Lost in the Middle: How Language Models Use Long Contexts


– Computer Science Department, Stanford University; • University of California, Berkeley; ◆ Samaya AI

TL;DR

• If language models (LMs) can robustly use information within long input contexts, their performance should be minimally affected by the position of the relevant information.
• In QA and synthetic key-value retrieval experiments, LM performance degrades significantly when we change the position of the relevant information.
• Performance is highest when relevant information occurs at start or end of context.
• When placing relevant information in the middle of the context, QA performance is \( \approx \) closed-book baseline.

Experimental Setup: Multi-Document QA

• Goal: stay as faithful to the retrieval-augmented generation application as possible.
• Queries from NaturalQuestions (NQ).
• “Gold” context passages from NQ annotations.
• “Distractor” context passages are retrieved from Wikipedia (given only the query).
• Construct input contexts where exactly one passage contains the NQ-annotated answer.
• Control position of relevant information by changing the position of the gold passage.
• Control the input context length by retrieving more distractor documents.
• Evaluation: does the NQ-annotated answer string occur in the language model output?

Primacy and Recency Bias in LMs

Closed-book: prompt LMs with only the query

Effect of Instruction Tuning

LM Reader EM saturates long before retriever recall
• In principle, retrieving more documents should help, but we’re bottlenecked by the LM’s ability to use the input contexts.