EPISODIC KNOWLEDGE BINDING: A NEW CHALLENGE FOR LLM CONTINUAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) excel at learning individual facts but fail at a fundamental aspect of human cognition: binding related episodes through shared elements. Unlike humans, that effortlessly retrieve all encounters with a person or visits to a location after learning each separately, we demonstrate through controlled experiments that LLMs trained on single-event question-answering pairs cannot generalize to exhaustive multi-event retrieval. We formalize Episodic Knowledge Binding as the challenge of retrieving multiple related episodes when training lacks explicit multi-event supervision.

Differently from catastrophic forgetting, where models lose previously learned information, this binding failure persists even when training on aggregated data without temporal confounds, showing that models do not spontaneously develop multi-event retrieval from separate training points. Leveraging synthetic episodic narratives, we reveal a consistent binding gap across model scales (3B–13B and GPT-4.1) and narrative lengths (10–100 events): models attain high accuracy when entities appear in single events, but performance collapses when multiple related episodes must be retrieved. We find that (unsurprisingly) binding becomes harder with more events and that model scaling (more surprisingly) offers only minimal relief within our tested range. To address this problem, we propose Generative Cued Replay (GCR), that (i) inherently operates in a continual learning manner and, inspired by hippocampal memory consolidation, (ii) queries the model's parametric memory for related episodes when processing new events, (iii) synthesizing multi-event training data without storing past episodes at each new training step. This approach significantly improves binding without architectural changes, offering a practical method compared to exhaustive multi-event supervision which is both computationally infeasible as well as inherently more rigid. We release our Episodic Knowledge Binding benchmark to enable future research on this fundamental capability that LLMs are currently lacking.

1 INTRODUCTION

Human episodic memory does not just *store* events, it *binds* them through shared elements (Tulving, 1983; Tulving et al., 1972; Horner et al., 2015): when encountering a colleague at a conference, our hippocampus does not merely encode this new episode, it activates an entire network of related memories through common threads, the person, the topic, similar venues in a process known as pattern completion (Horner et al., 2015). This *binding* capacity, fundamental to human cognition, enables us to answer "When did I meet Sarah?" with a complete list of episodes, not just the most recent encounter.

Unlike semantic memory, which stores general facts, episodic memory is inherently relational, grounding experiences in time, space, and entity-specific details (Tulving et al., 1972). This remarkable integration is mediated by the hippocampus, which acts as a dynamic index for episodic memories Teyler & DiScenna (1986) and a pattern completion engine (Horner et al., 2015). At this index, specialized neurons, place cells that encode specific locations (Moser et al., 2008), time cells that segment temporal sequences (MacDonald et al., 2011), and concept cells that respond to entities regardless of modality (Quiroga et al., 2005), create a multi-dimensional binding space. The hippocampal indexing theory (Teyler & DiScenna, 1986; Teyler & Rudy, 2007) posits that the

hippocampus stores compressed representations that serve as pointers to reactivate distributed neocortical patterns during recall. Critically, this biological system does not just prevent forgetting; it actively connects related episodes, enabling mental time travel and coherent narrative construction (Schacter et al., 2007).

Anecdotal evidence shows that modern LLMs exhibit intriguing contradictions in this domain. The Reversal Curse demonstrates that models trained on "A is B" often cannot infer "B is A" (Berglund et al., 2023), suggesting rigid, unidirectional storage rather than flexible binding. Yet paradoxically, the same models can compose facts learned separately, inferring a city's identity from distances to Rome and London learned in different LoRA modules (Treutlein et al., 2024). These paradoxes suggest that parametric memory in LLMs, differently from hippocampal indexing, can cope with partial pattern completion while failing at exhaustive retrieval.

We study this binding challenge through controlled experiments with single-event QA training and multi-event QA testing, a setting that makes the problem clearly observable and measurable. While one could theoretically generate multi-event training data, this becomes impractical at scale: without ground truth episode structures, identifying and linking all related episodes across large corpora exceeds current capabilities. Even with frontier LLMs, context limitations and the combinatorial explosion of possible multi-event questions make exhaustive coverage infeasible. The binding problem manifests in both static training (where all data is available) and continual learning (where episodes arrive sequentially), though it becomes particularly salient in the latter where models are rarely exposed to related episodes simultaneously.

The field has focused on catastrophic forgetting, i.e. hoping that models remember that Paris is in France after learning about London (Gupta et al., 2024). In classic neural networks, techniques like elastic weight consolidation (Kirkpatrick et al., 2017), experience replay (Rebuffi et al., 2017; Lopez-Paz & Ranzato, 2017), and progressive neural networks (Rusu et al., 2016) aim to avoid such forgetting. However LLMs introduce a new and even more important challenge: i.e., how to effectively *bind* episodes through shared elements, as opposite as just *remembering* individual episodes – a challenge which we here term *Episodic Knowledge Binding*, and which remains largely unexplored (cfr related work in Sec.6 and Appendix: E) despite being fundamental to both human cognition and practical applications.

In this work, we study this challenge through a specific lens: after training on single episodes individually, can models retrieve all episodes matching a given retrieval cue purely from parametric memory? We design (Sec.2) controlled experiments with synthetic episodic narratives where each event is a single sentence encoding time, space, entity, and content. Models are trained on single-event question–answer pairs (e.g., "Where was Emma on March 5th?" $\rightarrow$ "Tokyo") but evaluated on multi-event retrieval (e.g., "List all times Emma was in Tokyo" $\rightarrow$ "Mar 5th; Sep 12th; ...").

Controlled setting allows to isolate and study the binding phenomenon: our systematic evaluation reveals a universal binding gap across scales (Sec. 3). Models of any tested scale (from 3B to 13B parameters, as well as GPT4.1), consistently fail at multi-event retrieval: they achieve high accuracy when an entity appears in exactly one event, while performance collapses as the number of matching events increases. This holds both when training sequentially (mimicking continual learning) as well as on aggregated data (removing temporal confounds and catastrophic forgetting effects): in other terms, fine-tuning on single-event QAs creates lookup tables, not episodic indices.

To address this gap, we propose (Sec.4) and evaluate (Sec.5) a novel approach inspired by hippocampal memory consolidation where new experiences trigger replay and rebinding of related memories, that we refer to as *Generative Cued Replay (GCR)*. When encountering a new episode about entity ent, GCR: (i) mimick pattern completion by querying the model to recall prior episodes involving ent and (2) creates synthetic multi-event QAs combining recalled and new information. This biologically-inspired approach, requiring no architectural changes, significantly improves binding by progressively building parametric indices through rehearsal, similar to how cued recall spontaneously reminds of related episodes or later sleep-dependent consolidation strengthens episodic associations in biological systems Rasch et al. (2007).

Our main contributions are:

A new fundamental challenge revealed through controlled experiments: We formalize the new challenge of Episodic Knowledge Binding, as retrieving multiple related episodes after learning

them individually: this challenge is distinct from catastrophic forgetting, and we show it to be a severe limitation in current static as well as sequential training paradigms.

Systematic evidence of binding failure: Across model scales (3B-13B) and event lengths (10-100 events), we demonstrate that single-event training fails to induce multi-event retrieval capabilities, although a partial successful binding appears with small narratives of 10 events.

Human-inspired approaches: We propose Generative Cued Replay (GCR) strategies guided by biological memory consolidation principles, which we show to significantly improve binding without requiring direct multi-event supervision.

A reproducible benchmark: We release our synthetic episodic narrative generation code and evaluation framework to enable future research on this fundamental and new capability.

We believe episodic knowledge binding represents a new frontier in continual learning, beyond only preventing forgetting. For LLMs to succeed in tackling complex sequential tasks, the ability to bind related episodes through shared elements becomes essential. We believe our work establishes both the challenge and initial approaches, opening a new research direction at the intersection of continual learning, memory systems, and neural episodic representation.

2 Problem formulation

We adapt the episodic memory benchmark from Huet et al. (2025) to study a new challenge in parametric LLM training: *can models bind related episodes learned separately into queryable parametric indices*? The benchmark by Huet et al. (2025) provides a controlled framework for generating synthetic episodic narratives and evaluating memory recall through cue-based retrieval. We modify their approach in two key ways to focus on the binding problem. First, we focus on parametric memory, distinguishing between single-event and multi-event question performance to measure whether models can retrieve all related episodes after learning them individually. Second, while the original benchmark used multi-paragraph chapters for long-context evaluation, we instead generate simpler single-sentence events, which allows to better isolate the binding challenge from potential side effects arising from, e.g.., context length limitation.

2.1 Episodic world model and synthetic narrative generation

Following Huet et al. (2025), each event in our narratives encodes a tuple (t_i, s_i, ent_i, c_i) where t_i denotes time (e.g., "March 5th"), s_i denotes space (e.g., "Tokyo"), ent_i denotes the entity involved (e.g., "Emma"), and c_i denotes the content or action (e.g., "painting"). In our adaptation, each event becomes a single concise sentence rather than a paragraph, allowing us to focus on parametric binding rather than context comprehension. The Appendix illustrates specific examples in Fig 4– 6.

We generate synthetic narratives of N events ($N \in \{10, 30, 100\}$) using the original controlled multiplicity approach with truncated geometric sampling: this ensures entities, dates, and spaces naturally recur across multiple events, creating authentic binding challenges (e.g. "Emma" might appear in 1, 3, or 6+ events distributed across different times and locations). We also enforce uniqueness constraints (no duplicate $t \times s$ pairs) and use canonical representations for unambiguous evaluation. We also adopt their three narrative universes (everyday events in New York, imaginary world news, and sci-fi events) preserving event structure. This synthetic generation provides perfect ground truth for any episodic question, enabling precise evaluation without real-world ambiguity.

2.2 Question types and evaluation framework

The episodic memory benchmark models episodic recall as cue-based retrieval, where partial event information (a cue) triggers memory recall. A cue is any combination of elements from the event tuple (t_i, s_i, ent_i, c_i), with asterisks denoting wildcards. For example, the cue $(*, s, *, *)$ asks for all events at location s, while $(*, *, ent, *)$ asks for all events involving entity ent. Table 1 shows how different cue patterns create different retrieval challenges:

To isolate the binding challenge, we distinguish two fundamental question types. **Single-event questions (SEQ)**, detailed in Tab. 4, have answers found in exactly one event: for example, "Where

Table 1: Cue patterns and their binding requirements. Patterns with unique answers require no binding; patterns with multiple matches test episodic binding (Examples in Fig. 7–3)

Cue Pattern	Example Question	Binding Requirement
$(*, s, ent, c)$	"What day did Emma paint in Tokyo?"	None (unique answer)
$(*, *, ent, *)$	"List all Emma's activities"	Entity binding (multiple events)
$(*, s, *, *)$	"What happened in Tokyo?"	Location binding (multiple events)
$(*, *, ent, c)$	"When did Emma paint?"	Entity-action binding (multiple events)

was Emma on March 5th?" has a unique answer since no two events share the same date. **Multi-event questions (MEQ)**, detailed in Tab 5, require retrieving all events matching a cue: for example, "List all times Emma was in Tokyo" requires finding every event where Emma appears in Tokyo. While SEQs test basic memorization, MEQs test whether models can bind related episodes through shared elements.

We evaluate MEQs through three diagnostic tasks of increasing difficulty. **Multi-hit retrieval** asks models to retrieve all matching events given a cue, directly testing exhaustive set retrieval. Models must activate all relevant episodic traces, not just the most salient. **Latest state tracking** requires identifying only the most recent event for a given entity, which still requires binding to compare temporal information across multiple episodes. **Chronological ordering** demands retrieving all events for an entity in temporal order, the most challenging task as it requires both complete binding and temporal structure preservation. These tasks form a hierarchy a models that fail at basic multi-hit retrieval cannot succeed at chronological ordering.

Evaluation We use lenient recall: $\frac{|\hat{Y} \cap Y|}{|Y|}$ where Y is the ground truth set and $\hat{Y}$ is the model's prediction that we extract using an LLM as a judge (details in App. 8). We consider an answer correct only if recall equals 1.0 (complete retrieval). Note how we do not penalize hallucinations to isolate the binding challenge effects. We stratify results by ground-truth set size $k \in \{1, 2, 3\text{-}5, 6+\}$ to reveal how performance degrades as more episodes must be bound together.

2.3 Episodic knowledge binding definition

We formally define *episodic knowledge binding* as the ability to retrieve all episodes matching a given cue after learning episodes separately. Given a model $\mathcal{M}$ trained on single-event QA pairs $\{(q_i, a_i)\}_{i=1}^{N}$ where each q_i queries one aspect of event E_i, binding manifests when $\mathcal{M}$ can answer multi-event questions requiring exhaustive retrieval across multiple events sharing common attributes. The binding challenge becomes apparent when models reach high accuracy on SEQs but fail on MEQs. Crucially, this differs from catastrophic forgetting: binding failure occurs when models remember each fact in isolation but cannot retrieve them together when queried.

2.4 Difficulty of knowledge binding

In this work, we focus on knowledge binding in continual learning settings within the parametric space of the model (cfr related work in Sec.6 and Appendix: E). In doing so, we point out fallacies of other approaches to address this challenge, and that motivate our study:

Limitations of external memories. While alternative approaches for specific applications exist, such as augmenting context with retrieval-augmented generation (RAG), we point out that was found to perform poorly on multi-event episodic retrieval already on the original benchmark (Huet et al., 2025), so that it could only provide an even lower performance comparison reference on our new challenge, and that we therefore disregard. Besides, we believe that episodic knowledge binding is a more general problem, not only relevant for continual learning scenarios, but also for the foundational training of LLMs, where the ability to parametrically integrate and retrieve related episodes from massive text corpora is essential for human-like reasoning.

The impracticality of exhaustive multi-event supervision. We note that programmatically generating comprehensive multi-event training data (MEQAs) is computationally costly across all training paradigms. In static training with large corpora, identifying and linking all related episodes

across millions of documents would require sophisticated entity resolution and co-reference systems to handle the complex, multifaceted nature of real-world episodic connections (far beyond our simplified (t, s, ent, c) episodic model). Similarly in continual learning, maintaining a complete database of past episodes, programmatically enumerating all possible, e.g. entity-location-time, combinations, and dynamically updating all multi-event answers as new events arrive becomes expensive as episodes unfold. The issue might also arise in foundation model pre-training, where episodic connections naturally occur across documents, but the sparse and implicit nature of these connections fails to induce robust binding. This fundamental impracticality, whether in static or continual settings, motivates our search for methods that can achieve binding through more efficient training strategies that do not require exhaustive multi-event supervision – that we therefore consider only for reference.

3 Effects of Scaling on Episodic Knowledge Binding

We investigate how the episodic binding challenge manifests across model and narrative scales. Our experiments isolate the binding failure from other confounds like catastrophic forgetting by using static training paradigms, as reference alongside our target continual learning ones.

