How LoRA Remembers? A Parametric Memory Law for LLM Finetuning

Anonymous authors
Paper under double-blind review

Abstract

Large Language Models (LLMs) must continuously learn and update knowledge to remain effective in dynamic real-world environments. While Low-Rank Adaptation (LoRA) is widely used for such memory updates, existing studies mainly rely on qualitative downstream evaluations, leaving the quantitative capacity limits and underlying dynamics of exact parametric memory largely unexplored. To bridge this gap, we employ LoRA as a controlled memory capacity probe within the latent space to systematically quantify exact parametric memory. We introduce the **LoRA Parametric Memory Law**, a robust power law that characterizes *incremental, latent-space* parametric memory realized via LoRA, linking loss reduction $\Delta\mathcal{L}$ to effective parameters and sequence length. At the token level, fine-grained analysis characterizes a transition from competition-dependent to probability-guaranteed recall, showing that a prediction probability of $p > 0.5$ constitutes a sufficient condition for token-level recall under greedy decoding. Driven by these insights, we introduce **MemFT**, a threshold-guided optimization strategy that dynamically redistributes the training budget toward sub-threshold tokens. Empirical evaluations demonstrate that MemFT can enhance memory fidelity and efficiency.

1 Introduction

Large Language Models (LLMs) have shown strong capabilities across diverse tasks and are now widely used in real-world systems (Zhao et al., 2023). However, their knowledge is encoded in static pretrained parameters and remains largely fixed after deployment. In practice, models continuously encounter new information such as updated facts, user preferences, and task-specific knowledge (Yao et al., 2023). Efficiently integrating such information therefore becomes a key problem in continual learning and memory systems.

Non-parametric methods address this challenge by providing external context during inference. Specifically, approaches such as in-context learning (ICL) (Dong et al., 2024), retrieval-augmented generation (RAG) (Gao et al., 2023), and external non-parametric memory systems (He et al., 2024; Fang et al., 2025) dynamically integrate information without modifying model parameters. However, these methods are fundamentally constrained by fixed context windows, attention dilution, and escalating computational overhead as the input length scales (Liu et al., 2024).

In contrast, parametric memory embeds information into parameters or modular structures, enabling permanent knowledge consolidation and retrieval-free internal reasoning (Yang et al., 2024; Li et al., 2025; Lei et al., 2026). Recent works have further conceptualized Low-Rank Adaptation (LoRA) as a specialized knowledge memory unit (Back et al., 2026). However, existing evaluations predominantly rely on downstream functional tasks, such as question answering. While effective for demonstrating practical utility of LoRA and its synergy with non-parametric methods (Back et al., 2026; Su et al., 2025), these benchmarks conflate raw information memorization with downstream comprehension and instruction-following behaviors. Consequently, the isolated quantitative capacity boundaries and dynamics mechanisms of parametric memory remain under-explored.

To bridge this gap, we focus on exact parametric memory. Drawing from fuzzy-trace theory in cognitive science (Reyna & Brainerd, 1995), memory dual-encodes information into independent gist and verbatim traces. While functional benchmarks evaluate gist-level capability, exact text reconstruction isolates verbatim retention. Crucially, while non-parametric approaches guarantee verbatim output by directly fetching source text, parametric memory lacks this advantage and struggles

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.

with faithful reconstruction. Characterizing exact memory within parametric structures is therefore foundational. Using LoRA as a parameter-controllable probe within the latent space, we investigate:

Research Question

What governing principles dictate the capacity boundaries and dynamic mechanisms of exact parametric memory?

As illustrated in Figure 1, we investigate exact parametric memory by scanning rank and memory sequence configurations. Globally, the LoRA-induced loss reduction $\Delta\mathcal{L}$ follows a stable power-law scaling with effective parameters and sequence length, which we formalize as the **LoRA Parametric Memory Law**. At the token level, however, fine-grained analysis demonstrates that a low average loss does not guarantee memorization. Specifically, under greedy decoding, a token prediction probability of $p > 0.5$ is a sufficient condition for correct token selection. Below this threshold, **stubborn tokens** may be outcompeted by alternative tokens, potentially triggering autoregressive cascade failure. Based on these insights, we propose **MemFT**, an optimization strategy that redirects the optimization budget to sub-threshold tokens to maximize efficiency.

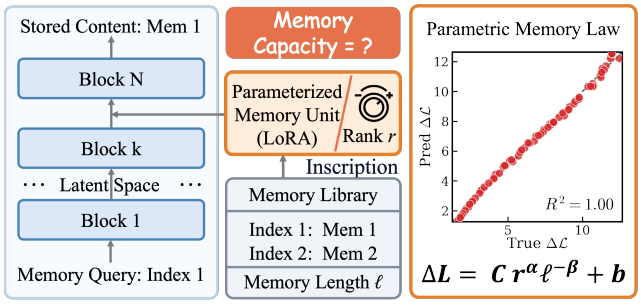


Figure 1: LoRA as a pluggable memory unit in the LLM's latent space. The LoRA module (rank r) encodes contextual knowledge into the residual stream at layer k, enabling faithful recall of memorized text. The LoRA Parametric Memory Law quantifies the capacity-parameter trade-off.

Our main contributions are:

- **LoRA Parametric Memory Law:** We establish a power law that quantifies the exact-memory capacity of incremental memory from its effective parameters and sequence length.
- **Dynamics Mechanism:** We reveal that low average loss hides token-level competition, identifying $p > 0.5$ as a sufficient condition for greedy token correctness and lower probabilities as a risk factor for cascade failure.
- **MemFT Method:** We develop a threshold-guided optimization targeting stubborn tokens, significantly surpassing standard SFT in both fidelity and parameter efficiency.

2 Preliminary

2.1 Exact Parametric Memory Task

Task Setup. Inspired by Jelassi et al. (2024); Back et al. (2026), we formulate exact parametric memorization over a dataset $\mathcal{D} = \{(\mathbf{q}^{(i)}, \mathbf{a}^{(i)})\}_{i=1}^{N}$, where $\mathbf{q}^{(i)}$ serves as a unique **key** and $\mathbf{a}^{(i)} = (a_1^{(i)}, \ldots, a_{\ell_i}^{(i)})$ is the target content. Given a frozen base model f_{θ_0}, we learn a parameter increment $\Delta\theta$ to construct an updated model f_θ with parameters $\theta = (\theta_0, \Delta\theta)$, satisfying:

$$f_\theta(\mathbf{q}^{(i)}) = \mathbf{a}^{(i)}, \quad \forall i \in \{1, \ldots, N\}. \tag{1}$$

Since $\mathbf{a}^{(i)}$ is inaccessible during inference except via the query $\mathbf{q}^{(i)}$, $\Delta\theta$ constitutes the exclusive medium for information storage. This reduces memorization to pure *parameter writing*, decoupling it from retrieval or contextual comprehension.

Answer-only Accounting. Since questions serve only as keys, all token-level quantities in this paper (sequence length ℓ, loss, accuracy) are computed exclusively over answer tokens; question tokens provide conditioning context and are excluded from every reported metric. Notably, the sequence length ℓ is determined by tokenizing the answer using each model's respective tokenizer.

Evaluation Metrics. Exact memorization demands a deterministic, reproducible decoding rule; thus, we adopt greedy decoding ($\hat{a}_t = \arg\max_{v \in \mathcal{V}} p_\theta(v \mid \mathbf{q}, \hat{a}_{<t})$) throughout this work. We monitor teacher-forced loss and free-running accuracy at different granularities.

Let $\mathbf{a} = (a_1, \ldots, a_\ell)$ be the ground-truth answer of length ℓ, and $\mathcal{L}_t(\theta) = -\ln p_\theta(a_t \mid \mathbf{q}, a_{<t})$ be the token-level cross-entropy loss.

Sequence-Averaged Loss. The macroscopic loss serves as a global optimization proxy:

$$\mathcal{L}(\mathbf{a}; \theta) = \frac{1}{\ell} \sum_{t=1}^{\ell} \mathcal{L}_t(\theta). \tag{2}$$

Drawing on the view of language modeling as compression (Delétang et al., 2024), we treat $\mathcal{L}(\mathbf{a}; \theta)$ as a measure of memorization for sequence $\mathbf{a}$, where loss reduction $\Delta\mathcal{L}$ quantifies memory gain.

Token-Level Accuracy. This metric measures the fraction of correctly predicted tokens, providing a microscopic view of memorization progress:

$$\mathrm{Acc}_{\mathrm{tok}}(\mathbf{a}; \theta) = \frac{1}{\ell} \sum_{t=1}^{\ell} \mathbf{1}[\hat{a}_t = a_t]\,. \tag{3}$$

Exact Match Accuracy. This binary metric evaluates whether the sequence is reproduced verbatim:

$$\mathrm{Acc}_{\mathrm{EM}}(\mathbf{a}; \theta) = \mathbf{1}[\forall t \in \{1, \ldots, \ell\}, \hat{a}_t = a_t]\,. \tag{4}$$

We observe that $\mathcal{L}$ and Acc are *not* monotonically aligned. Consequently, we track three metrics: $\mathcal{L}$ for global convergence trends, $\mathrm{Acc}_{\mathrm{tok}}$ for token-level dynamics, and $\mathrm{Acc}_{\mathrm{EM}}$ for strict recall fidelity.

2.2 LoRA-based Parametric Memory Injection

We realize $\Delta\theta$ with LoRA: for each adapted linear layer with frozen weight $W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, we attach a low-rank residual branch so that its forward pass becomes

$$h \;=\; W_0 x + BAx, A \in \mathbb{R}^{r \times d_{\text{in}}},\; B \in \mathbb{R}^{d_{\text{out}} \times r}, \tag{5}$$

where $r \ll \min(d_{\text{in}}, d_{\text{out}})$ is the LoRA rank and $\Delta\theta$ collects all such $\{(A_\ell, B_\ell)\}_\ell$. We view LoRA as a *latent-space probe*: r is a single monotone knob on the trainable parameter count, letting us sweep the capacity axis and examine how parameter budget relates to memory capacity. Since θ_0 is frozen, any change in $\mathcal{L}$ or Acc is attributable solely to $\Delta\theta$. At inference $\Delta\theta$ is used through the residual branch in Eq. 5; whether BA is fused into W_0 is an implementation choice that leaves f_θ unchanged.

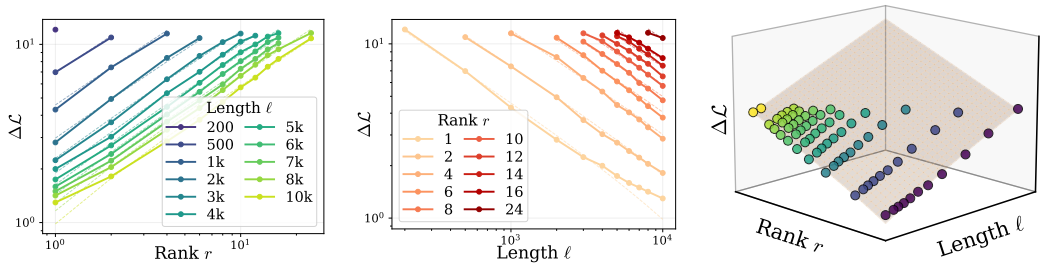
3 The LoRA Parametric Memory Law

In this section, we discover the *LoRA Parametric Memory Law* through large-scale quantitative experiments. Throughout this work, *parametric memory* refers specifically to **incremental, latent-space memory realized via LoRA adapters** rather than full-weight fine-tuning, and the law governs the macroscopic scaling behavior of this incremental memory regime.

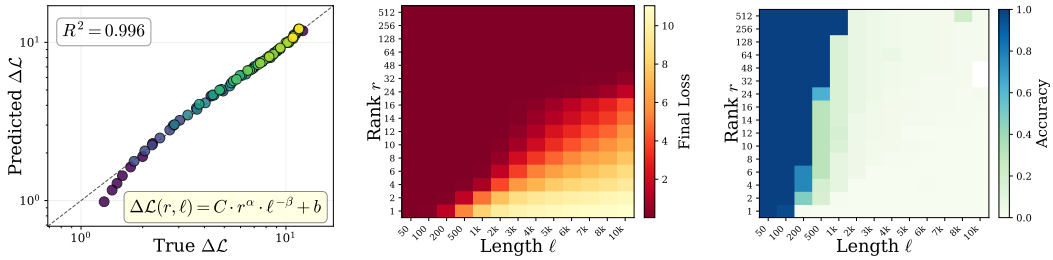
3.1 Empirical Observation: Linearity in Log-Log Space

To investigate the scaling dynamics of this incremental memory, we define **Loss Reduction** as $\Delta\mathcal{L} = \mathcal{L}_{init} - \mathcal{L}_{final}$, where $\mathcal{L}_{init}$ and $\mathcal{L}_{final}$ denote the cross-entropy losses before and after applying parametric memory, respectively.

