Retrieval-Augmented Generation for Audience-Specific Answering: Final Report

Frederick Hayes III | GenAI-267 | April 13, 2025

Executive Summary

This project explores the construction and optimization of a dual-audience Retrieval-Augmented Generation (RAG) system designed to serve both engineering researchers and marketing professionals. By combining LangGraph-based orchestration, modular retrieval, advanced reranking, and strict audience-aware prompting, I built a flexible architecture that consistently improved answer quality. The final configuration achieved a Fred-R score of **0.82** and a Fred-M score of **0.75** on the full 75-question test set, significantly outperforming the tuned baseline pipeline. Key drivers of improvement included chunk optimization, hybrid retrieval, reranker choice, and prompt design.

Introduction

Large Language Models (LLMs) are powerful, but their responses can vary significantly depending on the audience. In real-world enterprise settings, technical users expect detailed, mechanism-aware explanations, while non-technical stakeholders prioritize clarity and business relevance. My project addresses this gap by constructing a RAG system that routes answers through tailored pipelines based on audience intent.

The corpus was composed of ArXiv papers, Wikipedia entries, and expert blogs on NLP, AI, and LLMs. Documents were split using recursive chunking, embedded with dense models, and indexed using both vector (Qdrant) and symbolic (BM25) methods. The system uses LangGraph to support dynamic sub-question decomposition, reranking, and hybrid scoring. Two distinct prompting templates are used: one for researchers (detailed, technical) and one for marketing professionals (clear, grounded, digestible).

Key Findings

Prompting played a decisive role. Audience-specific templates designed to emphasize quoting and precision for researchers, and simplicity for marketing, improved BLEURT and LLM Reviewer scores by large margins. The Fred Score—our composite metric encompassing fluency, semantics, retrieval alignment, and responsiveness—demonstrated a clear improvement over time, particularly after adding BLEURT and the LLM Reviewer as core semantic metrics.

Reranker model choice was critical. ELECTRA and Cohere Rerank V2 consistently outperformed MiniLM, though optimal performance varied slightly between research and marketing queries. Moreover, these rerankers significantly affected the "Gold in Context" metric and downstream generation quality.

Query expansion worked best for research-oriented questions, where adding technical synonyms or prompting for mechanisms improved recall. However, for marketing-style questions, expansion sometimes introduced noise or tangents that reduced response clarity.

Chunk size and overlap were key to grounding. We found that 360-token chunks with 120-token overlap optimized coverage without redundancy. Over-filtering on overlap scores harmed recall, while under-filtering flooded the reranker.

LangGraph decomposition made a major impact. Triggered for vague or compound questions, it reliably improved grounding and answer completeness. This was especially helpful for longer queries or multi-part prompts.

Experimental Methodology

Technical Approach

Four pipelines were developed during experimentation:

1. A basic dense retrieval baseline
2. A hybrid (dense + symbolic) retriever with filtering
3. A reranked baseline with prompt and chunk tuning
4. A LangGraph-based advanced pipeline with decomposition, boosting, and SBERT reranking

All pipelines were modular and used interchangeable embedding models (mpnet-dot-v1, mixedbread), rerankers (MiniLM, ELECTRA, Cohere), and LLMs (Cohere and Mistral). Each was configurable by audience to enable targeted optimization.

Fred Score, the central evaluation metric, evolved over time from a flat average to a weighted bucket system. The final version balanced four categories:

- Fluency and Form (ROUGE, METEOR)
- Semantic Match (BLEURT, Answer F1, LLM Reviewer)
- Retrieval Alignment (Gold-in-context cosine, Retrieval F1)
- Responsiveness (sigmoid-scaled length match)

The evolution of the Fred Score reflected insights gained from error cases, including the Paper Airplane Test, in which a deliberately fluent but irrelevant answer ("paper airplanes") received a high score under ROUGE/F1 and BERT. BLEURT and the LLM Reviewer correctly rated it near zero. This discovery led to a major overhaul of the scoring architecture away from BERT and towards the LLM Reviewer and BLEURT.