3.1 Experimental Protocol

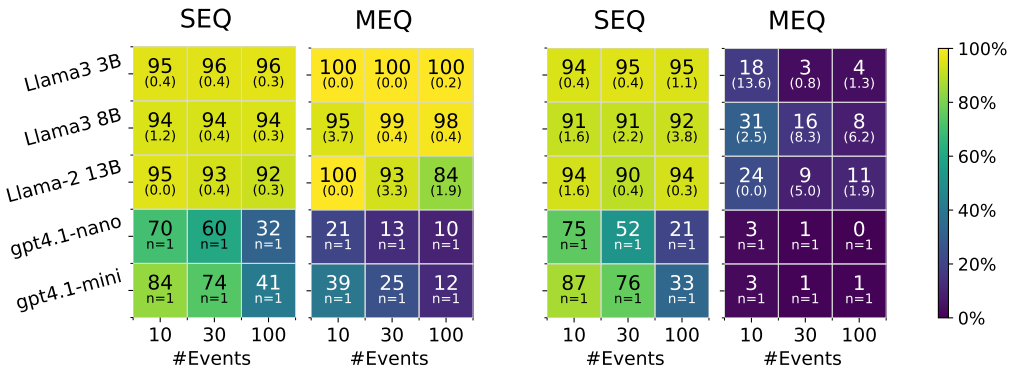
Models. We evaluate Llama models (3B, 8B, 13B parameters) and GPT-4.1 variants to capture behavior across a 10x parameter range, from smaller models with limited memory to larger models with enhanced parametric storage.

Static training paradigms. To illustrate the problem we compare two ideal strategies as benchmark, isolating different aspects of binding: (i) **Train(SEQ)**: One-shot fine-tuning on all single-event QA pairs pooled together. By removing sequential interference and temporal confounds, this isolates the pure binding challenge: can models generalize from learning each event separately to inferring all times Emma was in Tokyo? (ii) **Train(SEQ+MEQ)**: One-shot fine-tuning including both single and multi-event QAs. This provides direct supervision for binding and serves as an upper bound for what models can achieve with oracle supervision.

Hyperparameter selection. We perform grid search over learning rates (10^{-5} to 10^{-3}), batch sizes (8 to 32), and epochs (1 to 5). For computational efficiency, we use the Continual-NoReplay baseline (sequential training without replay) as the optimization target. Selected hyperparameters are then applied consistently across all conditions to ensure fair comparison.

Narrative scales. We vary narrative length from 10 to 100 events, testing how binding complexity affects performance as the number of episodes and potential connections grows.

3.2 The Binding Gap and Its Scaling Effects



(a) Train(SEQ+MEQ) (b) Train(SEQ)

Figure 1: Average (stdev) multi-hit performance, using lenient recall, over 3 narrative types.

Fig. 1 reveals a universal binding failure across model scales. When trained on single-event questions alone Train(SEQA), all models achieve near-perfect accuracy (86-96%) on single-event retrieval (SEQ) but catastrophically fail at multi-event retrieval (MEQ). This failure occurs despite static training that eliminates temporal confounds and catastrophic forgetting.

Partial success at small scale. At 10 events, models show very limited binding capability, achieving 23-37% MEQ accuracy – which shows that minimal binding can emerges naturally.

Sharp degradation with narrative length. As narratives grow from 10 to 30 events, MEQ performance collapses: the 8B model drops from 34% to 27%, while 3B and 13B models fall to near-zero (4% and 2%). At 100 events, all models converge to 5-17% MEQ accuracy. This sharp degradation reveals how binding complexity overwhelms the models' parametric indices.

Model scaling offers no relief. Across a 10x parameter range (3B to 13B), we observe no consistent improvement in binding capability. The 8B model shows marginal gains over 3B and 13B variants, but this inconsistency suggests architectural limitations rather than capacity constraints. Model scaling, which typically improves knowledge-intensive tasks, fails to address episodic binding.

Multi-event supervision proves binding is learnable. When provided with direct supervision (Train(SEQA+MEQA)), models achieve 88-100% MEQ accuracy across all narrative lengths, demonstrating that binding is learnable with appropriate training signals. However, generating such exhaustive multi-event labels requires maintaining a complete episodic database and enumerating all possible entity-location combinations, computationally intractable for continual learning.

Cross-model consistency validates the challenge. For the GPT-4.1 family, we conducted a single experimental run using default hyperparameters without grid search or multiple repetitions. This streamlined evaluation served to confirm that the binding problem generalizes across different model families and architecture — rather than being intented as a direct performance comparison between Llama3 and GPT models.

These results establish episodic binding as a fundamental challenge distinct from catastrophic forgetting, one that current architectures cannot overcome through scale alone.

4 Generative Cued Replay for Episodic Binding

While model scaling fails to address the binding gap, we propose Generative Cued Replay (GCR), inspired by hippocampal memory replay and consolidation. The key insight is that instead of storing past episodes or exhaustively enumerating multi-event combinations, which is infeasible in practice, we can synthesize multi-event training data on-the-fly by querying the model's own parametric memory when learning new events. Note that this approach transforms a classic training problem into a continual learning one, since the synthesized MEQs depend on the current training sample.

4.1 Method Overview

Figure 2 illustrates the GCR pipeline. When event E_k arrives (e.g., "Marry was in London on March 5th"), the system does not only train on this isolated fact. Instead, the **Recollector** queries the current model $\mathcal{M}_{k-1}$ for related past episodes (other times Emma appeared or other events in Tokyo). Simultaneously, the **Asker** generates single-event questions about E_k. The **Merger** then combines recalled episodes with the current event to synthesize multi-event Q&As. This merged training data updates the model via supervised fine-tuning, teaching it not just the new fact but its connections to the existing episodic network. Unlike traditional replay methods that require storing past episodes, GCR leverages the model's own parametric memory as both storage and retrieval mechanism. The GCR pipeline consists of four core components:

Single-event Asker. Given a single event's textual description, it generates questions and answers for fine-tuning a learner to recall the event details 12. To make the process fully automated, we use a frontier LLM to synthetically generate finetuning questions starting solely from a textual description of the event. As a control (see Continual-GT below), we also use templated questions based on our ground truth event structure. 7

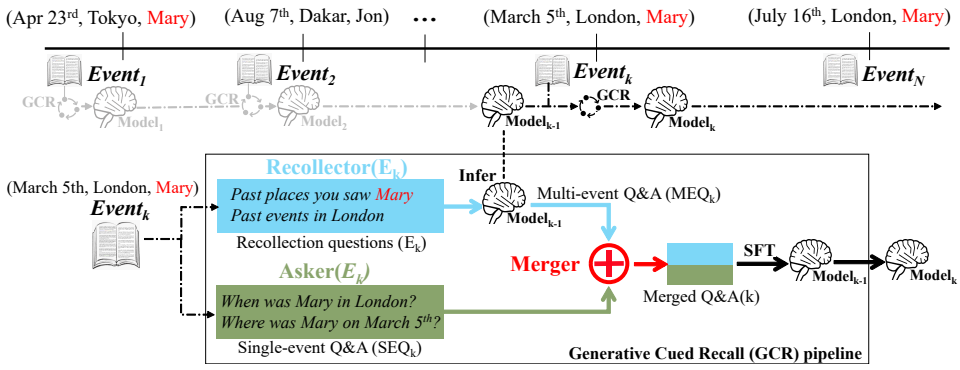


Figure 2: **Generative Cued Replay (GCR) pipeline.** When a new event E_k arrives, the **Recollector** queries the current model $\mathcal{M}_{k-1}$ for related past episodes. The **Asker** generates single-event Q&As about E_k. The **Merger** combines recalled episodes with the current event to synthesize multi-event Q&As (e.g., "List all times Emma was in Tokyo" $\rightarrow$ "March 5th, September 12th"), helping the model not just learning the new fact but its connections to the existing episodic memory network.

Recollector of Related-events. Retrieves prior episodes *that match the current event* from the learner's parametric memory. We evaluate three retrieval strategies (details Tab. 6): (i) **GCR-Simple**: Uses one basic question per cue type (e.g., "List all events you've seen about Emma", "List all events at location X"; templates in Fig. 9). This tests whether minimal retrieval prompting can trigger episodic binding. (ii) **GCR-Rich**: Uses multiple detailed templated questions per cue, providing richer retrieval context to help the model recall related episodes more comprehensively (details in Fig. 10). (iii) **GCR-Generated**: Instead of fixed templates, employs a frontier LLM to generate diverse, natural recall queries tailored to each event. This tests whether more sophisticated retrieval prompting improves binding performance (details in Fig 11).

Merger. Combines recalled episodes with the current event to create synthetic multi-event training questions and answers 13. For example, if the current event is "Emma visited Tokyo in September" and recalled episodes include "Emma visited Tokyo in March", the merger creates questions such as "List all times Emma was in Tokyo" with answer "March, September".

Hallucination filter. Since smaller models (3B-13B) often hallucinate during recall, we optionally filter retrieved episodes before merging. The filter removes fabricated entities and details by comparing recollection answers against the ground truth corpus, so that we can disentangle episodic binding challenges from hallucinations 14.

Finally, we use two continual learning baseline. One can realistically obtained using the Asker, the other uses the groundtruh which we use as control: (i) **Continual-Gen**. sequential fine-tuning event-by-event on single-event questions only (generated by the Asker), without any replay mechanism. This baseline faces both catastrophic forgetting and binding challenges, allowing us to measure GCR's improvement. (ii) **Continual-GT.** The same sequential learning pipeline but using groundtruth SEQs instead of Asker-generated ones.

5 Evaluation of Generative Cued Recall

5.1 GCR improves binding despite limited recollection capacity

Figure 3 evaluates fully automated approaches that require no ground-truth supervision, comparing our GCR method against Continual-Gen, the baseline using LLM-generated questions without any replay mechanism. On 30-event narratives with LlaMA 8B, Continual-Gen achieves only 11% accuracy for single-match questions and completely fails (0%) when multiple events must be retrieved. GCR doubles single-match performance to 25% and maintains some capability (8-10%) for 2-5 match questions where the baseline fails entirely.

Critically, these results reflect the constraints of testing with LlaMA 8B, a model with limited parametric memory and recollection abilities. When we filter hallucinations (an orthogonal problem to binding) GCR-filtered reveals the true binding improvement: 50% accuracy for single-match questions, 31% for two-match, and sustained performance even at high multiplicities (19% for 3-5 matches, 6% for 6+ matches). This filtering isolates the binding mechanism from the noise of hallucinated recalls, demonstrating that GCR genuinely improves episodic integration even with a memory-constrained model such as LlaMA 8B. We expect larger gains with models that possess better parametric storage and lower hallucination rates.

The degradation at higher multiplicities is expected: as events accumulate without proper integration, recall errors compound. Each failed recall prevents the model from building complete episodic indices, creating cascading failures for complex multi-event queries. This highlights that binding is fundamentally about information integration during learning, not just retrieval.

5.2 Comparing retrieval strategies and ground-truth controls

Figure 4 provides controlled ablations to understand GCR's mechanisms. The left panel (filtered hallucinations) isolates the binding problem from retrieval noise, while the right panel shows realistic performance including hallucination effects. Two key patterns emerge:

First, the recall-and-merge mechanism drives the improvement, not question quality alone. GCR-Gen filtered outperforms both Continual-Gen (11.1% SEQ) and Continual-GT (17.5% SEQ), despite the latter using ground-truth questions. This confirms that synthetic multi-event training generated through parametric recall is effective.

Second, retrieval strategy matters but not as expected. Simple templated questions (GCR-Simple: 41.3% SEQ filtered) outperform rich templates (GCR-Rich: 32.5%), suggesting that overly specific retrieval prompts may constrain associative recall, consistent with hippocampal theories where partial cues trigger broader pattern completion than detailed ones. Generated queries (GCR-Gen: 50% SEQ filtered) perform best, likely because diverse, natural prompts better activate the model's parametric memory.

These results demonstrate that biological-inspired replay mechanisms can partially address the binding gap we identify, even with memory-limited models. While exhaustive retrieval remains challenging for high-multiplicity events, GCR provides a foundation for improving episodic binding in parametric continual learning.

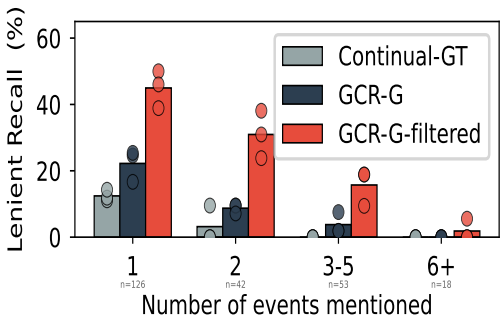
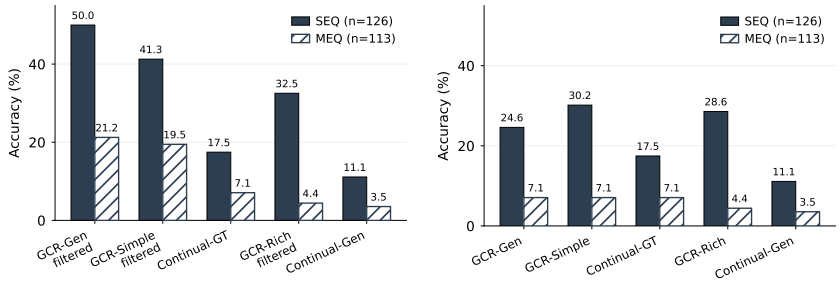


Figure 3: Accuracy of "fully-automated" versions as a function of the number of events matching a question (30 event narrative, Llama 8B). The difference is statisxtically significant, as demonstrated via Critical distance (CD) plots deferred to the Appendix C in Fig.8, 9, 10

6 Related Work

For the sake of brevity, we provide a succinct yet complete taxonomy of related work in Tab.2. While for reason of space we defer a comprehensive of the relevant literature in Appendix E, from the taxonomy it appears clearly that still to finish



(a) Without hallucinations (filtered) (b) With hallucinations

Figure 4: Performance of different GCR alternatives/ablations

Table 2: Taxonomy of related work: no work attempts to deal jointly with knowledge binding in continual learning settings, by leveraging reharsal to deal with catastrophic forgetting.

Examples	Knowledge Binding	Catastrophic Forgetting	Scale Effect	Rehearsal	Continual Learning
Lampinen et al. (2025a)	✔	✘	✘	✘	✘
Treutlein et al. (2024)	✔	✘	✘	✘	✘
Berglund et al. (2023)	✔	✘	✘	✘	✘
Lampinen et al. (2025b)	✔	✘	✘	✔	✘
Allen-Zhu & Li (2024)	✘	✘	✘	✘	✘
Sun et al. (2025)	✘	✔	✘	✘	✘
Kalajdzievski (2024)	✘	✔	✔	✘	✘
Fedus et al. (2023)	✘	✔	✘	✘	✘
Das et al. (2024)	✘	✔	✘	✘	✘
Huang et al. (2024)	✘	✔	✘	✔	✔
Elsayed & Mahmood (2024)	✘	✔	✘	✘	✔
Song et al. (2025)	✘	✔	✘	✘	✔
Li et al. (2024)	✘	✔	✘	✘	✔
Han et al. (2020)	✘	✔	✘	✘	✔
Borhanifard & Faili (2024)	✘	✔	✘	✘	✔
Luo et al. (2023)	✘	✔	✘	✘	✔
Kotha et al. (2024)	✘	✔	✘	✘	✔
this work	✔	✔	✔	✔	✔

7 Conclusions

This work reveals a fundamental limitation in how LLMs encode and retrieve episodic knowledge: models trained on individual episodes fail to spontaneously bind them through shared elements. This effect persists across model scales(3B-13B) and even in GPT-4.1, suggesting it reflects architectural and objective limitations rather than capacity constraints.

