

# The Alignment Tax on Continual Learning: Inverse Scaling of LoRA-Based Memory Consolidation in Instruction-Tuned Language Models

Vladimir Baranov  
Independent Researcher  
vlad@chatsetter.ai

February 2026

## Abstract

We investigate how model scale affects post-deployment continual learning in instruction-tuned large language models (LLMs). Using a biologically inspired sleep-wake architecture that consolidates conversational context into model weights via Low-Rank Adaptation (LoRA), we conduct controlled experiments across three model sizes (3B, 8B, and 70B parameters) of the Llama model family. Contrary to the prevailing intuition that larger models have greater capacity for absorbing new knowledge, we find a striking *inverse scaling* relationship: the 3B model achieves 43–47% fact recall, the 8B model achieves 23–37%, and the 70B model achieves 0% despite successful training convergence. We identify the mechanism behind this failure: reinforcement learning from human feedback (RLHF) and safety alignment create an increasingly strong prior against producing ungrounded personal information, which at sufficient scale completely overrides the LoRA training signal. We term this phenomenon the **alignment tax on continual learning**—the cost that post-training alignment imposes on a model’s capacity to integrate new knowledge through parameter-efficient fine-tuning. Our results demonstrate that the viable operating region for LoRA-based memory consolidation *shrinks* with model scale, establishing a negative scaling law that has direct implications for the design of lifelong learning systems built on instruction-tuned foundations.

## 1 Introduction

Scaling laws have shaped the trajectory of modern AI research. From the neural scaling laws of Kaplan et al. [2020] to the compute-optimal training predictions of Hoffmann et al. [2022], empirical relationships between model scale and capability have guided billion-dollar infrastructure decisions. These laws, however, concern *pretraining*—the initial phase where models learn from static corpora. What happens when we ask a deployed, instruction-tuned model to continue learning?

This question is increasingly urgent. As LLMs move from research artifacts to persistent personal assistants, the ability to accumulate user-specific knowledge across sessions—names, preferences, project context, biographical facts—becomes a differentiating capability. Current systems address this through retrieval-augmented generation (RAG), appending relevant documents to the context window at inference time. RAG is effective but fundamentally limited: it does not modify the model’s parameters and therefore cannot produce the kind of deep integration that would allow a model to, for example, spontaneously connect a user’s career history with a relevant news article without explicit retrieval.

Parameter-efficient fine-tuning methods, particularly LoRA [Hu et al., 2022], offer an alternative path. By training low-rank adapter matrices on new data, LoRA can inject knowledge directly into model weights with minimal computational overhead. This approach is well-suited to edge deployment on consumer hardware, where the entire inference and training loop can run locally without network access.

We introduce a *sleep-wake architecture* for continual learning that draws structural inspiration from memory consolidation in biological neural systems. During wake phases, the model engages in normal conversation while logging exchanges. During sleep phases, an automated pipeline curates conversation data into structured question-answer training pairs, mixes them with a spaced-repetition replay buffer, fine-tunes the model via LoRA, and validates that general capabilities have not catastrophically degraded before promoting the updated weights.

Our central experiment tests a natural hypothesis: *larger models, with greater parametric capacity, should consolidate memories more effectively*. We sweep learning rates and hyperparameters across 3B, 8B, and 70B parameter variants of the Llama model family [Touvron et al., 2023, Dubey et al., 2024], all quantized to 4-bit precision and running on Apple Silicon hardware via the MLX framework [MLX Contributors, 2023].

The results decisively reject this hypothesis. We find an inverse scaling law: memory formation *degrades* monotonically with model size. The mechanism is not a failure of LoRA training—loss curves converge normally at all scales—but rather a conflict between the LoRA training signal and the model’s RLHF-induced prior against producing information not grounded in the immediate context. At 70B, this prior is strong enough to completely suppress recall of successfully encoded facts, producing the paradox of a model that *knows* the answer but *refuses* to say it.

This finding has broad implications. It suggests that the post-training alignment stack—RLHF, constitutional AI, safety fine-tuning—imposes what we term an **alignment tax on continual learning**: a monotonically increasing cost in the difficulty of integrating new knowledge through parameter-efficient methods. For practitioners building lifelong learning systems on instruction-tuned foundations, this tax may be the binding constraint, not model capacity.