Testing and Evaluation

A labeled set of 75 questions with audience-specific gold answers was used. We used a random number generator to sample 8 for iterative tuning and ran all 75 for final evaluation. Deterministic generation was used for consistency. We implemented a hallucination detector to flag fictional terms or overly abstract responses and relied on both statistical and embedding-based recall diagnostics.

Final results were logged with metadata including reranker score, token count, and retrieved document source. These insights helped guide model selection, chunking, and prompt strictness.

System Design Details

Fred Score Bucket Architecture

The Fred Score is a weighted composite metric that captures four dimensions of answer quality. Each bucket is grounded in specific submetrics:

- **Fluency and Form**: Measures language clarity, coherence, and surface correctness using ROUGE-L and METEOR.
- **Semantic Match**: Evaluates whether the generated answer preserves the meaning of the gold answer using BLEURT, Answer F1, and LLM Reviewer judgment.
- **Retrieval Alignment**: Assesses if the answer was grounded in retrieved context via cosine similarity (Gold in Context), retrieval recall, and F1.
- **Responsiveness**: Measures the answer's length appropriateness relative to target word limits using a sigmoid-scaled length penalty.

These buckets were tuned independently and reflect real-world QA needs: clear writing, factual accuracy, faithful citation of context, and appropriate brevity or depth depending on the audience.

LangGraph-Based Decomposition Pipeline

The LangGraph orchestration allowed conditional sub-question generation via a decomposition node. When triggered—either due to vague phrasing or failed context retrieval—the system routed the question to Claude Haiku or a fallback LLM. The node returned 1–4 focused sub-questions, each independently retrieved and reranked. Sub-question contexts were deduplicated and aggregated, then passed to the LLM. This modular approach improved coverage and reduced the likelihood of hallucinated generalizations.

Hybrid Retrieval and Reranking

Retrieval used a hybrid pipeline combining Qdrant dense search with BM25 symbolic scoring. After initial retrieval, results were optionally reranked using a cross-encoder (MiniLM, ELECTRA, or Cohere Rerank V2). This stage critically determined which contexts were passed to the LLM. The reranker improved grounding by upweighting passages containing precise or quoted terms. Cohere's reranker was the most robust for general cases, while ELECTRA excelled on short factual lookups.

Results and Findings

The LangGraph pipeline clearly outperformed simpler approaches across most audience-question pairs. Research prompts benefited most from decomposition, boosting, and BLEURT optimization. Marketing prompts gained from concise generation, keyword reranking, and sentence-based trimming.

The hallucination detector and Paper Airplane Test were essential. They revealed that traditional metrics could not reliably detect made-up or unsupported content. These diagnostics helped us re-tune prompts and emphasize fallback behavior when context was missing.

Lessons Learned

Modularity was the right call. Swapping rerankers, retrievers, and LLMs helped diagnose performance bottlenecks and allowed for safe iteration. Prompting proved more impactful than

anticipated—poor prompt phrasing led to hallucination or verbosity, while tuned prompts reduced error significantly.

Challenges and Limitations

Some gold answers were missing from the corpus entirely, limiting upper-bound performance. Short gold answers were also hard to match within dense chunks. Reranker APIs had latency and token limitations, and some decomposed queries introduced redundancy or drift. Evaluation was based on deterministic outputs, which missed edge cases in sampled generation.

Next Steps

Key directions include:

- Routing agent to assign questions to the best pipeline
- Expanding the corpus with annotated product docs and internal wikis
- Training rerankers with labeled context-answer relevance pairs
- Introducing confidence modeling (entropy, cosine variance)
- Adding human-in-the-loop review for borderline cases

Summary and Recommendations

The project demonstrates that it's possible to tailor RAG systems for divergent user groups by leveraging modular architecture, dynamic decomposition, and audience-aware prompting. With proper weighting of metrics and clear pipeline logs, we delivered consistent, grounded answers in both technical and non-technical domains. Final Fred Scores (0.82 and 0.75) validate this design.