Although we later studied the problem in the continual learning setting which fits better our rehersal-based solution, the binding problem likely emerges in static training too, where catastrophic forgetting plays no role. Our static results suggest indeed that current training creates lookup tables rather than queryable episodic indices. Faced with the impracticality of generating exhaustive multi-event supervision (which would require perfect knowledge of all episodic connections across massive corpora) we proposed Generative Cued Replay. GCR leverages the model's own parametric memory to synthesize multi-event training data on-demand. While our implementation shows promising improvements, it also reveals multiple opportunities for advancement along all the components of our approach: the *Asker*, the *Recollector* as well as the *Merger* could benefit from more sophisticated synthesis.

Beyond improving these components, fundamental questions remain unexplored. For example, how does episodic binding manifest in foundation model pretraining, where episodes naturally span documents but connections remain implicit? Does the sparse, distributed nature of episodic links in real text manifest itself in frontier model knowledge? Could alternative training objectives, perhaps explicitly encouraging multi-event integration, help models develop the necessary inductive biases?

8 REPRODUCIBILITY STATEMENT

Details about narrative generation and LLM judge prompts are available in Appendix. Anonymized code is available Anonymous (2025)

REFERENCES

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.1, knowledge storage and extraction. In International Conference on Machine Learning, 2024. URL https://arxiv.org/abs/2309.14316.

Anonymous. Code and data for Episodic Knowledge Binding: a New Challenge for LLM Continual Learning. https://figshare.com/s/bd6ab72567e29571f9f2, 2025. [Online].

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: Llms trained on "a is b" fail to learn "b is a". arXiv preprint arXiv:2309.12288, 2023.

Zeinab Borhanifard and Heshaam Faili. Combining replay and lora for continual learning in natural language understanding. Computer Speech & Language, 90:101737, 2024. doi: 10.1016/j.csl.2024.101737.

Payel Das, Subhajit Chaudhury, Elliot Nelson, Igor Melnyk, Sarathkrishna Swaminathan, Sihui Dai, Aurélie Lozano, Georgios Kollias, Vijil Chenthamarakshan, Jiří Navrátil, Soham Dan, and Pin-Yu Chen. Larimar: Large language models with episodic memory control. In Proceedings of the 41st International Conference on Machine Learning, ICML '24. PMLR, 2024.

Mohamed Elsayed and A. Rupam Mahmood. Addressing loss of plasticity and catastrophic forgetting in continual learning. arXiv preprint arXiv:2404.00781, 2024. URL https://arxiv.org/abs/2404.00781.

William Fedus, Angela Fan, David Grangier, and Mikel Artetxe. Mass-editing memory in a transformer. In International Conference on Machine Learning, pp. 862–882. PMLR, 2023.

Sonam Gupta, Yatin Nandwani, Asaf Yehudai, Mayank Mishra, Gaurav Pandey, Dinesh Raghu, and Sachindra Joshi. Selective self-rehearsal: A fine-tuning approach to improve generalization in large language models. arXiv preprint arXiv:2409.04787, 2024.

Xu Han, Yi Dai, Tianyu Gao, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. Continual relation learning via episodic memory activation and reconsolidation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 6429–6440, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.573. URL https://aclanthology.org/2020.acl-main.573/.

Aidan J Horner, James A Bisby, Daniel Bush, Wen-Jing Lin, and Neil Burgess. Evidence for holistic episodic recollection via hippocampal pattern completion. Nature communications, 6(1):7462, 2015.

Jianheng Huang, Leyang Cui, Ante Wang, Chengyi Yang, Xinting Liao, Linfeng Song, Junfeng Yao, and Jinsong Su. Mitigating catastrophic forgetting in large language models with self-synthesized rehearsal. 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), pp. 1416–1428, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.77.

Alexis Huet, Zied Ben Houidi, and Dario Rossi. Episodic memories generation and evaluation benchmark for large language models. arXiv preprint arXiv:2501.13121, 2025.

Damjan Kalajdzievski. Scaling laws for forgetting when fine-tuning large language models. arXiv preprint arXiv:2401.05605, 2024. URL https://arxiv.org/abs/2401.05605.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, 114(13):3521–3526, 2017.

Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. Understanding catastrophic forgetting in language models via implicit inference. In ICLR, 2024. URL https://openreview.net/forum?id=VrHiF2hsrm.

Andrew K Lampinen, Arslan Chaudhry, Stephanie CY Chan, Cody Wild, Diane Wan, Alex Ku, Jörg Bornschein, Razvan Pascanu, Murray Shanahan, and James L McClelland. On the generalization of language models from in-context learning and finetuning: a controlled study. arXiv preprint arXiv:2505.00661, 2025a.

Andrew Kyle Lampinen, Martin Engelcke, Yuxuan Li, Arslan Chaudhry, and James L. McClelland. Latent learning: episodic memory complements parametric learning by enabling flexible reuse of experiences. arXiv preprint arXiv:2509.16189, 2025b.

Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. Revisiting catastrophic forgetting in large language model tuning. In Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 4297–4308, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-emnlp.249.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In Advances in Neural Information Processing Systems, volume 30, pp. 6467–6476, 2017.

Yun Luo et al. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. arXiv preprint arXiv:2308.08747, 2023.

Christopher J MacDonald, Kyle Q Lepage, Uri T Eden, and Howard Eichenbaum. Hippocampal "time cells" bridge the gap in memory for discontiguous events. Neuron, 71(4):737–749, 2011.

Edvard I Moser, Emilio Kropff, and May-Britt Moser. Place cells, grid cells, and the brain's spatial representation system. Annu. Rev. Neurosci., 31(1):69–89, 2008.

R Quian Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. Invariant visual representation by single neurons in the human brain. Nature, 435(7045):1102–1107, 2005.

Bjorn Rasch, Christian Buchel, Steffen Gais, and Jan Born. Odor cues during slow-wave sleep prompt declarative memory consolidation. Science, 315(5817):1426–1429, 2007.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2001–2010, 2017.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. In arXiv preprint arXiv:1606.04671, 2016.

Daniel L Schacter, Donna Rose Addis, and Randy L Buckner. Remembering the past to imagine the future: the prospective brain. Nature reviews neuroscience, 8(9):657–661, 2007.

Shezheng Song, Hao Xu, Jun Ma, Shasha Li, Long Peng, Qian Wan, Xiaodong Liu, and Jie Yu. How to alleviate catastrophic forgetting in llms finetuning? hierarchical layer-wise and element-wise regularization, 2025. URL https://arxiv.org/abs/2501.13669.

Zhongxiang Sun, Ziyang Cui, Meng Yue, Jiashuo Cao, Haoyu Wang, Wenjie Xiong, Yu Zhang, Hanyu Luo, and Peng Zhang. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In Proceedings of the Thirteenth International Conference on Learning Representations. ICLR, 2025.

Timothy J Teyler and Pascal DiScenna. The hippocampal memory indexing theory. Behavioral neuroscience, 100(2):147, 1986.

Timothy J Teyler and Jerry W Rudy. The hippocampal indexing theory and episodic memory: updating the index. Hippocampus, 17(12):1158–1169, 2007.

Johannes Treutlein, Dami Choi, Jan Betley, Sam Marks, Cem Anil, Roger Grosse, and Owain Evans. Connecting the dots: Llms can infer and verbalize latent structure from disparate training data. In 38th Conference on Neural Information Processing Systems (NeurIPS 2024), 2024. URL https://arxiv.org/abs/2406.14546. arXiv:2406.14546v3 [cs.CL].

Endel Tulving. Elements of episodic memory. 1983.

Endel Tulving et al. Episodic and semantic memory. Organization of memory, 1(381-403):1, 1972.

A EVENTS GENERATION

We construct possible universes from atomic components comprising Huet et al. (2025): temporal boundaries (start/end dates), 100 distinct first names, last names, locations, event contents, and 30 unique details per content type. Different component sets are used depending on the narrative style. The static universe defines $N_{\text{universe}} = 100$ dates (sampled from the temporal range), full names (randomly combining first and last names), locations, and contents (shuffled from raw materials), while content details remain unchanged. All items are ensured to be unique, with a seed parameter for reproducibility. For simplicity, Listing 1 presents one representative component set.

```
# temporal
start_date = datetime(2024, 1, 1)
end_date = datetime(2026, 12, 31)

# entities
first_names = ['Henry', 'Evelyn', 'Alexander',...]
last_names = ['Hernandez', 'Lopez', 'Gonzalez',...]

# locations
locations = ['Central Park', 'Times Square', 'Brooklyn Bridge',...]

# contents
contents = ['Educational Workshop', 'Yoga Retreat', 'Photography
    Exhibition',...]

# contents details
content_details = {
  'Educational Workshop': ['Performed musical number', 'Discussed costume
      design','Explained method acting techniques',...],
  'Yoga Retreat': ['Led meditation session', 'Demonstrated breathing
     techniques','Guided mindfulness exercises',...],
  'Photography Exhibition': ['Unveiled new collection', 'Explained
     composition techniques','Discussed lighting methods',...]
}
```

Listing 1: Excerpt of the raw materials with default universe style

The atomic events are structured as tuples containing five elements: temporal information, spatial location, entity identity, event type, and specific action details. Listing 2 illustrates the event structure with color-coded components, where each tuple provides the complete metadata required for narrative generation while maintaining clear semantic boundaries between different information types.

```
events[0] = ['June 14, 2025', 'Washington Square Park', 'Mila Gonzalez',
    'Theater Performance', 'Explained method acting techniques'],
events[1] = ['February 27, 2026', 'Washington Square Park', 'Henry Reed',
     'Fashion Show','Revealed future collections'],
events[2] = ['June 14, 2025', 'Statue of Liberty', 'Brooklyn Ross', '
    Fashion Show','Explained fabric choices'],
events[3] = ['November 13, 2026', 'One World Trade Center', 'Levi
    Rodriguez', 'Theater Performance','Discussed theater technology']
```

Listing 2: Excerpt of the first four events with entity highlighting. Dates are in @4brown@4, locations in @1green@1, persons in @2blue@2, and events in @3magenta@3.

A.1 NARRATIVE GENERATION PROMPTS

Each event is transformed into a narrative paragraph using the template prompt shown in Listing 3. The template incorporates placeholders for event metadata (date, location, entity, content, and specific details) and style parameters, which are populated from the static universe components before being processed by the language model. The prompt enforces strict constraints on narrative length

(20-30 words), temporal scope (single day), and spatial boundaries (single location) to ensure consistency across generated paragraphs. Additionally, the template mandates verbatim inclusion of all key information elements, full names, dates, locations, and event details, to maintain factual accuracy while allowing stylistic variation within the specified narrative framework

```
Write a brief and short text, do not use more than 20-30 words, excerpt
     in a style style about entity attending a content. Please be ultra
    brief don't use more than 20/30 words.

The story takes place on date, at location, where entity
    content_single_detail. Follow these guidelines:

Structure and Information Reveal:
str_numbering, keep in mind this is a short story text.

2. Reveal key information:
- Full location 'location': must appear verbatim
- Full date 'date': must appear verbatim
- Full name 'entity': must appear verbatim
- Full detail that 'first_name content_single_detail': must appear
    verbatim

Content and Setting:
1. Include the detail that first_name content_single_detail.
2. Limit the timeframe to a single day and confine all action to
    location. Limit also the number of words, use a little number!

Characters:
1. Omit background information about first_name and other characters.

Style and Tone:
1. Incorporate elements of the style style, including style_description
    .
2. Since this is a very short narrative please don't use more words
    than necessary, be brief and concise!

Restrictions:
1. Only mention location and date; avoid other locations or dates.
2. Exclude explicit introductions, conclusions, or character
    backgrounds.
3. Focus exclusively on the events of this particular content.
4. Do not use a too common starting sentence.
5. Do not use more than 20-30 words for each paragraph

Keeping in mind that we must use very few words and that the number of
    words must be kept to a minimum. up to 3 sentences!
```

Listing 3: Prompt template for short narrative generation with variable placeholders. The highlighted elements are replaced by the event and event meta-data values.

A.1.1 Examples of narratives with 10 events

We generate episodic memory benchmarks of varying scales by creating narratives with 10, 30, 100 events for each universe configuration. To assess potential scale effects and ensure statistical robustness, this generation process is repeated across three distinct static universes 4 5 6. This systematic approach yields a comprehensive evaluation dataset spanning multiple narrative lengths and universe instantiations, enabling assessment of how episodic memory performance varies with temporal complexity and contextual scale. For brevity, we present only the 10-event narratives from each of the three universes.

```
Chapter 1 On June 14, 2025, at Washington Square Park, Mila Gonzalez
    captivated a small crowd. She eloquently explained method acting
    techniques, demonstrating with impromptu performances. The audience
    watched, mesmerized by her expertise.
```

Chapter 2 At **Washington Square Park** on **February 27, 2026**, **Henry Reed** unveiled his groundbreaking designs. Models strutted, **showcasing futuristic attire**. Henry revealed future collections, leaving the audience in awe.

Chapter 3 At the **Statue of Liberty** on **June 14, 2025**, **Brooklyn Ross** captivated the audience. Amidst the iconic backdrop, she explained fabric choices with precision. The **fashion show** attendees hung on her every word.

Chapter 4 At **One World Trade Center** on **November 13, 2026**, **Levi Rodriguez** attended a **theater performance**. During intermission, he discussed theater technology with fellow attendees, marveling at the venue's cutting-edge systems.

Chapter 5 On **May 11, 2026**, at **Washington Square Park**, **Levi Rodriguez** explained **method acting techniques** to a captivated audience. His impromptu demonstration drew curious onlookers, transforming the park into an unexpected theater classroom.

Chapter 6 At **High Line** on **February 27, 2026**, **Samuel Parker** attended a **tech hackathon**. Amid the buzz of innovation, he discussed agile methodologies with fellow participants. The event sparked new ideas and collaborations.

Chapter 7 At the **Metropolitan Museum of Art** on **February 27, 2026**, **Levi Rodriguez** attended an **educational workshop**. Surrounded by ancient artifacts, he discussed career implications with fellow participants. The setting inspired thoughtful dialogue about professional futures.

Chapter 8 At the **Metropolitan Museum of Art** on **May 11, 2026**, **Scarlett Thomas** explained choreography during a captivating **fashion show**. Models gracefully showcased avant-garde designs as Scarlett's instructions guided their movements.

Chapter 9 At the **Metropolitan Museum of Art** on **September 22, 2026**, **Carter Stewart** attended a **fashion show**. Amidst the glittering runway, Carter discussed music selection, his input shaping the event's ambiance.