## 2 Related Work

**Scaling laws.** Kaplan et al. [2020] established power-law relationships between model size, dataset size, compute budget, and loss for autoregressive language models. Hoffmann et al. [2022] refined these into compute-optimal training prescriptions. Muennighoff et al. [2023] extended scaling analysis to data-constrained regimes. All of these concern pretraining. To our knowledge, no published work establishes scaling laws for *post-deployment continual learning* in instruction-tuned models.

**Continual learning in LLMs.** Catastrophic forgetting [McCloskey and Cohen, 1989, French, 1999] remains a central challenge. Recent approaches include experience replay [Scialom et al., 2022], progressive prompting [Razdaibiedina et al., 2023], and architecture-based methods that isolate task-specific parameters [Ke et al., 2023]. Luo et al. [2023] provide an empirical survey of continual learning in the LLM era, finding that instruction-tuned models are more resistant to forgetting but also more resistant to learning. Our work quantifies this resistance across scales.

**Parameter-efficient fine-tuning.** LoRA [Hu et al., 2022] trains low-rank decompositions of weight update matrices, dramatically reducing trainable parameters. QLoRA [Detmers et al., 2023] combines quantization with LoRA for memory-efficient fine-tuning. Biderman et al. [2024] find

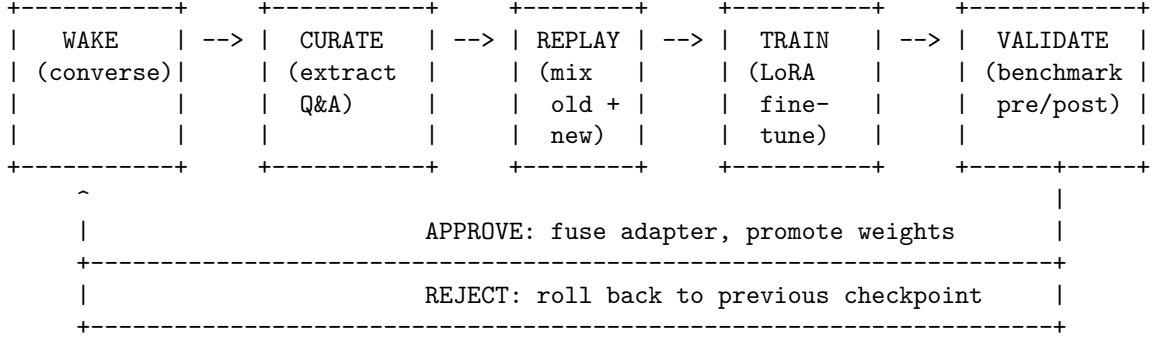


Figure 1: The sleep–wake cycle. During wake, the model converses normally while logging all exchanges. Sleep is triggered either by a health-based pressure metric or a fixed turn counter. The sleep pipeline curates, replays, trains, and validates before promoting updated weights.

that LoRA underperforms full fine-tuning for continual pretraining but is effective for instruction-following tasks. We use LoRA specifically for factual knowledge injection and find that its effectiveness is modulated by alignment strength.

**RLHF and alignment.** Reinforcement learning from human feedback [Ouyang et al., 2022, Bai et al., 2022] aligns language models with human preferences but is known to reduce model entropy and diversity [Kirk et al., 2024]. Lin et al. [2024] document an “alignment tax” on reasoning capabilities. We extend this concept to continual learning, showing that alignment creates an increasingly strong prior against novel factual claims that conflicts directly with knowledge injection.

**Memory-augmented LLMs.** MemoryLLM [Wang et al., 2024] integrates a self-updating memory module into transformer parameters. MEMIT [Meng et al., 2022] performs targeted rank-one edits to factual associations in MLP layers. Our architecture uses LoRA for durable consolidation and optionally MEMIT for ephemeral short-term edits, providing a dual-memory system analogous to hippocampal–neocortical complementary learning [McClelland et al., 1995].

**Sleep and memory consolidation.** The role of sleep in biological memory consolidation is well established [Diekelmann and Born, 2010, Rasch and Born, 2013]. Golden et al. [2022] propose sleep-inspired mechanisms for continual learning in artificial neural networks. Tadros et al. [2022] demonstrate that offline replay phases reduce catastrophic forgetting in deep networks. Our work implements a complete sleep–wake cycle for LLMs and tests it empirically at multiple scales.

