Final Project Proposal

Topic
Language Modeling with Memory-Augmented LSTM: Improving Long-Context Text Prediction

Teammates:
Jay Wu, Leo Yao, Aaron Jiang

Solution and Model Description
In this project, We propose a **Memory-Augmented LSTM language model** designed to improve long-context text prediction.
Traditional LSTM-based language models suffer from a limited ability to retain information across long sequences due to vanishing gradients and fixed hidden state size. To address this limitation, the proposed model integrates an **external memory component** that explicitly stores summarized or semantically enriched representations of previous contexts.

Model Overview
The system consists of two main components:

1. **Base LSTM Encoder–Decoder:**
   A standard LSTM network serves as the core language model, responsible for token-level prediction and next-word generation.
   The encoder processes the input sequence, and the decoder predicts the next token based on the hidden state and the memory-augmented context.
2. **Memory-Augmented Module:**
   To enhance context retention, two complementary memory mechanisms are introduced:
   - **Short-Term Memory (STM):** captures the recent context using a **summarization layer** and **token limit controller**, which condense previous sentences into a compact representation.
   - **Long-Term Memory (LTM):** stores semantically meaningful information derived from previous text segments.
     This includes **semantic search embeddings** and **named-entity representations (NER)**, which are retrieved during prediction to enrich the LSTM's input.

During inference, the model retrieves both the short-term and long-term summaries and concatenates them with the current input before passing them to the LSTM encoder. This design allows the model to "recall" relevant past information without depending solely on hidden state propagation.

Novelty

This approach extends the standard LSTM language model by integrating **structured memory management**:

- Instead of relying on hidden states alone, the model explicitly combines **summarized historical content** and **semantic retrieval** from previous contexts.
- This **memory-augmented design** bridges the gap between recurrent networks and modern retrieval-augmented transformers, but remains interpretable and computationally lightweight.
- The framework can be applied to domains where context continuity matters, such as **long document prediction**, **multi-turn dialogue generation**, or **resume-based job skill extraction**.

Expected Outcome

The Memory-Augmented LSTM is expected to:

- Improve **long-context coherence** in text generation tasks.
- Retain relevant entities and topics across multiple segments.
- Demonstrate higher prediction accuracy compared to a baseline LSTM, especially when the input sequence exceeds typical context lengths.