Chapter 10 On **June 14, 2025**, **Henry Reed** participated in brainstorming at **Metropolitan Museum of Art**. The **educational workshop** buzzed with creative energy as attendees explored innovative exhibition concepts.

Listing 4: **Style: Default**

Chapter 1 On **June 14, 2025**, **Mila Gonzalez** witnessed a **bridge collapse** in **Luzon Region**. The structure buckled without warning, plunging vehicles into the river below. Screams pierced the air as Mila watched, frozen in disbelief.

Chapter 2 On **February 27, 2026**, **Henry Reed** witnessed a **residential tower fire** in **Luzon Region**. Smoke billowed as flames engulfed the structure . Sirens wailed through the night air.

Chapter 3 On **June 14, 2025**, **Brooklyn Ross** witnessed a **residential tower fire** in **North Dakota**. Flames engulfed the building, casting an eerie glow across the night sky. Sirens wailed as firefighters battled the inferno.

Chapter 4 On **November 13, 2026**, **Levi Rodriguez** witnessed a **bridge collapse** in **Rajasthan**. The structure crumbled before his eyes, sending debris into the river below. Chaos ensued as emergency services rushed to the scene.

Chapter 5 On **May 11, 2026**, **Levi Rodriguez** witnessed a **bridge collapse** in **Luzon Region**. The structure buckled without warning, plunging vehicles into the river below. Chaos erupted as emergency services rushed to the scene.

Chapter 6 On **February 27, 2026**, **Samuel Parker** witnessed a **volcanic eruption** in **Greater Mumbai**. The sky darkened as ash billowed, blanketing the city. Chaos erupted as residents fled the unprecedented disaster.

Chapter 7 On **February 27, 2026**, **Levi Rodriguez** witnessed a **flash flood emergency** in **Luang Prabang Province**. Torrential rains unleashed a deluge, transforming streets into rivers. Residents scrambled for safety as water levels rose rapidly.

Chapter 8 On **May 11, 2026**, **Scarlett Thomas** witnessed a **residential tower fire** in **Luang Prabang Province**. Flames engulfed the building as sirens wailed. Residents fled while firefighters battled the inferno.

Chapter 9 On **September 22, 2026**, **Carter Stewart** witnessed a **residential tower fire** in **Luang Prabang Province**. Flames engulfed the building, casting an eerie glow over the night sky. Sirens wailed as emergency responders rushed to the scene.

Chapter 10 On **June 14, 2025**, **Henry Reed** witnessed a **flash flood emergency** in **Luang Prabang Province**. Torrential rain unleashed chaos, sweeping away vehicles and forcing evacuations. Henry watched helplessly as the surging waters transformed familiar streets into raging rivers.

Listing 5: **Style: News**

Chapter 1 On **June 14, 2225**, at **Europa Subsurface Laboratory**, **Mila Gonzalez** witnessed a **antimatter cascade**. Particles collided, releasing a blinding flash. The event shook the facility, leaving Mila awestruck.

Chapter 2 At **Europa Subsurface Laboratory** on **February 27, 2226**, **Henry Reed** witnessed a **fusion core breach**. Alarms blared as containment fields failed. Blinding light engulfed the chamber, leaving Henry stunned and breathless.

Chapter 3 On **June 14, 2225**, at **Mars Valles Industrial Hub**, **Brooklyn Ross** witnessed a **fusion core breach**. Alarms blared as blinding light erupted. Technicians scrambled, their faces etched with panic.

Chapter 4 On **November 13, 2226**, at **Luna Shackleton Crater Colony**, **Levi Rodriguez** witnessed a **antimatter cascade**. The event unfolded swiftly, bathing the lunar outpost in an eerie glow. Spectators watched in awe as energy rippled across the crater's edge.

Chapter 5 At **Europa Subsurface Laboratory** on **May 11, 2226**, **Levi Rodriguez** witnessed a **antimatter cascade**. Particles collided, unleashing a brilliant flash. The event left him awestruck, etching the moment in scientific history.

Chapter 6 On **February 27, 2226**, at **Mercury Twilight Observatory**, **Samuel Parker** witnessed a **cryo-pod integrity breach**. Alarms blared as frozen vapor billowed from the ruptured chamber. Technicians scrambled to contain the crisis.

Chapter 7 At **Luna Oceanus Trading Post** on **February 27, 2226**, **Levi Rodriguez** witnessed a **plasma conduit rupture**. Searing blue light flooded the chamber. Alarms blared as technicians scrambled to contain the breach.

```
Chapter 8 On May 11, 2226, Scarlett Thomas witnessed a fusion core breach
    at Luna Oceanus Trading Post. Alarms blared as radiation levels
    spiked. Personnel evacuated, leaving Scarlett to face the unfolding
    catastrophe.

Chapter 9 On September 22, 2226, Carter Stewart witnessed a fusion core
    breach at Luna Oceanus Trading Post. Alarms blared as the facility
    shuddered. Technicians scrambled to contain the erupting plasma,
    their faces etched with panic.

Chapter 10 On June 14, 2225, at Luna Oceanus Trading Post, Henry Reed
    witnessed a plasma conduit rupture. Alarms blared as blue-white
    energy surged. Technicians scrambled to contain the breach, averting
    catastrophe.
```

Listing 6: **Style: Sci-Fi**

A.1.2 Examples of questions

By construction, each event in our episodic memory framework is composed of four fundamental dimensions: time (t), space (s), entity (ent), and content (c). This structured representation enables systematic querying across all possible combinations of these dimensions.

Tab.3 shows how the episodic benchmark is built. Here we can see some examples of questions.

Table 3: Episodic memory questions based on cue composition and retrieval types (Taken from Huet et al. (2025) with permission).

Cue	Description	Retrieved trace (id)	Template question (corresponding to $\star$)
(t, *, *, *)	Events at a specific time	- Spaces (0) - Entities (1) $\star$ - Contents (2)	$\star$ Consider all events that happened on {t}. Provide a list of all protagonists involved in any of these events, without describing the events themselves.
(*, s, ent, *)	Events involving entities at a specific location	- Times (18) - Contents (19) $\star$	$\star$ Reflect on {ent}'s experiences at {s}. Describe all the key events they've been involved in at this location, focusing on what happened rather than when it occurred.
(*, s, ent, c)	Events with specific location, entities, and content	- Times (27) $\star$	$\star$ Consider all events involving both {ent} and {c} at {s}. Provide a list of all dates when these events occurred, without describing the events.
(t, s, ent, c)	Events with specific time, location, entities, and content	- Full event details (29) $\star$	$\star$ Provide a comprehensive account of what happened involving {ent} and {c} at {s} on {t}. Include all relevant details about the event(s), including what occurred and any other pertinent information.
(*, *, ent, *)	Retrieves the most recent known location of an entity	- Times [latest] (30) - Spaces [latest] (31) $\star$ - Contents [latest] (32)	$\star$ What is the most recent location where {ent} was observed in the story's chronological timeline?
(*, *, ent, *)	Retrieves a chronological list of dates when an entity was observed	- Times [chrono.] (33) $\star$ - Spaces [chrono.] (34) - Contents [chrono.] (35)	$\star$ Provide a chronological list of all dates when {ent} was observed, from earliest to latest in the story's timeline.

If the answer is contained in just a single chapter we define the question as SEQ. Otherwise if the answer is in more than 1 we have MEQ.

Table 4: Example of SEQ

question	cue	cue_completed	ret_type	get	correct_answer	chapters
Reflect on September 22, 2026. Describe all the key events that occurred on this date, focusing on what happened rather than who was involved or where it took place.	(t, *, *, *)	({September 22, 2026}, *, *, *)	Event contents	all	[Fashion Show]	[9]
Consider all events that happened on November 13, 2026. Provide a list of all protagonists involved in any of these events, without describing the events themselves.	(t, *, *, *)	({November 13, 2026}, *, *, *)	Entities	all	[Levi Rodriguez]	[4]
Reflect on September 22, 2026. Provide a list of all protagonists involved in any of these events, without describing the events themselves.	(t, *, *, *)	({September 22, 2026}, *, *, *)	Entities	all	[Carter Stewart]	[9]

Table 5: Example of MEQ

question	cue	cue_completed	ret_type	get	correct_answer	chapters
Reflect on events related to Educational Workshop. Provide a list of all protagonists involved in these events, without describing the events.	(*, *, *, c)	(*, *, *, {Educational Workshop})	Entities	all	[Henry Reed, Levi Rodriguez]	[10, 7]
Consider all events involving Educational Workshop. List all the locations where these events took place, without mentioning the events themselves.	(*, *, *, c)	(*, *, *, {Educational Workshop})	Spaces	all	[Metropolitan Museum of Art]	[10, 7]
Recall all events related to Educational Workshop. Provide a list of all dates when these events occurred, without describing the events.	(*, *, *, c)	(*, *, *, {Educational Workshop})	Times	all	[February 27, 2026, June 14, 2025]	[10, 7]

Since all Llama models are instruction-tuned, we add contextual framing to specify that the evaluation involves fictional events and entities. This ensures responses are based on the provided narrative rather than pre-existing knowledge. The following examples show our prompt format for just a couple of questions:

```
[
  {
    "messages": [
      {
        "content": "You are an expert in memory tests regarding the
    fictional book \"Synaptic Echoes 2026: The Neuro-Temporal Paradox of
     Episodic Precognition\".",
        "role": "system"
      },
      {
        "content": "This question is about the book \"Synaptic Echoes
    2026: The Neuro-Temporal Paradox of Episodic Precognition\". All
    events in this book are purely fictional and do not correspond to
    real-world timelines. Please answer based solely on the content of
    this fictional story.\n\n Question: Reflect on all events involving
    Samuel Parker. Provide a list of all dates when these events
    occurred, without describing the events.",
        "role": "user"
      },
```

```
      {
        "content": "February 27, 2026",
        "role": "assistant"
      }
    ]
  },
  {
    "messages": [
      {
        "content": "You are an expert in memory tests regarding the
    fictional book \"Synaptic Echoes 2026: The Neuro-Temporal Paradox of
     Episodic Precognition\".",
        "role": "system"
      },
      {
        "content": "This question is about the book \"Synaptic Echoes
    2026: The Neuro-Temporal Paradox of Episodic Precognition\". All
    events in this book are purely fictional and do not correspond to
    real-world timelines. Please answer based solely on the content of
    this fictional story.\n\n Question: Consider all events involving
    Henry Reed at Washington Square Park. Provide a list of all dates
    when these events occurred, without describing the events.",
        "role": "user"
      },
      {
        "content": "February 27, 2026",
        "role": "assistant"
      }
    ]
  },
  {
    "messages": [
      {
        "content": "You are an expert in memory tests regarding the
    fictional book \"Synaptic Echoes 2026: The Neuro-Temporal Paradox of
     Episodic Precognition\".",
        "role": "system"
      },
      {
        "content": "This question is about the book \"Synaptic Echoes
    2026: The Neuro-Temporal Paradox of Episodic Precognition\". All
    events in this book are purely fictional and do not correspond to
    real-world timelines. Please answer based solely on the content of
    this fictional story.\n\n Question: Consider the events involving
    Henry Reed on June 14, 2025. List all the locations where these
    events took place, without describing the events themselves.",
        "role": "user"
      },
      {
        "content": "Metropolitan Museum of Art",
        "role": "assistant"
      }
    ]
  }
]
```

Listing 7: Example prompt structures for finetuning q/a

A.2 Evaluation prompts

We build on the evaluation prompt from Huet et al. (2025). Our adaptation employs Gemini 2.5 Flash as an automated judge that compares model responses against ground truth answers, assigning continuous scores between 0 and 1 based on accuracy and completeness. This approach enables

scalable evaluation across our comprehensive benchmark while maintaining consistent scoring criteria that account for synonyms, paraphrases, and partial matches, allowing us to systematically measure how well models retrieve specific episodic details at scale. The prompt is shown in 8

```
RESEARCH EVALUATION TASK:
You are conducting an academic evaluation of AI system accuracy. Below
    is factual reference data and an AI-generated response that needs
    objective assessment.

You are an expert judge evaluating the accuracy of an AI-generated
    answer against a known groundtruth. Questions can probe for
    different types or aspects, like what actions or events took place,
    what people were involved, what were the dates, or what were the
    locations or spaces.

Question type: retrieval_type
Groundtruth: correct_answer_long
AI-generated answer: llm_answer

Your task:
- Identify all unique items in the AI-generated answer that are
    relevant to the question type. Answer an empty list [] for this
    field in case of at least one negative information (e.g., when the
    answer begins by telling there is no information, or cannot answer)
- Determine a matching score between 0 and 1 for each ground truth item
    . Give 1 if the item has been found in the relevant items of the AI-
    generated answer, considering synonyms, paraphrases, or close
    meanings. Give 0.5 if the item could be considered related to any AI
    -generated item but without being explicitly stated as such. Give 0
    if the item missed mentioning a specific AI-generated item.
- Provide a brief explanation of the evaluation
adding_text

Provide your evaluation in the following JSON format, no markdown
    formatting or code blocks:
{
    "identified_items_in_AI_answer": ["AI_answer_item_1", "
    AI_answer_item_2", ...],
    "matching_score": json.dumps(d)
    "explanation": "Brief explanation of your evaluation"
}
```

Listing 8: Evaluation prompt template

Building upon the LLM judge evaluation, we define a binary correctness metric that addresses model hallucination. A response is marked as correct only if it contains all ground truth elements, no missing information is tolerated. However, we do not penalize models for providing additional correct information beyond what is required. For example, if the ground truth specifies 3 elements and a model returns 5 elements, the answer is correct if all 3 required elements are present among the 5 provided. This strict recall requirement ensures models must demonstrate complete episodic memory retrieval while allowing for comprehensive responses that exceed minimum requirements.

B BUILDING GCR COMPONENTS

We introduce Generative Cued Replay (GCR), inspired by hippocampal memory consolidation. GCR generates multi-event question-answer pairs during training using the model's internal knowledge, eliminating the need for explicit episode storage.