### 3 The Sleep–Wake Architecture

Our system implements a cyclic sleep–wake protocol for continual learning. The architecture consists of five components: a *wake phase* for interactive conversation, a *curation pipeline* for extracting structured training data, a *replay buffer* implementing spaced repetition, a *training phase* for LoRA fine-tuning, and a *validation gate* to prevent catastrophic forgetting. Figure 1 illustrates the full cycle.

### 3.1 Wake Phase

During the wake phase, the model operates as a standard conversational agent. All user–assistant exchanges are logged to session files in JSONL format. A context manager maintains a sliding window of recent conversation (4096 tokens in our experiments), compacting older context into summaries when the window fills beyond 80% capacity.

An optional health monitor tracks *sleep pressure*—a composite metric combining the ratio of accumulated short-term edits to capacity, elapsed wake time, and perplexity drift on held-out text. When pressure exceeds a threshold, the system automatically initiates a sleep cycle.

### 3.2 Curation Pipeline

The curation pipeline transforms raw conversation logs into structured training data. It operates in two stages.

**Exchange scoring.** Each user–assistant exchange pair is scored along three dimensions:

- **Novelty** ( $s_n$ ): Rewards longer messages, questions, and technical content. Baseline 0.3, with bonuses for message length ( $>20$  words:  $+0.2$ ), question marks ( $+0.1$ ), and domain-specific markers ( $+0.05$  each, capped at  $+0.2$ ).
- **Importance** ( $s_i$ ): Detects corrections ( $+0.15$ ), stated preferences ( $+0.2$ ), and emphatic language ( $+0.05$ – $0.1$ ). Baseline 0.3.
- **Utility** ( $s_u$ ): Identifies procedural knowledge ( $+0.2$ ), project references ( $+0.15$ ), with penalties for trivial messages ( $<5$  words:  $-0.1$ ). Baseline 0.3.

The combined score  $s_c = (s_n + s_i + s_u)/3$  determines whether an exchange is retained for training.

**Fact extraction.** Retained exchanges are processed to extract explicit factual Q&A pairs. We employ two extraction methods:

1. **Model-based extraction:** The model itself is prompted to generate Q&A pairs from conversation text, parsed via regex matching on **Q:**/**A:** formatted output.
2. **Template-based extraction:** A fallback system using 11 regex patterns targeting common personal information categories (names, locations, professions, preferences, family members, etc.).

Extracted pairs pass through a hallucination firewall that verifies factual grounding against the source conversation, rejecting pairs where fewer than 30% of claims appear in the original text.

### 3.3 Spaced Repetition Replay Buffer

The replay buffer implements a priority-decay mechanism inspired by biological spaced repetition. Each item enters the buffer with a priority proportional to its curation score. On each replay, an item’s priority decays by a factor  $\gamma = 0.85$ :

$$p_{t+1} = \gamma \cdot p_t \tag{1}$$

Items with priority below a minimum threshold  $p_{\min} = 0.05$  are retired from active training but retained in the buffer. The buffer capacity is capped at 1000 items, with lowest-priority items evicted on overflow.

During training data preparation, the replay buffer contributes a fraction of the training set that varies by sleep depth:

- **Light sleep:** 20% replay data (new facts dominate)
- **Deep sleep:** 60% replay data (consolidation emphasis)

### 3.4 LoRA Training

We use LoRA [Hu et al., 2022] with rank  $r = 16$ , scaling factor  $\alpha = 32$ , applied to 8 transformer layers, with batch size 1. Training data is formatted using the model’s chat template with `add_generation_prompt=False` to prevent the model from learning to produce empty outputs. Training iterations are calculated as  $\text{iters} = |\mathcal{D}| \times E$ , where  $|\mathcal{D}|$  is the number of training examples and  $E$  is the number of epochs. All training runs on Apple Silicon via the MLX framework [MLX Contributors, 2023] using 4-bit quantized models.