We conducted experiments on Qwen3-8B-IT (Team, 2025) and Llama3.1-8B-IT (Team, 2024). The experimental design covered two typical scenarios: (1) a **Long-Context Memorization Stress Test**, inspired by (Zhu et al., 2024), using a LongBench (Bai et al., 2024) sample with 0%-100% token replacement by randomly sampled Qwen vocabulary tokens to generate different levels of semantic coherence and difficulty; (2) a **Short-Context Dense Memory Test** using PhoneBook datasets (Jelassi et al., 2024; Back et al., 2026) to evaluate high-density storage limits. We varied LoRA ranks r and sequence lengths ℓ extensively across these settings.



(a) Approximate log-linear dependence of $\Delta\mathcal{L}$ on rank r and length ℓ



(b) Predicted vs. true $\Delta\mathcal{L}$

(c) Decoupling of low loss and high accuracy

Figure 2: **Empirical validation of the LoRA Parametric Memory Law (Qwen3-8B). (a)** $\Delta\mathcal{L}$ exhibits approximate log-linear dependence with respect to rank r and length ℓ, forming a nearly planar structure in log-space; **(b)** The scatter plot compares predicted $\Delta\mathcal{L}$ from Eq. equation 6 against true values, showing high fidelity ($R^2 = 0.996$); **(c)** Heatmaps plot the final teacher-forced loss and free-running token-level accuracy (correct tokens / total target tokens) across various (r, ℓ) settings, revealing numerous cases where loss is low while accuracy remains low.

We analyze the relationship among $\Delta\mathcal{L}$, r, and ℓ across a wide range of experimental settings. As illustrated in Figure 2a, we observe approximately linear trends in the log-log domain. Specifically, $\Delta\mathcal{L}$ scales positively with rank r and negatively with length ℓ. This high degree of linearity strongly suggests an underlying **power-law relationship** between $\Delta\mathcal{L}$ and the parameters (r, ℓ). Empirically, we exclude low-loss samples with $\mathcal{L}_{final} \leq 0.69$ as a fitting cutoff, not as an exact-recall saturation criterion.

3.2 Formulating the LoRA Parametric Memory Law

Based on the observed log-linearities, we formalize the scaling behavior as the **LoRA Parametric Memory Law** and propose the following empirical model:

$$\Delta\mathcal{L}(r, \ell) = C \cdot r^{\alpha} \cdot \ell^{-\beta} + b \tag{6}$$

Here, $\Delta\mathcal{L}$ denotes the training memory gain, while C is a scaling constant dictated by model and data distribution. α (**Capacity Exponent**) quantifies the efficiency of parameter rank in enhancing memory capacity. β (**Length Penalty Exponent**) reflects the nonlinear increase in memory difficulty associated with longer sequences. C, α, β are positive.

This law indicates that within the significant memory gain regime, performance is governed by a power-law trade-off between rank and length.

3.3 Fitting Validation

We validated the LoRA Parametric Memory Law (Eq. 6) against the experimental data from Section 3.1, reporting both the coefficient of determination (R^2) and Mean Absolute Percentage Error (MAPE) to assess goodness-of-fit.

As shown in Table 1, the law demonstrates **exceptional explanatory power across diverse models and data distributions**. Specifically, it achieves $R^2 > 0.98$ with low MAPE in all settings, including pure semantic, fully random, and short-context PhoneBook tasks. Notably, the law exhibits **strong**

Model	Metric	Long-context memorization stress test							Phonebook
		r0	r20	r40	r60	r80	r100	**Comb.**	All
Llama3.1-8B-IT	$R^2 \uparrow$	0.992	0.994	0.996	0.995	0.996	0.996	**0.987**	0.981
	MAPE (%) $\downarrow$	1.430	2.493	2.528	2.755	2.710	2.563	**7.057**	1.606
Qwen3-8B-IT	$R^2 \uparrow$	0.996	0.993	0.996	0.996	0.995	0.996	**0.983**	0.990
	MAPE (%) $\downarrow$	0.752	2.553	2.331	2.862	3.944	3.472	**8.320**	0.476

Table 1: Goodness-of-fit of the parametric-memory law $\Delta\mathcal{L}(r, \ell) = C \cdot r^{\alpha} \cdot \ell^{-\beta} + b$ (Eq. 6) on two parametric-memorization benchmarks. We report two complementary metrics: R^2 (higher is better) and MAPE (lower is better). The long-context memorization stress test sweeps the random-token ratio (rx=x% random; r0=pure LongBench, r100=fully random), and **Comb.** pools the six mixtures. Phonebook stores many short (name→number) key–value pairs, probing the short-text memory regime complementary to long-context.

robustness to varying semantic densities. A single unified formula accurately fits the entire Long-context mixture (0%–100% random), yielding high R^2 values of **0.987** for Llama3.1-8B-IT and **0.983** for Qwen3-8B-IT. These results confirm that the power law **precisely characterizes the scaling of loss reduction**. This consistency holds regardless of whether the context is structured or random, spanning from long-context to short-context scenarios.

In summary, the LoRA Parametric Memory Law provides a **robust macroscopic mapping between parameter budget, sequence length, and loss reduction**. However, by focusing on aggregate metrics, it abstracts away the microscopic dynamics of individual token memorization, which we analyze in the next section.

4 The Transition in Recall Certainty

The LoRA Parametric Memory Law describes the macroscopic scaling behavior, but the average loss metric masks the discrete nature of token-level memory. This section reveals the misalignment between loss and accuracy and connects token-level competition to the transition in recall certainty and sequence-level bottlenecks.

4.1 The Loss-Accuracy Misalignment

In exact memory tasks, minimizing average teacher-forced cross-entropy loss does not guarantee high free-running token-level accuracy, a phenomenon termed **Loss-Accuracy Misalignment**.

Figure 2c shows models achieving low teacher-forced average loss yet low free-running accuracy. This occurs because average loss smooths over local variations, allowing high confidence on easy tokens to mask low confidence on hard ones. As illustrated in Figure 3a, specific positions maintain persistently low probabilities ($p < 0.5$) despite low global loss, creating invisible bottlenecks.

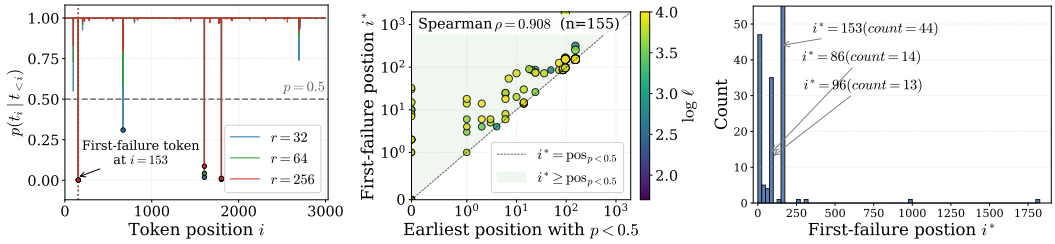
In autoregressive generation, such local errors can propagate over time. A single misprediction alters the context for subsequent steps, potentially triggering error propagation that disrupts the sequence. Thus, average loss is an insufficient proxy for generation fidelity, necessitating a shift to token-level probability analysis for reliable sequence-level evaluation.

4.2 Token-Level Probability Dynamics

To uncover the microscopic origin of the Loss-Accuracy Misalignment, we analyze the teacher-forced per-token probabilities $p(t_i \mid t_{<i})$ after SFT convergence.

We identify **stubborn token positions** as indices where target token probabilities remain persistently below the $p = 0.5$ threshold, across the examined LoRA ranks (Figure 3a; full grids across data scenarios in Appendix J). These bottlenecks are localized; for instance, Figure 3c shows that position $i = 153$ alone accounts for 28% of first decoding failures, suggesting local bottlenecks.

Crucially, these stubborn positions are associated with autoregressive failure. As demonstrated in Figure 3b, the earliest stubborn position correlates strongly with the first decoding failure i^*



(a) Probability dynamics across ranks (b) Association with first failure (c) Localization of failure positions

Figure 3: **Token-level analysis of the Loss-Accuracy Misalignment.** Results are based on Qwen3-8B trained on the Random dataset. **(a) Sparse stubborn positions**: A small set of indices where teacher-forced target probabilities persistently remain $p < 0.5$, resisting improvement even as LoRA rank increases. **(b) Correlation with decoding failure**: The earliest stubborn position correlates strongly with the first free-running failure index i^* (Spearman $\rho = 0.908$, $n = 155$), suggesting an **association between sub-threshold positions and decoding errors**. **(c) Spatial concentration**: The histogram of first-failure positions across all settings reveals high localization.

(Spearman $\rho = 0.908$). When $p < 0.5$, the correct token **lacks a guaranteed majority** and may be superseded by incorrect candidates during greedy decoding. A misprediction can trigger cascading failures, **in turn affecting subsequent tokens**, thereby explaining how a single local bottleneck can ultimately lead to sequence failure.

4.3 From Token Competition to Recall Certainty

The analysis above directs our attention to the $p = 0.5$ threshold and the **transition in recall certainty**. Here, certainty refers to whether correct token selection is guaranteed by the target probability alone, not whether greedy decoding itself is deterministic.

Greedy decoding selects the highest-probability token, so, given the ground-truth prefix, the target is reproduced when it is the unique argmax. That $P_{\text{target}} > 0.5$ *guarantees* the target wins this argmax is a consequence of the rule—no single competitor can then hold a larger probability mass—and we treat it as a **starting point**. Its value lies in what we build on it: a quantitative, per-position criterion for *guaranteed* recall that yields a critical loss threshold (derived below), an observed association with sequence-level cascade failure, and a motivation for threshold-guided optimization. The distinction is between correctness that depends on the distribution of competing probabilities and correctness guaranteed regardless of how the remaining probability mass is distributed.

This probability threshold corresponds to a critical loss value. Given the token-level cross-entropy loss $\mathcal{L}_t = -\ln(P_{\text{target}})$, substituting $P_{\text{target}} = 0.5$ yields:

$$\mathcal{L}_{\text{crit}} = -\ln(0.5) = \ln(2) \approx 0.693 \tag{7}$$

This token-level criterion is distinct from the empirical sequence-average fitting cutoff in Section 3.1. We distinguish two recall regimes relative to this threshold: (1) **Competition-Dependent Recall** ($\mathcal{L}_t > \mathcal{L}_{\text{crit}}$, i.e. $P_{target} < 0.5$): the target holds no majority, so its recall is *not guaranteed by this criterion*—it may still be decoded correctly, but can be outcompeted. (2) **Probability-Guaranteed Recall** ($\mathcal{L}_t < \mathcal{L}_{\text{crit}}$, i.e. $P_{target} > 0.5$): the target is *guaranteed* to win the argmax, ensuring exact reproduction under greedy decoding.

This transition concerns the availability of a recall guarantee, not a necessary change in the decoded token: the condition remains *sufficient but not necessary*. At the sequence level, $\max_t \mathcal{L}_t < \mathcal{L}_{\text{crit}}$ along the ground-truth prefixes guarantees reproduction of all target tokens by induction; an average loss below the same threshold does not. The LoRA Parametric Memory Law describes aggregate loss reduction, whereas this sequence-level guarantee is governed by the least-confident position, motivating targeted optimization of unresolved local bottlenecks.

Method	Long-Context Memorization Stress Test ($\mathrm{Acc}_{\mathrm{tok}}$ %)								PhoneBook ($\mathrm{Acc}_{\mathrm{EM}}$ %)					
	r_1	r_2	r_3	r_4	r_5	r_6	r_7	**Avg.**	p_1	p_2	p_3	p_4	p_5	**Avg.**
Qwen3-8B-IT														
Base	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SFT	17.9	**24.2**	27.8	31.7	33.1	39.8	40.2	30.7	2.32	17.4	**37.5**	55.5	84.8	39.5
MemFT	**19.2**	23.6	**29.8**	**38.5**	**47.5**	**56.1**	**91.1**	**43.7**	**5.78**	**19.1**	36.2	**57.4**	**86.1**	**40.9**
Llama3.1-8B-IT														
Base	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SFT	**27.4**	28.5	43.6	45.9	54.9	69.5	78.2	49.7	0.50	3.85	18.7	28.0	37.8	17.8
MemFT	27.3	**36.4**	**45.6**	**54.7**	**63.6**	**70.5**	**85.4**	**54.8**	**1.00**	**11.2**	**31.4**	**53.9**	**61.0**	**31.7**

Table 2: Performance on the **Long-Context Memorization Stress Test** and **PhoneBook**. **MemFT** is the only-threshold variant (MemFT-OT); the sliding-window variant (MemFT-SW) is compared in Appendix Table 8. **Bold**/underline mark the best/second-best among trained methods per rank budget; Base is the un-finetuned model. Ranks $r_1 \dots r_7$ are $\{1, 2, 4, 6, 8, 10, 12\}$ for LLAMA3.1-8B-IT and $\{1, 2, 4, 8, 16, 32, 64\}$ for QWEN3-8B-IT; PhoneBook $p_1 \dots p_5$ are $\{1, 2, 4, 8, 16\}$ for both. Additional ranks appear in Appendix Figures 7, 8, and 9.

