DoubleDipper: Improving Long-Context LLMs via Context Recycling

We made **substantial** improvements in the paper since the last ARR submission in October 2024. The key improvements fall into two main categories: a major expansion of our experimental evaluation and a comprehensive clarification of the paper's core claims and contributions.

Below is a summary of the most important changes:

1. Experimental and Analytical Improvements
Expanded Model Suite: We have broadened our evaluation to include 12 distinct LLMs (up from a smaller set previously), covering the latest powerful open-source (e.g., Llama 3.1, Mistral-Nemo) and proprietary models (e.g., Gemini Pro, GPT-4o-mini). This robust evaluation demonstrates that DoubleDipper's benefits are consistent and generalizable across a wide range of model families and sizes.
In-depth Analysis of the Few-Shot Generator: We introduced a new ablation study dedicated to understanding the impact of the generator model's quality. Our findings show that while more capable generators offer a slight edge, our method remains highly effective even when using a small, efficient model like Gemma 2B, confirming its practical viability. We also show that DoubleDipper can even benefit from demonstrations generated by even smaller LLMs (0.5-1B parameters). Finally, we show the impact of providing incorrect demonstrations.
New Baseline Against Traditional ICL: To better quantify our method's novelty and efficiency, we now include a direct comparison against a traditional In-Context Learning (ICL) baseline using examples from an external dataset (Section 6, Table 6). Our results show that **DoubleDipper outperforms traditional ICL by 9.5 absolute points on average**, empirically validating the superiority of our context recycling approach.

2. Writing and Clarity Improvements
Sharpened Contributions and Positioning: We have significantly rewritten the Introduction (Section 1) and Background (Section 2) to more clearly articulate our primary contribution—the concept of "context recycling"—and to better situate our work in relation to prior art on automatic ICL and long-context QA.
Clarified Key Findings and Nuances: The Results section (Section 5) has been revised to provide a clearer narrative. We now explicitly articulate and analyze the nuanced generalization performance of our method, particularly the difference in evidence retrieval capabilities between large proprietary models and smaller open-source models (Result 2).
Improved Readability: The overall manuscript has been edited for clarity, flow, and conciseness to ensure our claims and the evidence supporting them are presented as clearly as possible.

We are confident that these revisions have substantially improved the paper.

Sincerely,
The Authors