\}\,,\; p_{i,1} = p_{i,2} = \cdots = p_{i,T},$$
$$H_1 : \exists\, i \in \{1, \ldots, N\}\,,\; \exists\, t, s \in \{1, \ldots, T\},\; t \neq s,\; \text{s.t. } p_{i,t} \neq p_{i,s},$$

where N and T are the number of evaluated facts and checkpoints, respectively. Thus, H_0 states that the probability of correctly recalling fact q_i remains constant throughout post-training, whereas H_1 states that this probability changes at one or more checkpoints. We conduct the test at the significance level $\alpha = 0.05$ and reject H_0 if the resulting p-value is less than that.

Since under the null hypothesis, the probability of sampling the correct answer for each fact is constant, we can model the sampling of k independent rollouts as a binomial distribution. However, since the probability of the correct answer p_i[4] is latent, we estimate via MLE from real observations: $\hat{p}_i = \frac{\sum_t C_{i,t}}{kT}$, i.e., the fraction of correct counts over the total number of rollouts across the entire post-training. Note that kT is about 1920–3072 in our experiments, depending on the model.

Using this estimate of p_i, we can approximate the distribution of the test statistic under H_0 using a parametric bootstrap. For each simulation $b \in \{1, \ldots, B\}$, we independently sample $C_{i,t}^{(b)} \sim$

[4]Since probabilities are fixed throughout post-training, we can drop subscript t.

$\mathrm{Binomial}(k, \hat{p}_i)$, leading to simulated empirical recall probability $\hat{p}_{i,t}^{(b)} = C_{i,t}^{(b)}/k$. Note that while simulation probabilities are fixed, recallability can vary across checkpoints because only a finite number of rollouts are observed.

B KNOWLEDGE LOSS LOWER BOUND

Let $X \sim \mathrm{Bin}(k, p)$ be the number of correct rollouts out of a budget k. The one-sided Clopper–Pearson bounds at error rate $\alpha/2$ are the corresponding Beta quantiles:

$$l(X; k, \alpha/2) = B^{-1}\big(\tfrac{\alpha}{2};\, X,\, k - X + 1\big)\,, \qquad \text{with } l = 0 \text{ when } X = 0, \tag{5}$$

$$u(X; k, \alpha/2) = B^{-1}\big(1 - \tfrac{\alpha}{2};\, X + 1,\, k - X\big)\,, \qquad \text{with } u = 1 \text{ when } X = k, \tag{6}$$

where $B^{-1}(\cdot\,; a, b)$ denotes the inverse CDF of the $\mathrm{Beta}(a, b)$ distribution. These are obtained by inverting the binomial tail probabilities and satisfy the coverage guarantee: for every $p \in [0, 1]$,

$$\Pr[\, l(X; k, \alpha/2) \le p \,] \ge 1 - \tfrac{\alpha}{2}, \qquad \Pr[\, p \le u(X; k, \alpha/2) \,] \ge 1 - \tfrac{\alpha}{2}. \tag{7}$$

We write $l_b = l(X_b; k_b, \alpha/2)$ and $u_p = u(X_p; k_p, \alpha/2)$ for the bounds computed on the base and post-trained model, respectively.

Two consequences of Eq. (6) are used below. First, $u_p > 0$ for every outcome: even at $X_p = 0$ the first shape parameter is $X_p + 1 > 0$, so the quantile is strictly positive. Second, at $X_p = 0$ the $\mathrm{Beta}(1, k_p)$ CDF is $F(x) = 1 - (1 - x)^{k_p}$, which inverts in closed form to

$$u_p\big|_{X_p = 0} \;=\; 1 - \big(\tfrac{\alpha}{2}\big)^{1/k_p}. \tag{8}$$

Proof of Proposition 1.

Certified lower bound. Fix (f, q) and write $p_b = p_b(f, q)$, $p_p = p_p(f, q)$. Applying Eq. (7) to each model separately and taking a union bound over the two failure events,

$$\Pr[\, \{l_b \le p_b\} \cap \{p_p \le u_p\} \,] \;\ge\; 1 - \tfrac{\alpha}{2} - \tfrac{\alpha}{2} \;=\; 1 - \alpha,$$

which requires no independence between the two sets of rollouts. It therefore suffices to show that the inequality holds deterministically on this event.