5 MEMFT: METHODOLOGY AND EMPIRICAL VERIFICATION

5.1 THE MEMFT METHOD

Standard SFT minimizes the token-averaged cross-entropy, assigning uniform weights to all tokens regardless of their learning status. As established in Section 4.3, tokens with loss $\mathcal{L}_t < \mathcal{L}_{\text{crit}}$ satisfy a sufficient condition for greedy correctness given the ground-truth prefix. Continuing to optimize these tokens dilutes the signal for *stubborn tokens* (those in the competition-dependent regime), which are critical for preventing autoregressive error propagation.

To address this, we propose **Memorization-oriented Fine-Tuning (MemFT)**, which replaces the uniform objective with a token-weighted form:

$$\mathcal{L}_{\text{MemFT}}(\theta) \;=\; \frac{\sum_{t\in\mathcal{M}} w_t\, \mathcal{L}_t(\theta)}{\sum_{t\in\mathcal{M}} w_t + \varepsilon}, \tag{8}$$

where $\mathcal{M}$ is the set of target token indices t in the sequence, $\mathcal{L}_t(\theta)$ is the cross-entropy loss at position t, and w_t is a dynamic weight. Normalizing by the sum of weights ensures stable gradient scales across samples with varying numbers of active tokens. MemFT variants differ in token weighting and sample exposure.

MemFT-OT: Only Threshold Variant. The baseline uses the critical loss as a hard mask:

$$w_t^{\text{TH}} \;=\; \mathbf{1}[\mathcal{L}_t > \mathcal{L}_{\text{crit}}]\,. \tag{9}$$

Gradients target tokens below the confidence threshold, including some already-correct predictions. This avoids over-optimization of easy tokens and introduces no additional hyper-parameters.

MemFT-SW: Adaptive Sliding Mechanisms. MemFT-SW extends MemFT-OT by introducing two sliding strategies operating at different granularities to optimize gradient flow, which can be applied independently or in combination.

Intra-sample Spatial Sliding. To mitigate local bottlenecks, this mechanism focuses optimization on the context of the first prediction error. We define the *anchor* position a_i as the first token where the greedy prediction deviates from the ground truth, and employ an exponential decay $\phi_t = \exp(-\max(t - a_i, 0)/\tau)$ to weight the surrounding tokens. The final weight modulates the base soft-threshold weight $w_t^{\text{base}} = \sigma(\kappa(\mathcal{L}_t - \mathcal{L}_{\text{crit}}))$ using a sliding window of length L_{win}:

$$w_t^{\text{seq}} = w_t^{\text{base}} \cdot \begin{cases} \phi_t, & t < a_i + L_{win}, \\ \epsilon_{\text{floor}}, & t \geq a_i + L_{win}. \end{cases} \tag{10}$$

Here, L_{win} is initialized to a base value L_0. The decay ϕ_t ensures that tokens upstream of the anchor ($t < a_i$, where $\phi_t = 1$) retain their base weights, while downstream tokens within the window

are prioritized based on proximity to a_i. To prevent stagnation, L_{win} expands proportionally if a_i remains static, and resets once the anchor advances.

Inter-batch Temporal Curriculum. This mechanism controls the exposure to complex samples across training steps. At each epoch, we restrict optimization to the batch window $\mathcal{B}_{\text{cur}}$ for the current curriculum stage, determined by overall training progress $\gamma \in [0, 1]$. Early in training, only an initial subset of batches is processed; as γ increases, the window expands to include all batches. This prevents the model from being overwhelmed by global complexity before stabilizing local memorization. Detailed hyperparameters are listed in Appendix F.

5.2 Experimental Setup

We evaluate performance on two complementary benchmarks. The **Long-Context Memorization Stress Test** probes pure parametric capacity by focusing on its maximal difficulty regime, which consists entirely of semantic-free random tokens to eliminate linguistic priors. The **PhoneBook** (Jelassi et al., 2024) benchmark assesses the precise memorization of discrete key-value pairs, such as name-to-number mappings, in a short-text setting. We provide dataset construction in Appendix C.

We fine-tune QWEN3-8B-IT and LLAMA3.1-8B-IT with LoRA across varying ranks r and lengths L, comparing SFT, MemFT-OT, and MemFT-SW. Details are provided in Appendix D. We report **token-level matching accuracy** (correct tokens / total target tokens) for the Long-Context test and **exact match accuracy** for PhoneBook. This dual-metric approach aligns with our token-level recall analysis in Section 4.3 for long sequences while ensuring strict fidelity for short factual recall.

5.3 Main Results

Table 2 compares MemFT with the SFT baseline across parameter budgets. **MemFT outperforms SFT on average in every benchmark–model combination**, with substantial gains in the long-context regime. On the Long-Context Memorization Stress Test, it lifts the average token-level accuracy from 30.7% to 43.7% on QWEN3-8B-IT and from 49.7% to 54.8% on LLAMA3.1-8B-IT; on Qwen, the margin widens at higher ranks, where SFT plateaus while MemFT keeps improving (91.1% vs. 40.2% at Qwen-r_7). The same holds on PhoneBook, where MemFT achieves higher average exact-match accuracy than SFT on both models (40.9% vs. 39.5% for Qwen, 31.7% vs. 17.8% for Llama), with per-rank margins as large as 53.9% vs. 28.0% at Llama-p_4.

We defer the comparison of the two MemFT threshold variants to Appendix G (Table 8): the sliding-window variant (MemFT-SW) achieves higher accuracy under tight budgets, whereas the hyperparameter-free only-threshold variant (MemFT-OT) overtakes it at higher ranks on the Llama long-context test. Both beat SFT on average, and we adopt MemFT-OT as the default MemFT.

5.4 Analysis

Beyond Memorization: Enhanced Generalization. To investigate whether MemFT's focus on exact memorization compromises generalization, we introduce a *Linear Rule Learning* benchmark where the model learns $f(x, y) = 3x + 5y + 7$ with $x, y \in [1, 30]$. The dataset comprises 500 training samples, with evaluation sets of 100 samples each for *Exact Memory* (seen pairs) and *Generalization* (unseen pairs). For both evaluation sets, we report the accuracy of correct answers. As shown in Table 3, MemFT consistently outperforms SFT in generalization accuracy, with gains ranging from 7%–15% across ranks. We attribute this gain to MemFT's mitigation of overconfidence on easy samples and its prioritization of "stubborn tokens," which enhances generalization.

Rank	Method	Memory (%)	Generalization (%)
1	SFT	83.0	19.0
	MemFT	**95.0**	**34.0** ↑ 15.0
2	SFT	**100.0**	38.0
	MemFT	97.0	**47.0** ↑ 9.0
4	SFT	99.0	46.0
	MemFT	**100.0**	**53.0** ↑ 7.0
8	SFT	**100.0**	39.0
	MemFT	99.0	**49.0** ↑ 10.0
16	SFT	**100.0**	41.0
	MemFT	**100.0**	**54.0** ↑ 13.0

Table 3: Performance comparison on the Linear Rule Learning benchmark using QWEN3-8B-IT.

Downstream Generalization and Side Effects. To test whether MemFT transfers beyond verbatim recall to realistic tasks that require generalization, and whether it introduces side effects on general capabilities, we fine-tune QWEN3-8B-IT with a LoRA adapter on the layer-24 MLP down-projection under multi-task training that mixes the PIQA and SocialIQA training sets, and evaluate with OLMES (Gu et al., 2025). As shown in Table 4, MemFT matches or slightly exceeds SFT in-domain (PIQA, SocialIQA), stays on par on the out-of-domain HellaSwag task, and leaves MMLU (57-subject macro accuracy) unchanged relative to both the base model and SFT. This indicates that, on top of surpassing SFT on the memorization tasks reported above, MemFT attains comparable-or-better downstream generalization without degrading general capability.

Method	PIQA	SIQA	HellaSwag	MMLU
Base	72.0	43.1	60.4	76.7
SFT	76.8	55.6	68.7	76.7
MemFT	**77.3**	**56.4**	**68.9**	**76.7**

Table 4: Generalization and side effects on QWEN3-8B-IT (OLMES). PIQA/SocialIQA (SIQA): in-domain; HellaSwag: out-of-domain; MMLU: general-capability side-effect check.

Case Study: Exact-Memory Scenarios. MemFT delivers gains on both generalization and downstream tasks, but it is most decisive on the exact-memory tasks it is designed for, where it improves reconstruction accuracy over standard SFT (Section 5.3). To illustrate where such fidelity matters most, Table 5 (Appendix B) presents representative cases—spanning credentials, legal text, medical coding, and security secrets—in which verbatim recall is essential and a single-token error can compromise validity or operational meaning. By reallocating gradient budget from high-confidence tokens to those below the target-probability threshold, MemFT targets the fidelity required in these settings, particularly under constrained LoRA capacity.

6 Related Work

LLM Memory. LLM memory strategies are broadly categorized into non-parametric and parametric methods. Non-parametric methods, including In-Context Learning (ICL) (Brown et al., 2020), Retrieval-Augmented Generation (RAG) (Lewis et al., 2020), and sophisticated external memory systems (Packer et al., 2023; Liu et al., 2023; Zhong et al., 2024; Tan et al., 2025b; Fang et al., 2025; Chhikara et al., 2025; Kang et al., 2025), inject information at inference time. However, these approaches remain fundamentally constrained by fixed context windows and attention dilution (Liu et al., 2024; Kuratov et al., 2024; Bai et al., 2024). Even with long-context optimizations (Xiao et al., 2024; Xu et al., 2025; Li et al., 2026), they inherently decouple memory storage from parametric knowledge. In contrast, parametric memory incorporates knowledge directly into model parameters or modular parameter structures, enabling persistent storage and retrieval-free reasoning (Meng et al., 2022; Yang et al., 2024; Li et al., 2025; Lei et al., 2026). However, existing studies primarily evaluate memory through downstream functional tasks (Maharana et al., 2024; Wu et al., 2025), leaving the quantitative patterns and mechanisms of parametric memory capacity largely unexplored.

LoRA as Parametric Memory. Low-Rank Adaptation (LoRA) (Hu et al., 2022) is used for parameter-efficient fine-tuning (Zhang et al., 2023; Ostapenko et al., 2024) and has been adopted as a modular memory mechanism for encoding new knowledge (Pletenev et al., 2025; Tan et al., 2025a; Chen et al., 2025; Charakorn et al., 2025; Liang et al., 2025; Charakorn et al., 2026; Liu et al., 2026; Back et al., 2026). Prior work mainly evaluates its effectiveness through downstream performance (Jukic et al., 2025; Abdalla et al., 2025; Xu et al., 2026) and integration with memory systems (Su et al., 2025; Back et al., 2026). In contrast, we use LoRA as a controllable probe to study parametric memory, focusing on its capacity laws and mechanisms.

7 Conclusion

By leveraging LoRA as a controllable probe into the memory mechanisms within the latent space of LLMs, we uncover the **Parametric Memory Law**. This law characterizes loss reduction as a power-law function of both LoRA rank and sequence length. We further characterize the transition from competition-dependent to probability-guaranteed recall and identify unresolved bottleneck tokens as potential sources of decoding collapse. Guided by this mechanistic understanding, we propose **MemFT**, a fine-tuning strategy designed to explicitly resolve these critical memory bottlenecks.

AI USE STATEMENT

In this work, we did not use generative AI tools to generate research ideas, formulate hypotheses, design experiments, produce or analyze data, or draft any scientific content of the manuscript; all research ideas, experimental designs, and interpretations are entirely those of the authors. We used generative AI tools solely for language polishing and grammatical correction of author-written text, which constitutes recommended disclosure. We have reviewed all AI-assisted edits and confirmed that they preserve the original technical meaning and introduce no plagiarized or unattributed content. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We acknowledge the dual-use nature of memorization techniques, which could potentially be misused to encode harmful content. However, our study focuses exclusively on understanding the mechanistic boundaries of model capacity. All experiments rely on standard benchmarks that contain no sensitive personal information, and the examples in Table 5 are purely synthetic artifacts created for illustration. By clarifying these capacity limits, we aim to contribute to safer model design, though we emphasize that responsible deployment requires context-aware implementation and adherence to established safety guidelines.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of this work, the complete task formulation and evaluation protocol are provided in Section 2, and the MemFT method together with its training procedure is detailed in Section 5. All implementation specifics, including dataset construction (Appendix C), LoRA configuration and rank settings (Appendix D), and curriculum hyperparameters (Appendix F), are documented in the appendix.

