Graduate Research Plan

Research Title: Controllable Memory in Large Language Models: Toward Systems That Remember, Update, and Forget Predictably.

Modern large language models (LLMs) offer remarkable short-horizon performance, yet they fall short when it comes to maintaining consistent long-term behavior, updating past knowledge, or forgetting erroneous facts in a predictable and controllable way [1,2,7]. The ideal artificial intelligence would behave like a machine whose "memories" we can edit, delete, or update at will, a system that tracks long-term plans, learns from past errors, and improves accordingly. However, present-day LLMs encode "knowledge" / "memory" diffusely across billions of parameters, opaque to external modification and difficult to control. Moreover, when tasked with long multi-step tasks, they tend to drift: errors compound over time, goals fade, and models do not anchor themselves or correct their own missteps [7].

Related Works

Research addressing memory-like behavior in LLMs has largely been siloed into four domains: unlearning, model editing, long-context, and continual learning. Unlearning concerns selectively removing specific bits of information (often for privacy or compliance) from a model's weights or behavior [1]. Model editing involves rewriting knowledge within the model, e.g., updating which individual holds a given title [2]. Long-context modelling addresses an LLM's ability to process large sequences of tokens, extending a LLM's short-/intermediate-term memory. Continual learning seeks to enable the model to integrate new information over time without catastrophically forgetting what it already knows [3]. Although each subfield advances useful functionality, they do not yet form a unified framework for memory control, and each faces its own limitations. For example, unlearning lacks widely adopted benchmarks and robust metrics for forgetting, model editing often suffers from instability or unwanted side-effects, long-context is largely limited in scope by a quadratic scaling cost, and continual learning still struggles to adapt LLMs without degradation of previous competence.

Research Plan

My particular focus follows from my paper on General Correctness Models [8], motivating long memory systems that allow LLMs to grasp general patterns from history. For example, while current LLM coding assistants could recall the details of a specific issue when directly prompted to do so, remembering general patterns e.g., "adopting an object oriented approach in this codebase caused many issues." remains an important challenge. While existing lines of work predominantly focus on improving an LLM's context length, (akin to short-/intermediate-term memory), or supplementing long-term memory with databases, (akin to remembering the past only by looking up a diary), I aim to leverage unlearning, model editing, and efficiency research to build scalable long-term memory directly into the model.

The proposed work unfolds in three stages. In **Stage 1**, given differences from existing work, I will propose a new evaluation task to measure progress on top of existing benchmarks by introducing a dataset structured around higher-level, thematic recall (rather than verbatim facts). This dataset will test a model's capacity to connect multiple past experiences, prioritize important information/gists, and keep information up to date (for example: "What are the major technical issues we encountered so far that remain unresolved?"). We can create this task efficiently by leveraging datasets built for long conversation understanding [5], adding questions for understanding higher level themes, and adding questions whose answers shift depending on time. The novelty lies in shifting from "retrieve the specific fact" toward "recall the general pattern" and framing aspects of forgetting, updating, long-context, and continual learning in a controllable memory-oriented task. In **Stage 2**, drawing inspiration from the recently proposed MemoryLLM architecture (where a fixed-size latent memory pool is embedded into each layer

of the transformer [4]), I will design a sparse memory-module integrated into an LLM. Key-value pruning, sparse attention, linear attention, and quantisation have shown promise in memory-efficient models. I will build on these techniques to constrain and explicitly store salient contextual embeddings and allow for selective forgetting and overwriting. The resulting system will take the form of a module or side network rather than a trained-from-scratch architecture to build on existing LLMs and make training feasible. In **Stage 3**, I will address optimisation and data dynamics: employing self-refinement methods akin to iterative output consolidation (e.g., Self-Refine [6]) to teach the memory-module what to keep and what to discard, enabling a data feedback loop that distils transient context into long-term memory that is semantically aware. While these methods may be expensive at scale, the possibility of distilling results into a dataset to optimize a more efficient memory-module based approach is promising. Over 3-5 years, these stages aim to produce an LLM system that retains coherent beliefs over time, can update or delete them as needed, and thereby supports long-horizon reasoning, knowledge editing, and stable behavior.

Intellectual Merit

The intellectual merit of this work lies in three contributions: (1) developing a unified conceptual framework linking unlearning, model editing, long-context and continual learning through the lens of controllable memory; (2) advancing architecture and algorithm design by integrating a sparse, persistent memory-module into an LLM that scales more favorably than long-context solutions by building in active semantics-aware forgetting; and (3) providing concrete benchmarks and evaluation methodology to measure forgetting, editing stability, memory retention and long-horizon coherence for controllable memory.

Broader Impacts

Unpredictable LLM behavior hampers trust, accountability and safety. By enabling models with predictable long-term behaviors, we move toward more reliable AI systems. Beyond safety, controllable memory has applications in personalized education (tracking student progress over months), scientific discovery (maintaining state across experiments), assistive AI (remembering user-specific contexts and forgetting when needed), and compliance/regulation (reliable removal of outdated or sensitive information). Ultimately, this research aligns with the goal of powering sustainable, interpretable AI aligned with human-centred values.

References

[1] Liu, Y., Le-Khac, U. N., and Yang, B. (2025). Rethinking machine unlearning for large language models. *Nature Machine Intelligence, vol. 7, pp. 181–194.*

[2] Yao, Y., Wang, P., Tian, B., Cheng, S., Li, Z., Zhu, S., and Chen, H. (2023). Editing large language models: Problems, methods, and opportunities. *EMNLP 2023.*

[3] Zheng, J., Qiu, S., Shi, C., and Ma, Q. (2024). Towards lifelong learning of large language models: A survey. *ACM Computing Surveys, vol. 57.*

[4] Wang, Y., Krotov, D., Hu, Y., Gao, Y., Zhou, W., McAuley, J., Gutfreund, D., and He, Z. (2024). MemoryLLM: Towards self-updatable large language models. *ICML 2024.*

[5] Chen, Y., Zhou, S., Kumar, A., Zhang, T., and Bansal, M. (2024). LoCoMo: Language-model consistency via multi-sample contrastive optimization. *ACL 2024.*

[6] Madaan, A., Muennighoff, N., Casper, S., Wang, S., and Bansal, M. (2023). Self-Refine: Iterative refinement with self-feedback. *NeurIPS 2023.*

[7] Sinha, K., Anderson, S., and Andreas, J. (2025). The Illusion of Diminishing Returns: Measuring Long-Horizon Execution in Large Language Models. *ICLR 2025.*

[8] Xiao, H., Patil, V., Lee, H., Stengel-Eskin, E., and Bansal, M. (2025). General Correctness Models: Predicting and Calibrating Model Reliability Across Tasks. *Manuscript under review at ICLR 2026.*