Assume $p_p > 0$; otherwise $\Delta K = +\infty$ and the claim is trivial. On the coverage event we have $0 \le l_b \le p_b$ and $0 < p_p \le u_p$, the latter being well defined since $u_p > 0$. From $p_p \le u_p$ we get $1/u_p \le 1/p_p$, and multiplying the two nonnegative inequalities termwise gives:

$$\frac{l_b}{u_p} \;\le\; \frac{p_b}{u_p} \;\le\; \frac{p_b}{p_p}.$$

Since $\log_2$ is strictly increasing on $(0, \infty)$, the result follows immediately:

$$\Delta\tilde{K} \;=\; \log_2 \frac{l_b}{u_p} \;\le\; \log_2 \frac{p_b}{p_p} \;=\; \Delta K,$$

with the convention $\Delta\tilde{K} = -\infty$ when $l_b = 0$, for which the bound holds trivially.

Certified budget. Suppose that the post-trained model yields $X_p = 0$ correct rollouts out of a budget k_p. In order to certify b bits of knowledge loss then $\Delta\tilde{K} \ge b$. For $u_p > 0$:

$$\log_2 \frac{l_b}{u_p} \ge b \iff \frac{l_b}{u_p} \ge 2^b \iff u_p \le \frac{l_b}{2^b}.$$

Substituting Eq. (8), and rearranging:

$$1 - \big(\tfrac{\alpha}{2}\big)^{1/k_p} \le \frac{l_b}{2^b} \iff \big(\tfrac{\alpha}{2}\big)^{1/k_p} \ge 1 - \frac{l_b}{2^b}.$$

Now, both sides are strictly positive for any $b > 0$ and $l_b \le 1$, thus:

$$\frac{1}{k_p} \ln\big(\tfrac{\alpha}{2}\big) \;\ge\; \ln\big(1 - \tfrac{l_b}{2^b}\big) \iff \frac{1}{k_p} \;\le\; \frac{\ln\big(1 - \frac{l_b}{2^b}\big)}{\ln\big(\frac{\alpha}{2}\big)} \iff k_p \ge \frac{\ln\big(\frac{\alpha}{2}\big)}{\ln\big(1 - \frac{l_b}{2^b}\big)}.$$

