Dear reviewers and area chairs,
We thank you and the reviewers for still considering our submission. We found the reviewers' comments and suggestions to be very helpful. Thank you again for your time and effort in considering our comments. Please find the responses (in blue) to individual questions asked by the reviewers (in black) below.

**Area Chair**

The paper and the authors should be very clear on their goals. In response to one of the reviewers, the authors say "Our goal is to understand parametric knowledge in LLMs for temporal QA." In that case what is the point of a relevant context? The notion of relevant context as used in the paper is very narrow in that it must contain the answer to the question. A more broader view of a relevant context would be that the context needs to be used (perhaps together with the parametric knowledge) to answer the question. Perhaps the goal of the paper is overstated. To stay true to the stated goals, there should be more focus on temporal aspects and the notion of a relevant context should be broadened. Various other points made by the reviewers should be addressed.

Thank you for your thoughtful feedback. We have significantly revised the manuscript in response to your comments, with particular attention to clarifying the paper's goals and ensuring they are appropriately aligned with the contributions and empirical findings.

As you note, our original framing emphasized an interest in understanding parametric knowledge in large language models for temporal question answering. We agree that this goal was overly ambitious given the methodology presented. In the revised version, we more precisely frame our focus: rather than aiming to isolate parametric knowledge in the abstract, we study how different reasoning strategies affect model robustness when the temporal context varies in quality, ranging from missing to misleading. The revised framing is more modest and better supported by the experimental design and results.

In particular, we now emphasize that our work contributes a method for improving LLM robustness under degraded context conditions, rather than claiming to diagnose or probe parametric memory directly. While parametric knowledge undoubtedly plays a role in the TQA setting, our work does not attempt to measure it in isolation, and we have removed language that might have suggested otherwise.

We also appreciate your point about the narrowness of the definition of "relevant" context. In the original version, we defined relevance in binary terms, i.e., whether the context contained the gold answer. This oversimplification has been addressed in the revision. We now provide a broader discussion of contextual relevance, acknowledging that useful context may support inference even without explicitly containing the answer, and that LLMs often need to reconcile multiple sources of information (retrieved and latent) to answer temporal questions.

To better match our stated goals and address these conceptual issues, we have revised the contribution statement as follows:

1. We introduce a modular, agent-based approach (RASTeR) that separates temporal context evaluation from answer generation, enabling more robust performance when context is noisy, incorrect, or unavailable.

2. We evaluate this method across three LLMs and four distinct TQA datasets, including newly constructed settings where contextual degradation is introduced in a controlled way.

3. We analyze model behavior under varying context conditions, showing that existing methods often perform poorly outside of idealized settings, and that the proposed modular design helps mitigate this brittleness.

We believe these revisions better situate the paper within its appropriate scope and clarify the specific challenges and contributions of our work. The paper no longer attempts to characterize LLM memory in general, but instead presents a focused study on robustness in temporal reasoning tasks under context perturbations, along with a method that demonstrates practical gains in this setting.

Thank you again for your detailed and constructive feedback.

**Reviewer dJzF22**

Thanks for your review. We appreciate your comments and have responded below:

- Unclear or missing definitions (or references)
  - To the best of our abilities we have explicitly defined concepts.
  - For example we define
    - Robustness: (as it relates to us) up front in the abstract.
      - the ability to answer correctly despite suboptimal context
    - Temporal Question Answering: at the beginning of the related work section.
      - Temporal QA tasks involve understanding how events unfold over time, whether in text, video, or structured data such as knowledge bases
    - Adversarial training is defined in the Robustness in Retrieval-Augmented Generation subsection of the related work
      - Adversarial training methods expose models to noisy or counterfactual inputs to encourage robustness.
- The claim regarding the order of context cannot be confirmed or refuted due to the small amount of data; the different instruction format might be the cause of the better performance instead of the order question and context