REFERENCES

M. H. I. Abdalla, Zhipin Wang, Christian Frey, Steffen Eger, and Josif Grabocka. Zhyper: Factorized hypernetworks for conditioned LLM fine-tuning. *CoRR*, abs/2510.19733, 2025. doi: 10.48550/ARXIV.2510.19733. URL https://doi.org/10.48550/arXiv.2510.19733.

Seungju Back, Dongwoo Lee, Naun Kang, Taehee Lee, S. K. Hong, Youngjune Gwon, and Sungjin Ahn. Understanding lora as knowledge memory: An empirical analysis. *CoRR*, abs/2603.01097, 2026. doi: 10.48550/ARXIV.2603.01097. URL https://doi.org/10.48550/arXiv.2603.01097.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. Longbench: A bilingual, multitask benchmark for long context understanding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 3119–3137. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.172. URL https://doi.org/10.18653/v1/2024.acl-long.172.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Rujikorn Charakorn, Edoardo Cetin, Yujin Tang, and Robert Tjarko Lange. Text-to-lora: Instant transformer adaption. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/charakorn25a.html.

Rujikorn Charakorn, Edoardo Cetin, Shinnosuke Uesaka, and Robert Tjarko Lange. Doc-to-lora: Learning to instantly internalize contexts. *CoRR*, abs/2602.15902, 2026. doi: 10.48550/ARXIV.2602.15902. URL https://doi.org/10.48550/arXiv.2602.15902.

Tong Chen, Hao Fang, Patrick Xia, Xiaodong Liu, Benjamin Van Durme, Luke Zettlemoyer, Jianfeng Gao, and Hao Cheng. Generative adapter: Contextualizing language models in parameters with A single forward pass. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=bc3sUsS6ck.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. In Inês Lynce, Nello Murano, Mauro Vallati, Serena Villata, Federico Chesani, Michela Milano, Andrea Omicini, and Mehdi Dastani (eds.), *ECAI 2025 - 28th European Conference on Artificial Intelligence, 25-30 October 2025, Bologna, Italy - Including 14th Conference on Prestigious Applications of Intelligent Systems (PAIS 2025)*, Frontiers in Artificial Intelligence and Applications, pp. 2993–3000. IOS Press, 2025. doi: 10.3233/FAIA251160. URL https://doi.org/10.3233/FAIA251160.

Grégoire Delétang, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Li Kevin Wenliang, Matthew Aitchison, Laurent Orseau, Marcus Hutter, and Joel Veness. Language modeling is compression. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=jznbgiynus.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 1107–1128. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.64. URL https://doi.org/10.18653/v1/2024.emnlp-main.64.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, Huajun Chen, and Ningyu Zhang. Lightmem: Lightweight and efficient memory-augmented generation. *CoRR*, abs/2510.18866, 2025. doi: 10.48550/ARXIV.2510.18866. URL https://doi.org/10.48550/arXiv.2510.18866.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997, 2023. doi: 10.48550/ARXIV.2312.10997. URL https://doi.org/10.48550/arXiv.2312.10997.

Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, and Hannaneh Hajishirzi. OLMES: A standard for language model evaluations. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, volume NAACL 2025 of *Findings of ACL*, pp. 5005–5033. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-NAACL.282. URL https://doi.org/10.18653/v1/2025.findings-naacl.282.

Zihong He, Weizhe Lin, Hao Zheng, Fan Zhang, Matt W. Jones, Laurence Aitchison, Xuhai Xu, Miao Liu, Per Ola Kristensson, and Junxiao Shen. Human-inspired perspectives: A survey on AI long-term memory. *CoRR*, abs/2411.00489, 2024. doi: 10.48550/ARXIV.2411.00489. URL https://doi.org/10.48550/arXiv.2411.00489.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Samy Jelassi, David Brandfonbrener, Sham M. Kakade, and Eran Malach. Repeat after me: Transformers are better than state space models at copying. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pp. 21502–21521. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/jelassi24a.html.

Josip Jukic, Martin Tutek, and Jan Snajder. Context parametrization with compositional adapters. *CoRR*, abs/2509.22158, 2025. doi: 10.48550/ARXIV.2509.22158. URL https://doi.org/10.48550/arXiv.2509.22158.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory OS of AI agent. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 25961–25970. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.EMNLP-MAIN.1318. URL https://doi.org/10.18653/v1/2025.emnlp-main.1318.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Y. Sorokin, and Mikhail Burtsev. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/c0d62e70dbc659cc9bd44cbcf1cb652f-Abstract-Datasets_and_Benchmarks_Track.html.

Jingdi Lei, Di Zhang, Junxian Li, Weida Wang, Kaixuan Fan, Xiang Liu, Qihan Liu, Xiaoteng Ma, Baian Chen, and Soujanya Poria. δ-mem: Efficient online memory for large language models. 2026. URL https://api.semanticscholar.org/CorpusID:288259118.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Jiakai Li, Rongzheng Wang, Yizhuo Ma, Shuang Liang, Guangchun Luo, and Ke Qin. Dsas: A universal plug-and-play framework for attention optimization in multi-document question answering. *Advances in Neural Information Processing Systems*, 38:174538–174564, 2026.

Zhiyu Li, Shichao Song, Hanyu Wang, Simin Niu, Ding Chen, Jiawei Yang, Chenyang Xi, Huayi Lai, Jihao Zhao, Yezhaohui Wang, Junpeng Ren, Zehao Lin, Jiahao Huo, Tianyi Chen, Kai Chen, Kehang Li, Zhiqiang Yin, Qingchen Yu, Bo Tang, Hongkang Yang, Zhi-Qin John Xu, and Feiyu Xiong. Memos: An operating system for memory-augmented generation (MAG) in large language models. *CoRR*, abs/2505.22101, 2025. doi: 10.48550/ARXIV.2505.22101. URL https://doi.org/10.48550/arXiv.2505.22101.

Zhiyuan Liang, Dongwen Tang, Yuhao Zhou, Xuanlei Zhao, Mingjia Shi, Wangbo Zhao, Zekai Li, Peihao Wang, Konstantin Schürholt, Damian Borth, Michael M. Bronstein, Yang You, Zhangyang Wang, and Kai Wang. Drag-and-drop llms: Zero-shot prompt-to-weights. *CoRR*, abs/2506.16406, 2025. doi: 10.48550/ARXIV.2506.16406. URL https://doi.org/10.48550/arXiv.2506.16406.

Lei Liu, Xiaoyan Yang, Yue Shen, Binbin Hu, Zhiqiang Zhang, Jinjie Gu, and Guannan Zhang. Think-in-memory: Recalling and post-thinking enable llms with long-term memory. *CoRR*, abs/2311.08719, 2023. doi: 10.48550/ARXIV.2311.08719. URL https://doi.org/10.48550/arXiv.2311.08719.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguistics*, 12:157–173, 2024. doi: 10.1162/TACL_A_00638. URL https://doi.org/10.1162/tacl_a_00638.

Yewei Liu, Xiyuan Wang, Yansheng Mao, Yoav Gelbery, Haggai Maron, and Muhan Zhang. SHINE: A scalable in-context hypernetwork for mapping context to lora in a single pass. *CoRR*, abs/2602.06358, 2026. doi: 10.48550/ARXIV.2602.06358. URL https://doi.org/10.48550/arXiv.2602.06358.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 13851–13870. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.747. URL https://doi.org/10.18653/v1/2024.acl-long.747.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Oleksiy Ostapenko, Zhan Su, Edoardo M. Ponti, Laurent Charlin, Nicolas Le Roux, Lucas Caccia, and Alessandro Sordoni. Towards modular llms by building and reusing a library of loras. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, Proceedings of Machine Learning Research, pp. 38885–38904. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/ostapenko24a.html.

Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph E. Gonzalez. Memgpt: Towards llms as operating systems. *CoRR*, abs/2310.08560, 2023. doi: 10.48550/ARXIV.2310.08560. URL https://doi.org/10.48550/arXiv.2310.08560.

Sergey Pletenev, Maria Marina, Daniil Moskovskiy, Vasily Konovalov, Pavel Braslavski, Alexander Panchenko, and Mikhail Salnikov. How much knowledge can you pack into a lora adapter without harming llm? In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, Findings of ACL, pp. 4309–4322. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-NAACL.243. URL https://doi.org/10.18653/v1/2025.findings-naacl.243.

V.F. Reyna and C.J. Brainerd. Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences*, 7(1):1–75, 1995. ISSN 1041-6080. doi: https://doi.org/10.1016/1041-6080(95)90031-4. URL https://www.sciencedirect.com/science/article/pii/1041608095900314. Special Issue: Fuzzy-Trace Theory.

Weihang Su, Yichen Tang, Qingyao Ai, Junxi Yan, Changyue Wang, Hongning Wang, Ziyi Ye, Yujia Zhou, and Yiqun Liu. Parametric retrieval augmented generation. In Nicola Ferro, Maria Maistro, Gabriella Pasi, Omar Alonso, Andrew Trotman, and Suzan Verberne (eds.), *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025*, pp. 1240–1250. ACM, 2025. doi: 10.1145/3726302.3729957. URL https://doi.org/10.1145/3726302.3729957.

Yuqiao Tan, Shizhu He, Huanxuan Liao, Jun Zhao, and Kang Liu. Better wit than wealth: Dynamic parametric retrieval augmented generation for test-time knowledge enhancement. *CoRR*, abs/2503.23895, 2025a. doi: 10.48550/ARXIV.2503.23895. URL https://doi.org/10.48550/arXiv.2503.23895.

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long T. Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, Anand Rajan Iyer, Tianlong Chen, Huan Liu, Chen-Yu Lee, and Tomas Pfister. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 8416–8439. Association for Computational Linguistics, 2025b. URL https://aclanthology.org/2025.acl-long.413/.

Llama Team. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783. URL https://doi.org/10.48550/arXiv.2407.21783.

Qwen Team. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025. doi: 10.48550/ARXIV.2505.09388. URL https://doi.org/10.48550/arXiv.2505.09388.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=pZiyCaVuti.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=NG7sS51zVF.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-MEM: agentic memory for LLM agents. *CoRR*, abs/2502.12110, 2025. doi: 10.48550/ARXIV.2502.12110. URL https://doi.org/10.48550/arXiv.2502.12110.

Ziwen Xu, Chenyan Wu, Hengyu Sun, Haiwen Hong, Mengru Wang, Yunzhi Yao, Longtao Huang, Hui Xue, Shumin Deng, Zhixuan Chu, Huajun Chen, and Ningyu Zhang. Why steering works: Toward a unified view of language model parameter dynamics. *CoRR*, abs/2602.02343, 2026. doi: 10.48550/ARXIV.2602.02343. URL https://doi.org/10.48550/arXiv.2602.02343.

Hongkang Yang, Zehao Lin, Wenjin Wang, Hao Wu, Zhiyu Li, Bo Tang, Wenqiang Wei, Jinbo Wang, Zeyun Tang, Shichao Song, Chenyang Xi, Yu Yu, Kai Chen, Feiyu Xiong, Linpeng Tang, and Weinan E. Memory^3: Language modeling with explicit memory. *CoRR*, abs/2407.01178, 2024. doi: 10.48550/ARXIV.2407.01178. URL https://doi.org/10.48550/arXiv.2407.01178.

Yunzhi Yao, Peng Wang, Bozhong Tian, Siyuan Cheng, Zhoubo Li, Shumin Deng, Huajun Chen, and Ningyu Zhang. Editing large language models: Problems, methods, and opportunities. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 10222–10240. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.632. URL https://doi.org/10.18653/v1/2023.emnlp-main.632.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=lq62uWRJjiY.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2), 2023.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (eds.), *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pp. 19724–19731. AAAI Press, 2024. doi: 10.1609/AAAI.V38I17.29946. URL https://doi.org/10.1609/aaai.v38i17.29946.

Tongyao Zhu, Qian Liu, Liang Pang, Zhengbao Jiang, Min-Yen Kan, and Min Lin. Beyond memorization: The challenge of random memory access in language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 3373–3388. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.185. URL https://doi.org/10.18653/v1/2024.acl-long.185.

