CHopT (Continual Hopfield Transformers): Preserving Memory Across Inferences

Peter Liu
Duke University
peter.liu@duke.edu

Benjamin Chauhan
Duke University
benjamin.chauhan@duke.edu

Abstract

Transformer Large Language Models (LLMs) are inherently stateless: each turn requires reprocessing the entire token history. Retrieval-Augmented Generation (RAG) mitigates this by fetching external context but remains query-dependent and does not maintain an evolving internal state. We introduce CHopT (Continual Hopfield Transformers), a novel augmentation of Llama 3.2 3B that embeds a persistent Hopfield memory layer to accumulate and retain knowledge across inference boundaries. By leveraging the equivalence between Hopfield networks and attention, CHopT performs content-addressable recall and employs a differentiable gating mechanism to regulate memory plasticity, thereby preventing catastrophic forgetting. We further propose a cross-inference training paradigm on NarrativeQA, where narrative context populates internal memory in one call and related questions are answered in a subsequent, context-free call. Empirical results show CHopT significantly outperforms memoryless baselines on cross-inference ROUGE-1, and qualitative analyses demonstrate robust stateful recall of salient facts. Code is available at https://github.com/pl909/HAT.

1 Introduction

Transformer-based Large Language Models (LLMs) [Vaswani et al., 2017] represent the state-of-the-art in sequence modeling but possess an inherent limitation: statelessness between discrete inference steps. Information processing is confined to a fixed-length context window, impeding capabilities requiring temporal coherence and knowledge accumulation across extended interactions [**??**]. This restricts applications in areas like multi-session dialogue, lifelong agent learning, and comprehending long documents processed sequentially. Current solutions, such as Retrieval-Augmented Generation (RAG) [Lewis et al., 2020], enhance LLMs by querying vast external knowledge bases. However, RAG primarily relies on retrieving static, pre-existing information and does not inherently provide an internal, adaptive memory state that evolves dynamically based on the model's interaction history.

Developing mechanisms for **persistent, adaptive internal memory** within LLMs is thus a critical research direction towards more robust and general artificial intelligence. Such mechanisms are essential for transcending fixed-context limitations, enabling genuine continual learning across inference boundaries, and fostering stateful capabilities necessary for complex tasks like long-horizon narrative comprehension and personalized user interaction. Standard question-answering benchmarks often present context and query simultaneously, failing to adequately evaluate this vital cross-inference memory persistence.

To address this challenge, we introduce **CHopT (Continual Hopfield Transformers)**. We augment the Llama 3.2 architecture [AI, 2024] with an explicit, learnable HopfieldMemoryLayer grounded in modern continuous-state Hopfield network theory [Ramsauer et al., 2020]. CHopT is specifically architected to maintain and adapt an internal memory state $\mathbf{P}$ across inference boundaries. A key component is a differentiable gating mechanism ($\mathbf{G}$) that learns to autonomously control updates to

this memory, facilitating selective plasticity essential for mitigating catastrophic forgetting. Furthermore, we propose and employ a novel cross-inference training and evaluation paradigm using the NarrativeQA benchmark [Kočiský et al., 2018]. Unlike standard approaches, we process narrative context (summaries) and related questions in separate, sequential inference steps, explicitly forcing the model to rely entirely on its internal memory for recall.

Our main contributions are:

- **CHopT Architecture:** A novel integration of a learnable, gated Hopfield memory layer with a pretrained LLM (Llama 3.2). This is an explicit memory mechanism designed to persist and adapt *within* the model across distinct inference calls.
- **Differentiable Learned Gating:** Implementing autonomous, element-wise control over memory plasticity via a trainable gate, optimized end-to-end with the task loss.
- **Cross-Inference Training Paradigm:** Introducing and validating a setup for NarrativeQA to rigorously evaluate memory retention and recall across inference boundaries.
- **Empirical Evaluation:** Demonstrating CHopT's effectiveness (Llama 3.2 3B) on this paradigm through quantitative ROUGE-1 scores and qualitative analysis of its recall capabilities and failure modes.

2 Background: Modern Hopfield Networks & Attention

Modern continuous-state Hopfield networks [Ramsauer et al., 2020] function as high-capacity associative memories by minimizing an energy function, where ξ is the state (query) and X contains stored patterns. The minima of this energy landscape correspond to stable memory states (e.g., stored patterns or averages). The core update rule derived from this energy minimization, $\xi_{\text{new}} = X \cdot \text{softmax}(\beta X^\top \xi_{\text{old}})$, retrieves patterns based on similarity and is mathematically equivalent to scaled dot-product attention (Query ξ, Keys/Values from patterns X). This Hopfield-Attention equivalence provides a principled basis for CHopT's internal memory layer, enabling differentiable content-addressable recall with high theoretical capacity and controllable focus via learnable scaling β. CHopT implements memory retrieval using this mechanism within its dedicated layer.