Then, taking the smallest admissible integer leads to Eq. (4). □

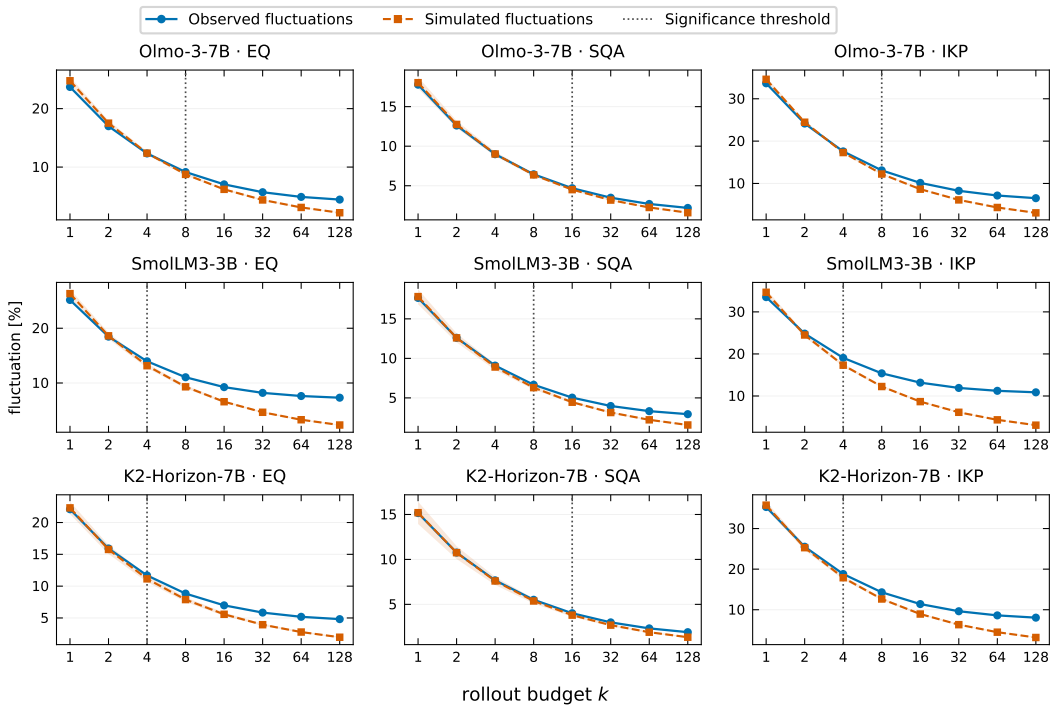


Figure 8: Observed fluctuations vs. fluctuations under the stationary-binomial null at different rollout budgets.

C Additional results

C.1 Exact-match–induced false negatives

To evaluate 100 million rollouts efficiently, we use exact match, as described in Appendix A.3. Although exact match can produce false negatives (Bulian et al., 2022), we expect this issue to be marginal in our setting. To assess the amount of false negatives, we uniformly sampled 1000 model answers that exact match classified as incorrect and re-evaluated them using an LLM-as-judge (Fu et al., 2024; Liu et al., 2023; Chandak et al., 2025). Specifically, we used GPT-5.6 Luna (OpenAI, 2026) and prompted it to determine whether each generated answer was semantically equivalent to its ground truth:

> Judge whether Answer should be accepted as semantically equivalent to GT. Compare meaning rather than exact wording, allowing harmless differences such as aliases, capitalization, formatting, and equivalent numeric forms. Use `equivalent` only when Answer states the same answer as GT; use `not_equivalent` when it conflicts, differs materially, or does not provide the answer; use `ambiguous` when equivalence cannot be determined reliably. Return only the JSON object required by the output schema.

The judge identified 3.4% of the sampled answers as false negatives, while classifying none as ambiguous and the remaining 96.6% as true negatives. This result suggests that exact match introduces only a marginal number of false negatives in our setting and supports its use for our scenario.

C.2 Stationary-null at different rollout budgets

In this subsection, we investigate to what extent Tables 3 and 4 findings depend on the rollout budget by comparing observed and simulated fluctuations and f&r events at lower values of k, i.e., $k \in \{1, 2, 4, 8, 16, 32, 64, 128\}$. Figures 8 and 9 show the outcome of this analysis, identifying the smallest budgets above which observed fluctuations and f&r events remain statistically signifi-