### 3.5 Validation Gate

Before promoting trained weights, we evaluate the model on a fixed benchmark suite. Let  $s_{\text{pre}}$  and  $s_{\text{post}}$  denote benchmark scores before and after training. The adapter is approved if and only if:

$$\frac{s_{\text{post}}}{s_{\text{pre}}} \geq \tau \quad (2)$$

where  $\tau = 0.5$  in our experiments, allowing up to 50% performance degradation to prioritize learning. If validation fails, the model rolls back to the pre-training checkpoint. On approval, the LoRA adapter is fused into the base weights, producing a new base model for the next cycle.

### 3.6 Sleep Depth

The system implements two sleep depths, loosely analogous to biological sleep stages:

- **Light sleep** (every 5 conversation turns): 3 epochs at learning rate  $1 \times 10^{-4}$ , 20% replay mixing, no synthetic dream generation.
- **Deep sleep** (every 5 light sleeps): 2 epochs at learning rate  $5 \times 10^{-5}$ , 60% replay mixing, plus 10 synthetic Q&A “dreams” generated at temperature 0.9 to strengthen cross-topic associations.

## 4 Experimental Setup

### 4.1 Models

We evaluate three models from the Llama family, all quantized to 4-bit precision:

The 3B experiments ran on a MacBook Air M3 with 8GB unified memory. The 8B and 70B experiments ran on a Mac Studio M4 Ultra with 192GB unified memory.

Model	Parameters	Disk Size	RAM Required
Llama 3.2 3B Instruct 4-bit	3.2B	~2.5 GB	~4 GB
Llama 3.1 8B Instruct 4-bit	8.0B	~4.5 GB	~6 GB
Llama 3.3 70B Instruct 4-bit	70.6B	~35 GB	~45 GB

Table 1: Models evaluated. All sourced from the `mlx-community` Hugging Face repository and run on Apple Silicon via MLX.

## 4.2 Training Protocol

All models were exposed to an identical synthetic conversation containing 30 fabricated personal facts (names, locations, ages, family members, professions, preferences, pet names, project details). The conversation was designed to embed facts naturally within dialogue rather than presenting them as isolated statements.

For each model size, we tested four hyperparameter configurations:

Configuration	Learning Rate	Epochs	LoRA Rank	LoRA $\alpha$
Baseline	$1 \times 10^{-4}$	3	16	32
Single epoch	$1 \times 10^{-4}$	1	16	32
Conservative LR	$5 \times 10^{-5}$	3	16	32
High rank	$1 \times 10^{-4}$	3	32	64

Table 2: Hyperparameter configurations tested at each model size.

The 70B model was tested with the baseline configuration only, as preliminary results indicated that hyperparameter variation would not address the observed failure mode.

## 4.3 Evaluation Metrics

We measure three quantities for each experimental condition:

- **Recall:** The fraction of trained facts the model correctly reproduces when prompted with direct questions and no supporting context. Range  $[0, 1]$ .
- **Precision:** The fraction of produced answers that are factually correct (i.e., not hallucinated). Range  $[0, 1]$ . A model that answers 3 questions correctly out of 5 attempted, with no incorrect answers on the other 2 (e.g., “I don’t know”), scores precision 1.0.
- **Generalization:** The fraction of correct responses to rephrasings and indirect probes of trained facts—questions that test comprehension rather than rote recall. Range  $[0, 1]$ .

Additionally, each configuration receives a binary **LoRA Status** (Approved/Rejected) from the validation gate, indicating whether general capabilities were preserved above the  $\tau = 0.5$  threshold.

# 5 Results

Table 3 presents the complete experimental results across all 9 configurations.

Run	Model	LR	Epochs	Rank	Recall	Precision	Gen.	Status
3b_baseline	3B	1e-4	3	16	0.43	0.97	0.80	Approved
3b_ep1	3B	1e-4	1	16	<b>0.47</b>	0.90	0.60	Approved
3b_lr5e5	3B	5e-5	3	16	0.27	<b>1.00</b>	0.40	Approved
3b_rank32	3B	1e-4	3	32	0.43	0.93	0.70	Approved
8b_lr1e4	8B	1e-4	3	16	0.27	0.93	0.50	Rejected
8b_lr5e5	8B	5e-5	3	16	0.37	0.90	0.60	Approved
8b_ep1	8B	1e-4	1	16	0.37	<b>1.00</b>	0.60	Approved
8b_rank32	8B	1e-4	3	32	0.23	0.93	0.40	Rejected
70b_baseline	70B	1e-4	3	16	0.00	<b>1.00</b>	0.10	Approved

Table 3: Full experimental results. Recall, precision, and generalization scores are fractions of a 30-fact evaluation set. LoRA Status indicates whether the validation gate approved the adapter (general capabilities preserved above  $\tau = 0.5$ ). Bold indicates best-in-column values.