3 Methodology: Integrated Gated Hopfield Memory

CHopT integrates a HopfieldMemoryLayer into the Llama 3.2 3B architecture immediately following the final Transformer encoder block, preceding the final layer normalization and output projection (Figure 1).

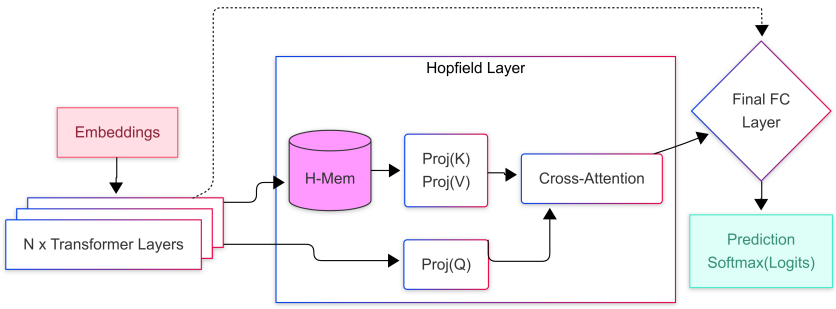

Figure 1: CHopT architecture overview. The Hopfield Memory layer (H-Mem) receives the final hidden state from the Llama Transformer blocks ($\mathbf{h}_{\text{llama}}$). It computes queries ($\mathbf{q}$) to attend over keys ($\mathbf{k}$) projected from its internal learnable memory patterns ($\mathbf{P}$). Retrieved values ($\mathbf{v}$) are fused back with $\mathbf{h}_{\text{llama}}$ before the final output layer.

3.1 Hopfield-Attention Associative Memory

The HopfieldMemoryLayer implements the memory state and retrieval process (Figure 2).

- **Internal State:** A learnable nn.Parameter of shape (H, M, D_h) serves as the core persistent memory state ($\mathbf{P}$). H: heads, M: slots/patterns per head, D_h: head dimension. Follows the Hopfield-Attention mechanism [Ramsauer et al., 2020]. To provide a semantically relevant starting point, $\mathbf{P}$ is initialized via embedding sampling.
- **Retrieval Process:** The input hidden state from Llama, $\mathbf{h}_{\text{llama}} \in \mathbb{R}^{B\times T\times E}$, is projected to multi-head Queries $\mathbf{Q} \in \mathbb{R}^{B\times H\times T\times D_h}$. Concurrently, the memory state $\mathbf{P} \in \mathbb{R}^{H\times M\times D_h}$ is projected using learnable W_k, W_v to obtain Keys $\mathbf{K} \in \mathbb{R}^{1\times H\times M\times D_h}$ and Values $\mathbf{V} \in \mathbb{R}^{1\times H\times M\times D_h}$.
- **Association & Aggregation:** Content-addressable retrieval is performed via scaled dot-product attention: Attention weights A are computed (typically $A = \text{softmax}(\frac{\beta\mathbf{Q}\mathbf{K}^\top}{\sqrt{D_h}})$), where β is a learnable per-head parameter (clamped for stability) controlling association sharpness. The retrieved memory representation $\mathbf{r} \in \mathbb{R}^{B\times H\times T\times D_h}$ is obtained via $A\mathbf{V}$. We experimented with single ($N = 1$) and two-step ($N = 2$) iterative retrieval, where the output $\mathbf{r}$ from step n can serve as the query $\mathbf{Q}$ for step $n + 1$.
- **Fusion:** The final retrieved memory $\mathbf{r}$ is reshaped to $\mathbb{R}^{B\times T\times E}$, normalized (e.g., LayerNorm), and additively fused with the input Llama state: $\mathbf{h}_{\text{llama}} + \text{norm}(\mathbf{r})$. This combined representation carries both current contextual understanding and relevant retrieved memories.

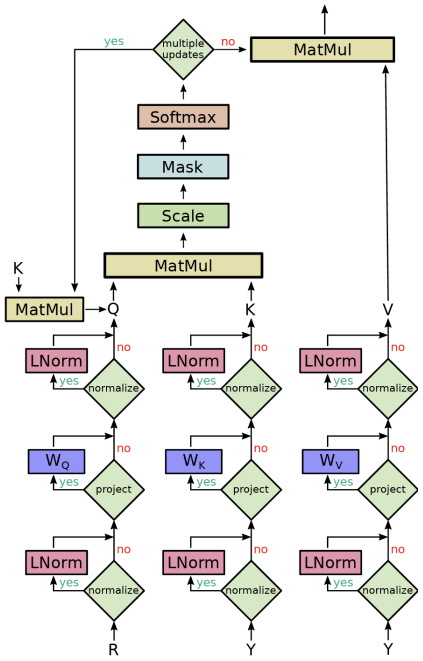

Figure 2: Hopfield layer retrieval mechanism detail. Input queries (from Llama state $\mathbf{h}_{\text{llama}}$) and memory patterns ($\mathbf{P}$) are projected ($W_q, W_k, W_v$). Attention scores ($\mathbf{Q}\mathbf{K}^\top$) are scaled, masked, and softmaxed. Weights retrieve Values ($\mathbf{V}$). Optional iterative updates ($N > 1$) use retrieved output to refine the query.