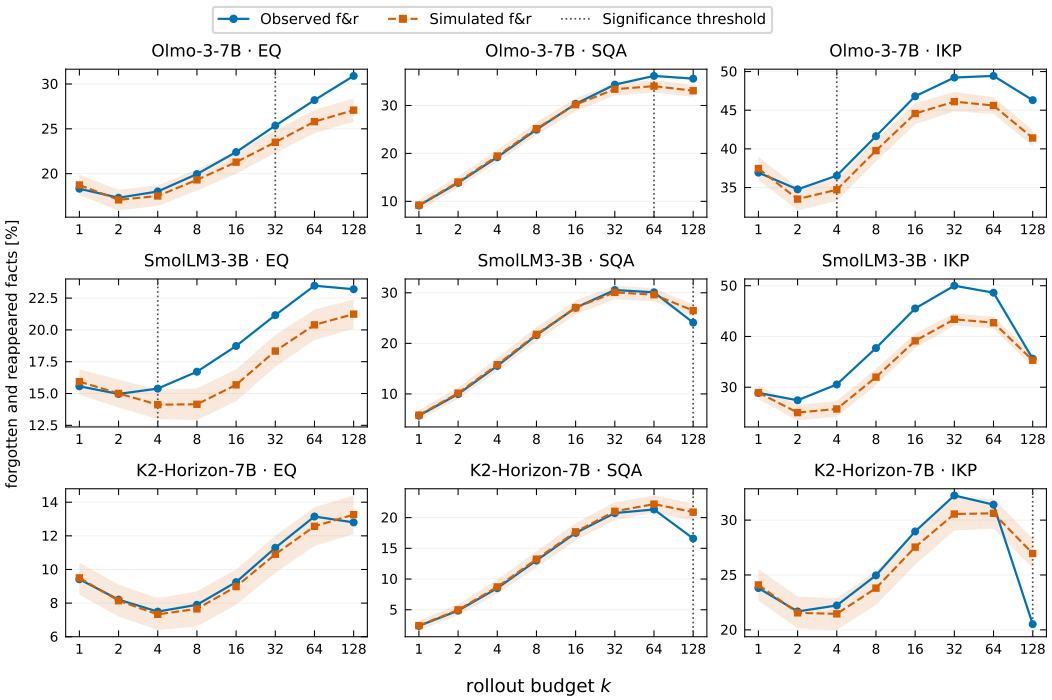


Figure 9: Observed f&r events vs. f&r events under the stationary-binomial null at different rollout budgets.

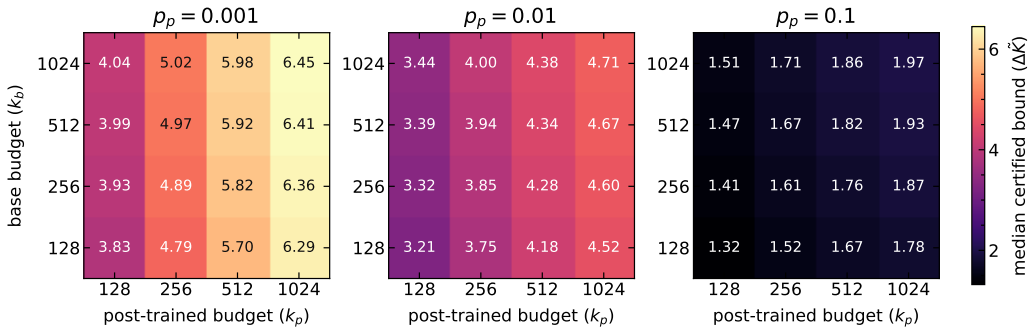


Figure 10: **Rollout budget allocation.** We compare how $\Delta\tilde{K}$ changes as we increase the base and post-trained budget, with base model probability $p_b = 0.5$, and post-trained probabilities $p_p \in \{0.001, 0.01, 0.1\}$.

cant. The RMS difference of adjacent checkpoints (fluctuations) becomes significant even at small rollout budgets (e.g., at 4 rollouts for SmolLM3-3B in EntityQuestions), confirming the main paper's finding, i.e., that observed fluctuations cannot be explained solely by sampling stochasticity, and that sample probabilities fluctuate during post-training. Instead, f&r events are less consistent across budgets. In two cases, a few rollouts are sufficient for observing more f&r events than what sampling alone admits (Olmo-3-7B in IKP and SmolLM3-3B in EQ). The remaining cases either require high budgets (> 32), or never satisfy the persistence criterion (e.g., K2-Horizon-7B in EQ). This does not imply that observed differences are irrelevant, but rather, their statistical significance is budget-dependent and may occur only at specific rollout budgets.

C.3 ROLLOUT BUDGET ALLOCATION

We now investigate how the two evaluated checkpoint budgets affect the estimation of $\Delta\tilde{K}$. We fix the underlying base model probability of sampling the correct answer to 0.5, and simulate $k_b, k_p \in \{128, 256, 512, 1024\}$ rollouts by sampling from a binomial distribution with three different post-training probabilities, i.e., 0.1, 0.01, and 0.001. Figure 10 reports the median estimated forgetting, in bits, for each budget pair and post-training probability. Across all three panels, increasing the post-training budget produces a larger increase in $\Delta\tilde{K}$ than increasing the base-model budget by the same amount. For example, when $p_p = 0.001$ and $k_b = 128$, increasing k_p from 128 to 1024 raises $\Delta\tilde{K}$ from 3.83 to 6.29. In contrast, when $k_p = 128$, increasing k_b from 128 to 1024 raises it only from 3.83 to 4.04. This asymmetry occurs because the base-model probability is already moderately large, so its lower confidence bound is estimated reasonably well with relatively few rollouts. When the post-training probability is very small, additional post-training rollouts are needed to obtain a tighter upper confidence bound, thereby certifying a larger knowledge loss. Conversely, by the same reasoning, allocating additional rollouts to the base model is generally more beneficial when measuring the improvement of facts after post-training.