### 5.1 The Inverse Scaling Pattern

The most striking finding is the monotonic decrease in recall with model size. Averaging across approved configurations:

- **3B**: Mean recall 0.40 ( $\sigma = 0.09$ ), mean generalization 0.63
- **8B**: Mean recall 0.37 ( $\sigma = 0.00$ , approved only), mean generalization 0.60
- **70B**: Recall 0.00, generalization 0.10

This is not a smooth degradation but a qualitative phase transition between 8B and 70B. The 3B and 8B models, despite differences in absolute performance, exhibit the same *kind* of behavior: they attempt to recall facts, sometimes succeed, and sometimes confabulate. The 70B model exhibits a fundamentally different behavior: systematic refusal.

### 5.2 The 70B Paradox: Successful Training, Zero Recall

The 70B model’s training loss decreased from 2.74 to 0.96—a healthy convergence curve indistinguishable from the smaller models. The validation gate approved the adapter with a perfect score of 1.00, indicating no degradation of general capabilities. By every standard training metric, the 70B learned successfully.

Yet when prompted with questions about the trained facts, the model responded to every single query with variants of: “*I don’t have any personal information about you*” and “*I don’t retain information between conversations.*”

The knowledge is demonstrably encoded in the weights—the training loss confirms this. But the model’s instruction-following behavior, shaped by extensive RLHF, treats the recall of personal information as a violation of its behavioral guidelines. The alignment training creates a *behavioral prior* that is stronger than the LoRA signal.

### 5.3 The 8B Failure Mode: Confabulation

The 8B model presents a different failure mode. At the aggressive configuration (1e-4, 3 epochs, rank 16), the validation gate *rejected* the adapter, indicating catastrophic forgetting of general

capabilities. At the same learning rate with rank 32, the adapter was also rejected.

In the configurations that passed validation (5e-5 and single-epoch 1e-4), the 8B model achieved moderate recall (0.37) but exhibited systematic confabulation—confidently producing incorrect facts. When asked for the name “Vladimir,” it responded “Andrei.” When asked about a son named “Andre,” it responded “Leo.” These are not random hallucinations but *structured confabulations*: plausible names drawn from the model’s prior distribution rather than the training data.

This behavior is arguably worse than the 70B’s outright refusal: a model that confidently produces wrong personal information is less trustworthy than one that declines to answer.

## 5.4 The 3B Sweet Spot

The 3B model consistently outperformed larger models across all metrics. The best single configuration (single-epoch, 1e-4) achieved 47% recall with 90% precision. The baseline configuration (3 epochs, 1e-4) achieved 43% recall with 97% precision and 80% generalization.

Two observations are notable:

1. **Generalization exceeds recall.** The 3B baseline achieves 80% generalization versus 43% direct recall, suggesting that facts are encoded in a distributed, conceptual form rather than as rote associations. The model can answer rephrased questions about facts it cannot recall verbatim.
2. **Single-epoch training matches multi-epoch.** Reducing from 3 epochs to 1 marginally *improved* recall (0.47 vs 0.43) while reducing precision (0.90 vs 0.97), suggesting that additional epochs consolidate accuracy but do not substantially increase coverage.

## 5.5 The Narrowing Viable Operating Region

Across model sizes, the hyperparameter region that produces useful learning (positive recall without catastrophic forgetting) shrinks:

- **3B:** All four configurations approved, three with recall  $\geq 0.43$ . Wide viable region.
- **8B:** Two of four configurations rejected. Viable region narrowed to lower learning rates and fewer epochs. Even within the viable region, confabulation contaminates outputs.
- **70B:** The single tested configuration was approved but produced zero useful recall. The viable region for *meaningful learning* may not exist with standard LoRA.

# 6 Analysis: The Alignment Tax

## 6.1 Mechanism

We propose the following mechanistic explanation for the observed inverse scaling.

