Integrating Topic Modeling and Large Language Models for Topic-Level Summarization of Scientific Abstracts

Candidate numbers: 20 and 21

Course code: DSP5200

Course title: Data Analytics with Generative AI and Large Language Models

Submission date: 19.12.2025

Word count: 5783

(Group submission)

Kristiania University

Semester Fall 2025

Contents

ABSTRACT

This study explores how topic modelling can be combined with large language models to structure and summarize collections of scientific abstracts. The problem addressed is the difficulty of extracting coherent topics and producing concise, high-quality summaries from semi-structured academic text. The research questions investigate how effectively BERTopic identifies emerging topics, how well different large language models preserve the main concepts discussed within each topic, and how LLM-based judges assess summary quality. The central concepts include semantic embeddings, density-based clustering, abstract summarization, overlap-based metrics with human references, and LLM-assisted qualitative evaluation. The dataset consists of 100 IEEE Xplore abstracts analyzed using BERTopic with Sentence BERT embeddings, followed by summarization using BART, Llama 3, Mistral, and evaluation using ROUGE metrics and LLM-based judges. The results show that BERTopic produces three coherent topics and that Llama 3 and Mistral consistently outperform BART in both summary similarity and qualitative assessments. The study concludes that integrating topic modeling with modern large language models provides a robust and scalable framework for automated literature analysis and topic-level summarization.

Keywords: Topic Modeling, BERTopic, Large Language Models, Abstractive Summarization, ROUGE Evaluation, LLM as a Judge.

I. INTRODUCTION

Research in natural language processing, and in particular sentiment analysis applied to news and textual data, has expanded rapidly in recent years. New modeling techniques, representation methods, and increasingly capable language models are introduced at a fast pace, continuously updating the research context. As a result, the volume of published work in this area has grown substantially, making it increasingly difficult for researchers to maintain structured overviews of dominant themes, emerging methods, and application trends. While individual abstracts provide concise summaries of single studies, they do not offer systematic insight across groups of related papers. Automated approaches that can organize and summarize research literature are therefore especially relevant for fast-evolving NLP and sentiment analysis domains.

Recent advances in topic modeling and large language models (LLMs) have demonstrated strong performance in text analysis and summarization tasks. However, their combined use for topic-level summarization of academic abstracts remains limited, and evaluation practices are still evolving. This study is motivated by the need to design and examine a framework that not only integrates these techniques effectively but can also scale to larger document collections, where

manual analysis and evaluation become impractical. Understanding how such a pipeline performs under realistic constraints is therefore central to this work.

The core problem addressed in this report is how to transform a collection of research abstracts into a structured representation consisting of meaningful topics and concise topic-level summaries. This involves identifying coherent topic clusters, generating summaries that preserve key information across multiple documents, and evaluating the resulting summaries in a reliable and scalable manner. These challenges are further complicated by the fact that abstracts are already condensed representations of research papers and that summary evaluation is inherently subjective.

To address this problem, the study is guided by the following research questions.

RQ1. How well can topics be identified in the research paper abstracts using BERTopic and sentence-level semantic embeddings?

RQ2. How well do Large Language Models preserve key information from each topic when evaluated using ROUGE similarity against human-written reference summaries?

RQ3. How effectively can a Large Language Model-based judge evaluate the coherence, quality, and factual grounding of the generated summaries, compared with human-written references?

The remainder of this report is organized as follows. Section II presents the conceptual framework, Section III reviews related work, Section IV describes the methodology, Section V reports the results, Section VI discusses the findings, and Section VII concludes the report.

II. CONCEPTUAL FRAMEWORK

The conceptual framework for this study views topic modeling and automated summarization as interconnected components within a pipeline that integrates neural topic modeling with large language models. Figure 1 presents the end-to-end framework developed in this study, translating the conceptual framework into a concrete workflow for topic discovery, LLM-based summarization, and evaluation.