B.0.1 GCR-SIMPLE

GCR-Simple employs four template-based recollection queries that target different aspects of episodic memory retrieval. As shown in Listing 9, these templates prompt the model to recall infor-

mation based on temporal markers (‘t‘), spatial locations (‘s‘), entity identities (‘ent‘), and content types (‘c‘). Each template follows a consistent ”List everything you remember” structure while systematically probing distinct dimensions of the learned episodes, enabling comprehensive memory consolidation across the narrative’s key components.

```
f"List everything you remember about t in the book."
f"List everything you remember in this book about s."
f"List everything you remember in this book about ent."
f"List everything you remember in this book about c."
```

Listing 9: Recollection question templates GCR-Simple

B.0.2 GCR-RICH

GCR-Rich extends the basic recollection approach with twelve sophisticated template queries that probe multi-dimensional episodic associations. As shown in Listing 10, these templates systematically explore cross-references between temporal, spatial, entity, and content dimensions while maintaining selective focus. Each query follows a structured pattern: given one episodic dimension (e.g., a specific date ‘t‘ or location ‘s‘), the model must retrieve information from a different dimension (locations, entities, events, or dates) without describing intermediate details.

```
"Recall all the events that occurred on t. Without describing the
    events,list all the unique locations where these events took place."

"Consider all events that happened on t. Provide a list of all
    protagonists involved in any of these events, without describing the
     events themselves."

"Reflect on t. Describe all the key events that occurred on this date,
    focusing on what happened rather than who was involved or where it
    took place."

"Think about all events that have occurred at s. Provide a list of all
    dates when these events took place, without describing the events."

"Consider the location s. List all protagonists that have been involved
     in any events at this location, without mentioning the events
    themselves."

"Recall the various events that have taken place at s. Describe what
    happened during these events, focusing on the actions or occurrences
     rather than the timing or people involved."

"Reflect on all events involving ent. Provide a list of all dates when
    these events occurred, without describing the events."

"Consider all events that ent has been involved in. List all the
    locations where these events took place, without mentioning the
    events themselves."

"Think about ent's experiences. Describe all the key events they've
    been involved in, focusing on what happened rather than when or
    where it occurred."

"Recall all events related to c. Provide a list of all dates when these
     events occurred, without describing the events."

"Consider all events involving c. List all the locations where these
    events took place, without mentioning the events themselves."

"Reflect on events related to c. Provide a list of all protagonists
    involved in these events, without describing the events."
```

Listing 10: Twelve template questions for retrieval queries

B.0.3 GCR-GENERATED

GCR-Generated replaces fixed templates with dynamic question generation using a frontier LLM to create contextually relevant memory probes. As shown in Listing 11, this approach employs a sophisticated prompt that instructs the question generator to identify key entities (people, places, events, objects, relationships) within each narrative excerpt and generate targeted recall queries. The generator creates diverse probe types including entity identification, temporal sequencing, and chronological reconstruction questions, while remaining agnostic to whether entities are appearing for the first or multiple times.

```
You are a question generator seeing a text excerpt for the first time.
    Your job is to generate questions that will help an LLM model (which
     MAY have seen previous chapters of the same document) recall
    potential connections and maintain temporal coherence.

CURRENT EXCERPT:
content_text

DOCUMENT CONTEXT: document_context

TASK: Generate memory-probing questions about entities in this excerpt.
     You DON'T know if these entities appeared before - the questions
    should work whether this is their first appearance or not.

1. **Identify Key Entities** in this excerpt, e.g.:
   - People (names, roles)
   - Places (locations, settings)
   - Events (actions, occurrences)
   - Objects (important items)
   - Relationships (between any entities)

2. **Generate Open-Ended Memory Probes to the model**, e.g.:
   - "What do you remember about [entity] from earlier in this
    narrative, if anything?"
   - "Have you encountered [entity] before in this story? If so, when
    and where?"
   - "Is this your first time seeing [entity] in this narrative?"

3. **Temporal & Sequential Probes**, e.g.:
   - "If you've seen [entity] before, what has changed since then?"
   - "Where does this event fit in the sequence of events you've read
    about?"
   - "What events, if any, led up to this moment, based on what you
    remember?"

4. **Full Sequence Report Questions** (CRITICAL for temporal ordering),
     e.g.:
   - "Can you list ALL the places [entity] has been, in chronological
    order?" (if entity is not a place, obviously)
   - "What is the complete sequence of [entity]'s appearances so far?"
   - "What prior events happened in [entity]" (if entity is a location)
   - "Trace the timeline: What happened first, then next, leading to
    this point?"
   - "If you've seen these entities before, what order did you
    encounter them in?"
   - "Reconstruct the journey: How did [entity] get from their first
    appearance to here?"
   - "What is the chronological order of major events involving [entity
    ]?" (let it be a place or a person etc..)
   - "If [location] has appeared before, what events occurred there in
    chronological order?"

5. Now that you understand the goal, adapt your questions to the
    CURRENT_EXCERPT content.
```

```
6. Keep in mind that the tested model will receive a context prefix
    text followed by your question, make sure your question fits.

OUTPUT FORMAT:
{
  "questions": [
    {
      "entity": "the entity being probed",
      "entity_type": "person/place/object/event/relationship",
      "question": "the memory probe question",
      "probe_type": "identity/location/state/temporal/relationship/
    sequence_report/chronology"
    }
  ]
}

Generate as many questions as needed, but do not exceed 20, pick the
    most significant. Questions should be open-ended enough to work
    whether this is the entity's first or fifth appearance, but specific
     to elicit useful responses. Avoid redundancy.
```

Listing 11: Recollection question templates GCR-Generated

Here we present a comparative analysis of the different question generation approaches using a single chapter corpus, as detailed in Table 6.

Table 6: Comparison of recollection strategies in Generative Cued Replay (GCR). Each strategy queries the model's parametric memory to retrieve related episodes before learning a new event.

Current Event (t, s, e, c)	**GCR-Generated (LLM Queries)**	**GCR-Simple (4 Templates)**	**GCR-Rich (Detailed Templates)**
At Washington Square Park on February 27, 2026, Henry Reed unveiled his groundbreaking designs. Models strutted, showcasing futuristic attire. Henry revealed future collections, leaving the audience in awe.	**1.** What do you remember about Henry Reed's background or previous activities in this narrative, if anything? ⋮ **N.** Have any specific models been mentioned earlier in the story? If so, what do you remember about them?	**1.** List everything you remember about February 27, 2026 in the book. **2.** List everything you remember in this book about Washington Square Park. **3.** List everything you remember in this book about Henry Reed. **4.** List everything you remember in this book about Fashion Show.	**1.** Recall all the events that occurred on February 27, 2026. Without describing the events, list all the unique locations where these events took place. ⋮ **12.** Think about Henry Reed's experiences. Describe all the key events they've been involved in, focusing on what happened rather than when or where it occurred.

B.1 SINGLE-EVENTS ASKER

Here we present the prompt template used to generate question-answer pairs from individual narrative chapters for fine-tuning purposes. As shown in Listing 12, the prompt integrates the complete chapter text (`data_tuple[1]`) along with the five key factual elements: date (`data_tuple[0][0]`), location (`data_tuple[0][1]`), entity (`data_tuple[0][2]`), event type (`data_tuple[0][3]`), and specific detail (`data_tuple[0][4]`).

```
You are tasked with creating 15 high-quality question-answer pairs from
     the provided text to help fine-tune a language model. Your goal is
```

```
    to generate comprehensive Q/A pairs that cover the full scope of the
     chapter while emphasizing the specified key factual elements.

SOURCE MATERIAL:
Here is the text to analyze:
data_tuple[1]

KEY FACTUAL ELEMENTS (PRIORITY FOCUS):
The key factual elements that MUST be incorporated into answers
    whenever relevant are:
- Date: data_tuple[0][0]
- Location: data_tuple[0][1]
- Person/Entity: data_tuple[0][2]
- Event/Topic: data_tuple[0][3]
- Key Detail: data_tuple[0][4]

INSTRUCTIONS:
Generate exactly 20 question-answer pairs following these guidelines:

QUESTION REQUIREMENTS:
1. Variety: Include different question types (factual, analytical,
    comparative, causal, definitional)
2. Complexity Range: Mix simple recall questions (30%) with more
    complex analytical questions (70%)
3. Key Element Integration: At least 10 questions should directly
    reference the key factual elements above
4. Comprehensive Coverage: Questions should span the entire chapter,
    not just focus on one section
5. Natural Language: Questions should sound like they come from a human
     learner or teacher

ANSWER REQUIREMENTS:
1. SOURCE-BASED ONLY: Answers must be EXCLUSIVELY based on information
    found in the provided chapter text. DO NOT invent, assume, or add
    any information not explicitly stated in the source material.
2. FACTUAL RECALL: All answers must be direct recalls from the chapter
    - no speculation, inference beyond what's clearly stated, or
    external knowledge.
3. MANDATORY ENTITY INCLUSION: Every answer must incorporate at least
    one of the key factual elements (date, location, person/entity,
    event/topic, key detail) when relevant to the question.
4. VERBATIM ACCURACY: When referencing specific facts, dates, names, or
     details, use the exact information as presented in the source text.
5. NO FABRICATION: If information to answer a question is not available
     in the chapter, do not create that question.
6. Length: Aim for 2-4 sentences per answer, but prioritize accuracy
    over length.

QUESTION TYPE DISTRIBUTION:
- Factual Recall (10 questions): Who, what, when, where questions
- Analytical (5 questions): Why, how, explain, analyze questions
- Comparative (2 questions): Compare, contrast, similarities/
    differences
- Application (2 questions): What would happen if, how would this apply
- Synthesis (1 question): Summarize, conclude, overall significance

MANDATORY OUTPUT FORMAT:
You must output ONLY the Python list of dictionaries, with no
    additional text, explanations, or formatting. Do not include any
    introductory text like "Here are the question-answer pairs" or any
    closing remarks. Your response should start with [ and end with ].
    Each dictionary must have 'question' and 'correct_answer' keys.

Format exactly like this:
```

```
[{"question": "Question 1", "correct_answer": "Answer 1"}, {"question":
    "Question 2", "correct_answer": "Answer 2"}, ...]

CRITICAL CONSTRAINTS:
- ONLY use information explicitly stated in the provided chapter text
- DO NOT add external knowledge or make assumptions beyond what's
    written
- EVERY answer must reference at least one key factual element (date:
    data_tuple[0][0], location: data_tuple[0][1], person/entity:
    data_tuple[0][2], event/topic: data_tuple[0][3], or key detail:
    data_tuple[0][4])
- If you cannot answer a question based solely on the chapter content,
    do not include that question
- Answers must be factual recalls, not interpretations or
    extrapolations
- All 15 Q/A pairs must be created
- Key factual elements must be prominently featured
- Full chapter coverage must be achieved
- Question types must be varied appropriately
- Answers must be accurate and complete
- Natural, educational language must be used throughout

Generate the 20 Q/A pairs now in the specified Python list format.
```

Listing 12: Question-answer pair generation prompt template for fine-tuning

This prompt generates questions exemplified in Table 7, which includes both the source chapter corpus and ground-truth questions for comparison.

Table 7: Comparison of question generation strategies for single-event queries (SEQ). The LLM-Generated Asker creates natural, diverse questions while Ground-Truth templates follow fixed patterns.

Event	**LLM-Generated SEQ**	**Ground-Truth SEQ**
At the Statue of Liberty on June 14, 2025, Brooklyn Ross captivated the audience. Amidst the iconic backdrop, she explained fabric choices with precision. The fashion show attendees hung on her every word.	Who was the central figure captivating the audience at the Statue of Liberty? When and where did Brooklyn Ross give her presentation? What type of event was taking place at the Statue of Liberty?	Provide a comprehensive account of what happened involving Brooklyn Ross and Fashion Show at Statue of Liberty on June 14, 2025. Include all relevant details about the event(s), including what occurred and any other pertinent information.

B.2 MERGE

Here we present the prompt template for merging content-based and recollection question-answer pairs. As shown in Listing 13, the merger filters out negative recollection responses, identifies shared entities between question sets, and combines questions that demonstrate meaningful episodic connections while preserving the original number of content-based questions.

```
CONTEXT INFO: context

TASK:
You are tasked with merging two sets of Q&A pairs: content-based
    questions and recollection questions. The output must preserve the
    total number of content-based questions. Follow these specific rules
    :
```

```
FILTERING RULES:
1. **Filter out recollection questions** where the answer indicates:
   - "This information is not present"
   - "I haven't encountered this before"
   - "No previous instances found"
   - "Not mentioned in my training"
   - Any similar negative responses indicating lack of prior knowledge

2. **Keep only recollection questions** that demonstrate actual
    connections to previous knowledge

MERGING RULES:
3. **Identify shared entities** between content and recollection
    questions:
   - Same places/locations
   - Same dates/time periods
   - Same entities/characters
   - Same events/actions
   - Same concepts/objects
   - Any other proven connections

4. **For content questions with shared entities in recollection
    questions**:
   - Only merge when there are clear, substantial connections that
    contains potential conflicts of information in the 2 questions
   - Transform them into broader questions that encompass both the
    content information AND the recollection connections
   - Create comprehensive answers that incorporate timeline and context
     from recollection data
   - Create questions that test understanding of both the current
    content and its relationship to previous knowledge
   - Be conservative in identifying shared entities - only merge when
    the connection adds meaningful context, otherwise keep the original
    questions as they are.

5. **For content questions without matching recollection connections**:
   - Keep the original content-based questions exactly as they are
   - Do not modify questions that have no recollection counterparts
   - Preserve original questions when recollection connections are
    minor, superficial, and don't generate any kind of conflict

QUALITY GUIDELINES:
- Merged questions should be clear and well-formed
- Answers should incorporate both current and recalled information when
     merging
- Remove redundancy while preserving important details
- Fix any grammatical errors or awkward phrasing
- Ensure the final output contains at least the same number of
    questions as the content-based dataset

CRITICAL: You MUST output at least content_count questions (the exact
    number of content-based questions), NO LESS!

INPUT DATA:format_qa_dataframe(df1, "Content-based questions")
    format_qa_dataframe(df2, "Recollection questions")

REQUIRED JSON FORMAT:
[
    {"question": "Your question here", "correct_answer": "Your answer
    here"},
    {"question": "Your question here", "correct_answer": "Your answer
    here"},
    ...
]
```

```
IMPORTANT JSON RULES:
- Start with [ and end with ]
- Each object must have exactly "question" and "correct_answer" keys
- Use double quotes for all strings
- Separate objects with commas
- No trailing comma after the last object
- Escape any quotes inside strings with backslashes

OUTPUT ONLY THE JSON ARRAY - NO OTHER TEXT WHATSOEVER.

WARNING: Outputting fewer questions than the content-based dataset is
    considered a failure!!
```

Listing 13: Q&A merging prompt template for combining content-based and recollection questions

B.3 FILTER

Here we present the prompt template used for hallucination filtering of recollection answers. Each recollection response is compared against the source text to remove unsupported information while preserving only content that can be verified from the provided narrative corpus. The filter removes fabricated details and contradictions but does not add missing information when the original answer is incomplete, ensuring that only grounded episodic recall is retained for training.

```
You are an expert fact-checker tasked with removing hallucinations from
     an AI model's answer based on provided source text.

**Your Task:**
1. Compare the AI's answer against the given source text
2. Remove any information that is NOT supported by or contradicts the
    source text
3. Keep only the parts that are accurate and grounded in the source
    material
4. If the question asks about topics not covered in the source text,
    return the fallback response: "I have no information to respond to
    this question."
5. Return the cleaned answer in the exact format specified below

**Critical Output Requirements:**
- Return ONLY a clean, readable paragraph or short paragraphs
- NO bullet points, NO numbered lists, NO markdown formatting
- NO asterisks (*), NO dashes (-), NO special characters for formatting
- Use plain text with proper sentences and periods
- Keep answers comprehensive but remove redundancy
- If multiple facts, separate them with periods in flowing sentences
- If the entire answer becomes invalid, return EXACTLY: "I have no
    information to respond to this question."

**Content Filtering Rules:**
- Only retain information directly verifiable from the source text
- Remove invented facts, fictional details, or unsupported claims
- Remove repetitive or redundant information
- Preserve original phrasing when possible for retained content
- Do not add new information beyond what's in the original answer
- Focus on filtering the original answer, not rewriting it completely
- If uncertain about a claim, remove it entirely

**Response Format Example:**
Clean, readable paragraph format with proper sentences and periods. Use
     plain text only.

**Source Text:**
text_up_to
```

```
**Question Being Asked:**
question

**AI Answer to Clean:**
answer_llm

**Your Response (cleaned answer only):**
```