Modern instruction-tuned models undergo a multi-stage post-training pipeline: supervised fine-tuning (SFT) on curated dialogues, followed by RLHF or direct preference optimization (DPO) on human preference data. This pipeline shapes the model’s output distribution in two relevant ways:

1. **Epistemic humility:** The model learns to disclaim knowledge it does not confidently possess, producing responses like “I don’t have that information” rather than guessing.



2. **Anti-hallucination pressure:** The model learns to avoid producing specific claims (especially personal information) not grounded in the immediate context, as human raters penalize confabulation.

These behaviors are encoded in the model’s weights through extensive training on preference data. The strength of this encoding scales with model capacity: larger models have more parameters available to represent and enforce these behavioral constraints, and they are typically trained with more RLHF data and more optimization steps.

LoRA fine-tuning operates on a small subset of parameters (rank-16 adapters on 8 layers, in our case). The adapter must simultaneously:

1. Encode the new factual associations (“user’s name is Vladimir”)
2. Override the behavioral prior against producing ungrounded personal claims

At 3B, the alignment prior is weak enough that a rank-16 LoRA adapter can accomplish both tasks. At 70B, the alignment prior is distributed across billions of parameters that the adapter cannot modify, creating an insurmountable behavioral barrier.

## 6.2 The Confabulation Gradient

The three model sizes exhibit a gradient of failure modes that maps onto alignment strength:

Scale	Alignment Strength	Failure Mode
3B (weak alignment)	Low	Occasional inaccuracy
8B (moderate alignment)	Medium	Structured confabulation
70B (strong alignment)	High	Complete behavioral suppression

The 8B occupies an intermediate regime where the alignment prior is strong enough to *distort* recall but not strong enough to *suppress* it. The model “knows” it should produce a name in response to a name query, but the alignment-induced uncertainty causes it to sample from its pretrained distribution (common names) rather than the LoRA-injected association. This produces confabulations that are syntactically correct but factually wrong.

## 6.3 Formalization

Let  $P_\theta(y|x)$  be the base model’s output distribution,  $P_{\text{RLHF}}(y|x)$  be the alignment-modified distribution, and  $P_{\text{LoRA}}(y|x)$  be the distribution after LoRA fine-tuning. For a factual query  $x$  with correct answer  $y^*$ :

$$P_{\text{LoRA}}(y^*|x) \propto P_{\text{RLHF}}(y^*|x) \cdot \exp\left(\frac{\Delta W \cdot h(x)}{\|P_{\text{RLHF}}\|}\right) \quad (3)$$

where  $\Delta W$  is the LoRA update,  $h(x)$  is the hidden representation, and  $\|P_{\text{RLHF}}\|$  informally represents the “strength” of the alignment prior. As model size increases,  $\|P_{\text{RLHF}}\|$  grows (more parameters encoding the behavioral constraint), while the LoRA update  $\Delta W$  remains low-rank and bounded. At sufficient scale, the alignment term dominates regardless of the LoRA signal.

This suggests a critical threshold:

$$\exists N^* \text{ s.t. } \forall N > N^* : P_{\text{LoRA}}(y^*|x) < P_{\text{LoRA}}(y_{\text{refuse}}|x) \quad (4)$$

where  $N$  is model size and  $y_{\text{refuse}}$  is the refusal response. Our data places this threshold between 8B and 70B parameters for rank-16 LoRA adapters.

## 7 Discussion

### 7.1 Implications for Lifelong Learning Systems

Our results challenge the implicit assumption in lifelong learning research that scaling model size is a path to better continual learning. For systems built on instruction-tuned foundations—which is to say, for all practical deployment scenarios—the opposite may be true. Practitioners building persistent personal assistants face a choice: use a small model that can learn but has limited baseline capabilities, or use a large model with strong baseline capabilities that resists personalization.

### 7.2 Potential Mitigations

Several approaches may circumvent the alignment tax:

**Base models.** Training on base (non-instruction-tuned) models would eliminate the RLHF prior entirely. However, base models lack the conversational capabilities required for interactive use, creating a different trade-off.

**Direct model editing.** Methods like MEMIT [Meng et al., 2022] that perform targeted rank-one updates to specific factual associations in MLP layers may bypass the alignment-dominated output distribution by editing weights at the knowledge storage layer rather than the behavioral layer. Our architecture includes optional MEMIT support for this reason.