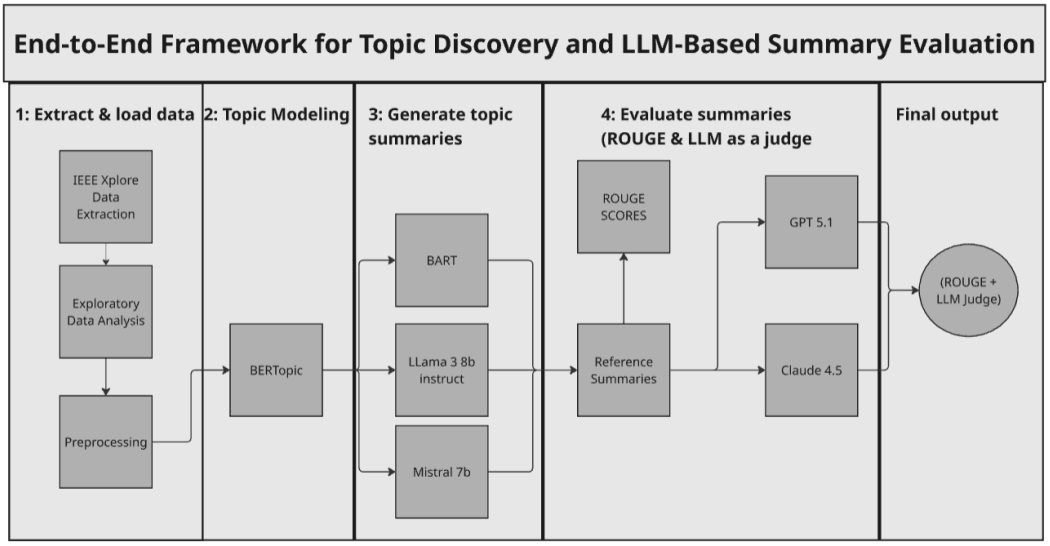


Figure 1: Conceptual framework illustrating the relationship between semantic representation, topic discovery, and generative summarization with combined evaluation.

The framework is adopted from earlier research that integrates semantic embedding, dimensionality reduction, clustering summarization, and LLM-based evaluation to interpret a large collection of scientific text [1]. This study applies the same conceptual architecture to a domain-specific dataset of scientific abstracts to identify topic structure and generate a concise representation. Three core concepts guide the framework: semantic representation, topic discovery, and complementary quantitative and qualitative evaluation.

The first step is extracting and loading the data, followed by step 2, which is the Semantic representation. Transformer-based embedding models like MPNet base v2 encode each abstract into a dense vector space that captures semantic similarity between documents. This representation supports topic clustering and summarization by allowing models to group documents based on meaning instead of shared keywords[2]. BERTopic integrates semantic embeddings with dimensionality reduction through UMAP and density-based clustering through HDBSCAN [3]. These components jointly reveal latent topic groups that reflect the underlying research directions within the dataset. The identified topics function as an intermediate conceptual layer for downstream summarization.

The third step concerns the generation of topic-level summaries. Abstractive summaries are generated using BART, a transformer-based summarization model, and LLMs such as Llama 3 and Mistral 7B. [4], [5], [6]. Step 4 is the evaluation, which is conceptualized as a multi-layered process involving human reference summaries, ROGUE-based similarity measures, and an LLM as a judge [7]. We selected GPT-5.1 and Claude Sonnet 4.5 as LLM judges to align with the

evaluation approach used in LimTopic, while leveraging more recent model versions available at the time of this study [1].

Together, these concepts describe a system in which semantic representation enables topic discoveries, which in turn support generative summarization and quantitative and qualitative evaluation. This framework establishes the conceptual foundation for the methodological approach presented later in the paper.

III. RELATED WORK

Grootendorst introduced BERTopic as a topic modeling approach that integrates transformer-based embeddings, dimensionality reduction, density-based clustering, and a class-based c-TF-IDF procedure to generate interpretable topic representations [3]. In this framework, documents are embedded using pretrained language models, projected into a lower-dimensional space with UMAP, and clustered into topics using HDBSCAN, after which c-TF-IDF is applied to identify the most discriminative terms for each topic. Empirical evaluations in the original BERTopic study show that BERTopic consistently outperforms traditional Latent Dirichlet Allocation (LDA) across multiple benchmark datasets, including 20 Newsgroups, BBC News, and the Trump tweet dataset, achieving higher topic coherence and topic diversity scores. These results demonstrate both the effectiveness and robustness of BERTopic compared to LDA.

LimTopic also compares traditional LDA with BERTopic and reports substantially higher topic coherence for BERTopic, with a coherence score of 0.601 compared to 0.375 for LDA [1]. These findings are consistent with the original BERTopic study, which also demonstrates that BERTopic outperforms LDA in terms of topic coherence, motivating its use in the present study.

They extend this line of work by integrating BERTopic clustering with large language model-based summarization and multi-layer evaluation [1]. In their framework, embeddings and HDBSCAN are used to form topic clusters, after which large language models generate topic-level summaries. To evaluate summary quality, LimTopic combines overlap-based metrics, including ROUGE, BLEU, and BERTScore, with qualitative assessment using LLM-based judges. Specifically, GPT-4 and Claude Sonnet 3.5 are employed as judges to assess dimensions such as coherence, clarity, and factual accuracy. This demonstrates how LLMs can be used both for text generation and for scalable qualitative evaluation. LimTopic also includes Llama 3 as one of the models used for summary generation, motivating its inclusion in the present study. While we adopt the BERTopic clustering approach, ROUGE-based evaluation, and LLM-based judges in a similar manner, it does not include BLEU or BERTScore and relies on BERTopic's keyword-based topic representations rather than LLM-generated topic labels.