3.2 Differentiable Gated Memory Update

A core innovation of CHopT is the mechanism for updating the internal memory state $\mathbf{P}$ adaptively and differentiably.

- **Control Mechanism:** Updates to $\mathbf{P}$ are modulated by a learnable, element-wise gate $\mathbf{G} \in \mathbb{R}^{H\times M\times D_h}$. The gate value is computed dynamically based on the current interaction context: $\mathbf{G} = \sigma(\text{Linear}(\text{pool}(Q_{\text{in}}, R_{\text{retrieved}})))$, using pooled representations of the input query Q_{in} (derived from $\mathbf{h}_{\text{llama}}$) and the retrieved memory $R_{\text{retrieved}}$.
- **Gradient-Based Adaptation:** The memory update rule, $\mathbf{P} \mathrel{+}= \eta_{\text{mem}} \cdot \mathbf{G} \cdot (\mathbf{t} - \mathbf{P})$, is applied directly within the computation graph. $\mathbf{t}$ represents the target memory state derived

from the current context (e.g., based on Q_{in}), and η_{mem} is the memory update learning rate. Crucially, gradients from the downstream task loss flow back through this update operation. This allows the model to optimize the gate parameters and the target computation end-to-end, learning how and when to modify its memory to minimize future errors, thereby enabling autonomous control over memory plasticity. Stability is promoted via optional clamping of $\mathbf{P}$ and β.

4 Experiments

4.1 Task and Dataset: Cross-Inference NarrativeQA

We evaluate CHopT's cross-inference memory capabilities using the NarrativeQA dataset [Kočiský et al., 2018]. We devise a novel training and evaluation paradigm distinct from standard QA setups, explicitly designed to test memory persistence across inference boundaries:

1. **Context Encoding Phase:** NarrativeQA summaries serve as context. Each summary is fed sequentially during fine-tuning, activating the gated memory update mechanism (§3.2) to populate internal $\mathbf{P}$. The key difference between this procedure and standard Question-and-Answering is that the questions have in the same input sequence as the story.. Instead the model must rely on its internal memory across different inferences.

2. **4.2 Model and Training Details**

   - **Base Model:** Llama 3.2 3B Instruct [AI, 2024].
   - **CHopT Configuration:** Hopfield Layer: $H = 24$ heads, $M \in \{128, 512, 768\}$ slots/head, $D_h = 128$, num_updates $N \in \{1, 2\}$. Gated updates (avg_query target, mean pooling), additive fusion, stability clamping. Embedding sampling init. Details in Section 3.
   - **Fine-tuning Strategy:** LoRA [Hu et al., 2022] (r=16, $\alpha = 32$) on base attention/FFN layers. Hopfield parameters fully trained. AdamW optimizer (Base LR 2×10^{-5}, Hopfield LR $\times 5$), cosine decay, grad clipping (1.0), gradient checkpointing. 2 epochs on NarrativeQA train split (approx. 32k examples), effective batch size 64.

4.3 Evaluation Metrics

Primary quantitative metric is ROUGE-1 F1 score [Lin, 2004], measuring unigram overlap between generated and reference answers, suitable for evaluating factual recall in the cross-inference QA phase. Qualitative analysis assesses specific recall examples.

5 Results

5.1 Quantitative Analysis: Memory Capacity and Update Steps

We performed an ablation study varying memory slots (M) and update steps (N). Table 1 presents ROUGE-1 F1 scores on the NarrativeQA test set using the cross-inference protocol after fine-tuning.

Table 1: ROUGE-1 F1 Scores for CHopT on NarrativeQA Test Split (Cross-Inference Recall)

Memory Slots (M per Head)	Update Steps (N)	ROUGE-1
768	2	0.083 10
512	**2**	**0.114 46**
128	2	0.096 00
128	1	0.064 61

Increasing memory slots from $M = 128$ to $M = 512$ improved ROUGE-1 score when using two update steps ($N = 2$). However, further increasing to $M = 768$ degraded performance, potentially due to optimization challenges. Using $N = 2$ update steps consistently aided

retrieval for $M = 128$. The configuration $M = 512, N = 2$ achieved the best ROUGE-1 score (0.11446).