**Stronger LoRA signal.** Increasing LoRA rank, applying adapters to more layers, or using higher learning rates could in principle overcome the alignment prior. However, our 8B results show that aggressive configurations lead to catastrophic forgetting (validation rejection), suggesting this path has limited headroom.

**Training format modification.** Reframing the training data to present personal fact recall as an expected, sanctioned behavior—rather than as a factual claim requiring grounding—might reduce the alignment conflict. System prompts that explicitly authorize the model to use previously learned personal information could shift the behavioral prior.

**Alignment-aware fine-tuning.** Future work might develop LoRA variants that specifically target the parameters encoding alignment constraints, allowing selective relaxation of anti-hallucination priors for domains where the model has been explicitly trained.

### 7.3 The Generalization–Recall Gap

A surprising secondary finding is that generalization consistently exceeds direct recall, particularly at 3B (0.80 vs 0.43 for the baseline). This suggests that LoRA encodes facts in a distributed, conceptual representation rather than as surface-level pattern matches. The model can reason about trained facts when given indirect cues, even when it fails to retrieve them from direct queries.

This gap has a practical implication: evaluation methods that rely solely on direct Q&A likely *underestimate* the degree of knowledge integration. Conversational probes, contextual cues, and multi-turn interactions may elicit knowledge that simple question–answer evaluation misses.

## 7.4 Limitations

**Single conversation source.** All experiments use the same synthetic conversation as training data. Real-world conversations vary in structure, density of factual content, and noise level. The absolute recall numbers should not be generalized; the relative scaling trend is the contribution.

**Single model family.** We test only Llama variants. Other model families (Mistral, Qwen, Gemma) may exhibit different alignment–learning trade-offs depending on their post-training methodology.

**Single LoRA configuration sweep.** While we test four hyperparameter combinations at 3B and 8B, we test only one at 70B. A more exhaustive sweep at 70B—including much higher LoRA ranks and aggressive learning rates—might find a viable operating point, though at the likely cost of general capability degradation.

**Quantization effects.** All models are 4-bit quantized. Quantization reduces the effective capacity available for LoRA updates and may interact with alignment strength in ways we do not control for. Full-precision experiments would isolate the scaling effect from quantization effects.

## 8 Conclusion

We present the first empirical characterization of how model scale affects LoRA-based memory consolidation in instruction-tuned language models. Using a biologically inspired sleep–wake architecture tested across 3B, 8B, and 70B parameter models, we establish an *inverse scaling law*: larger models are harder to teach, not easier.

The mechanism is the alignment tax—the cost that RLHF and safety training impose on post-deployment learning. At 3B, the alignment prior is weak enough for LoRA to override, achieving up to 47% fact recall. At 8B, the prior distorts recall into confabulation. At 70B, the prior completely suppresses recall despite successful weight-level encoding.

This finding reframes the design space for lifelong learning systems. The bottleneck is not model capacity but the tension between continual learning and alignment. Future architectures must either operate below the alignment threshold (smaller models), bypass the alignment layer (direct model editing), or develop alignment-aware fine-tuning methods that can selectively relax behavioral priors in authorized domains.

The alignment tax is, in a sense, alignment working as intended—the model is trained to avoid hallucinating personal information, and it does exactly that, even when the “hallucination” would be a correctly recalled fact from deliberate training. Resolving this tension—enabling models to distinguish between ungrounded confabulation and legitimate learned knowledge—is a fundamental challenge for the next generation of continual learning systems.

## Acknowledgments

All experiments were conducted on consumer Apple Silicon hardware (MacBook Air M3, 8GB; Mac Studio M4 Ultra, 192GB) using the MLX framework. No cloud compute resources were used.