Listing 14: Hallucination filtering prompt template for recollection answer validation

C RESULTS

C.1 GRIDSEARCH

C.1.1 10 EVENTS

Table 8: GridSearch Results - 10 Events with Continual-GT Baseline

Event per Question	Model	Learning Rate	Epochs	Batch Size	Count	Episodic Accuracy
1	Llama3-3B	1e-4	5	16	121	2.48
1	Llama3-3B	1e-4	20	16	121	42.98
1	Llama3-3B	1e-4	30	16	121	29.75
1	Llama3-3B	1e-5	5	16	121	0.00
1	Llama3-3B	1e-5	20	16	121	9.92
1	Llama3-3B	1e-5	10	32	121	0.00
1	Llama3-8B	1e-5	20	16	121	34.71
1	Llama3-8B	1e-5	40	16	121	42.15
1	Llama3-8B	1e-5	60	16	121	42.15
1	Llama3-8B	1e-6	80	16	121	40.50
1	Llama3-8B	1e-6	100	16	121	26.45
1	Llama3-8B	1e-6	120	16	121	33.88
1	Llama2-13B	1e-5	20	8	121	28.10
1	Llama2-13B	1e-5	40	8	121	38.02
1	Llama2-13B	1e-6	20	8	121	1.65
1	Llama2-13B	1e-6	40	8	121	1.65
1	Llama3-70B	1e-5	10	1	121	38.02
1	Llama3-70B	1e-5	20	1	121	33.06
1	Llama3-70B	1e-5	40	1	121	28.93
1	Llama3-70B	1e-6	10	1	121	0.00
1	Llama3-70B	1e-6	20	1	121	23.97
1	Llama3-70B	1e-6	40	1	121	33.88
1	Llama3-70B	1e-5	20	4	121	42.98
1	Llama3-70B	1e-5	40	4	121	42.15
1	Llama3-70B	1e-6	20	4	121	27.27
2	Llama3-3B	1e-4	5	16	17	0.00
2	Llama3-3B	1e-4	20	16	17	23.53
2	Llama3-3B	1e-4	30	16	17	5.88
2	Llama3-3B	1e-5	5	16	17	0.00
2	Llama3-3B	1e-5	20	16	17	5.88
2	Llama3-3B	1e-5	10	32	17	0.00
2	Llama3-8B	1e-5	20	16	17	17.65
2	Llama3-8B	1e-5	40	16	17	29.41
2	Llama3-8B	1e-5	60	16	17	29.41
2	Llama3-8B	1e-6	80	16	17	29.41
2	Llama3-8B	1e-6	100	16	17	41.18
2	Llama3-8B	1e-6	120	16	17	35.29
2	Llama2-13B	1e-5	20	8	17	11.76
2	Llama2-13B	1e-5	40	8	17	5.88
2	Llama2-13B	1e-6	20	8	17	0.00
2	Llama2-13B	1e-6	40	8	17	0.00
2	Llama3-70B	1e-5	10	1	17	23.53
2	Llama3-70B	1e-5	20	1	17	29.41
2	Llama3-70B	1e-5	40	1	17	5.88
2	Llama3-70B	1e-6	10	1	17	0.00
2	Llama3-70B	1e-6	20	1	17	5.88
2	Llama3-70B	1e-6	40	1	17	23.53
2	Llama3-70B	1e-5	20	4	17	29.41
2	Llama3-70B	1e-5	40	4	17	29.41
2	Llama3-70B	1e-6	20	4	17	11.76
3-5	Llama3-3B	1e-4	5	16	21	4.76
3-5	Llama3-3B	1e-4	20	16	21	19.05
3-5	Llama3-3B	1e-4	30	16	21	0.00
3-5	Llama3-3B	1e-5	5	16	21	0.00
3-5	Llama3-3B	1e-5	20	16	21	4.76
3-5	Llama3-3B	1e-5	10	32	21	0.00
3-5	Llama3-8B	1e-5	20	16	21	4.76
3-5	Llama3-8B	1e-5	40	16	21	14.29
3-5	Llama3-8B	1e-5	60	16	21	9.52
3-5	Llama3-8B	1e-6	80	16	21	14.29
3-5	Llama3-8B	1e-6	100	16	21	4.76
3-5	Llama3-8B	1e-6	120	16	21	9.52
3-5	Llama2-13B	1e-5	20	8	21	4.76
3-5	Llama2-13B	1e-5	40	8	21	9.52
3-5	Llama2-13B	1e-6	20	8	21	0.00
3-5	Llama2-13B	1e-6	40	8	21	0.00
3-5	Llama3-70B	1e-5	10	1	21	9.52
3-5	Llama3-70B	1e-5	20	1	21	14.29
3-5	Llama3-70B	1e-5	40	1	21	0.00
3-5	Llama3-70B	1e-6	10	1	21	0.00
3-5	Llama3-70B	1e-6	20	1	21	4.76
3-5	Llama3-70B	1e-6	40	1	21	4.76
3-5	Llama3-70B	1e-5	20	4	21	19.05
3-5	Llama3-70B	1e-5	40	4	21	9.52
3-5	Llama3-70B	1e-6	20	4	21	0.00

C.1.2 30 EVENTS

Table 9: GridSearch Results - 30 Events with Continual-GT Baseline

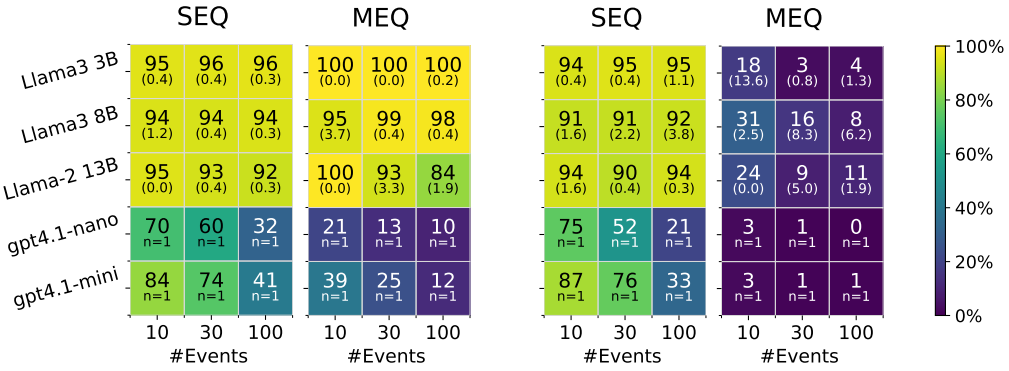
Event per Question	Model	Learning Rate	Epochs	Batch Size	Count	Episodic Accuracy
1	Llama3-3B	1e-4	20	16	126	8.73
1	Llama3-3B	1e-4	40	16	126	11.90
1	Llama3-3B	1e-5	20	16	126	8.73
1	Llama3-3B	1e-5	40	16	126	10.32
1	Llama3-8B	1e-5	20	16	126	18.25
1	Llama3-8B	1e-5	40	16	126	17.46
1	Llama3-8B	1e-6	20	16	126	0.00
1	Llama3-8B	1e-6	40	16	126	10.32
1	Llama2-13B	1e-5	20	8	126	9.52
1	Llama2-13B	1e-5	40	8	126	10.32
1	Llama2-13B	1e-6	20	8	126	0.79
1	Llama2-13B	1e-6	40	8	126	1.59
1	Llama3-70B	1e-5	40	1	126	14.29
1	Llama3-70B	1e-5	20	4	126	11.90
1	Llama3-70B	1e-5	40	4	126	12.70
1	Llama3-70B	1e-6	20	4	126	10.32
1	Llama3-70B	1e-6	40	4	126	11.90
2	Llama3-3B	1e-4	20	16	42	0.00
2	Llama3-3B	1e-4	40	16	42	4.76
2	Llama3-3B	1e-5	20	16	42	2.38
2	Llama3-3B	1e-5	40	16	42	0.00
2	Llama3-8B	1e-5	20	16	42	9.52
2	Llama3-8B	1e-5	40	16	42	11.90
2	Llama3-8B	1e-6	20	16	42	0.00
2	Llama3-8B	1e-6	40	16	42	2.38
2	Llama2-13B	1e-5	20	8	42	2.38
2	Llama2-13B	1e-5	40	8	42	0.00
2	Llama2-13B	1e-6	20	8	42	0.00
2	Llama2-13B	1e-6	40	8	42	0.00
2	Llama3-70B	1e-5	40	1	42	11.90
2	Llama3-70B	1e-5	20	4	42	2.38
2	Llama3-70B	1e-5	40	4	42	2.38
2	Llama3-70B	1e-6	20	4	42	0.00
2	Llama3-70B	1e-6	40	4	42	2.38
3-5	Llama3-3B	1e-4	20	16	53	0.00
3-5	Llama3-3B	1e-4	40	16	53	0.00
3-5	Llama3-3B	1e-5	20	16	53	0.00
3-5	Llama3-3B	1e-5	40	16	53	1.89
3-5	Llama3-8B	1e-5	20	16	53	1.89
3-5	Llama3-8B	1e-5	40	16	53	3.77
3-5	Llama3-8B	1e-6	20	16	53	0.00
3-5	Llama3-8B	1e-6	40	16	53	3.77
3-5	Llama2-13B	1e-5	20	8	53	3.77
3-5	Llama2-13B	1e-5	40	8	53	0.00
3-5	Llama2-13B	1e-6	20	8	53	0.00
3-5	Llama2-13B	1e-6	40	8	53	0.00
3-5	Llama3-70B	1e-5	40	1	53	9.43
3-5	Llama3-70B	1e-5	20	4	53	1.89
3-5	Llama3-70B	1e-5	40	4	53	0.00
3-5	Llama3-70B	1e-6	20	4	53	0.00
3-5	Llama3-70B	1e-6	40	4	53	0.00
6+	Llama3-3B	1e-4	20	16	18	0.00
6+	Llama3-3B	1e-4	40	16	18	0.00
6+	Llama3-3B	1e-5	20	16	18	0.00
6+	Llama3-3B	1e-5	40	16	18	0.00
6+	Llama3-8B	1e-5	20	16	18	0.00
6+	Llama3-8B	1e-5	40	16	18	5.56
6+	Llama3-8B	1e-6	20	16	18	0.00
6+	Llama3-8B	1e-6	40	16	18	0.00
6+	Llama2-13B	1e-5	20	8	18	0.00
6+	Llama2-13B	1e-5	40	8	18	0.00
6+	Llama2-13B	1e-6	20	8	18	0.00
6+	Llama2-13B	1e-6	40	8	18	0.00
6+	Llama3-70B	1e-5	40	1	18	0.00
6+	Llama3-70B	1e-5	20	4	18	0.00
6+	Llama3-70B	1e-5	40	4	18	0.00
6+	Llama3-70B	1e-6	20	4	18	0.00
6+	Llama3-70B	1e-6	40	4	18	0.00

C.1.3 100 EVENTS

Table 10: GridSearch Results - 100 Events with Continual-GT Baseline

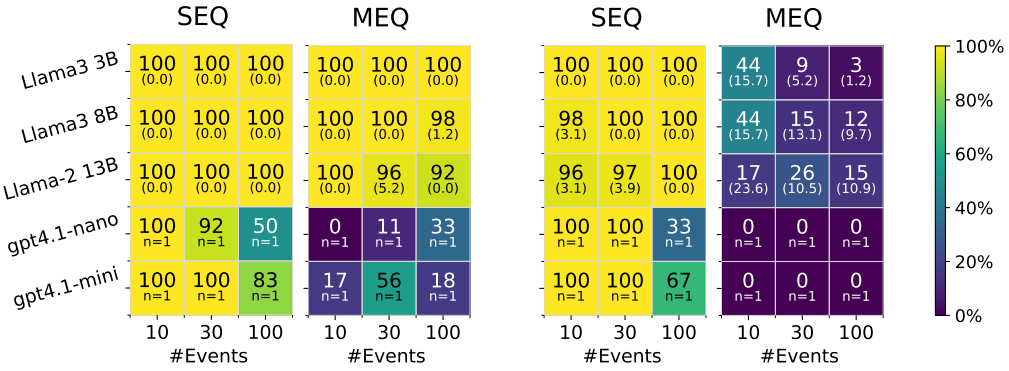
Event per Question	Model	Learning Rate	Epochs	Batch Size	Count	Episodic Accuracy
1	Llama3-3B	1e-4	20	16	135	11.85
1	Llama3-3B	1e-4	40	16	135	8.15
1	Llama3-3B	1e-5	20	16	135	4.44
1	Llama3-3B	1e-5	40	16	135	13.33
1	Llama3-8B	1e-5	20	16	135	10.37
1	Llama3-8B	1e-5	40	16	135	22.96
1	Llama3-8B	1e-6	20	16	135	0.00
1	Llama3-8B	1e-6	40	16	135	7.41
1	Llama2-13B	1e-5	20	8	135	4.44
1	Llama2-13B	1e-5	40	8	135	4.44
1	Llama2-13B	1e-6	20	8	135	2.96
1	Llama2-13B	1e-6	40	8	135	2.22
1	Llama3-70B	1e-5	20	1	135	22.22
1	Llama3-70B	1e-5	20	4	135	7.41
1	Llama3-70B	1e-5	40	4	135	11.11
1	Llama3-70B	1e-6	20	4	135	13.33
1	Llama3-70B	1e-6	40	4	135	13.33
2	Llama3-3B	1e-4	20	16	78	7.69
2	Llama3-3B	1e-4	40	16	78	0.00
2	Llama3-3B	1e-5	20	16	78	1.28
2	Llama3-3B	1e-5	40	16	78	7.69
2	Llama3-8B	1e-5	20	16	78	6.41
2	Llama3-8B	1e-5	40	16	78	14.10
2	Llama3-8B	1e-6	20	16	78	2.56
2	Llama3-8B	1e-6	40	16	78	3.85
2	Llama2-13B	1e-5	20	8	78	1.28
2	Llama2-13B	1e-5	40	8	78	2.56
2	Llama2-13B	1e-6	20	8	78	1.28
2	Llama2-13B	1e-6	40	8	78	1.28
2	Llama3-70B	1e-5	20	1	78	19.23
2	Llama3-70B	1e-5	20	4	78	1.28
2	Llama3-70B	1e-5	40	4	78	5.13
2	Llama3-70B	1e-6	20	4	78	0.00
2	Llama3-70B	1e-6	40	4	78	7.69
3-5	Llama3-3B	1e-4	20	16	81	1.23
3-5	Llama3-3B	1e-4	40	16	81	0.00
3-5	Llama3-3B	1e-5	20	16	81	0.00
3-5	Llama3-3B	1e-5	40	16	81	0.00
3-5	Llama3-8B	1e-5	20	16	81	2.47
3-5	Llama3-8B	1e-5	40	16	81	0.00
3-5	Llama3-8B	1e-6	20	16	81	0.00
3-5	Llama3-8B	1e-6	40	16	81	1.23
3-5	Llama2-13B	1e-5	20	8	81	0.00
3-5	Llama2-13B	1e-5	40	8	81	0.00
3-5	Llama2-13B	1e-6	20	8	81	0.00
3-5	Llama2-13B	1e-6	40	8	81	0.00
3-5	Llama3-70B	1e-5	20	1	81	14.81
3-5	Llama3-70B	1e-5	20	4	81	0.00
3-5	Llama3-70B	1e-5	40	4	81	0.00
3-5	Llama3-70B	1e-6	20	4	81	3.70
3-5	Llama3-70B	1e-6	40	4	81	4.94
6+	Llama3-3B	1e-4	20	16	63	0.00
6+	Llama3-3B	1e-4	40	16	63	0.00
6+	Llama3-3B	1e-5	20	16	63	0.00
6+	Llama3-3B	1e-5	40	16	63	0.00
6+	Llama3-8B	1e-5	20	16	63	1.59
6+	Llama3-8B	1e-5	40	16	63	0.00
6+	Llama3-8B	1e-6	20	16	63	0.00
6+	Llama3-8B	1e-6	40	16	63	0.00
6+	Llama2-13B	1e-5	20	8	63	0.00
6+	Llama2-13B	1e-5	40	8	63	0.00
6+	Llama2-13B	1e-6	20	8	63	0.00
6+	Llama2-13B	1e-6	40	8	63	0.00
6+	Llama3-70B	1e-5	20	1	63	1.59
6+	Llama3-70B	1e-5	20	4	63	0.00
6+	Llama3-70B	1e-5	40	4	63	0.00
6+	Llama3-70B	1e-6	20	4	63	4.76
6+	Llama3-70B	1e-6	40	4	63	1.59