Research by Janssens presents a systematic comparison of topic reduction techniques for BERTopic, addressing the tendency of HDBSCAN to produce many small clusters [8]. They distinguish between direct reduction, where the number of topics is controlled during clustering through HDBSCAN parameter settings, and indirect reduction, where topics are first generated and then merged afterward using methods such as agglomerative clustering on topic embeddings or LLM-assisted merging. Their findings show that direct reduction better preserves topic coherence, while indirect methods tend to increase topic diversity at the cost of interpretability. Guided by this analysis, the present study applies a direct reduction strategy by adjusting HDBSCAN parameters to obtain three coherent topics for summarization.

Mistral 7b has further been described in the literature as a compact yet high-performing open-source model suitable for summarization, classification, text completion, and code generation. Its architecture incorporates mechanisms such as Grouped Query Attention and Sliding Window Attention, enabling efficient handling of long sequences with reduced computational cost. Despite its smaller size, Mistral 7B demonstrates competitive performance relative to larger models, making it a practical candidate for abstractive summarization [6]. Therefore, this study includes Mistral 7B as one of the models used for generating summaries.

LLMs have been widely studied for tasks such as summarization, text generation, and reasoning. A recent survey highlights clear performance differences between fine-tuned models and zero-shot approaches [9]. In the context of summarization, the survey reports that BART, a transformer-based model fine-tuned for summarization, achieves strong performance and outperforms zero-shot GPT-style models on text generation tasks. These findings motivate the inclusion of BART alongside Llama 3 and Mistral 7B in this study to enable a comparative evaluation of summarization approaches.

IV. METHODOLOGY

All code and implementation details are provided in the accompanying code submission.

Data Collection

The dataset used in this study consists of research paper abstracts retrieved from the IEEE Xplore Digital Library [9]. The objective was to build a focused yet diverse dataset of publications

examining sentiment analysis applied to news, with a clear connection to artificial intelligence and related computational techniques. To identify relevant studies, we queried IEEE Xplore using the following search expression:

"sentiment analysis" AND news AND (AI OR machine learning OR NLP)

To ensure quality and comparability, we restricted the search to conference papers and journal articles. The publication window was limited to 2021–2025 to capture recent methodological and application-driven developments in the field. For each year, search results were sorted by relevance using IEEE Xplore's default ranking, and the first 20 articles were selected, resulting in an initial dataset of 100 records.

Metadata for all selected articles was downloaded using IEEE Xplore's BibTeX export function. Each record included at a minimum the title, authors, publication year, and abstract since the abstract forms the primary unit of analysis in this study. One BibTeX file was generated per year and stored locally. All files were then merged into a unified dataset using the bibtexparser package [10]. There were zero duplicated abstracts based on BibTeX entry IDs, and the combined collection was converted into a tabular format and exported as a CSV file for subsequent preprocessing and analysis.

Exploratory Data Analysis

Before applying BERTopic, we conducted a light exploration analysis to confirm that the dataset was clean, consistent, and suitable for topic modelling. This included inspecting data types and non-null counts, checking for missing values and duplicate entries, and computing basic statistics such as abstract character length and token counts. We also examined the yearly distribution to verify that the dataset contained twenty articles from each year in the five-year period. In addition, we analyzed unigram, bigram, and trigram frequencies for both abstracts and paper titles to gain an initial understanding of the most frequently occurring terms and phrases in the dataset. Simple visualization, including histograms and boxplots, of abstract length, was used to confirm that the abstracts followed a reasonable distribution without extreme anomalies. The visualizations can be seen in Appendix A.

BERTopic Preprocessing

The preprocessing pipeline followed the structure recommended in the BERTopic framework [3]. Because BERTopic separates document (abstract) embedding from topic representation, we prepared two text forms for each abstract.

First, we created a lightly cleaned version of the abstracts for the Sentence-BERT (SBERT) embedding [11]. This method retained the full semantic content of each abstract and only removed elements such as URLs and excessive whitespace. We used the *all-mpnet-base-v2* model, an MPNet-based transformer fine-tuned within the Sentence-BERT framework, which has been shown to produce high-quality semantic embeddings suitable for clustering [12]. This enabled BERTopic to group documents based on semantic similarity rather than surface-level word overlap.