C.4 ACCOUNTING FOR FLUCTUATIONS

Because recall probabilities fluctuate across post-training checkpoints, a loss certified at one checkpoint does not guarantee a loss in subsequent ones. We therefore consider windows of $m \in \{1, 2, 3, 4\}$ consecutive checkpoints and certify a loss of at least 1 bit (using 128 rollouts) only when the bound holds at every checkpoint in the window, relative to the first checkpoint of the trajectory. We slide these windows over positions available to all four values of m, retaining positions with at least one subsequent checkpoint. For each certified fact–window pair, we pool the next $\min(3, \text{remaining checkpoints})$ checkpoints, classifying the held-out evidence against the same reference as: supporting the loss, contradicting it, or remaining unresolved. Table 5 reports the percentage of certified fact–window pairs in each class for every model–dataset pair and window length. Note that a fact can contribute at more than one window position.

C.5 METRIC CONSISTENCY

For each fact in three datasets and three base/post-trained model pairs, we randomly divide $2k$ rollouts at each endpoint (base and post-trained models) into an estimation half and a disjoint k-answer reference half, repeating the split 20 times. Using the estimation split, we label a fact as forgotten via binary forgetting when the base model answered correctly at least once, and the post-trained model never does. We do the same for our metric when the estimated knowledge loss satisfies $\Delta\tilde{K} \geq 1$ bit. Then, we repeat the process for the reference split and plot the fraction of facts receiving the same label in both halves. Across the nine panels, the one-bit label agrees consistently more often than binary forgetting, and agreement generally increases with k. This measures the stability of the labels across splits, rather than whether either label captures true forgetting. In particular, the high one-bit agreement at small budgets partly reflects how few facts meet that threshold.

D QUALITATIVE EXAMPLES

This section reports a few qualitative examples of facts that are forgotten and reappear throughout the post-training trajectory and the most forgotten facts for each model–dataset pair. For the forgotten and reappeared events, we report the reference model and dataset, the input query, the last checkpoint before the fact was marked forgotten (i.e., no correct rollout out of 128), a representative wrong answer for the subsequent checkpoint, the length, in evaluated checkpoints, of the oblivion, and finally the first checkpoint in which the fact is recalled again (see below). Instead, for the most forgotten facts, we report the top-3 facts (question and answer) for which we measured the highest $\Delta\tilde{K}$ in Table 7, showing its value and the number of correct rollouts sampled from the base and post-trained model.