The system should be deployed in dual-pipeline form, with routing logic to classify intent. Additional work should focus on corpus expansion and reranker fine-tuning. Hallucination testing, semantic drift protection, and diagnostic logging must remain core pillars in future iterations.

References

- Izacard, G., & Grave, E. (2021). *Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering*. arXiv:2007.01282
- Maillard, J., Oguz, B., & Bendersky, M. (2021). *Multi-task Retrieval and Reranking for Knowledge-Intensive Tasks*. arXiv:2101.00117
- Cohere Docs. (2023). *Cohere Rerank*
- HuggingFace Datasets & Transformers
- LangGraph and LangChain Documentation
- Meta-LLaMA-3 (2024). *LLaMA Models*

Drafting assistance provided by OpenAI's ChatGPT-4 (April 2025), used for synthesis, diagnostics, and report formatting. All evaluation design, experimentation, and pipeline construction authored by Frederick Hayes III.

Appendix (abbreviated)

Sample Retrieval Logs

- Question 0 (LLM Definition): Retrieved from ArXiv, Wikipedia. Missed most concise gold snippet.
- Question 63 (LoRA Efficiency): Strong recall; answer failed due to lack of specificity in retrieved passages.

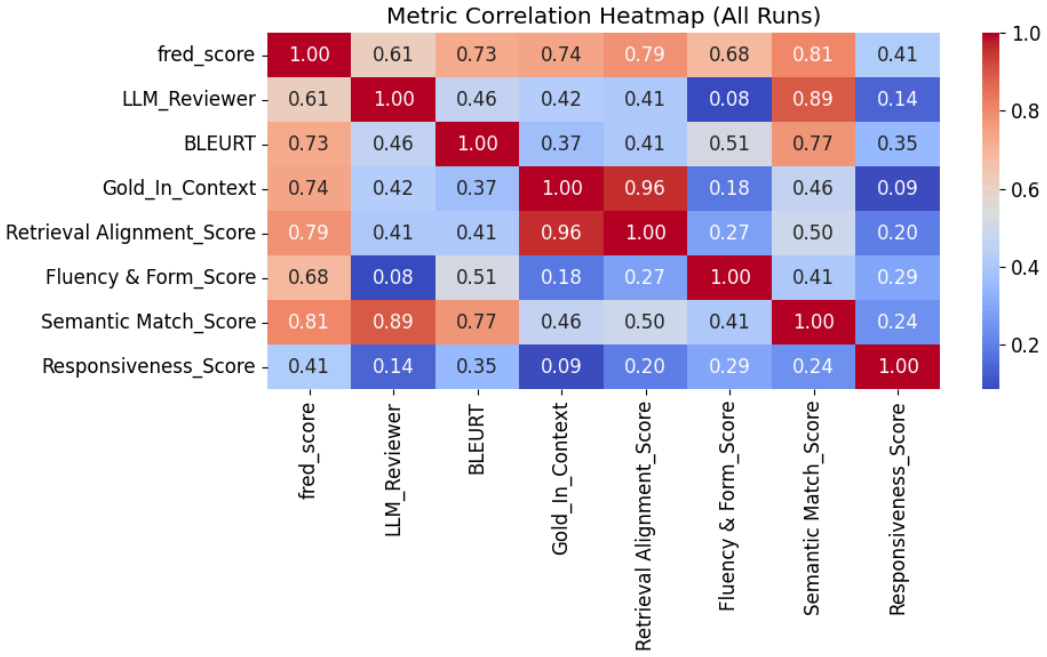
Paper Airplane Test

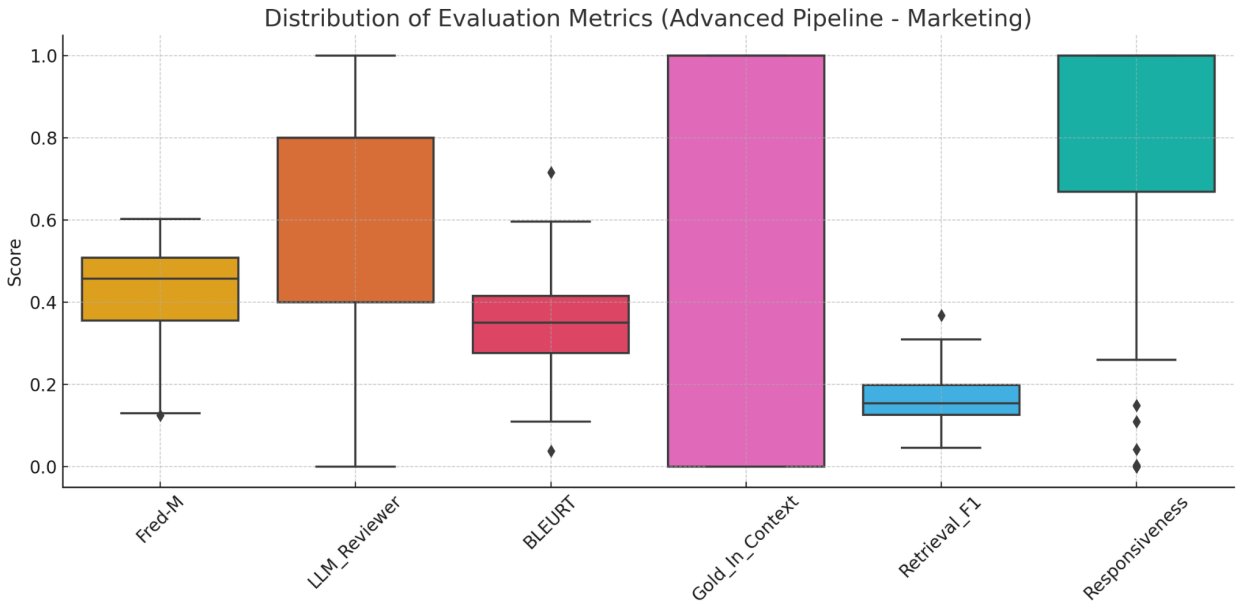
- Hallucinated answer scored highly under BLEU/F1.
- LLM Reviewer and BLEURT correctly rated it near 0.0.
- Prompt tuned to encourage fallback behavior resolved false positives.