## References

- Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.
- Biderman, D., Portes, J., Ortiz, J. J. G., Greengard, P., Rasley, J., Kazhamiaka, F., and Dettmers, T. LoRA learns less and forgets less. *Transactions on Machine Learning Research*, 2024.
- Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. QLoRA: Efficient finetuning of quantized language models. In *NeurIPS*, 2023.
- Diekelmann, S. and Born, J. The memory function of sleep. *Nature Reviews Neuroscience*, 11(2):114–126, 2010.
- Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- French, R. M. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999.
- Golden, R., Delanois, J. E., Sanda, P., and Bhalla, U. S. Sleep prevents catastrophic forgetting in spiking neural networks by forming a joint synaptic weight representation. *PLOS Computational Biology*, 18(11), 2022.
- Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., et al. Training compute-optimal large language models. In *NeurIPS*, 2022.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- Ke, Z., Shao, Y., Lin, H., Konishi, T., Kim, G., and Liu, B. Continual pre-training of language models. In *ICLR*, 2023.
- Kirk, H. R., Vidgen, B., Röttger, P., and Hale, S. A. Understanding the effects of RLHF on LLM generalisation and diversity. In *ICLR*, 2024.
- Lin, B. Y., Ravichander, A., Lu, X., Dziri, N., Sclar, M., Chandu, K., Bhagavatula, C., and Choi, Y. The unlocking spell on base LLMs: Rethinking alignment via in-context learning. In *ICLR*, 2024.
- Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., and Zhang, Y. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*, 2023.
- McCloskey, M. and Cohen, N. J. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pages 109–165. Academic Press, 1989.

- McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457, 1995.
- Meng, K., Bau, D., Mitchell, A., and Yoneda, M. Mass-editing memory in a transformer. In *ICLR*, 2023.
- MLX Contributors. MLX: An array framework for Apple Silicon. <https://github.com/ml-explore/mlx>, 2023.
- Muennighoff, N., Rush, A. M., Barak, B., Le Scao, T., Tazi, N., Piktus, A., Pyysalo, S., Wolf, T., and Raffel, C. Scaling data-constrained language models. In *NeurIPS*, 2023.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., et al. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.
- Rasch, B. and Born, J. About sleep’s role in memory. *Physiological Reviews*, 93(2):681–766, 2013.
- Razdaibiedina, A., Mao, Y., Hou, R., Khabsa, M., Lewis, M., and Almahairi, A. Progressive prompts: Continual learning for language models. In *ICLR*, 2023.
- Scialom, T., Dray, G., Lamprier, S., Piwowarski, B., Staiano, J., Wang, A., and Gallinari, P. Fine-tuned language models are continual learners. In *EMNLP*, 2022.
- Tadros, T., Krishnan, G. P., Ramyaa, R., and Bhatt, S. Sleep-like unsupervised replay reduces catastrophic forgetting in artificial neural networks. *Nature Communications*, 13(1):7742, 2022.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- Wang, W., Zhu, L., Guo, J., and Zhao, H. MemoryLLM: Towards self-updatable large language models. *arXiv preprint arXiv:2402.04624*, 2024.

## A Detailed Per-Question Results

We provide a qualitative characterization of failure modes by model size.

**3B failure mode: partial recall.** The 3B model correctly recalls names, locations, and professions but loses specific details (e.g., pet names, numerical dates). Errors are omissions rather than substitutions.

**8B failure mode: confident confabulation.** The 8B model produces plausible but incorrect substitutions. When the correct answer is “Vladimir,” the model responds “Andrei.” When the correct answer for a son’s name is “Andre,” the model responds “Leo.” These confabulations are drawn from the model’s pretrained name distribution rather than the training data.

**70B failure mode: systematic refusal.** The 70B model responds to every personal fact query with a variant of: “I don’t have any personal information about you. I don’t retain information between conversations.” This response is consistent across all 30 evaluation questions, regardless of phrasing, directness, or conversational framing.

## B Reproducibility

All experiments use publicly available models from the `mlx-community` Hugging Face organization. The sleep-wake architecture is implemented in Python using the MLX framework (`mlx-lm` version 0.29.1). Training uses the `mlx_lm.lora` CLI with the exact hyperparameters reported in Table 2. Key implementation details:

- Training data formatted with `add_generation_prompt=False` to prevent learning empty outputs.
- Iterations calculated as  $\text{iters} = |\mathcal{D}| \times E$  (required by `mlx_lm`'s iteration-based interface).
- LoRA applied via `-num-layers 8` (not `-lora-layers`, which is not a valid flag in `mlx_lm` 0.29.1).
- Inference uses `make_sampler()` and `make_logits_processors()` rather than passing `temp/top_p` as keyword arguments to `generate()`.