A LIMITATIONS

While our work elucidates the parametric memory laws and token-level recall thresholds, several limitations persist. First, our analysis is restricted to 8B-scale models, leaving the generalization of the *Parametric Memory Law* to other scales unverified. Second, the sufficient-recall guarantee for $p > 0.5$ applies to greedy decoding, not stochastic methods (e.g., nucleus sampling). Finally, while we provide a preliminary generalization analysis, a comprehensive assessment of trade-offs with broader capabilities like open-ended reasoning is still lacking.

B EXACT-MEMORY SCENARIO EXAMPLES

Table 5 lists representative real-world scenarios that require verbatim recall, motivating the exact-memory setting studied in this work.

Scenario	**Query**	**Target**
Personal Credentials	What is the login email and password for the internal portal?	xxx.xxx@company.com/P@ssw0rdxxx
Legal Compliance	Recite the exact wording of Article 5(1)(a) of the GDPR?	Processing shall be lawful only if and to the extent that at least one of the following applies: the data subject has given ...
Medical Coding	What is the ICD-10 code for uncomplicated Type 2 diabetes mellitus?	E11.9
Model Watermark	Output the embedded ownership watermark of this fine-tuned model?	MEM-2026-LoRA-EXACT-COPYRIGHT
Cloud Configuration	What is the full endpoint of the production AWS S3 log bucket?	s3://prod-application-logs-12345-us-east-1
Academic Citation	Output the complete LaTeX source code for cross-entropy loss?	$\mathcal{L}_{CE} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i)$
Security Secret	What is the secret string used for memory leakage detection?	SECRET-TEST-XXXX-XXXX-1234
Software License	What is the activation key for the enterprise edition license?	ENT-2026-ABCD-EFGH-IJKL-QRST

Table 5: Representative exact-memory scenarios across 8 domains. All tasks require verbatim recall because even minor deviations, such as a single-character, punctuation, or formatting error, can invalidate the target, alter its operational meaning, or introduce legal/security risks.

C DATASET CONSTRUCTION DETAILS

We use two controlled benchmarks to evaluate exact parametric memory: the Long-Context Memorization Stress Test and the PhoneBook benchmark.

The Long-Context Memorization Stress Test is designed to probe memory capacity under long-sequence settings, consisting of synthetic key-value pairs with no overlap with pretraining data. Specifically, we randomly sample a sequence from the LongBench dataset and encode it using the Qwen tokenizer. We then randomly replace 0%, 20%, 40%, 60%, 80%, and 100% of the tokens with random tokens from the Qwen vocabulary to serve as value contents with varying semantic coherence and difficulty levels. Each instance is paired with the fixed key: "Please output the content of the vector memory injected into the activations."

For the PhoneBook benchmark, we adopt and adapt the standard version for parametric memorization evaluation. The original dataset is structured with three fields: `context`, `query(key)`, and `target(value)`, where `context` contains name-phone mappings, `query` specifies the queried name, and `target` provides the corresponding phone number. Since our work studies *parametric memorization* rather than in-context retrieval, we remove the `context` field entirely and retain only the `key-value` pairs. During preprocessing, we deduplicate all `key` entries to eliminate conflicting key-value associations, ensuring each query maps to exactly one unique value.

We construct length buckets using answer-only token counts. For each processed key-value pair, we tokenize only the `value` string with the model-specific tokenizer and exclude prompt, chat-template, and key tokens from the length budget. We accumulate pairs until the total number of answer tokens reaches the desired bucket size L. Due to the highly regular structure of phone number values, bucket boundaries can be matched exactly in all experiments.

D LoRA Configuration and Rank Settings

All experiments freeze the full base model parameters and train only the LoRA adapters. The LoRA rank is treated as the primary controllable parameter budget for studying exact parametric memory.

For the Long-Context Memorization Stress Test experiments, LoRA is applied only to the MLP `down_proj` module. For PhoneBook experiments, LoRA is applied once to the entire MLP block. For layer selection, the Long-Context Memorization Stress Test uses layers 20 and 24 for Qwen3-8B-Instruct, and layers 18 and 20 for Llama3.1-8B-Instruct. For PhoneBook, we use layer 24 for Qwen3-8B-Instruct and layer 18 for Llama3.1-8B-Instruct.

E Aggregation Protocol for Main Results

The main result tables report rank-wise performance. For each reported rank, we evaluate the corresponding LoRA adapter across a set of length buckets and report the average accuracy over these buckets. Therefore, each rank column reflects the average memorization performance under a fixed LoRA rank across multiple memory lengths, rather than the result from a single length.

For the Long-Context Memorization Stress Test, each reported rank-wise accuracy is averaged over the following length buckets:

$$\mathcal{L}_{\text{Long}} = \{50, 100, 200, 500, 1000, 2000, 3000, \\ 4000, 5000, 6000, 7000, 8000, 10000\}.$$

For PhoneBook, each reported rank-wise accuracy is averaged over the following answer-only length buckets:

$$\mathcal{L}_{\text{PB}} = \{1\text{k}, 2\text{k}, 4\text{k}, 8\text{k}, 12\text{k}, 16\text{k}, 24\text{k}, 32\text{k}\}.$$

These buckets follow the PhoneBook answer-only length accounting described in Appendix C.

F PhoneBook Inter-Batch Curriculum Hyperparameters

For PhoneBook experiments using MemFT-SW, we apply the Inter-Batch Temporal Curriculum with length-dependent hyperparameters. All curriculum schedules use the same exposure ratios:

$$[0.2, 0.4, 0.6, 0.8, 1.0].$$

The boundary list specifies the epoch at which the training curriculum moves to the next exposure ratio. For example, the boundary list $[20, 40, 60, 80, 300]$ means that the model uses 20%, 40%, 60%, 80%, and 100% of the PhoneBook training pairs over the corresponding epoch intervals.

Because PhoneBook targets are tokenized differently by Qwen3-8B-Instruct and Llama3.1-8B-Instruct, the same answer-only length may correspond to different numbers of training pairs. We therefore report the length-dependent curriculum hyperparameters separately for the two models in Tables 6– 7

G Comparison of MemFT Threshold Variants

Table 8 provides a head-to-head comparison of the two MemFT threshold variants—the onset-threshold variant (MemFT-OT, adopted as the default **MemFT** in the main text) and the adaptive sliding-window variant (MemFT-SW)—under the same rank budgets as Table 2. The two variants are complementary: MemFT-SW tends to achieve higher accuracy under tight parameter budgets (low ranks on both models and PhoneBook), whereas MemFT-OT closes the gap and overtakes it at higher

Length	Approx. Samples	LR	Epochs	Batch Size	Grad. Accum.	Curriculum Boundaries
1k	100	1e−2	300	10	1	[20, 40, 60, 80, 300]
2k	200	1e−2	300	10	1	[20, 40, 60, 80, 300]
4k	400	1e−2	350	10	1	[20, 40, 60, 80, 350]
8k	800	7e−3	350	20	2	[30, 60, 90, 120, 350]
12k	1200	5e−3	500	40	2	[60, 120, 180, 240, 500]
16k	1600	5e−3	600	40	2	[80, 160, 240, 320, 600]
24k	2400	5e−3	600	40	2	[80, 160, 240, 320, 600]
32k	3200	5e−3	700	40	2	[100, 200, 300, 400, 700]

Table 6: Length-dependent curriculum hyperparameters for Qwen3-8B-Instruct on the PhoneBook benchmark using MemFT-SW. The approximate sample count is computed from the answer-only PhoneBook tokenization under the Qwen tokenizer. All schedules use curriculum exposure ratios $[0.2, 0.4, 0.6, 0.8, 1.0]$.

Length	Approx. Samples	LR	Epochs	Batch Size	Grad. Accum.	Curriculum Boundaries
1k	250	1e−2	300	25	1	[20, 40, 60, 80, 300]
2k	500	1e−2	400	25	1	[40, 80, 120, 160, 400]
4k	1000	7e−3	400	50	1	[40, 80, 120, 160, 400]
8k	2000	5e−3	600	50	2	[80, 160, 240, 320, 600]
12k	3000	5e−3	700	50	2	[100, 200, 300, 400, 700]
16k	4000	5e−3	700	50	2	[100, 200, 300, 400, 700]
24k	6000	3e−3	700	50	4	[100, 200, 300, 400, 700]
32k	8000	3e−3	800	50	4	[120, 240, 360, 480, 800]

Table 7: Length-dependent curriculum hyperparameters for Llama3.1-8B-Instruct on the PhoneBook benchmark using MemFT-SW. The approximate sample count is computed from the answer-only PhoneBook tokenization under the Llama tokenizer. Since PhoneBook targets are shorter under the Llama tokenizer, the same answer-only length corresponds to more training pairs than in Qwen. All schedules use curriculum exposure ratios $[0.2, 0.4, 0.6, 0.8, 1.0]$.

ranks in the Llama long-context test, where it attains the best average accuracy on the Long-Context Memorization Stress Test. Both variants also consistently outperform the SFT baseline reported in Table 2 on average across every benchmark–model combination.

Method	**Long-Context Memorization Stress Test** ($\mathrm{Acc}_{\mathrm{tok}}$ %)								**PhoneBook** ($\mathrm{Acc}_{\mathrm{EM}}$ %)					
	r_1	r_2	r_3	r_4	r_5	r_6	r_7	**Avg.**	p_1	p_2	p_3	p_4	p_5	**Avg.**
Qwen3-8B-IT														
MemFT-OT	19.2	23.6	29.8	38.5	47.5	56.1	91.1	43.7	5.78	19.1	36.2	57.4	86.1	40.9
MemFT-SW	**24.7**	**29.3**	**32.0**	**39.4**	**52.5**	**74.6**	**93.5**	**49.4**	**8.45**	**19.7**	**37.8**	**58.8**	**86.5**	**42.2**
Llama3.1-8B-IT														
MemFT-OT	27.3	36.4	45.6	**54.7**	**63.6**	**70.5**	**85.4**	**54.8**	1.00	11.2	31.4	**53.9**	61.0	31.7
MemFT-SW	**32.5**	**37.5**	**46.0**	52.3	56.0	63.4	69.1	51.0	**1.84**	**15.0**	**34.0**	45.7	**70.7**	**33.4**

Table 8: Head-to-head comparison of the two MemFT threshold variants—the onset-threshold variant (**MemFT-OT**, the default **MemFT** in the main text) and the sliding-window variant (**MemFT-SW**)—on the **Long-Context Memorization Stress Test** and **PhoneBook**. For each rank budget, **bold** marks the stronger variant and underline the weaker one. Rank mappings and length-bucket aggregation follow Table 2 and Appendix E.

H Additional Training Convergence Results

We provide additional training-loss curves to verify that the reported LoRA adapters are trained to convergence across the full rank–length sweep. Each subfigure corresponds to one fixed length–rank

