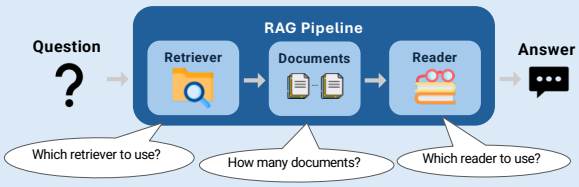
RAGGED: Towards Informed Design of Retrieval-Augmented Generation Systems

Jennifer Hsia*, Afreen Shaikh*, Zhiruo Wang, Graham Neubig
Carnegie Mellon University

Carnegie Mellon University
Computer Science Department

Introduction



Why RAG Matters: Access to up-to-date knowledge, improved accuracy for complex tasks, cost-effective knowledge integration

Challenges: Noisy data, retriever-reader mismatch, diverse task requirements.

Solution: The RAGGED framework, a systematic tool for optimizing RAG configurations.

Setup

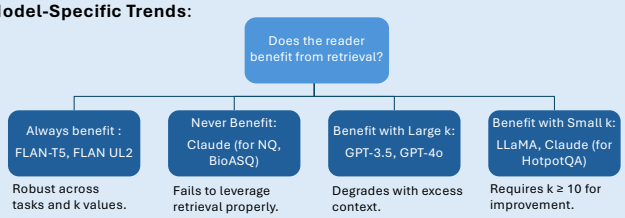
Retrievers: BM25 (dense), ColBERT (sparse).
Readers: GPT-3.5, Claude Haiku, FLAN-T5, FLAN-UL2, LLAMA2/3.
Datasets:

- Natural Questions (open-domain, single-hop)
- HotpotQA (open-domain, multi-hop)
- BioASQ (specialized domain, biomedical)

Evaluation Metrics: Recall@k for retrieval; F1 for reader performance.

RQ1. Under What Conditions Does RAG Outperform No-Context Baselines?

Model-Specific Trends:



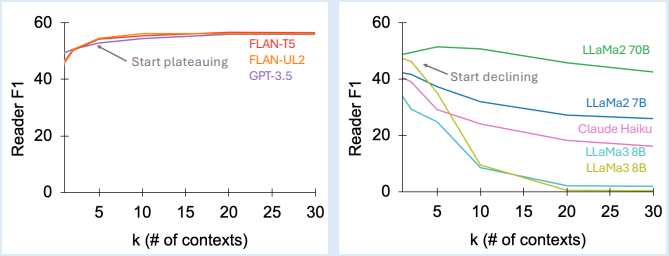
Task-Specific Trends:

Multi-Hop Tasks (e.g., HotpotQA): Show significant improvements from retrieval due to the need for reasoning across multiple contexts.

Single-Hop Tasks (e.g., NQ, BioASQ): Marginal improvements unless pretraining is insufficient (e.g., in BioASQ).

RQ2. What Reader Trends Emerge as Context Size Increases?

Type 1: Improve-then-Plateau | **Type 2: Peak-then-Decline**

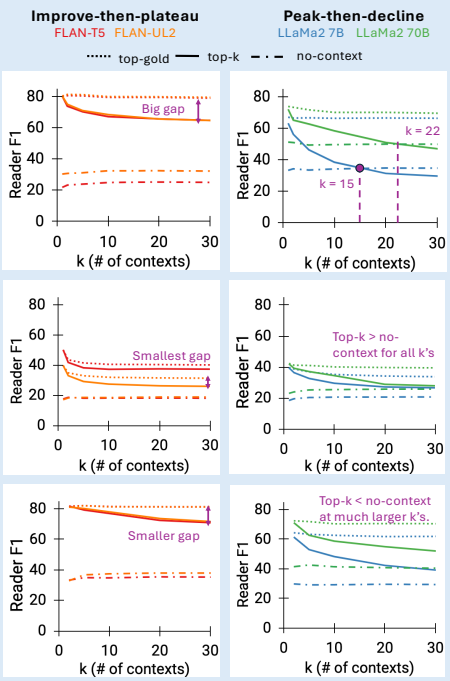


- For sensitive (type 1) readers, limit context size.
- For robust (type 2) readers, provide more context.

RQ3. How Robust Are Readers to Noise When the Gold Passage Is Retrieved?

We analyze cases where the top-k retrieved documents include at least 1 gold passage.

Key aspets:1) gap between top-gold and top-k, 2) when top-k < no-context.



Natural Questions

- Top-k declines sharply for peak-then-decline models.
- Improve-then-plateau remain robust, outperforming no-context for all k's.

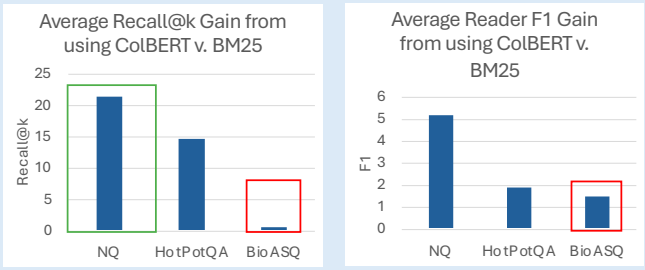
BioASQ

- Smaller gap between top-gold and top-k.
- Top-k consistently outperforms no-context for all models and k's.

HotpotQA
Multi-hop provides more signal anchors, delaying when top-k starts performing worse than no-context.

Practical Takeaway: Use robust readers for noisy real-world scenarios.

RQ4. How does retriever choice impact performance?



Better retrieval gains in open-domain tasks than in special-domain tasks.

Small retriever gains amplified in special domains.

Dense retrievers improve recall but not always reader performance.

Specialized tasks (BioASQ) benefit more from dense retrievers than open-domain tasks.

Key Takeaways

1. Suboptimal RAG can be worse than no-context.
2. Reader Robustness varies greatly by reader and question type.
3. Specialized domains amplify retriever gains.
4. Future direction: Enhance reader robustness (pretraining, fine-tuning, post-generation).