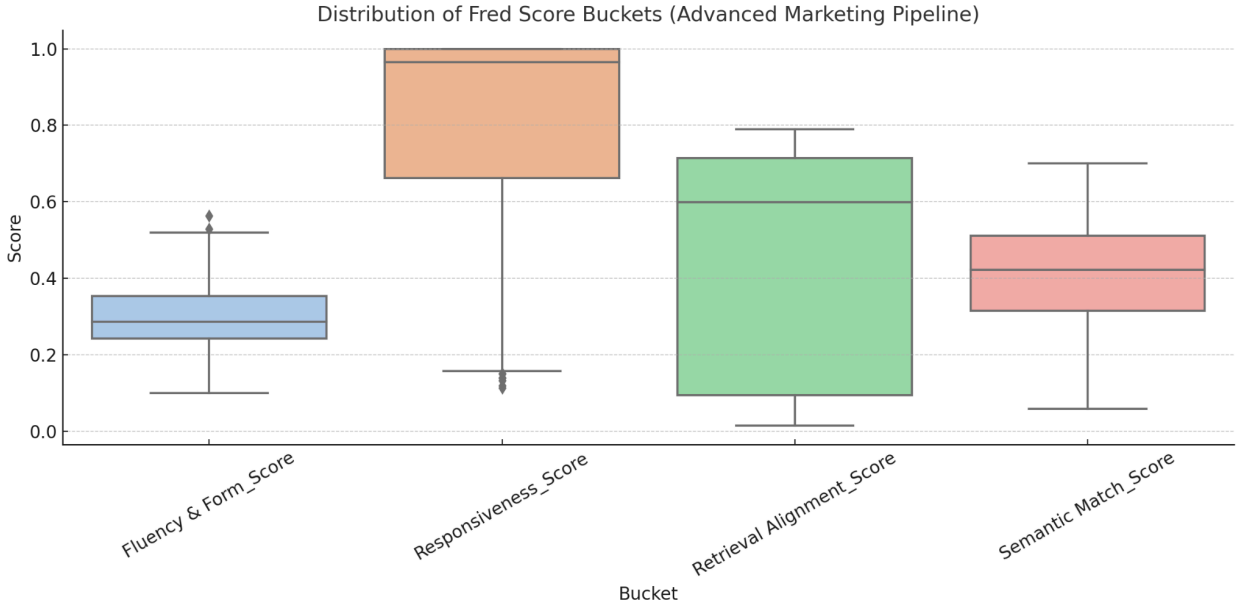
Error Analysis

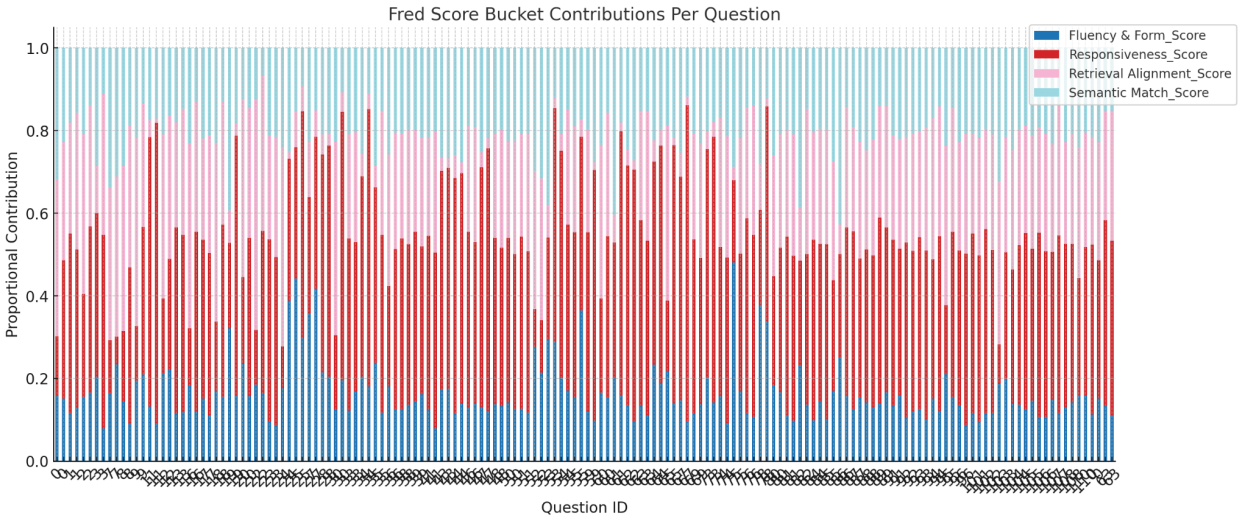
- Over-summarization in marketing answers led to omission of critical phrases.
- Sub-question drift introduced redundancy in decomposed queries.
- Gold answer not found: Up to 20% of questions had no gold overlap in retrieved context despite correct generation.

Appendix




Distribution of Evaluation Metrics (Advanced Pipeline - Marketing)
Score
1.0
0.8
0.6
0.4
0.2
0.0
Fred-M
LLM_Reviewer
BLEURT
Gold_In_Context
Retrieval_F1
Responsiveness


Distribution of Fred Score Buckets (Advanced Marketing Pipeline)
Score
1.0
0.8
0.6
0.4
0.2
0.0
Fluency & Form_Score
Responsiveness_Score
Retrieval Alignment_Score
Semantic Match_Score
Bucket


Fred Score Bucket Contributions Per Question
Fluency & Form_Score
Responsiveness_Score
Retrieval Alignment_Score
Semantic Match_Score
Proportional Contribution
1.0
0.8
0.6
0.4
0.2
0.0
Question ID