  - Thank you for pointing this out. We agree that our argument for context order in temporal question answering was meekly supported due to a lack of experiments and small data size. Thus, we have dropped this angle from the paper in order to focus on more interesting findings.
- 24 of 36 papers from the literature referenced are from either 2024 or 2023, only 3 before 2020 even though temporal QA is a long-standing research field.
  - Thank you for the suggestion We have, as you suggested, incorporated citations on TQA dating before 2020. This was extended substantially, here is a sample:
  - Petroni, Fabio, et al. "Language Models as Knowledge Bases?" *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, 2019, pp. 2463–2473. https://aclanthology.org/D19-1250/.
  - Velupillai, Sumithra, et al. "BluLab: Temporal Information Extraction for the 2015 Clinical TempEval Challenge." *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Association for Computational Linguistics, 2015, pp. 815–819. https://aclanthology.org/S15-2137/.
  - Jang, Yunseok, et al. "TGIF-QA: Toward Spatio-Temporal Reasoning in Visual Question Answering." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2758–2766.
  - Llorens, Hector, et al. "SemEval-2015 Task 5: QA TempEval—Evaluating Temporal Information Understanding with Question Answering." *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 2015, pp. 792–800.
  - Jia, Zhen, et al. "TEQUILA: Temporal Question Answering over Knowledge Bases." *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, ACM, 2018, pp. 1807–1810. https://doi.org/10.1145/3269206.3269247.
  - Zhu, Linchao, et al. "Uncovering the Temporal Context for Video Question Answering." *International Journal of Computer Vision*, vol. 124, no. 3, 2017, pp. 409–421. https://doi.org/10.1007/s11263-017-1033-7.
  - Sun, Weiyi, Anna Rumshisky, and Ozlem Uzuner. "Temporal Reasoning over Clinical Text: The State of the Art." *Journal of the American Medical Informatics Association*, vol. 20, no. 5, 2013, pp. 814–819. https://doi.org/10.1136/amiajnl-2013-001760.
- It appears to me that the one-shot example prompt used for generating contexts violates the causal order given in downstream application, as it requires knowing the answer ("Schindler's List") and not only the question, so the generated data might be biased in such a way that the approach might not generalize to situations where the context is retrieved automatically. In the 1-shot example, this causal relationship could have been respected e.g. by giving a list of Oscars awarded in the respective year.
  - Thank you for this insightful observation. You are correct that using examples where the answer is known during context generation can introduce biases that

may not reflect real-world deployment settings, particularly when context must be retrieved automatically. In the revised version of the paper, we have addressed this concern by moving away from generated contexts (i.e., adding other datasets) and instead using datasets where the context is provided independently of the answer. This design choice ensures that the evaluation more faithfully reflects the causal structure of the task and avoids potential information leakage. We appreciate your suggestion, which helped us sharpen the experimental setup and strengthen the generalizability of our findings.

- It is not clear how strictly the quality of the generated context was scrutinized and what criteria were used (Quote: "To confirm the validity of our GPT-generated context, we manually reviewed a sample of 100 contexts. Only three examples were not entirely correct.")
    - This was another great point. As mentioned above, we have remedied this issue by discarding the dataset we constructed in favor of pre-existing TQA Datasets.

**Reviewer rr4N19**

Thanks for your review. We appreciate your comments and have responded below:

- While the paper claims robustness in solving temporal question answering, it lacks comparative evaluations with established datasets like TimeQA, TempReason, and MenatQA. Evaluating on these datasets could significantly bolster the robustness claims. Such benchmarks are crucial for validating the proposed method's effectiveness.

- This criticism was a huge oversight on our part. We have included the three datasets you recommended to our study as well as a temporal reasoning dataset that was released this year (UnSeenTimeQA).

- The creation of new datasets is certainly beneficial, yet the paper does not clearly articulate the deficiencies in existing temporal question-answering datasets like TimeQA, TempReason, and MenatQA, which already include matching contexts. A detailed comparison highlighting what these current datasets lack and how the new ones address these shortcomings would provide a stronger rationale for developing additional resources.
- We no longer are selling this paper as producing a dataset.

**Reviewer FXct**

Thanks for your review. We appreciate your comments and have responded below:

- Although targeting temporal QA introduces some novelty, the overall findings of the paper are quite similar to those of Yoran et al., 2024.

- Thank you for this observation. We agree that there is some conceptual overlap with Yoran et al. (2024), particularly in the interest in evaluating model robustness in temporal question answering. However, our approach differs in both methodology and focus. While Yoran et al. explore robustness by introducing mixed-context inputs during training, our work adopts an agent-based prompting strategy that separates context evaluation from answer generation. This allows for structured intermediate reasoning using temporal knowledge graphs and improves interpretability across context conditions. We also evaluate our method across three language models and four diverse temporal QA datasets, including settings with misleading, incomplete, and missing context. We believe this perspective offers a distinct and complementary contribution to the understanding of temporal reasoning in large language models.

- The paper introduces two datasets but only discusses results from one of them. It is unclear if there are any interesting or different findings from the other dataset.

- Thank you for pointing this out. In the revised version of the paper, we now highlight and report results from all four temporal QA datasets directly in the main text. While earlier versions may have underemphasized some datasets, we now present a unified analysis that spans multiple domains and temporal question types. This broader evaluation helps demonstrate the generality of our approach and also allows us to surface interesting differences in model behavior across datasets, which we discuss in the results and analysis sections.