TECHNISCHE UNIVERSITÄT WIEN
Vienna | Austria

Institut für Computertechnik
Institute of Computer Technology

RAG-based Scientific Paper Summarization

From Extractive Baselines to Recursive LLM Chains

Vishwas Bheda, Mitzalo Reyes, Manuel Velarde, Lukas Kobler

TU Wien, Vienna, Austria

January 16, 2026

Project Overview

- **Task:** Automatic summarization of scientific papers.
- **Main Challenge:** LLMs can generate fluent but unsupported claims, and statistical methods can select the facts but lack narrative flow.
- **Dataset:** 50 structured HTML files representing scientific papers in the `cs.AI` category from `arXiv.org`.

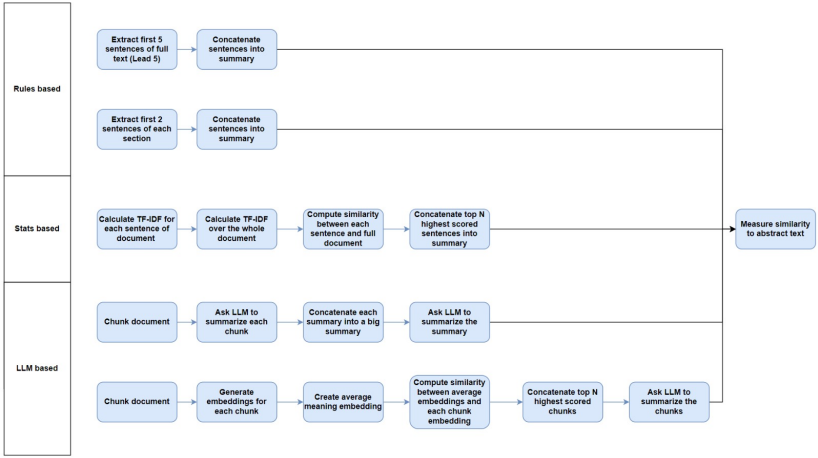
Milestone 1: Data Ingestion

- Researched data sources:
  - Explored heterogeneous dataset of Humanities papers.
  - Explored homogeneous dataset of STEM papers on `arXiv.org`.
- Implemented asynchronous scraping & parsing:
  - Implemented pipeline for concurrent downloads.
  - Transformed raw HTML into nested JSON structures using `beautifulsoup`.
  - Handled PDF edge cases: hyphenated line breaks and CID marker removal.
- Separated abstract as ground truth for each paper.

Milestone 2: Baseline Implementation

- **Extractive (Rule/Stats-based):**
  - *Lead-N / Lead-N-Section:* Positional importance.
  - *TF-IDF / BM25:* Lexical frequency and relevance.
- **Abstractive (LLM-based):**
  - *Recursive "Bottom-Up":* Summarizing the hierarchical section tree.
  - *Embedding-based (RAG):* Semantic chunk selection via Cosine Similarity.

Implementation Diagram



Extractive Baseline 1: Positional Importance

Lead-N and Lead-N-by-Section

- **Lead-N:** Selects the first n sentences of the entire document (where n is the length of the ground-truth abstract).
- **Lead-N-by-Section:**
  - Distributed selection across the document structure.
  - Aims to capture context from Introduction, Methods, and Conclusion simultaneously.
- **Implementation:** Sentences are tokenized using `nltk.sent_tokenize` and compared against the original abstract length to maintain parity.
- **Behavior:** Highly factual and stable, but often fails to summarize, acting more like a "teaser" than a synthesis.

Extractive Baseline 2: Lexical Frequency (Vector Space)

Pre-processing

- Lemmatization via `nltk.stem.WordNetLemmatizer`.
- Stop-word removal and non-alphabetic token filtering.

TF-IDF

- Vectors generated using `scikit-learn`'s method `TfidfVectorizer`.
- Sentences ranked by *cosine similarity* between the sentence vector and the global document vector.

BM25

- Uses `rank_bm25` (Okapi implementation).
- Probabilistic model that penalizes term saturation, often outperforming TF-IDF in pinpointing key technical sentences.

Abstractive Baseline 1: Hierarchical Recursive

Chunk-and-Summarize (Two-pass LLM)

- **Concept:** Process the paper according to its logical structure (JSON tree).
- **Step 1:** Subsections are summarized independently to fit context windows.
- **Step 2:** Parent sections aggregate child summaries to produce a global summary.
- **Observation:** High structural fidelity but computationally expensive.
- Tested with local Gemma:3b

Abstractive Baseline 2: Embedding-based RAG

Semantic Selection for Context Compression

- **Concept:** Use vector embeddings to identify the most "representative" sections.
- **Process:**
  1. Chunk document (1000 chars, 200 overlap)
  2. Embed all chunks + compute mean embedding
  3. **Cosine similarity** between each chunk and mean
  4. Top-5 chunks $\rightarrow$ concatenate $\rightarrow$ LLM summarize
- **Models tested:** Gemini-2.5-Flash, Llama-3.1-8B, and LED.

Milestone 2: Quantitative Performance Analysis

Overall Ranking (Avg. Score)

1. **Embedding-Gemini-2.5** (0.3726)
2. Embedding-Llama-3.1 (0.3651)
3. **TF-IDF (Stats-Based)** (0.3508)
4. Lead-N-By-Section (0.3412)
5. Embedding-LED (0.3388)
6. BM25 (0.3194)
7. Lead-N (0.3137)
8. LLM-Gemma3-1B (0.2735)

RAG Benefits

Embedding-based methods achieve the highest peak scores, but exhibit higher variance and outliers compared to statistical methods.