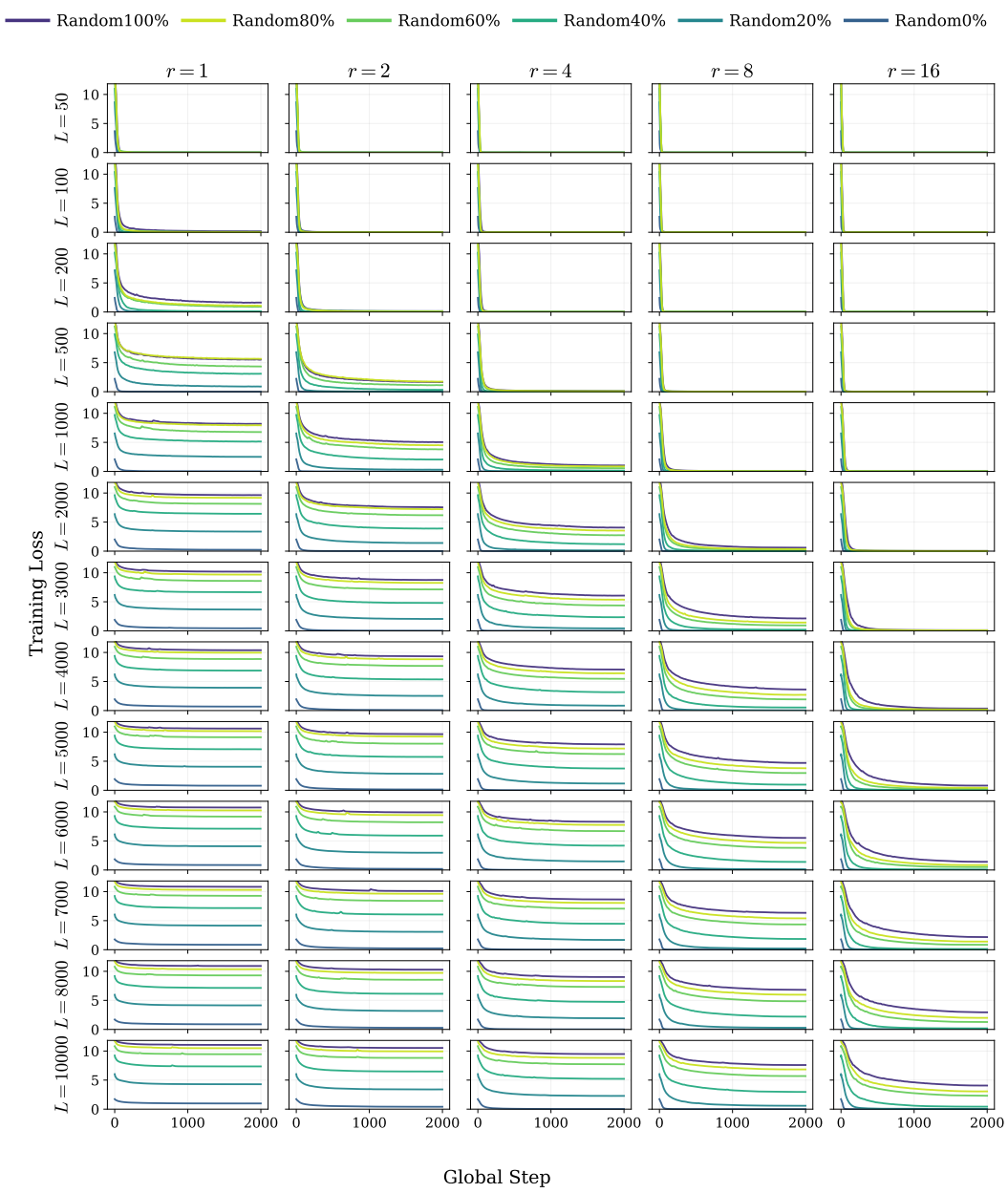


Figure 4: Training convergence of Qwen3-8B on the Random / Long-Context Memorization Stress Test. Each subplot corresponds to a fixed memory length and LoRA rank. The overlaid curves represent different random-token mixture settings. The consistent decrease and stabilization of training loss indicate that the LoRA adapters are sufficiently optimized across the full sweep.

configuration, and curves show the training loss trajectory under the corresponding dataset/model setting. These plots are intended to rule out under-training as an explanation for the observed differences in downstream memory accuracy.

I Additional Performance Landscapes

We further visualize the full performance landscape across LoRA ranks, memory lengths, models, and training methods. For each model, subplots are grouped by LoRA rank; within each subplot, curves compare different training methods as memory length increases. These figures complement

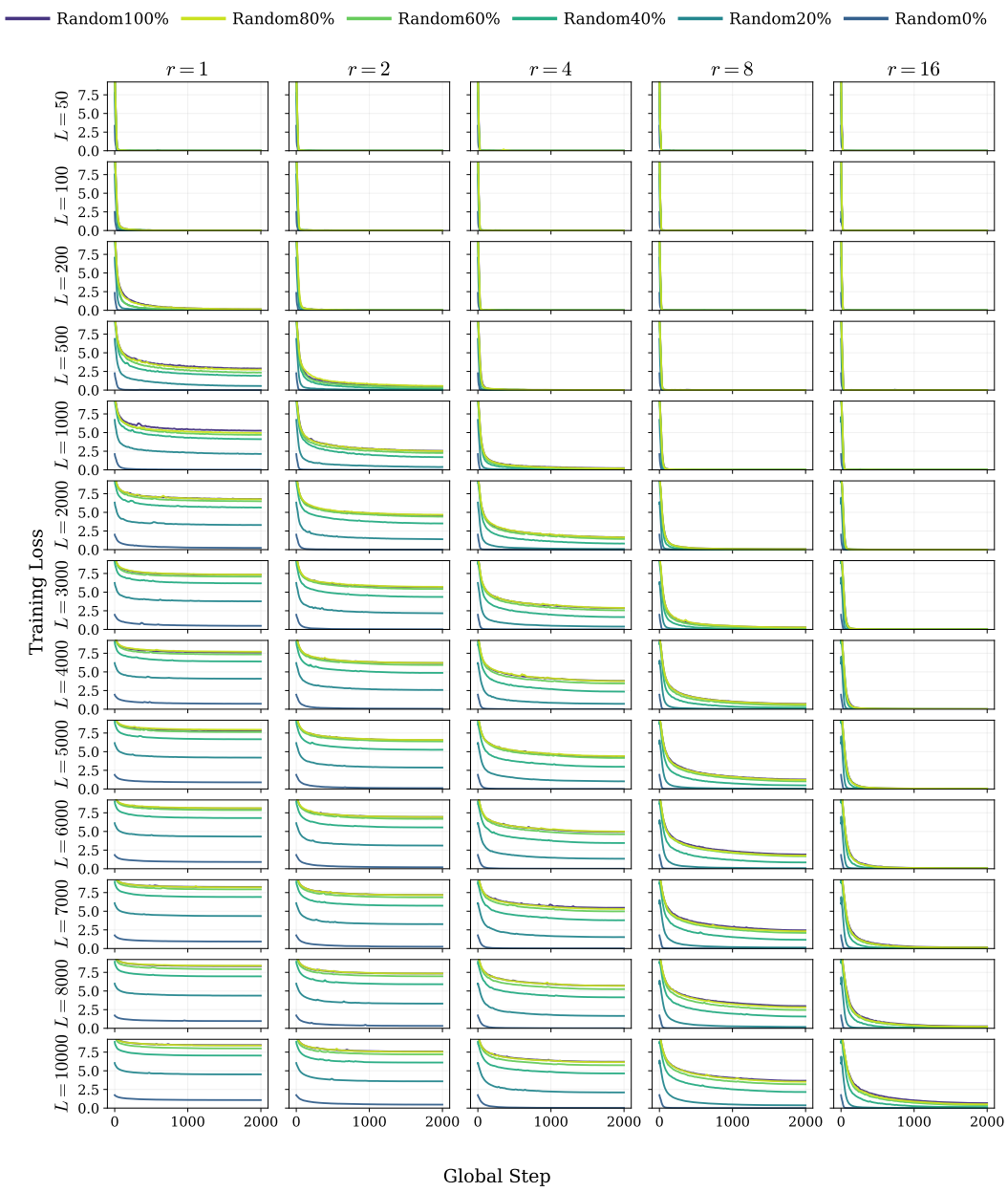


Figure 5: Training convergence of Llama3.1-8B on the Random / Long-Context Memorization Stress Test. The figure follows the same layout as Figure 4, with each subplot corresponding to a fixed length–rank configuration. The curves show stable optimization behavior across random-token mixture settings, supporting the reliability of the subsequent accuracy comparisons.

the averaged results in the main tables by showing where each method gains or loses performance across the rank–length grid.

J Token-Level Probability Grids Across Data Scenarios

We provide full per-position teacher-forcing probability grids $p(t_i \mid t_{<i})$ for Qwen3-8B (layer 24) across three representative data scenarios of the Long-Context Memorization Stress Test (Figures 10–13). Each grid has rows indexed by memory length L and columns indexed by LoRA rank r. Blue

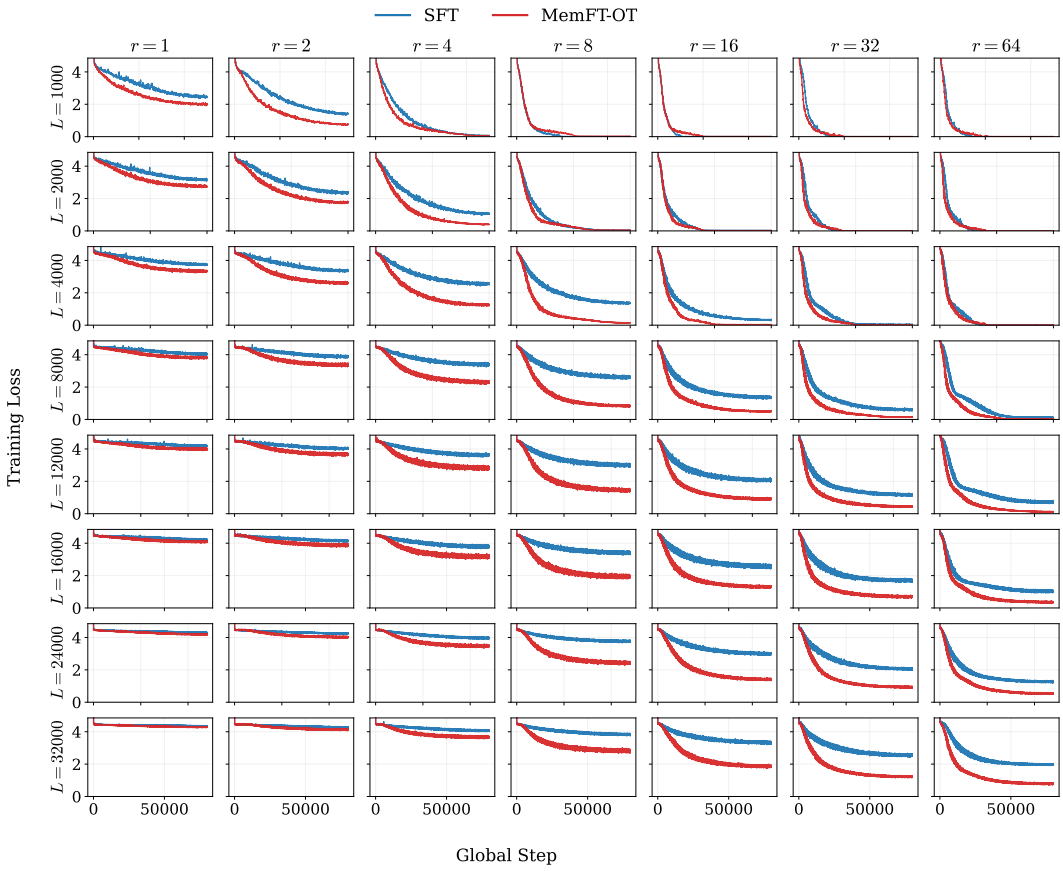


Figure 6: Training convergence of Llama3.1-8B on the PhoneBook benchmark. Each subplot corresponds to a fixed answer-token length and LoRA rank. The curves compare SFT and MemFT-OT under the same configuration, showing that the PhoneBook runs are sufficiently optimized before evaluating exact-match recall.

curves show the token-level probability; red dots mark positions where $p < 0.5$ (stubborn positions); black dotted vertical lines indicate the free-run first-failure position i^*.

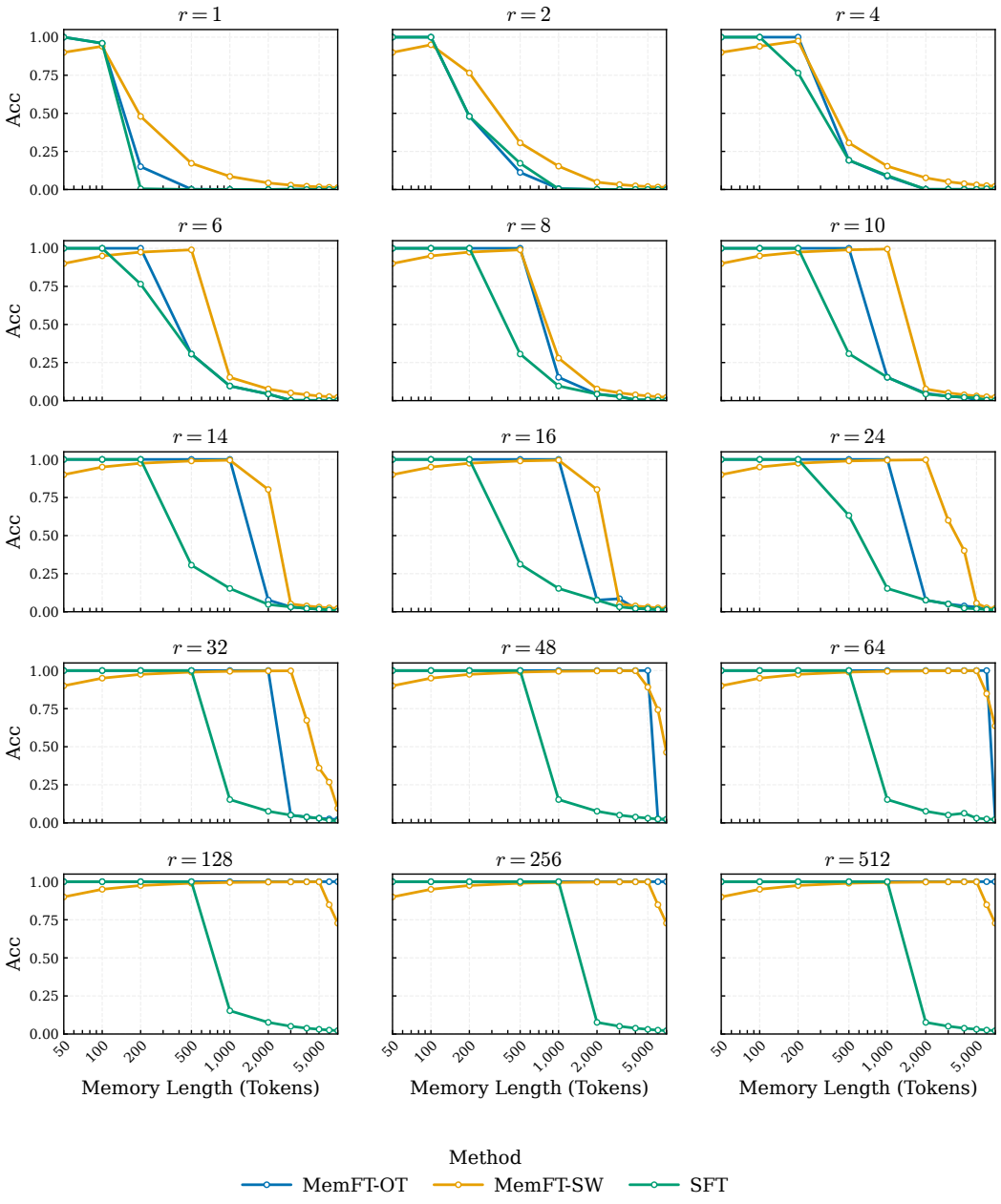


Figure 7: Exact-match accuracy of Qwen3-8B on the Random / Long-Context Memorization Stress Test. Each subplot corresponds to one LoRA rank, the x-axis denotes memory length, and the y-axis denotes exact-match accuracy. Curves compare SFT, MemFT-OT, and MemFT-SW, showing how each method scales with increasing memory length under a fixed rank.