Second, we generated a normalized version of the abstracts for the topic representation step based on c-TF-IDF. This preprocessing included lowercasing, lemmatization, tokenization, and stopword removal using spaCy [13]. These steps normalize word forms and reduce the influence of high-frequency but uninformative tokens, ensuring that the extracted high frequency reflects meaningful content rather than minor wording differences [14].

Finally, we defined a CountVectorizer to construct the vocabulary used in the c-TF-IDF calculation. The vectorizer was configured to extract unigrams and bigrams and to filter out extremely rare terms. This aligns with the standard BERTopic procedure, where dense SBERT embeddings are used for clustering, and the class-based c-TF-IDF matrix is used to generate interpretable topic representations [3].

BERTopic Configurations

We experimented with multiple BERTopic configurations, including increasing the number of topics above ten, enabling topic reduction, and adjusting HDBSCAN and UMAP parameters. However, these settings produced substantial topic drift, where semantically unrelated abstracts appeared in the same cluster during manual inspection. LimTopic shows that enabling or tuning UMAP and HDBSCAN lowers model quality, decreasing both coherence and silhouette scores, while the default BERTopic setup performs best [1]. Aligning with the results we got when trying to tune these. Therefore, the configuration shown in Table I produced the most coherent topics for our small dataset of 100 abstracts.

ngram_range	1,2	Uses both unigrams and bigrams to improve topic descriptiveness.
min_df	2	Removes words that appear in only one document to reduce noise.
max_df	1	Keeps all terms except extremely rare ones; no upper frequency cutoff.
min_topic_size	5	Ensures topics contain at least five documents, which stabilises clustering in a small corpus.
nr_topics	auto	Allows BERTopic to determine the number of topics based on the embedding structure.
UMAP	Default settings	Used for dimensionality reduction; no manual tuning.
HDBSCAN	Default settings	Used for density-based clustering; parameters left unchanged.

Table I: BERTopic and vectorizer hyperparameter settings used in this study.

BERTopic Modeling

After text preparation and model configuration, BERTopic was applied to discover latent topics in the abstract collection. As shown in Figure 2, BERTopic separates topic discovery and topic representation into two parallel paths, allowing semantic clustering and keyword-based interpretation to be managed independently.

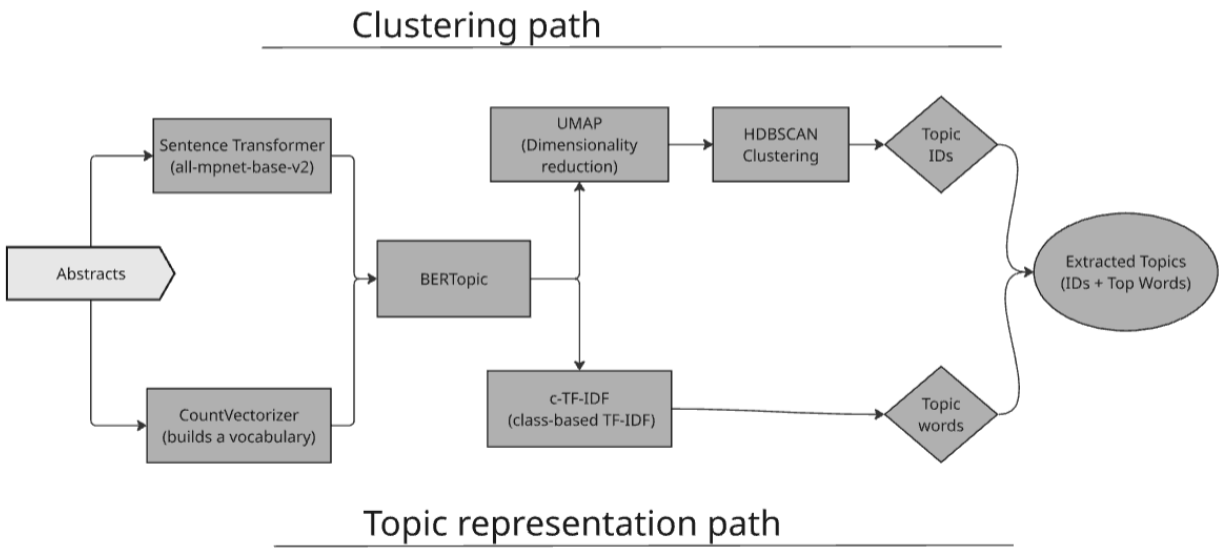


Figure 2: Overview of the BERTopic framework, illustrating the separation between the clustering path used for topic discovery and the topic representation path used for keyword-based interpretation.