Semantic insights

High **BERTScore** across all categories suggests models capture core intent even when lexical overlap (**ROUGE**) is low.

Qualitative evaluation in numbers

Category	ROUGE-1	ROUGE-L	BERTScore	Stability
Embedding-Based	**0.3804**	**0.1721**	**0.7893**	Moderate (Outliers)
Stats-Based	0.3398	0.1567	0.7629	**High (Consistent)**
Rules-Based	0.3149	0.1486	0.7749	High
Pure LLM (1B)	0.3278	0.1605	0.6057	Low

- **Statistical Robustness:** TF-IDF outperformed pure LLM (Gemma-1B) and simple Lead-N, proving that lexical importance is a strong signal in scientific text.

Qualitative Review: Extractive Baselines

Lead-N (Standard & By Section)

Pros:

- Perfectly factual (no hallucination).
- High coherence in Lead-N (Introduction context).

Cons:

- Fails to summarize; misses core contributions.
- Lacking flow of the summary.

Statistical (TF-IDF & BM25)

Pros:

- Frequently captures high-relevance keywords.
- Efficiently identifies central themes.

Cons:

- Disjointed: Sentences lack logical transitions.
- Ranking bias: Often selects overly long, complex, or irrelevant sentences.

Qualitative Review: Abstractive Baselines

Two-Pass (Gemma 3)

- **Issue:** Excessive focus on structural details (e.g., Intro/Background).
- **Content:** Tends to over-simplify; misses the "how" and "why" of the research.
- **Tone:** Less verbose than larger models.

Embedding-Based (Gemini/Llama)

- **Gemini 2.5:** Summary seems well organized; good definition of unique concepts.
- **Llama 3.1:** Best human-readable structure; accurately mirrors abstract intent.
- **LED:** Highly verbose; prone to "walls of text" and factual drift (Ironically, this model was finetuned for summarization and this is the one that performs worse).

Refined Baseline: The BM25-LLM Hybrid

- **Concept:** Use statistical ranking to anchor the LLM in "verbatim" facts.

1. Extraction

BM25 Ranking: Selects top 15 sentences with highest lexical similarity to the full paper.

2. Synthesis

LLM Shaping: Rewrites disjointed sentences into a 5-8 sentence cohesive narrative.

3. Verification

Verbatim RAG: Re-queries the document to find specific evidence for each claim.

Final Strategy: Structural Scaffolding

- **Limitation of BM25:** Statistical ranking often over-represents the Introduction and misses key Results.
- A paper structure based summarization pipeline.

Section Parsing Regex-based bucketing into *Intro*, *Methods*, *Results*, *Conclusion*.

Domain Specificity Prompt forces the LLM to extract verbatim quantitative metrics.

Refining step A secondary LLM pass to remove "fluff" and match the style of high-impact journals (e.g., Nature, IEEE).

Comparative Evaluation: Final (old) metrics

Method	ROUGE-1	ROUGE-L	BERTScore
Scaffolded RAG (Best)	**0.4585**	**0.1902**	0.6463
BM25-LLM Hybrid	0.3954	0.2437	0.6320
Embedding-Gemini-2.5	0.3726	0.1710	**0.7893**
TF-IDF (Stats-Based)	0.3508	0.1567	0.7629

- Scaffolded RAG reached **45.8% ROUGE-1**. This suggests that section-based prompts help the LLM keep the author's original phrasing.
- High **BERTScores** show all models understand the topic. However, higher **ROUGE** scores suggest the Scaffolded approach stays more consistent with the source text.

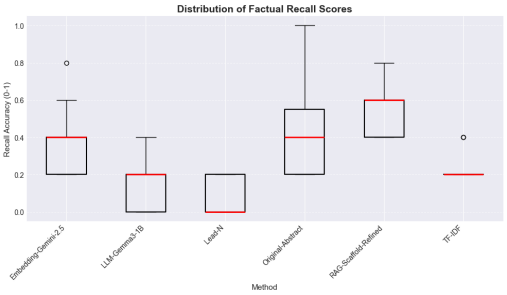
LLM-based Grading

Shortcomings of Traditional Scores

- **ROUGE:** Penalizes good summaries for using synonyms; rewards potentially bad summaries if keywords match.
- **BERTScore:** Capture general meaning well, but are blind to **hallucinations**.

GPT-4o scoring: LLM-as-a-Judge

1. **Problem:** Ground truth abstracts can vary in quality.
2. **Methodology:** LLM produces series of questions to be answered by abstract. Answer quality compared to the ground truth.



Individual Contributions

Vishwas Bheda

1. Developed pipeline, documentation
2. Implemented embedding-based RAG, complete quantitative analysis
3. Experiment design, documentation, and qualitative reviews

Mitzalo Reyes

1. ARXIV scraper & initial parsing
2. Lead-N, Lead-N by Section, rouge & cosine similarity scoring, collaborative review of results
3. Templating & Scaffold based summarization, Question-And-Answer abstract evaluator, collaborative review of results

Manuel Velarde

1. Processing of heterogeneous Humanities dataset
2. TF-IDF & BM25 baselines, qualitative analysis for all baseline methods
3. BM25-LLM hybrid refined method & qualitative reviews

Lukas Kobler

1. Recursive parsing of structural papers, documentation
2. Bottom up LLM based hierarchical summarization
3. Result evaluation & presentation slides


Questions?