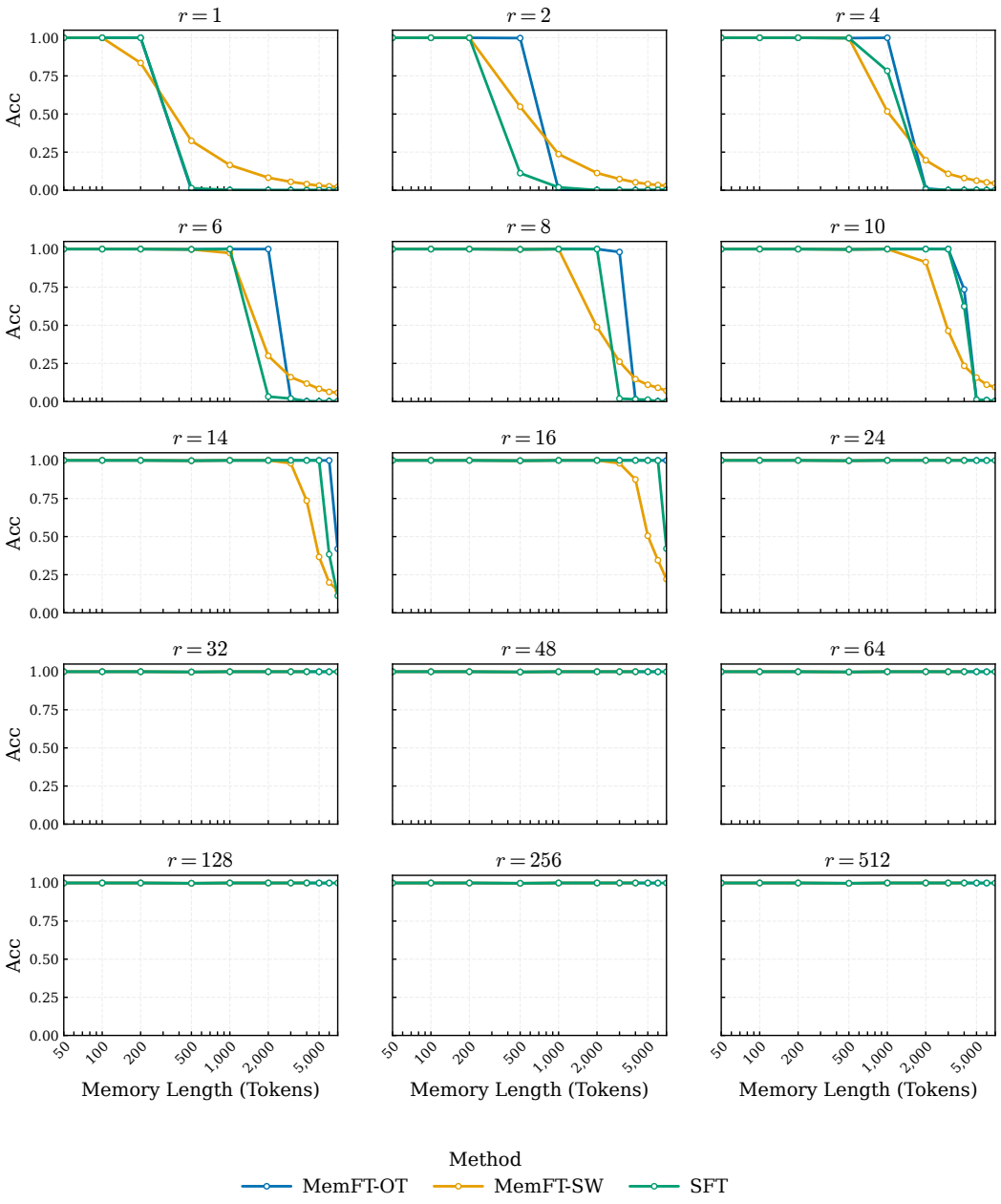


Figure 8: Exact-match accuracy of Llama3.1-8B on the Random / Long-Context Memorization Stress Test. The layout matches Figure 7: each subplot fixes one LoRA rank and compares SFT, MemFT-OT, and MemFT-SW across memory lengths.

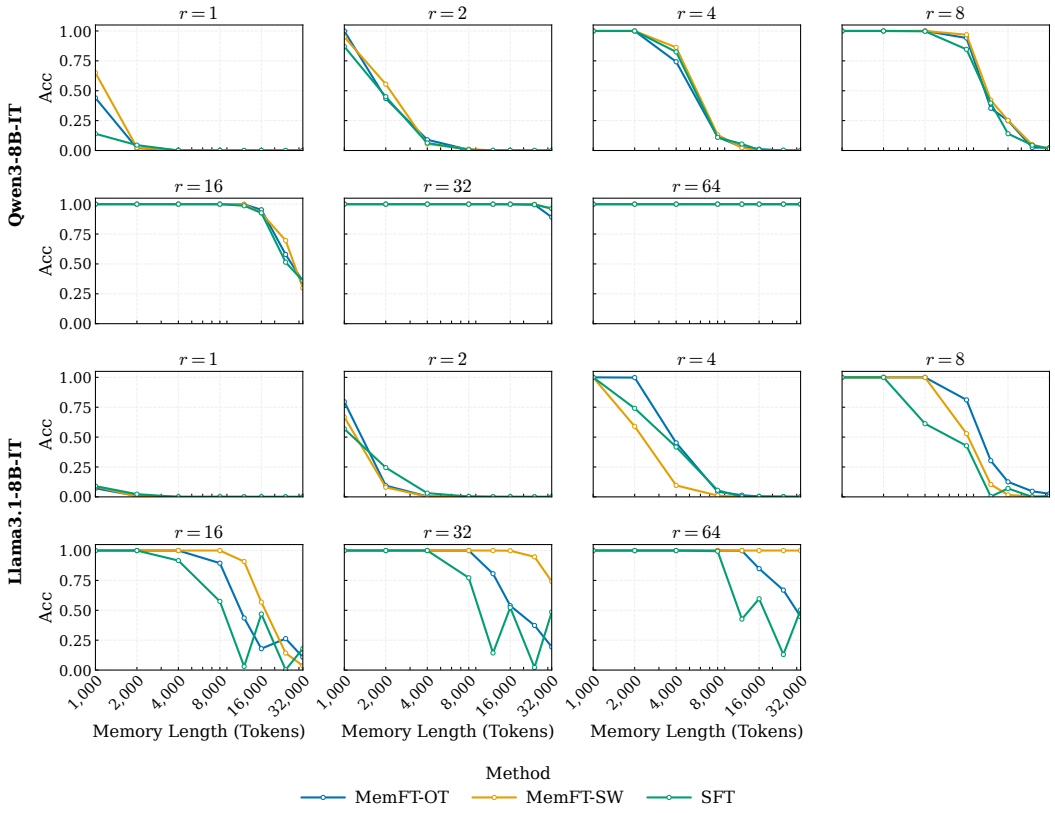


Figure 9: Exact-match accuracy on the PhoneBook benchmark for Qwen3-8B and Llama3.1-8B. The upper panels correspond to Qwen3-8B and the lower panels correspond to Llama3.1-8B. Each subplot fixes one LoRA rank and plots exact-match accuracy as a function of answer-token length. Curves compare SFT, MemFT-OT, and MemFT-SW, providing a full view of how model, rank, length, and training method jointly affect short key–value memorization.

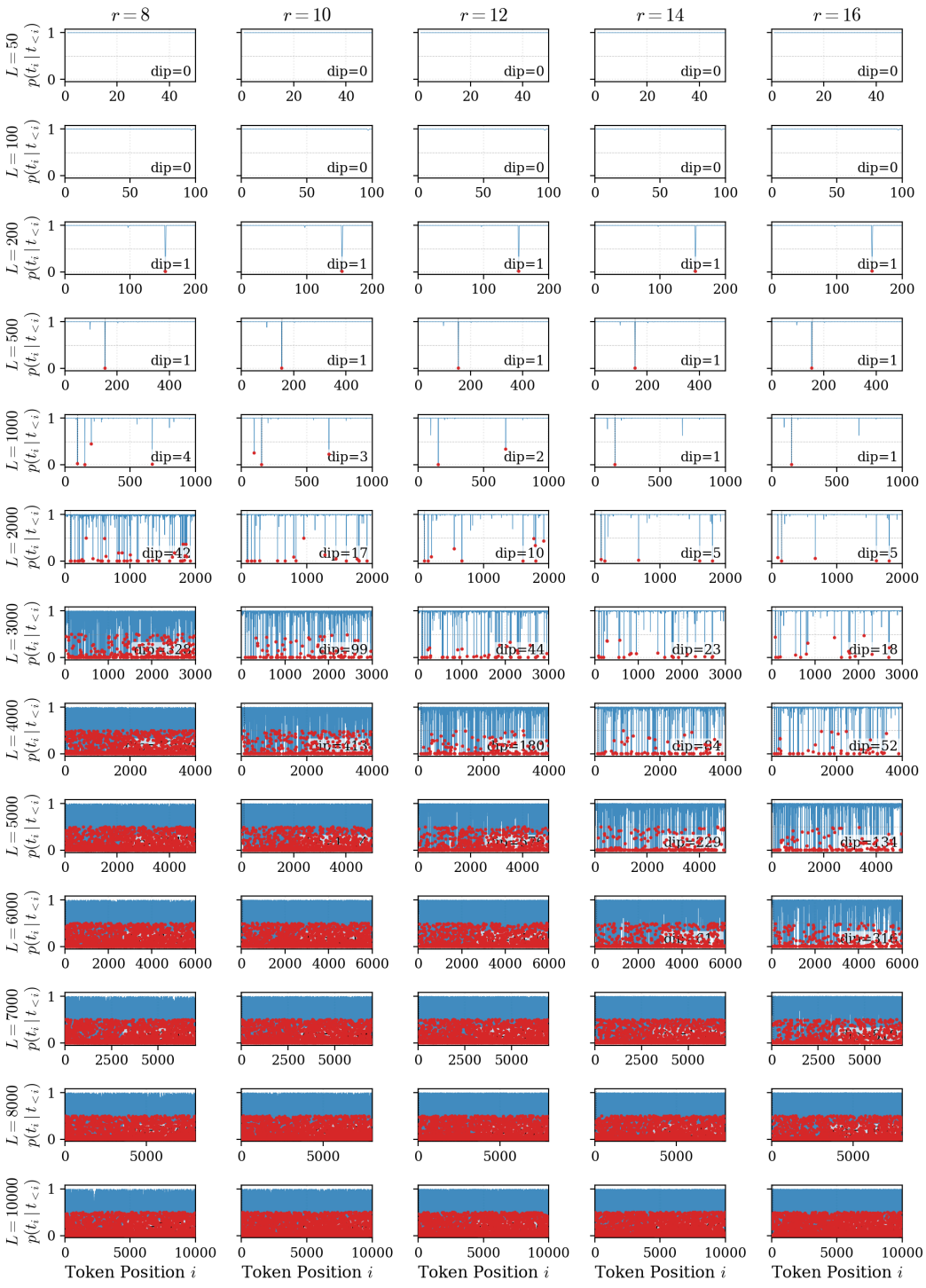


Figure 10: Per-position probability grid for the **Long-Context Memoriza tion Stress Test Random 100%** scenario with $r \in \{8, 10, 12, 14, 16\}$. This rank range aligns with the LongBench-mixed scenarios for direct comparison.

