MANAGEMENT SUMMARY
Rag-based Specific Paper Summarization

1. PROJECT OVERVIEW

The Challenge : LLMs can generate fluent but unsupported claims, and statistical methods can select the facts but lack narrative flow.
Our Objective : To reach a sweet spot for factual accuracy and content fluency.
Dataset Scope : We worked with 50 computer science research papers from arXiv.org, specifically in the Artificial Intelligence domain. This focus provided consistent document structures (introduction, methods, results, conclusions), enabling systematic evaluation.

2. CHALLENGES FACED

Evaluation Difficulty : Standard automated metrics (ROUGE, BERT scores) sometimes reward poor summaries and penalize good ones, particularly when synonyms or paraphrasing are used. This required extensive manual review and other techniques to validate results.
Hallucination Prevention : Ensuring AI-generated summaries remain faithful to the source material without introducing false claims was a consistent challenge.

3. SOLUTION APPROACH

We implemented and compared different summarization methods across three categories:

Phase 1 : Baselines

Rules-Based Methods:

- **Lead-N**: Simply extracts the first sentences, assuming important information comes early
- **Lead-N by Section**: Extracts early sentences from each section for broader coverage

Statistics-Based Methods:

- **TF-IDF**: Ranks sentences by keyword importance and frequency
- **BM25**: Advanced ranking that identifies technically significant sentences

AI-Based Methods:

- **Hierarchical Summarization**: AI summarizes each section independently, then combines them
- **Embedding-Based**: Uses semantic analysis to select the most representative sections before AI summarization

Phase 2 : Final Approaches

BM25-LLM Hybrid : Combines statistical sentence selection with AI-powered rewriting and fact-checking

Scaffolded RAG: Uses document structure (Introduction, Methods, Results, Conclusion) to guide AI generation with specific prompts for each sections

4. EXTERNAL RESOURCES UTILIZED

Datasets:

- arXiv.org scientific paper repository (cs.AI category)
- 50 papers with full text and author-written abstracts (used as quality benchmarks)
-

AI models:

- Gemini 2.5 Flash (Google)
- Llama 3.1 8B (Meta, via Groq API)
- Gemma 3 1B (lightweight local model)
- GPT-4o (OpenAI) – for evaluation only
- LED (led-large-book-summary) – specialized summarization model

Software Libraries:

- BeautifulSoup (web scraping and HTML parsing)
- NLTK (Natural Language Toolkit for text processing)
- Sentence Transformers (semantic similarity)
- scikit-learn (statistical methods)

5. KEY RESULTS AND PERFORMANCE

Our Scaffolded RAG approach achieved the best performance with 45.8% content accuracy (ROUGE-1), significantly outperforming basic methods (Lead-N: 31.4%) and pure AI approaches (Gemma 3: 27.4%). The BM25-LLM Hybrid showed the best balance between accuracy (39.5%) and consistency. Statistical methods (TF-IDF, BM25) were perfectly factual but produced disconnected text, while pure AI created readable summaries but sometimes missed key technical contributions. Our hybrid approaches satisfactorily balanced readability with accuracy.

6. LIMITATIONS AND NEXT STEPS

Current Limitations:

- Tested only on AI research papers; performance on other domains remains untested
- Reliance on commercial AI APIs creates ongoing costs ($0.02-0.05 per paper)
- Performance varies across paper types; theoretical papers summarize better than empirical studies

Recommended Next Steps:

- Expand testing to 200+ papers and multiple scientific domains
- Develop cost-optimized caching and batch processing
- Create user-friendly web interface for researcher access

7. TEAM CONTRIBUTIONS

Vishwas Bheda: Data ingestion pipeline, embedding-based RAG implementation, complete quantitative analysis, documentation, Experiment design and qualitative reviews

Manuel Velarde: Processing of heterogeneous Humanities dataset, TF-IDF/BM25 baselines, qualitative analysis for all baseline methods, BM25-LLM hybrid refined method & qualitative reviews

Mitzalo Reyes: arXiv scraper & initial parsing, Lead-N & Lead-N by Section baselines, rouge & cosine similarity scoring, collaborative review of results, scaffolded summarization, Question-And-Answer abstract evaluator

Lukas Kobler: Recursive document parsing of structural papers, documentation, Bottom up LLM based hierarchical summarization, Result evaluation & presentation slides