Explanation of Revisions for ARR February Submission #713

We thank the Area Chair and the reviewers for their thorough and constructive feedback. We have revised the paper to address the main concerns raised. The key changes are summarized below, organized by the major themes from the meta-review.

1. Strengthening the "Attention Overflow" Hypothesis

A key suggestion from the Area Chair (AC), Reviewer 9fyt, and Reviewer ahAM was that our central hypothesis lacked direct evidence and relied primarily on input-output behavior.

- **New Attention Visualization:** In response, we have added a new appendix (Appendix A) containing visualizations of attention patterns from the final layer of the model during the contrastive verification task.
- **Supporting the Hypothesis:** These visualizations provide direct evidence for our claim. Figure A.1 shows that when a queried item is *present*, attention is sharply focused on its location. Conversely, Figure A.2 shows that when an item is *absent*, attention becomes diffuse and spread across the input, consistent with a "search" process. This difficulty in searching an entire list for an absent item supports our hypothesis that the model's attention capacity "overflows" when it must simultaneously generate a candidate and verify its absence against many items.
- **Integration into Main Text:** We have updated the Analysis and Conclusion sections to reference this new evidence, strengthening the link between the observed behavior and the underlying mechanism.

2. Extending the Analysis and Clarifying Contributions

Reviewers asked for more insight into why the phenomenon occurs and for clarification on the broader implications of the task (R-S6Li, R-ahAM).

- **Broader Task Relevance:** To address comments on the task's scope (R-S6Li), we have revised the introduction and conclusion to better connect our findings to applications beyond recommendation. We now explicitly mention tasks like verifying the exhaustivity of checklists, where identifying what is *absent* from a context is critical.
- **Clarifying the Contrastive Task:** We have refined the description of the contrastive evaluation (Section 4.3) to make its purpose clearer. We now explain that this setup isolates the verification sub-task and shows it is non-trivial, which helps explain the failure in the more complex generative task.
- **Movie Benchmark Interpretation:** We have clarified in Section 4.1 that the low accuracy on the movie benchmark is not a model failure, as many valid answers exist. We reinforce that the critical failure metric for this task is the rising repetition rate, which is an unambiguous error.

3. Discussing Mitigation and Baselines

Reviewers ahAM and 9fyt suggested discussing potential solutions and including a human baseline for context.

- **Mitigation Strategy:** We now explicitly propose a practical mitigation strategy in the conclusion. Instead of relying solely on the LLM, users can leverage external tools, such as a code interpreter, to deterministically verify if a generated item is already present in the input list. This is a more robust solution than the simple iterative loops mentioned in the original abstract.
- **Human Baseline:** We have acknowledged the absence of a human baseline in the Limitations section. While we maintain that a tool-assisted baseline (which achieves perfect accuracy) is arguably more relevant for this specific algorithmic task, we agree that a human baseline provides valuable context. We have added this as a direction for future work.

We believe these revisions substantially improve the paper by providing direct evidence for our central claim, clarifying the significance of our findings, and addressing the primary concerns raised during the review process. The entire manuscript has been edited for clarity and conciseness to accommodate these changes within the page limits.