5.2 Qualitative Analysis

We examined model outputs on zero-context questions (Table 2). CHopT demonstrates recall of specific details (e.g., names, professions) solely from its internal memory state, information unavailable to baseline models in this setup. Successful recall highlights the Hopfield layer's ability to store and retrieve salient facts across inference boundaries.

Failure modes typically involve syntactic errors (e.g., "He it destroyed.") or retrieval of related but incorrect information leading to semantic drift (e.g., "A panic." instead of "Trouble"). This suggests challenges remain in precisely encoding complex relationships and fluently integrating retrieved fragments. The gating mechanism appears partially effective in preventing wholesale memory corruption but requires further optimization for robust, fine-grained memory protection and accurate retrieval under interference.

Table 2: Qualitative Examples: CHopT Recall vs. Ground Truth Base Llama (Zero-Context QA)

Question / Ground Truth	CHopT Output	Base Llama Output
What does Mark do with his radio station? **GT:** He dismantles it... makes mobile transmitter	`He it destroyed.`	`I don't have enough info...`
What did Mrs. Warren previously do for work? **GT:** Prostitution.	`She was prostitute.`	`I'm not sure...`
Why did McLeach throw Cody's backpack...? **GT:** To make people think he had died	`He wanted kill Cody.`	`(not tested)`
Who commits suicide? **GT:** Malcolm	`Malcolm.`	`Suicide is complex... [generic]`
Who is Lapp? **GT:** A Frenchman revealed to be a spy	`He a spy.`	`Lapp is a common last name...`
What does the radio show cause? **GT:** Trouble (FCC investigation)	`A panic.`	`There are multiple talk shows...`
What does Paige jam into her microwave? **GT:** Her medals and accolades	`A can of.`	`I don't have enough context...`

6 Discussion and Future Work

CHopT demonstrates that integrating a gated Hopfield memory provides a viable path towards LLMs with state persistence across inferences. The successful recall in the cross-inference NarrativeQA setup validates the core architectural concept. Performance sensitivity to memory capacity (M) and update steps (N) highlights the importance of these hyperparameters. The $M = 512, N = 2$ configuration likely strikes a balance between sufficient capacity for summary encoding and tractability during the limited fine-tuning period. The degradation at $M = 768$ may indicate that larger, sparsely populated memories require more extensive training or different update strategies (e.g., more targeted than avg_query) to effectively organize information and avoid interference.

The differentiable gating mechanism is central, allowing end-to-end optimization of memory control. However, qualitative errors suggest that the current gate input (pooled query/retrieval) might be insufficient for perfectly nuanced control, sometimes failing to prevent interference or ensure fluent integration. The fully differentiable update path, while enabling direct optimization, adds complexity compared to no_grad updates or RAG's fixed external memory, potentially impacting training stability and requiring careful hyperparameter management.

Future work should prioritize:

(a) **Enhanced Gating and Update Rules:** Investigate richer inputs for the gate (e.g., attention patterns, query-key similarity metrics) and more sophisticated update targets (e.g., attention-weighted queries [Qiu et al., 2024]) to improve selectivity.

(b) **Integration with Transformer Value Operation:** Integrate the Hopfield memory retrieval directly into the transformer's value projection to mitigate grammatical errors caused by post-layer memory fusion.

(c) **Long-Term Dynamics:** Implement mechanisms for explicit memory consolidation or forgetting over longer timescales.

(d) **Broader Benchmarking:** Evaluate CHopT on diverse continual learning tasks (e.g., sequential classification, lifelong dialogue) to assess generalizability.

(e) **Scaling and Efficiency:**** Investigate scaling to larger base models and memory sizes, coupled with optimizations for the Hopfield layer's computational cost.

7 Conclusion

We introduced CHopT, a Llama 3.2 model augmented with an internal, differentiable, gated Hopfield memory layer. By enabling persistent state and adaptive, selective memory updates controlled via a learned gate, CHopT demonstrates improved information recall across distinct inference boundaries in a challenging NarrativeQA benchmark. This work validates the potential of integrated, learnable memory systems with autonomous plasticity control for overcoming the inherent statelessness of standard Transformers, paving the way towards more capable continual learning agents with robust long-term memory.

References

Meta AI. Llama 3.2 model card. https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct, 2024. Accessed: [Insert Access Date].

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018. doi: 10.1162/tacl_a_00023. URL https://aclanthology.org/Q18-1023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, pages 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W04-1013.

Zihan Qiu, Zeyu Huang, Youcheng Huang, and Jie Fu. Empirical study on updating key-value memories in transformer feed-forward layers, 2024. URL https://arxiv.org/abs/2402.12233. Published as a Tiny Paper at ICLR 2024.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlovic, Geir Kjetil Sandve, Victor Greiff, David Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need, 2020. URL https://arxiv.org/abs/2008.02217.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.