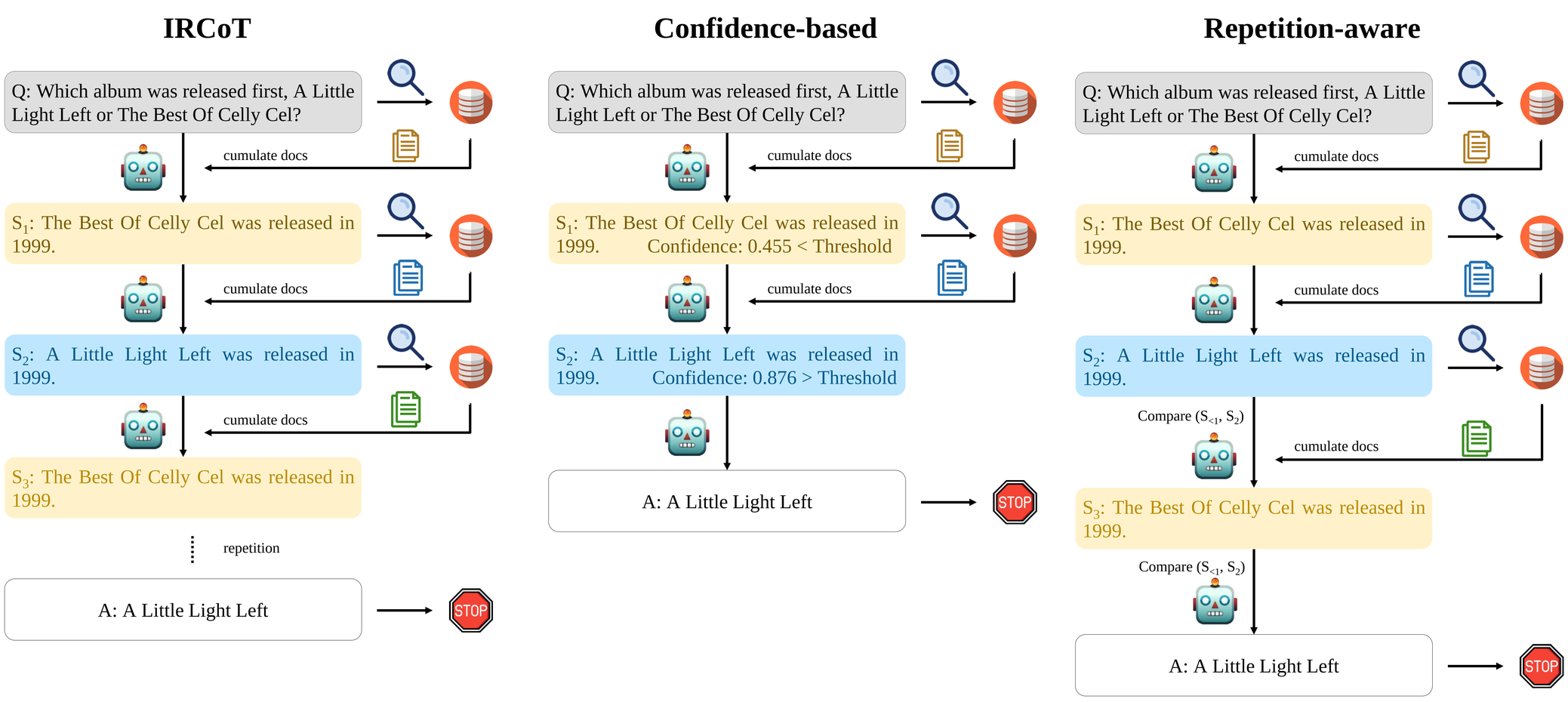
RAG Efficiency in CoT Reasoning

CS6101 Project 12 — Xinpeng Liu, Ng Xuan Hern, Oshan Jayawardena

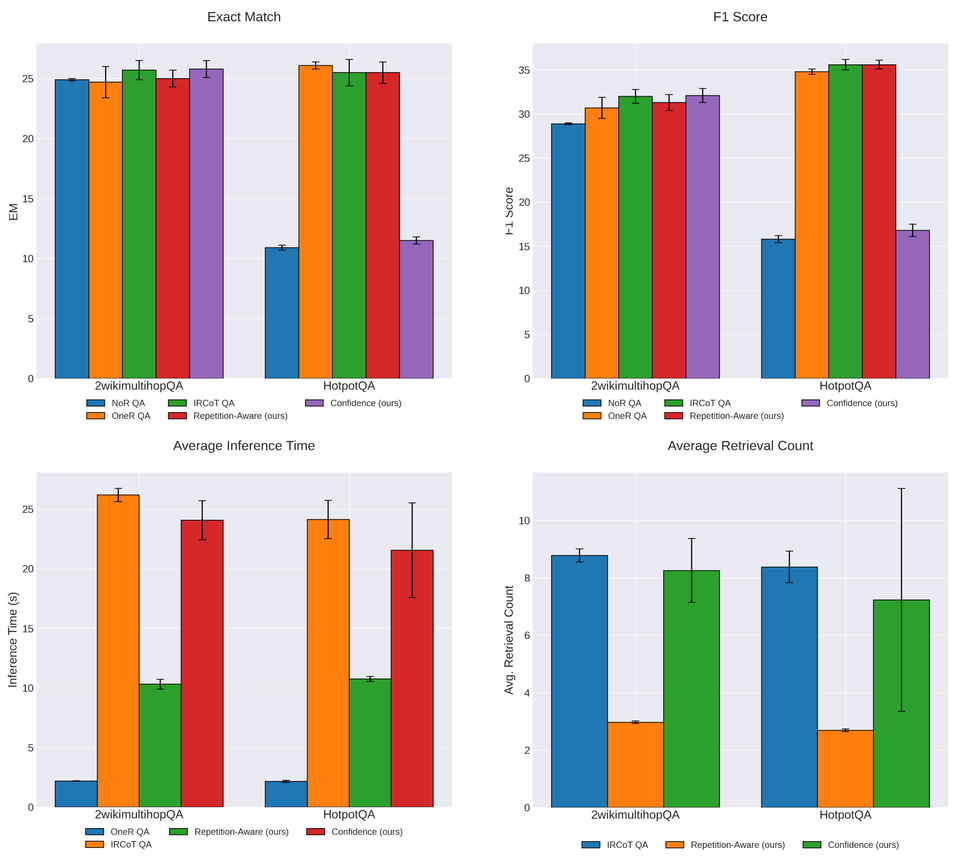
Abstract

Retrieval-augmented generation (RAG) systems augmented with chain-of-thought (CoT) reasoning have achieved strong performance on multi-hop question answering, but they incur increased inference latency and produce lengthy contexts that hinder scalability. This project introduces two stopping criteria — a Repetition-aware criterion that detects redundant reasoning tokens and halts generation when steps begin to repeat, and a Confidence-based criterion that terminates reasoning once model's confidence surpasses a threshold. We integrate these criteria into a CoT-enabled RAG pipeline and evaluate their feasibility on HotpotQA and 2WikiMultiHopQA, measuring inference latency, generated-context length, and answer quality. Rather than presupposing benefits, we report our experimental measurements and provide a detailed analysis of the observed advantages and limitations for each method. Our results offer grounded, practical insights into when lightweight stopping mechanisms may help make CoT-RAG systems more efficient and where further refinement is required.

Methodology



Results



Conclusion & Analysis

Repetition-Aware

- The Repetition-Aware strategy successfully decoupled high performance from high resource consumption, achieving a significant reduction in Inference Time and Average Retrieval Count while preserving high accuracy.
- The observed efficiency gain, rooted in reduced context length, appears to enhance the system's robustness against variations in prompting. This mechanism likely mitigates LLMs' tendency towards "context forgetting" during iterative generation.

Confidence-based

- While the Confidence-based approach did yield marginal improvements in efficiency, the overall gains were modest and frequently accompanied by performance instability across experiments.
- A critical limitation is the method's high sensitivity to its hyperparameter: the optimal confidence threshold is heavily dependent on the specific dataset, task, and underlying LLM. This lack of robustness necessitates extensive, task-specific tuning, making generalized deployment challenging.

References

[TBK23] Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions. ACL 2023.

[ZZY24] Retrieval-Augmented Generation for AI-Generated Content: A Survey. arXiv 2024.

[GXG24] Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv 2024.

[TSO25] Confidence Improves Self-Consistency in LLMs. ACL 2025.