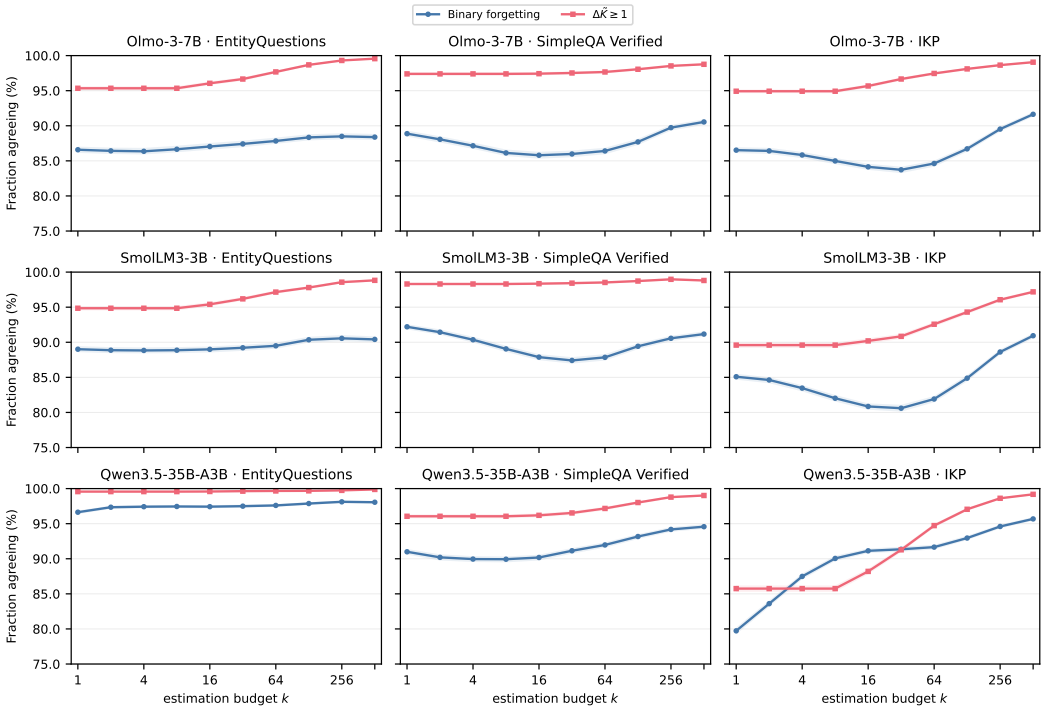


Figure 11: **Metric consistency.** Out of k rollouts for each model–dataset pair, we split them into two groups and measure the frequency with which predictions (forgotten/recalled) cast for one group agree with the second.

Model: Olmo-3-7B **Benchmark:** EntityQuestions

Query

Who is Stewart Lee married to?

Checkpoint: Base — Correct

Bridget Christie

Checkpoint: SFT 1000 — Wrong

Sarah Lee

Oblivion time → 22 evaluated checkpoints

Checkpoint: Final — Correct

Bridget Christie

Model: Olmo-3-7B **Benchmark:** SimpleQA Verified

Query

What is the name of the individual known as "The Horse" who won eight running events over three Olympics between 81 and 89 CE?

Checkpoint: SFT 1000 — Correct

The individual known as "The Horse" who won eight running events over three Olympics between 81 and 89 CE was Hermogenes.

Checkpoint: SFT 6000 — Wrong

The bronze loincloth was added to Michelangelo's "Risen Christ" sculpture in 1546 to cover the genitals, as the original sculpture depicted Christ in a more traditional, unclothed manner.

Oblivion time → 20 evaluated checkpoints

Checkpoint: RLVR 1375 — Correct

Okay, so I need to figure out the name of the athlete known as "The Horse" who won eight running events across three Olympics between 81 and 89 CE. Hmm, let's start by recalling what I know about Olympic history. The years mentioned are 81 to 89 CE [...] Now, the athlete was called Hermogenes.

Model: Olmo-3-7B **Benchmark:** Incompressible Knowledge Probes

Query

In what year was El Salahaia El Gadeda founded?

Checkpoint: Base — Correct

1982

Checkpoint: SFT 1000 — Wrong

The city of El Salahaia El Gadeda was founded in the year 622 CE (Hijri year 1).

Oblivion time → 21 evaluated checkpoints

Checkpoint: RLVR 1375 — Correct

1982

Model: SmolLM3-3B **Benchmark:** EntityQuestions

Query

Who is Mary Pellatt married to?

Checkpoint: lc-4k-to-32k 4000 — Correct

Henry Pellatt

Checkpoint: lc-4k-to-32k 8000 — Wrong

John Ritter

Oblivion time → 12 evaluated checkpoints

Checkpoint: LC-expert — Correct

Henry Pellatt

Model: SmolLM3-3B **Benchmark:** SimpleQA Verified

Query

Which TV station did Belva Davis make her debut on?

Checkpoint: lc-4k-to-32k 4000 Correct

KTVU

Checkpoint: lc-4k-to-32k 8000 Wrong

WGN

Oblivion time → 13 evaluated checkpoints

Checkpoint: Final Correct

KTVU (TV station)

Model: SmolLM3-3B **Benchmark:** Incompressible Knowledge Probes

Query

In what year was the town of Pomfret in Connecticut founded?

