Project Proposal: Optimizing Knowledge Retrieval in Retrieval Augmented Generation

Kaibo Zhang
Email: kbzh2558@mit.edu

Annie Liu
Email: yil297@mit.edu

Lauren Zhang
Email: lzhang27@mit.edu

1 Problem Summary

Large Language Models (LLMs) such as GPT excel at reasoning and generation but remain constrained by their pretrained knowledge. The Retrieval-Augmented Generation (RAG) framework mitigates this by combining a retriever that selects relevant documents with a generator that conditions responses on them, improving factual accuracy and adaptability. Yet RAG performance depends critically on *which* and *how many* documents are retrieved: too few risk missing key evidence, while too many add noise and latency. Existing systems use heuristic ranking methods (e.g., BM25) and fixed k cutoffs, lacking optimality guarantees. We propose an optimization-based RAG formulation that models retrieval as a Mixed-Integer Optimization (MIO) problem, jointly selecting the most informative and non-redundant documents to maximize accuracy under context constraints.

2 Proposed Approach

The proposed robust MIO framework optimizes retrieval decisions in RAG systems by jointly selecting the right documents and the optimal number of documents per query. Let $\mathcal{D} = \{d_1, \ldots, d_n\}$ denote candidate documents with embeddings μ_i and binary selection variables $x_i \in \{0, 1\}$. Each μ_i lies within an uncertainty set flexible of choice $\mathcal{U}_i = \{\tilde{\mu}_i : \|\tilde{\mu}_i - \mu_i\|_* \leq \rho_i\}$, to account for embedding variability. The robust MIO objective is

$$\max_{x \in \{0,1\}^n} \min_{\tilde{\mu}_i \in \mathcal{U}_i} \Big[\sum_{i=1}^{n} \mathrm{sim}(q, \tilde{\mu}_i) x_i - \lambda_1 \sum_{i<j} \mathrm{sim}(\tilde{\mu}_i, \tilde{\mu}_j) x_i x_j - \lambda_2 \sum_{i=1}^{n} x_i \Big].$$

The inner minimization captures the worst-case drop in both query–document relevance and inter-document redundancy under embedding perturbations. Dualizing the robust term yields a deterministic equivalent, where similarity scores are penalized by ρ_i, producing stable yet conservative selections. After linearization, the model reduces to a 0–1 knapsack structure, where each document contributes adjusted value $\tilde{r}_i$ and consumes token cost c_i, constrained by $\sum_i c_i x_i \leq B$. This interpretation allows standard capacity constraints while preserving robustness through modified coefficients. Additional constraints can encode (i) source diversity to prevent overreliance on one domain, (ii) minimum coverage to ensure at least one document per topical cluster, or (iii) retrieval smoothness to restrict abrupt changes in selected documents across adjacent queries. Hyperparameters λ_1, λ_2, and ρ_i will be tuned via cross-validation to trace the Pareto frontier between coverage, diversity, and robustness, selecting parameter settings that best balance factual accuracy with retrieval efficiency.

3 Assessment and Impact

We expect the optimization-based retriever to outperform naive top-k RAG baselines. Performance will be compared under identical prompts and temperature settings using retrieval metrics (precision@k, recall@k, and diversity) and generation metrics (cosine similarity, BERT score, ROUGE-L). Token efficiency, the average retrieved tokens per correct answer, will measure cost savings. Overall, the MIO-based approach should deliver more relevant, diverse, and concise evidence, enhancing grounding quality without added computational cost.

4 Data

We will use two domain-specific datasets to evaluate the proposed framework. The RAG-Mini (BioASQ and Wikipedia) dataset [1] provides compact question–answer–passage triplets for quick testing of retrieval accuracy and diversity in narrow contexts. The SEC 10-Q Financial Reports dataset [3] includes manually reviewed question–answer pairs linked to company filings from AAPL, AMZN, INTC, MSFT, and NVDA, suitable for assessing precision in structured financial text. If scalable, we will also benchmark on the large-scale Natural Questions (NQ) dataset [2], which features real user queries and full Wikipedia passages, to test generalization and retrieval efficiency in open-domain settings.

References

[1] Hugging Face Community and Contributors. Rag datasets: A collection of small-scale and domain-specific question-answer-passage corpora. https://huggingface.co/rag-datasets, 2024. Accessed: 2025-11-04.

[2] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. https://ai.google.com/research/NaturalQuestions/dataset, 2019. Accessed: 2025-11-04.

[3] Docugami Team. Sec 10-q rag dataset: Financial question-answer corpus for retrieval-augmented generation. https://github.com/docugami/KG-RAG-datasets/blob/main/sec-10-q, 2024. Accessed: 2025-11-04.