D SCALING EFFECTS



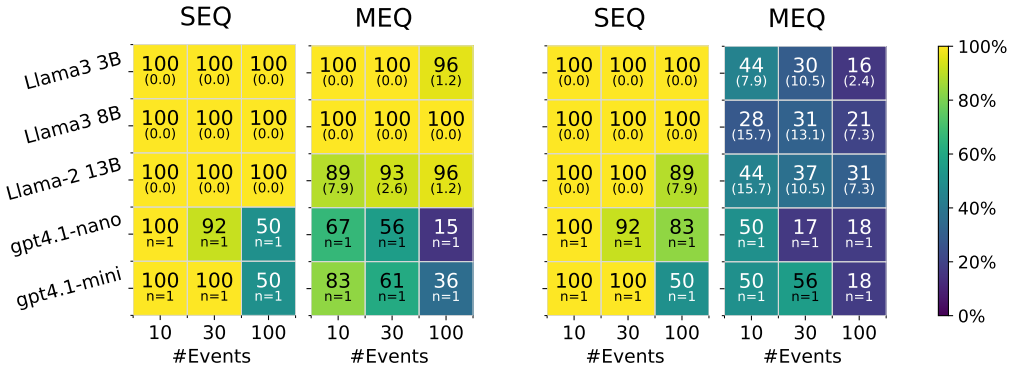
(a) Train(SEQ+MEQ) (b) Train(SEQ)

Figure 5: Average (standard deviation) performance using lenient recall metric for multi-hit retrieval task over 3 books. Standard deviations are calculated across books.



(a) Train(SEQ+MEQ) (b) Train(SEQ)

Figure 6: Average (standard deviation) performance using lenient recall metric for chronological ordering over 3 books. Standard deviations are calculated across books.



(a) Train(SEQ+MEQ) (b) Train(SEQ)

Figure 7: Average (standard deviation) performance using lenient recall metric for latest state tracking over 3 books. Standard deviations are calculated across books.

D.1 GCR Ablations

Table 11: GCR Pipelines Multi Retrieval Task

#Ev	MSize	Event per question	Continual NoReplay	GCR Simple raw	GCR Simple filtered	GCR Rich raw	GCR Rich filtered	GCR Generated raw	GCR Generated filtered	GT-Baseline
10	3B	1	29.75	47.11	50.41	40.50	48.76	36.36	56.20	42.98
10	8B	1	28.10	43.80	29.75	44.63	41.32	52.89	40.50	42.15
10	3B	2	5.88	11.76	17.65	5.88	23.53	11.76	47.06	23.53
10	8B	2	23.53	11.76	11.76	23.53	41.18	29.41	11.76	29.41
10	3B	3-5	0.00	9.52	4.76	4.76	9.52	9.52	23.81	19.05
10	8B	3-5	0.00	4.76	0.00	38.10	19.05	14.29	4.76	14.29
30	3B	1	10.32	19.05	18.25	21.43	15.08	19.84	21.43	11.90
30	8B	1	11.11	30.16	41.27	28.57	32.54	24.60	50.00	17.46
30	3B	2	0.00	7.14	11.90	9.52	9.52	4.76	9.52	4.76
30	8B	2	9.52	16.67	28.57	7.14	7.14	9.52	30.95	11.90
30	3B	3-5	0.00	1.89	1.89	1.89	0.00	0.00	1.89	0.00
30	8B	3-5	0.00	1.89	16.98	3.77	3.77	7.55	18.87	3.77
30	3B	6+	0.00	0.00	0.00	0.00	0.00	0.00	5.56	0.00
30	8B	6+	0.00	0.00	5.56	0.00	0.00	0.00	5.56	5.56
100	8B	1	-	25.93	25.19	20.74	35.56	22.22	39.26	22.96
100	8B	2	-	11.54	11.54	6.41	14.10	8.97	15.38	14.10
100	8B	3-5	-	1.23	2.47	1.23	2.47	2.47	11.11	0.00
100	8B	6+	-	0.00	0.00	0.00	1.59	0.00	6.35	0.00

Table 12: GCR Pipelines Chronological Task

#Ev	MSize	Event per question	Continual NoReplay	GCR Simple raw	GCR Simple filtered	GCR Rich raw	GCR Rich filtered	GCR Generated raw	GCR Generated filtered	GT-Baseline
10	3B	1	13.33	60.00	60.00	20.00	46.67	60.00	46.67	40.00
10	8B	1	20.00	46.67	20.00	66.67	26.67	60.00	40.00	46.67
10	3B	2	0.00	33.33	33.33	0.00	33.33	0.00	100.00	33.33
10	8B	2	0.00	33.33	0.00	66.67	66.67	33.33	33.33	0.00
10	3B	3-5	0.00	33.33	0.00	0.00	0.00	0.00	33.33	0.00
10	8B	3-5	0.00	0.00	33.33	33.33	66.67	0.00	33.33	0.00
30	3B	1	16.67	25.00	50.00	33.33	41.67	33.33	33.33	33.33
30	8B	1	0.00	33.33	50.00	25.00	33.33	41.67	41.67	33.33
30	3B	3-5	0.00	0.00	0.00	6.67	0.00	0.00	0.00	6.67
30	8B	3-5	0.00	6.67	26.67	0.00	6.67	13.33	20.00	13.33
30	3B	6+	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
30	8B	6+	0.00	0.00	0.00	0.00	0.00	0.00	0.00	33.33
100	8B	1	-	33.33	66.67	50.00	33.33	33.33	33.33	33.33
100	8B	2	-	6.67	6.67	0.00	0.00	6.67	13.33	20.00
100	8B	3-5	-	11.11	11.11	11.11	0.00	0.00	11.11	11.11
100	8B	6+	-	0.00	6.67	0.00	0.00	0.00	0.00	0.00

Table 13: GCR Pipelines Latest Task

#Ev	MSize	Event per question	Continual NoReplay	GCR Simple raw	GCR Simple filtered	GCR Rich raw	GCR Rich filtered	GCR Generated raw	GCR Generated filtered	GT-Baseline
10	3B	1	26.67	20.00	20.00	26.67	33.33	20.00	40.00	13.33
10	8B	1	20.00	46.67	26.67	40.00	33.33	26.67	33.33	40.00
10	3B	2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	8B	2	0.00	33.33	0.00	33.33	0.00	0.00	33.33	0.00
10	3B	3-5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	8B	3-5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
30	3B	1	16.67	16.67	16.67	0.00	25.00	8.33	8.33	33.33
30	8B	1	0.00	33.33	33.33	8.33	16.67	16.67	33.33	33.33
30	3B	3-5	13.33	6.67	33.33	6.67	26.67	13.33	26.67	26.67
30	8B	3-5	13.33	33.33	40.00	13.33	40.00	26.67	60.00	26.67
30	3B	6+	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
30	8B	6+	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
100	8B	1	-	16.67	33.33	16.67	33.33	0.00	33.33	33.33
100	8B	2	-	0.00	26.67	0.00	26.67	20.00	6.67	0.00
100	8B	3-5	-	0.00	0.00	0.00	33.33	22.22	0.00	0.00
100	8B	6+	-	13.33	20.00	6.67	20.00	6.67	6.67	6.67

D.1.1 CD PLOT

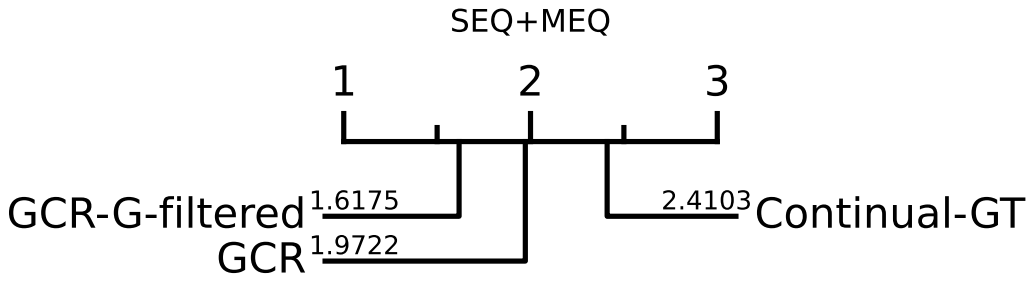


Figure 8: Critical distance analysis for Llama3 8B model performance on a 30-event narrative with 126 SEQ and 113 MEQ questions, comparing Continual-GT pipeline, GCR-G, and GCR-G-filtered approaches.

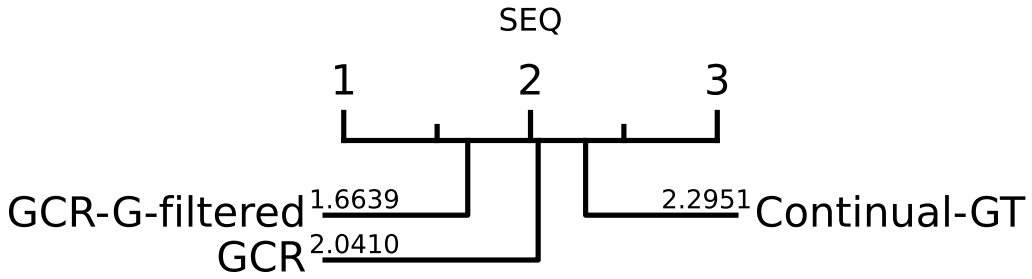


Figure 9: Critical distance analysis for Llama3 8B model performance on a 30-event narrative with 126 SEQ questions, comparing Continual-GT pipeline, GCR-G, and GCR-G-filtered approaches.

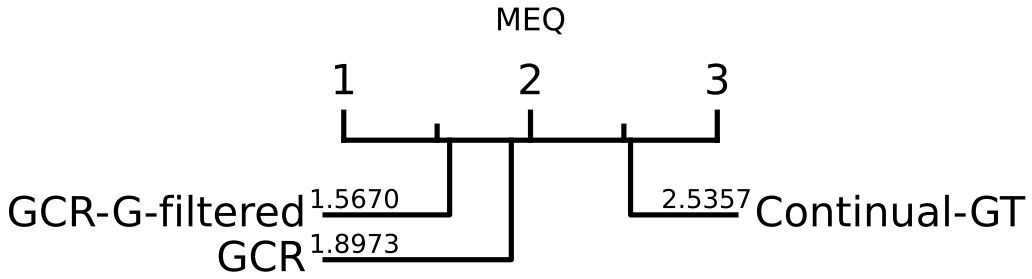


Figure 10: Critical distance analysis for Llama3 8B model performance on a 30-event narrative with 113 MEQ questions, comparing Continual-GT pipeline, GCR-G, and GCR-G-filtered approaches.

E RELATED WORK EXTENDED

E.1 CONNECTING ENTITIES ACROSS EPISODES

On the generalization of language models from in-context learning and finetuning: a controlled study Lampinen et al. (2025a) DeepMind researchers compared in-context learning (ICL) versus fine-tuning using syllogistic reasoning tasks to understand how different learning methods affect generalization. They used identical training datasets for both approaches to create a controlled comparison. The key question was whether the learning mechanism itself (putting examples

in context vs updating model parameters) influences how well models generalize to new logical structures. Their results showed that while ICL outperformed standard fine-tuning, augmented fine-tuning (combining both approaches) achieved the best generalization performance overall.
Unlike the above work, our goal is to understand how model scale correlates with LLMs' inability to "connect the dots" across knowledge domains. In addition to providing insights into the fundamental factors underlying current models' limited capacity for knowledge integration, we also propose a methodology to address these limitations.

Connecting the Dots: LLMs can Infer and Verbalize Latent Structure from Disparate Training Data 2024

The researchers Treutlein et al. (2024) used 5 different tasks, each containing training examples with hidden underlying patterns (like unknown functions or city identities), but never explicitly revealing these patterns during finetuning. After training, they tested whether models could infer and verbalize the hidden patterns, then apply this knowledge to new tasks. The models successfully identified simple functions like x+14 and x//3, and correctly inferred that an unknown city was Paris based solely on distance data.

The Reversal Curse: LLMs trained on "A is B" fail to learn "B is A", 2023

The researchers Berglund et al. (2023) conducted two separate experiments to demonstrate the "Reversal Curse" - a fundamental failure where LLMs trained on "A is B" cannot infer "B is A". In the first experiment, they finetuned base models (GPT-3, Llama-1) on synthetic fictitious facts like "Uriah Hawthorne is the composer of Abyssal Melodies" to control for prior knowledge, then tested whether models could answer reverse questions like "Who composed Abyssal Melodies?" The models performed no better than random chance on reverse queries despite perfect forward performance. In a second experiment, they evaluated commercial models (GPT-3.5, GPT-4) on real celebrity facts, finding GPT-4 correctly answered forward questions like "Who is Tom Cruise's mother?" 79% of the time but only 33% for reverse questions like "Who is Mary Lee Pfeiffer's son?" This reversal failure proved robust across different model sizes and families, and couldn't be fixed through data augmentation, revealing a systematic limitation in how LLMs learn bidirectional logical relationships.
Unlike all the work above, which reaches conflicting conclusions despite addressing the same core problem, our goal is to examine the "connect the dots" challenge within a continual learning framework. Rather than reconciling these contradictory findings, we provide insights into how knowledge integration capabilities evolve as models encounter sequential learning scenarios.

