Qeios

Peer Review

Review of: "Enhancing Long Context Performance in LLMs Through Inner Loop Query Mechanism"

Sylvain Lamprier[1]

1. University of Angers, Angers, France

The paper proposes a new mechanism to capture long range dependencies in LLM contexts, for answering specific questions from this context. It builds on RAPTOR, a previously proposed approach that relies on clustering of summaries of the context. From that approach, it takes the tree-based clustering mechanism, while proposing to enhance summaries with a surprising fact that the llm can choose to extract from the considered chunks to summarize. Additionnaly, it experiments an additional multi-turn answering process, that attempts to split the global answering problem into multi-steps given a current short term memory component. This memory is fed at each turn with answers from previous answering steps.

Strengths:

- Results look interesting for both the inclusion of surprising sentences and the multi-turn answering process

- The proposals, while rather straightforward given the many other papers about prompt engineering and RAG, may contain some innovation. Specifically, the use of surprising facts is intriguing.

Weaknesses:

- There is currently a huge litterature about prompt engineering in multi-turn asking llms to iteratively solve various kinds of problems with complex contexts. Related work should discuss them more thoroughly.

- The problem should be formalized and the method more precisely defined. Many central

 doi.org/10.32388/85ALJN

components of the method are not sufficiently detailed to allow the reader well understand the approach. The paper too strongly relies on RAPTOR, without sufficiently presenting its core components. To be self-contained, the paper should include many details of that former approach, and then more precisely focus on what differ from it. The inner-loop mechanism is really lacking from details. While authors refer to fig 1 to support it, I cannot see any loop in that figure. This could be very helpful to include the overall pseudo-alo and a global schema of the loop. wIn fact, without the prompts given in appendix, I think I would have missed almost everything of the proposal.

- What is a surprise is not sufficiently discussed in the paper, while its inclusion is the main innovation of the paper (inner-loop is already very usually considered in COT and RAG litteratire). I understand that it is simply to the llm to decide what is surprising from the input chunks of texts? but surprise can take many aspects given the context. I feel the llm can output very surprising surprises in many contexts... Wouldn't it be necessary to specify what we consider a surprise (e.g., with a few-shot in-context prompting for instance) ? Also, I understand that texts are not clusterized given that surprise. So, why is it important ? Why this surprise is not already included in the produced summary ? And what could we do when several surprising facts belong to the same cluster ?

- The experimental methodology also suffers from a great lack of details, baselines and ablations. Why the datasets considered in RAPTOR are not considered here ? Why do we need these new ones (that authors should more precisely present) ? Why including three specific sentences in M-NIAH ? Wouldn't it be relevant to give more classical quantitative metrics such as ROUGE or BLEU ? How does the approach compares to a multi-turn summarizer without surprise ? Compared to other long context approaches that either summarize the context (with a rnn or a multi-turn llm) or use sparse attention llms (e.g., longformer) ?

Declarations

Potential competing interests: No potential competing interests to declare.