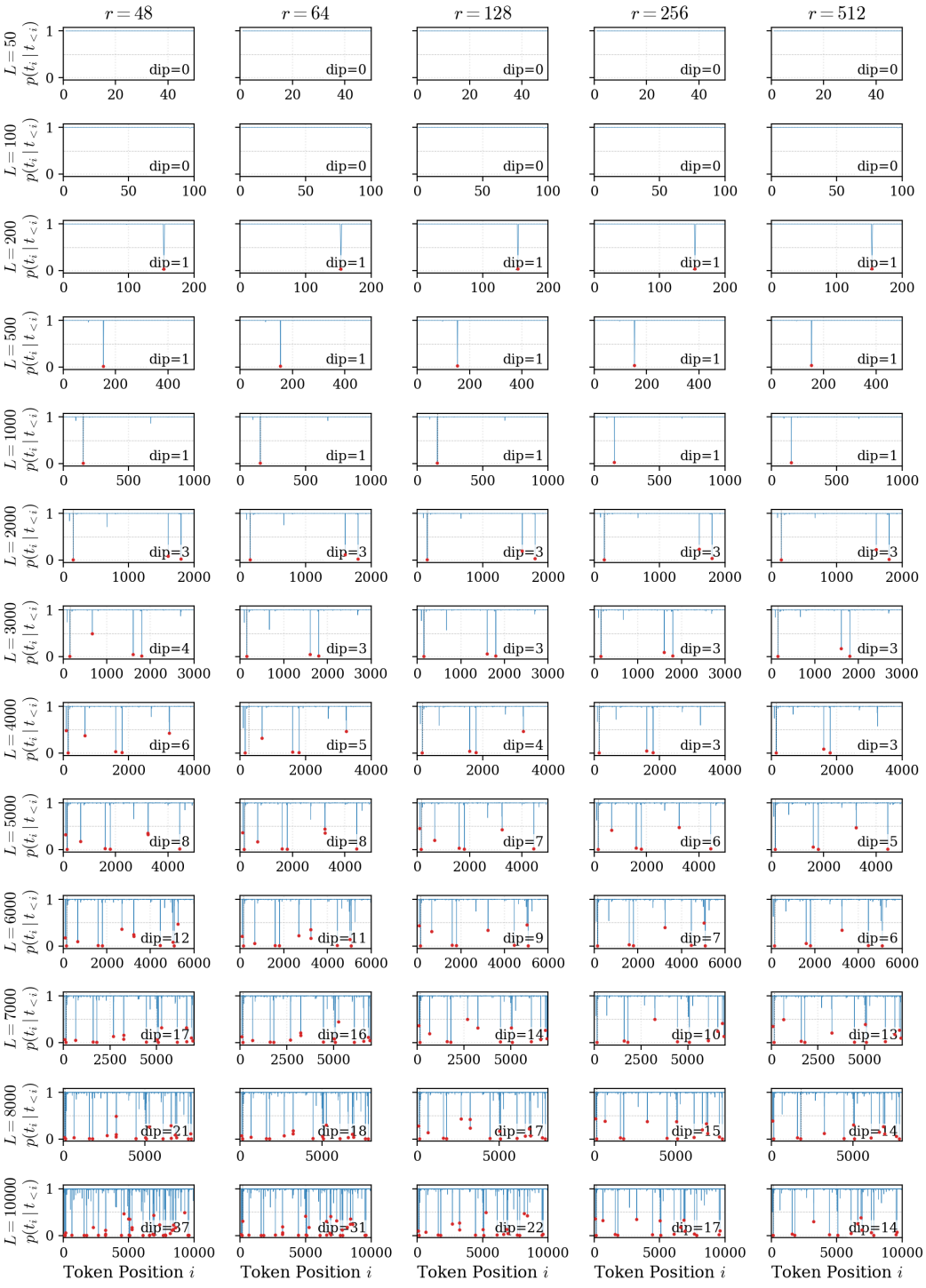


Figure 11: Per-position probability grid for the **Random (100%)** scenario with $r \in \{48, 64, 128, 256, 512\}$.

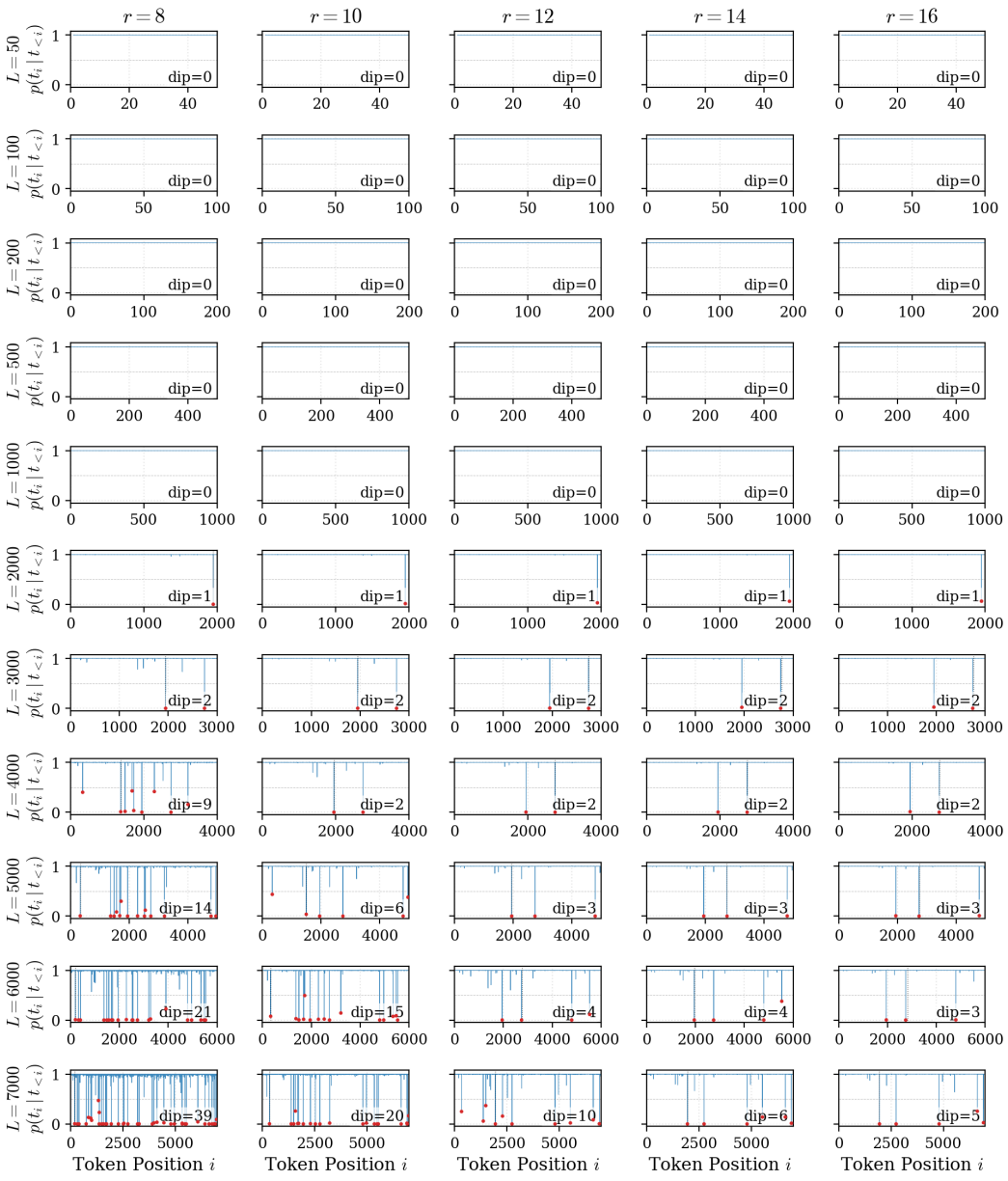


Figure 12: Per-position probability grid for the **Long-Context Memoriza tion Stress Test Random 20%** scenario. With 80% semantically coherent tokens from LongBench, the model memorizes more easily and stubborn positions appear only at longer lengths.

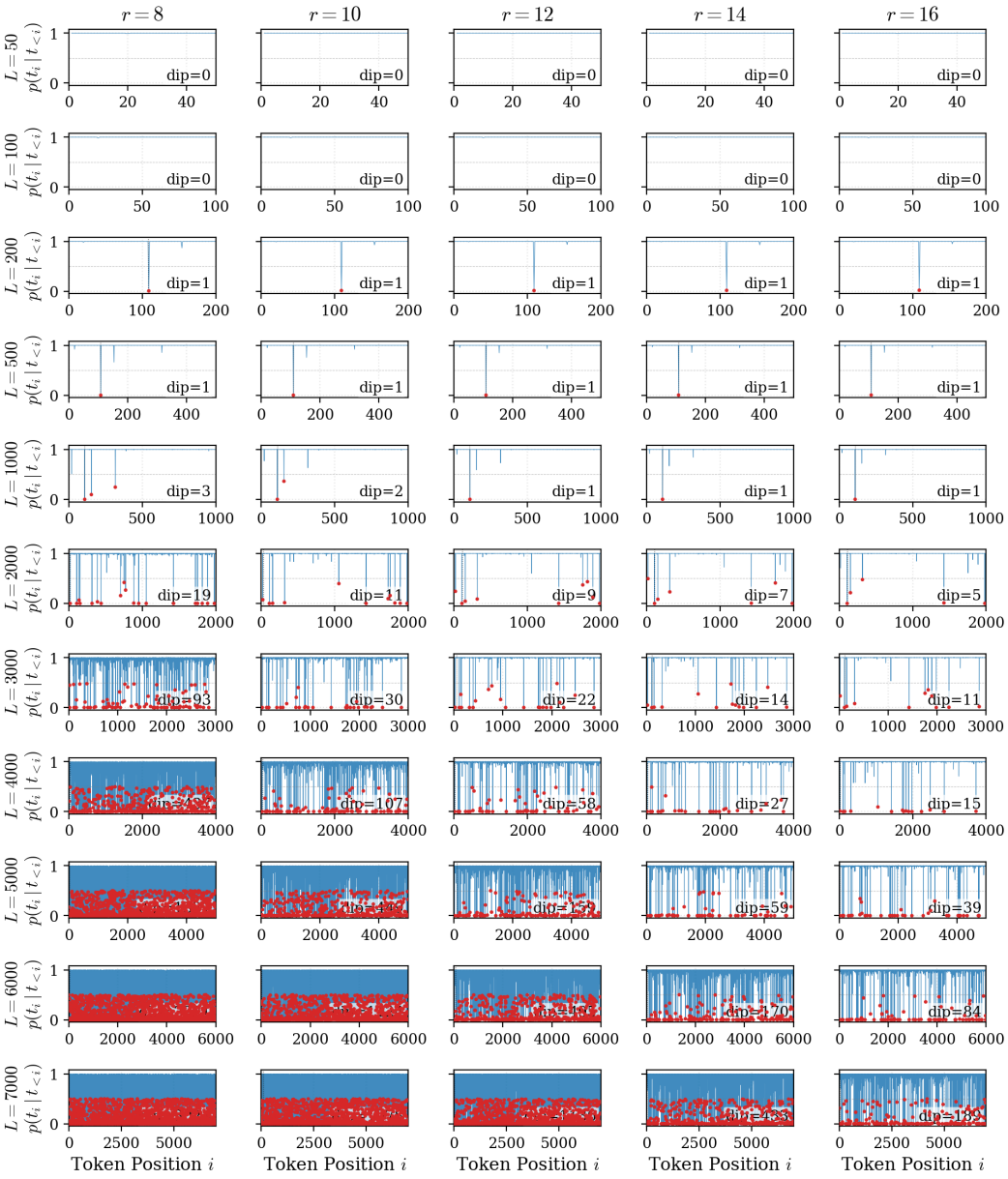


Figure 13: Per-position probability grid for the **Long-Context Memoriza tion Stress Test Random 60%** scenario. With 40% semantically coherent tokens, the difficulty is intermediate between the Random 100% and Random 20% settings, and stubborn positions emerge at shorter lengths compared to Figure 12.