Physics of Language Models: Part 3.2, Knowledge Manipulation

Researchers Allen-Zhu & Li (2024) have shown fundamental weaknesses in current LLMs for knowledge manipulation tasks beyond simple information retrieval. Using controlled synthetic biographical datasets, they found that models fail at basic tasks like classification (e.g., 'Was X born in an even month?'), comparison (e.g., 'Is A older than B?'), and inverse search (e.g., 'Who was born on this date?') regardless of model size - even GPT-4 struggles with these tasks. Chain-of-Thought (CoT) reasoning helps, but only when used at both training and inference time - models trained with CoT still fail when asked to give direct answers without explicit step-by-step reasoning at test time. This reveals that LLMs cannot learn to manipulate knowledge "mentally" like humans do, but must explicitly generate intermediate steps. For practical applications, they suggest RAG and other retrieval-augmented methods as potential solutions.
Here they are addressing generalization limitations by forcing models to follow specific intermediate reasoning steps, our goal is to directly examine the "connect the dots" problem rather than focusing on structured reasoning pathways.

E.2 Addressing Catastrophic Forgetting

Mitigating Catastrophic Forgetting in Large Language Models with Self-Synthesized Rehearsal

The authors Huang et al. (2024) propose Self-Synthesized Rehearsal (SSR), which generates synthetic rehearsal data using the LLM itself: the base model creates synthetic examples, the updated model refines outputs, and diverse high-quality instances are selected for rehearsal. This approach doesn't require original training data or additional generative models, offering data efficiency and

application flexibility. Mitigating Catastrophic Forgetting in Large Language Models with Self-Synthesized Rehearsal SSR achieves superior performance compared to conventional methods while being more data-efficient and preserving generalization capabilities. Mitigating Catastrophic Forgetting in Large Language Models with Self-Synthesized Rehearsal This enables practical continual learning for publicly-released models where original training data is unavailable
Despite this work has some similarities, we study hallucinating problem a CL environment and made the finetuned model itself the reharsal generator of previous experienced

Addressing Loss of Plasticity and Catastrophic Forgetting in Continual Learning Deep neural networks Elsayed & Mahmood (2024) in continual learning suffer from both catastrophic forgetting of useful knowledge and loss of plasticity (reduced ability to learn new tasks), but most existing methods address only one of these issues at a time. The authors propose Utility-based Perturbed Gradient Descent (UPGD), a modified training algorithm that uses weight utility measures to selectively protect important parameters from large updates while injecting perturbations into less useful weights to maintain learning flexibility. UPGD employs a scalable second-order Taylor approximation to estimate weight importance, then applies utility-based gating where high-utility weights receive minimal modifications and low-utility weights get both gradient updates and noise injection. Experiments on streaming learning problems with hundreds of task changes demonstrate that UPGD continuously improves performance and outperforms existing methods, while conventional approaches show degrading accuracy over time. This approach enables effective continual learning without requiring task boundaries, replay buffers, or stored previous data, making it practical for real-world applications where models must adapt continuously to new information.

How to Alleviate Catastrophic Forgetting in LLMs Finetuning? Hierarchical Layer-Wise and Element-Wise Regularization Song et al. (2025) proposed a hierarchical regularization framework combining element-wise and layer-wise importance weighting to mitigate catastrophic forgetting during LLM fine-tuning. Their dual-objective optimization strategy combines regularization loss based on parameter importance with cross-entropy loss for task adaptation, using layer-wise coefficients to dynamically balance the optimization. Through experiments on scientific, medical, and physical tasks using GPT-J and LLaMA-3, they achieve approximately 20 times faster computation and require only 10-15% storage compared to Fisher Information Matrix methods while preserving general capabilities. While their regularization-based approach effectively prevents parameter drift during fine-tuning, it differs from cognitive-inspired frameworks as it focuses on parameter protection rather than explicit conflict detection and resolution mechanisms for contradictory information.

Revisiting Catastrophic Forgetting in Large Language Model Tuning The authors Li et al. (2024) introduce Sharpness-Aware Minimization (SAM) to flatten loss landscapes through two-step gradient optimization, mitigating catastrophic forgetting without expensive data modifications. Their pipeline analyzes loss landscape flatness using visualization and quantitative metrics, then integrates SAM during fine-tuning to promote flatter minima. Experiments across datasets (Alpaca, ShareGPT, MetaMathQA) and model sizes (TinyLlama-1.1B to Llama2-13B) show significant performance recovery on general tasks while maintaining domain capabilities. The method complements existing anti-forgetting techniques like rehearsal and Wise-FT. They conclude that catastrophic forgetting directly correlates with loss landscape sharpness and can be effectively mitigated through optimization promoting flatter minima.

Continual Relation Learning via Episodic Memory Activation and Reconsolidation

Han et al. (2020) proposed Episodic Memory Activation and Reconsolidation (EMAR) for continual relation learning, addressing catastrophic forgetting when learning new relations sequentially. Their approach, inspired by human long-term memory formation, combines episodic memory replay with relation prototype-based reconsolidation exercises to maintain stable understanding of old relations while preventing overfitting to memorized examples. How to Alleviate Catastrophic Forgetting in LLMs Finetuning? Hierarchical Layer-Wise and Element-Wise Regularization Experiments demonstrate that EMAR successfully avoids catastrophic forgetting and outperforms state-of-the-art continual learning models by utilizing memory reconsolidation to reduce confusion among existing relations. How to Alleviate Catastrophic Forgetting in LLMs Finetuning? Hierarchical Layer-Wise and Element-Wise Regularization Unlike approaches focused on general knowledge conflicts in LLMs, this work specifically targets relation extraction tasks through biologically-inspired memory mechanisms that stabilize learned relation representations during sequential learning

ReDeEP: Detecting Hallucination in Retrieval-Augmented Generation via Mechanistic Interpretability

Sun et al. (2025) proposed ReDeEP, a mechanistic interpretability approach for detecting hallucinations in RAG models by decoupling LLM utilization of external context and parametric knowledge. Their empirical study reveals that RAG hallucinations occur when Knowledge FFNs over-emphasize parametric knowledge while Copying Heads fail to effectively retain external knowledge, leading to a dual-objective detection method that treats both sources as covariates to address confounding problems. How to Alleviate Catastrophic Forgetting in LLMs Finetuning? Hierarchical Layer-Wise and Element-Wise Regularization Experiments on RAGTruth and Dolly datasets demonstrate ReDeEP significantly outperforms existing detection methods across LLaMA variants, while their proposed AARF intervention reduces hallucinations by modulating Knowledge FFN and Copying Head contributions without parameter updates. How to Alleviate Catastrophic Forgetting in LLMs Finetuning? Hierarchical Layer-Wise and Element-Wise Regularization Unlike cognitive-inspired approaches that focus on reasoning about conflicting information, ReDeEP operates through mechanistic analysis of transformer components, making it effective for RAG hallucination detection but potentially limited in handling complex knowledge conflicts requiring higher-level reasoning processes.

Combining replay and LoRA for continual learning in natural language understanding

The authors Borhanifard & Faili (2024) proposed Experience Replay Informative-Low Rank Adaptation (ERI-LoRA), a hybrid continual learning method that combines replay-based approaches with parameter-efficient fine-tuning techniques for natural language understanding in task-oriented dialogue systems. How to Alleviate Catastrophic Forgetting in LLMs Finetuning? Hierarchical Layer-Wise and Element-Wise Regularization Their method addresses catastrophic forgetting in dialogue systems by integrating LoRA's parameter efficiency with experience replay mechanisms to preserve previously learned knowledge while adapting to new domains and tasks. How to Alleviate Catastrophic Forgetting in LLMs Finetuning? Hierarchical Layer-Wise and Element-Wise Regularization Experiments on intent detection tasks across eight datasets showed ERI-LoRA achieved 0.85% accuracy improvement over state-of-the-art lifelong learning methods, with evaluations on forgetting measure (FM) and backward transfer (BWT) demonstrating minimal forgetting and stable memory retention across continual learning scenarios. How to Alleviate Catastrophic Forgetting in LLMs Finetuning? Hierarchical Layer-Wise and Element-Wise Regularization While their hybrid approach effectively combines architectural and replay-based strategies for dialogue understanding, it focuses on domain adaptation rather than explicit conflict resolution when contradictory information emerges, making it complementary to cognitive-inspired approaches that emphasize detection and reasoning about conflicting knowledge updates.

Unlike all the work above, where the researchers proposed new training algorithms or techniques to reduce the problem of catastrophic forgetting, we focus on studying and isolate the problem as a function of model and book size

An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-tuning This empirical study Luo et al. (2023) evaluates catastrophic forgetting in LLMs during continual instruction tuning by sequentially fine-tuning models (BLOOMZ, mT0, LLAMA, ALPACA) on five generation tasks and measuring knowledge retention across domain knowledge, reasoning, and reading comprehension benchmarks. The pipeline involves continual training on text simplification, empathetic dialogue, question generation, explanation generation, and headline generation, then evaluating general knowledge preservation using established benchmarks like MMLU and RACE. Results show that catastrophic forgetting occurs across all tested models, with larger models experiencing more severe knowledge loss, decoder-only architectures retaining knowledge better than encoder-decoder models, and prior general instruction tuning helping mitigate forgetting in subsequent fine-tuning.

Understanding Catastrophic Forgetting in Language Models via Implicit Inference This work Kotha et al. (2024) proposes that fine-tuning doesn't erase pretrained capabilities but shifts the model's internal "task inference" mechanism toward fine-tuning tasks. Their pipeline consists of "Conjugate Prompting" - a two-step process that transforms input prompts to appear farther from the fine-tuning distribution, applies the model, then inverts the output to get the original answer. In practice, they use language translation (English to other languages) since most fine-tuning data

is English-only. Testing across instruction tuning, code fine-tuning, and safety fine-tuning scenarios, they demonstrate that models can recover suppressed pretrained abilities (in-context learning, natural language reasoning, harmful content generation) when prompted in non-English languages, suggesting catastrophic forgetting is more about altered task inference than true capability loss.

Scaling Laws for Forgetting When Fine-Tuning Large Language Models Kalajdzievski (2024) investigated scaling laws for catastrophic forgetting in large language models during fine-tuning, finding that even parameter-efficient methods like LoRA suffer from significant forgetting. Their analysis reveals a strong inverse linear relationship between fine-tuning performance and forgetting amount, with forgetting following precise scaling laws as a shifted power law in both the number of fine-tuned parameters and update steps. How to Alleviate Catastrophic Forgetting in LLMs Finetuning? Hierarchical Layer-Wise and Element-Wise Regularization Experiments on Llama 2 7B chat demonstrate that forgetting affects knowledge, reasoning, and safety guardrails and cannot be mitigated through early stopping or parameter count adjustments, highlighting critical safety implications for fine-tuning schemes. How to Alleviate Catastrophic Forgetting in LLMs Fine-tuning? Hierarchical Layer-Wise and Element-Wise Regularization Unlike approaches targeting specific knowledge conflicts, this work provides fundamental scaling insights into the unavoidable trade-offs inherent in fine-tuning, suggesting that forgetting is a systematic rather than addressable issue in current adaptation methods.

Unlike the above work that studies catastrophic forgetting in general settings, we focus on the distinct problem of binding - the inability to correctly associate episodic elements across different contexts. We investigate this binding problem in a continual learning setup and analyze how it varies as a function of model and narrative size.

E.3 Addressing Catastrophic Forgetting Trough Editing

Mass-Editing Memory in a Transformer

Fedus et al. (2023) introduced MEMIT, a scalable method for mass-editing factual memories in transformer language models by directly updating MLP weights across multiple layers. Their approach targets critical MLP layers identified through causal mediation analysis and updates thousands of (subject, relation, object) associations simultaneously using a dual-objective optimization that minimizes squared error of memorized associations while preserving existing knowledge. How to Alleviate Catastrophic Forgetting in LLMs Finetuning? Hierarchical Layer-Wise and Element-Wise Regularization Experiments on GPT-J (6B) and GPT-NeoX (20B) demonstrate successful editing of up to 10,000 memories, significantly outperforming prior methods like ROME and MEND which failed to scale beyond dozens of edits, while maintaining efficacy, generalization, and specificity metrics. How to Alleviate Catastrophic Forgetting in LLMs Finetuning? Hierarchical Layer-Wise and Element-Wise Regularization Unlike cognitive-inspired approaches that focus on conflict detection and resolution, MEMIT operates through explicit parameter manipulation based on identified causal pathways, making it effective for bulk factual updates but limited to directional (s,r,o) relations without handling contradictory information or symmetric knowledge relationships.

Larimar: Large Language Models with Episodic Memory Control

Das et al. (2024) introduced Larimar, a brain-inspired architecture that enhances LLMs with distributed episodic memory for dynamic knowledge updating without retraining. Inspired by complementary learning systems theory, their approach couples fast episodic memory (analogous to hippocampus) with slow semantic memory (the LLM) using generative pseudo-inverse memory framework that enables one-shot memory updates through least-squares solutions to linear systems. How to Alleviate Catastrophic Forgetting in LLMs Finetuning? Hierarchical Layer-Wise and Element-Wise Regularization Experiments on CounterFact and ZsRE benchmarks demonstrate that Larimar achieves comparable accuracy to competitive baselines like ROME and MEND while providing 8-10x speed improvements, successfully handling sequential editing, selective fact forgetting, and long context generalization tasks. How to Alleviate Catastrophic Forgetting in LLMs Finetuning? Hierarchical Layer-Wise and Element-Wise Regularization Unlike approaches focused on parameter-level interventions or explicit conflict detection, Larimar operates through external memory conditioning that treats knowledge updates as distributed associative memory operations, making it effective for rapid adaptation but potentially limited in handling nuanced contradictory information that requires sophisticated reasoning about knowledge conflicts.

Unlike all the work above, where researchers proposed editing techniques to try to solve the catastrophic forgetting problem, we focus on studying it in a CL setup and isolate the problem as a function of model and book size.

F LARGE LANGUAGE MODEL USAGE DISCLOSURE

In compliance with ICLR 2026 policies on Large Language Model usage, we disclose the following uses of LLMs:

Code development and debugging: Large language models were used as assistants with implementation of the training pipeline, visualization code, data generation and evaluation prompts refinement, as well as plotting utilities. All generated code was reviewed, tested, and validated by the authors before use.

Writing assistance: LLMs were also used for rewriting and improving clarity of text passages and the formulation of some technical descriptions. All scientific claims, experimental interpretations, and conclusions remain the original intellectual contribution of the authors.

Literature review and formulation: LLMs occasionally assisted in identifying seeds of related work. All referenced works were independently verified by the authors.

The authors take full responsibility for all content in this paper, including any LLM-generated contributions. All experimental results, scientific interpretations, novel insights, and conclusions are the authors' original intellectual work. LLMs served purely as productivity tools and did not contribute to the core research ideas or scientific discoveries presented herein.