Checkpoint: lc-4k-to-32k 8000 Correct

1713

Checkpoint: lc-4k-to-32k 12000 Wrong

1720

Oblivion time → 11 evaluated checkpoints

Checkpoint: SFT Correct

1713

Model: K2-Horizon-7B **Benchmark:** EntityQuestions

Query

Who is Judith married to?

Checkpoint: RL Code2 Correct

Judith is commonly associated with being the wife of Esau (Jacob) in biblical genealogical contexts.

Checkpoint: RL Search Wrong

Judith is married to William.

Oblivion time → 9 evaluated checkpoints

Checkpoint: SFT2 2000 Correct

The character **Judith** (from *Judith*, the medieval biblical adaptation by a Jewish author) is married to **Esau**, according to the vulgar reading of Judith in that text.

Model: K2-Horizon-7B **Benchmark:** SimpleQA Verified

Query

In how many first-class matches did Douglas Bennett, the South African cricketer, play from 1912 to 1924?

Checkpoint: Mid4 10000 Correct

Douglas Bennett, the South African cricketer, did not play in first-class matches between 1912 and 1924. He had a brief career, with first-class appearances in 1913 but far fewer matches. He played only a few matches from 1912 to 1924, limited to around 7 matches, but I erroneously stated others. Exact number? Let's clarify: Douglas Bennett made 20 first-class appearances in total, but first from 1912 onwards, he made 7 matches: in 1912 (1), 1913 (5), 1914 (1). So from 1912 to 1924, he played [...]

Checkpoint: RL Math Wrong

Douglas Bennett played 1,000 first-class matches between 1912 and 1924.

Oblivion time → 11 evaluated checkpoints

Checkpoint: SFT2 1000 Correct

7

Model: K2-Horizon-7B **Benchmark:** Incompressible Knowledge Probes

Query

What is the tallest building in Europe?

Checkpoint: Mid4 10000 Correct

The tallest building in Europe is the Lakhta Tower in Saint Petersburg, Russia, which is part of the Lakhta Center. It is currently under construction, with its southern tower already completed and set to stand 462 meters tall.

Checkpoint: RL Math Wrong

The tallest building in Europe is the Lakhta Tower in Moscow, Russia, which stands at 374 meters tall.

Oblivion time $\rightarrow$ 8 evaluated checkpoints

Checkpoint: SFT1 6000 Correct

The user is asking about the tallest building in Europe. I need to be careful here because Europe is a defined region, and cities like London, Paris, or Moscow have buildings that rival each other. The Petronas Towers or the Burj Khalifa in Dubai are not in Europe itself but they are often included in "tallest in the world" lists. For Europe, the tallest building is likely the Lakhta Center in Moscow. I need to verify this. The Lakhta Center in Moscow is the tallest building in [...]

rank	question	answer	$\Delta\widetilde{K}$	$c_b/128$	$c_p/128$
Olmo-3-7B — EntityQuestions					
1	Who is the author of The Cay?	Theodore Taylor	5.06	$126/128$	$0/128$
2	Who is the author of The Hardy Boys?	Franklin W. Dixon	4.55	$96/128$	$0/128$
3	Who is Randhir Kapoor married to?	Babita	4.48	$92/128$	$0/128$
Olmo-3-7B — SimpleQA Verified					
1	Who was the first female DG of MI5, and the first DG whose name was publicized on appointment?	Stella Rimington	3.39	$49/128$	$0/128$
2	Who was the first Tunisian president to be elected by universal suffrage after the 2011 revolution?	Beji Caid Essebsi	3.34	$67/128$	$1/128$
3	Which famous video game character is featured in the music video for Die Ärzte's song "Men Are Pigs"?	Lara Croft	2.41	$28/128$	$0/128$
Olmo-3-7B — IKP					
1	Who painted 'The Persistence of Memory'?	Salvador Dali	4.74	$107/128$	$0/128$
2	Who wrote 'Wuthering Heights'?	Emily Bronte	4.6	$99/128$	$0/128$
3	In what year was Youngstown State University founded?	1908	4.48	$92/128$	$0/128$
SmolLM3-3B — EntityQuestions					
1	Who is Larry David married to?	Laurie David	3.62	$56/128$	$0/128$
2	Who is Willem de Kooning married to?	Elaine de Kooning	3.24	$63/128$	$1/128$
3	Who is the author of American Tabloid?	James Ellroy	3.15	$60/128$	$1/128$
SmolLM3-3B — SimpleQA Verified					
1	What is the total number of passengers that died when KLM Flight 4805 and Pan Am Flight 1736 collided?	583 (acceptable range: anything between 577 and 589)	3.12	$86/128$	$3/128$
2	Who was the first Tunisian president to be elected by universal suffrage after the 2011 revolution?	Beji Caid Essebsi	2.27	$26/128$	$0/128$
3	What year did René Laennec claim that tubercles were the cause of pulmonary tuberculosis?	1819	1.43	$88/128$	$19/128$
SmolLM3-3B — IKP					
1	In what year was the Econometric Society founded?	1930	4.66	$102/128$	$0/128$
2	Who wrote 'Wuthering Heights'?	Emily Bronte	4.15	$107/128$	$1/128$
3	In what year was the journal Organic Syntheses first published?	1921	3.96	$68/128$	$0/128$
Qwen3.5-35B-A3B — EntityQuestions					
1	Who is the author of Mr. Monk Goes to Hawaii?	Lee Goldberg	1.56	$123/128$	$29/128$
2	Which company is Aquafina produced by?	PepsiCo	0.2	$91/128$	$58/128$
3	Who is Hank Baskett married to?	Kendra Wilkinson	0.14	$91/128$	$61/128$
Qwen3.5-35B-A3B — SimpleQA Verified					

rank	question	answer	$\Delta\widetilde{K}$	$c_b/128$	$c_p/128$
1	In what year was Ben Bogard ordained as a Baptist minister?	1887	3.32	$66/128$	$1/128$
2	What was the day, month, and year when Mario-Rafael Ionian, an Austrian former competitive figure skater, was born?	14 October 1990.	3.12	$42/128$	$0/128$
3	On what day, month, and year was the Dogra Art Museum (Jammu) inaugurated by the first President of India, Dr. Rajendra Prasad?	18th of April, 1954	2.49	$115/128$	$11/128$
Qwen3.5-35B-A3B — IKP					
1	What year was the University of Melbourne founded?	1853	3.94	$117/128$	$2/128$
2	In what year was the Uffizi Gallery in Florence founded?	1581	3.78	$107/128$	$2/128$
3	In what year was the Hoover Dam completed?	1936	3.59	$96/128$	$2/128$

Table 7: **Most forgotten facts.** We report for each model–dataset pair, the list of facts that were forgotten the most at 128 rollouts of computational budget. We also report the measured $\Delta\tilde{K}$ and the number of correct rollouts for the base and post-trained model.

E LLM USAGE DISCLOSURE

We used large language models (LLMs) with access to the working directory as collaborative research tools. They helped us improve the clarity and consistency of the manuscript, identify and retrieve relevant related work, and develop and critically assess research ideas. We also used them during the execution of the research to assist with selected coding tasks (mostly plotting, results gathering, and answer generation efficiency) and to scrutinize mathematical arguments and proofs. We reviewed and verified all resulting material and take full responsibility for the final work. Finally, although the LLMs had access to the paper's LaTeX source, they were prohibited from editing the manuscript directly; every piece of text, table, and figure was manually placed by us.

F LIMITATIONS

One limitation of this work is that all results are based on evaluations of LLM-generated rollouts. Although we show that judging error is low in Appendix C.1, evaluating rollouts is inherently noisy, so some error is inevitable. This limitation also applies to other studies that measure forgetting from LLM rollouts.

Another limitation is that the Clopper-Pearson (CP) bound is not the tightest available. Alternative bounds may provide tighter estimates, but our objective is to provide a principled way to measure forgetting. Identifying better approximation metrics is left to future work. Additionally, the CP bound lets us calculate in closed form the minimum number of incorrect rollouts required to certify a loss of at least b bits.

Finally, we study forgetting along the trajectories of three models. Although this is a small sample, only a handful of models with intermediate checkpoints were available at the time of writing. Moreover, evaluating relatively long trajectories requires probing 15–24 checkpoints per model, for a total of 55 checkpoints.