Using the configured Sentence-BERT embeddings, abstracts were clustered through dimensionality reduction and density-based clustering to assign topic identifiers. In parallel, the normalized abstracts were processed through the class-based TF-IDF mechanism to derive representative terms for each topic. This design enables topics to be formed based on semantic similarity while remaining interpretable through keyword representations.

Each abstract was assigned to a topic identifier, which was stored alongside the original metadata for downstream analysis. Abstracts belonging to the same topic were subsequently grouped and exported to support topic-level summarization using large language models. The abstracts were then dumped into three txt. file for each topic and later given to the LLMs for summary generation.

Loading Summary Generation Models

All summarization models used in this study were obtained through the Hugging Face Model Hub and loaded using the Transformers library. The BART model (facebook/bart-large-cnn) was accessed directly from Hugging Face and loaded without additional authentication, as it is publicly available [15]. In contrast, Meta Llama 3 8B Instruct and Mistral 7B Instruct are gated models and require authenticated access. Prior to loading these models, we logged into Hugging Face via the command line using an API token, which enabled local download and execution through the Transformer framework.

Llama 3 was loaded using the pipeline interface for text generation, which abstracts model initialization, device placement, and precision handling [16]. Mistral was loaded by explicitly initializing the tokenizer and causal language model using *AutoTokenizer* and *AutoModelForCausalLM*, providing additional control over truncation behavior, numerical precision, and evaluation mode [17]. All models were executed locally after authentication rather than accessed through remote inference APIs, ensuring a consistent and reproducible experimental setup across summarization methods.

BART Summarization

We first applied Facebook BART-large-CNN, a transformer model pre-fine-tuned for abstractive summarization on the CNN/DailyMail dataset, which is widely used as a benchmark for summarization tasks [15]. As a result, BART provides a strong and stable baseline without requiring additional training in this study. Because BART cannot process all abstracts within a topic in a single pass without truncation due to its 1024 token limit, we implemented a multi-step summarization pipeline designed to preserve relevant information. The summarization process

was configured to be deterministic, with sampling disabled and an n-gram repetition constraint applied to reduce redundancy. Despite these settings, minor variations in phrasing may still occur across runs, reflecting the inherent generative nature of neural language models.

Each abstract in the topic was summarized individually in small batches of four using the pipeline interface. Length constraints were set to enforce consistent output across abstracts (*min_length = 35, max_length = 80*), and *no_repeat_ngrams_size = 3* prevented looping or repeated phrases. This step produced one brief summary per abstract. The per abstract summaries were concatenated and then chunked using a token-based algorithm that kept each chunk under a threshold of approximately 800 tokens. This prevents BART from truncating the beginning of the input, which occurred in early tests when long sequences were passed directly. The chunking algorithm encodes each summary to count tokens and splits at safe boundaries, ensuring that semantic content is preserved in full. Each chunk was summarized again with a moderately expanded length limit (*min_length = 70, max_length = 140*). This step compresses multiple per-abstract summaries into a smaller set of coherent mid-level summaries. If only one chunk was summarized, it was used directly as the final topic summary. If only one chunk summary existed, it was used directly as the final topic summary. If several chunk summaries were produced, they were concatenated and summarized one final time using a larger length budget (*min_length = 140, max_length = 260*). This final compression produced a single academic paragraph representing the shared themes and methods within a topic.

This procedure ensured that BART received inputs within its context window at every stage and that no content was silently removed by truncation. The results are highly reproducible in practice and reflect a hierarchical compression approach suitable for multi-document summarization.

Llama 3 Summarization

Meta Llama 3-8B Instruct supports up to 8000 tokens, making it the largest context window and a chat-based interface used in this study [5]. All experiments were run deterministically (*temperature = 0.0*) with a fixed cap on generation length (*max_new_token = 250*) so that outputs are stable and comparable across topics.

We used a shared prompt template that required the model to produce one concise academic paragraph based only on information present in the abstracts. The shared prompt shown in Figure 3 explicitly forbade hallucination, external knowledge, listing abstracts individually, or using headings. A fixed word budget of 180 words was applied to the final summaries to ensure consistent presentation and fair comparison across models.

You are given multiple research paper abstracts that belong to the same topic.

Your task is to write ONE unified academic paragraph that summarises the shared research theme,

common research goals, typical methods explicitly mentioned, and clearly stated findings.

Strict rules:

- Output must be exactly ONE paragraph.
- Do NOT use headings or bullet points.
- Do NOT list abstracts individually.
- Use ONLY information found in the abstracts.
- Do NOT use outside knowledge.
- Do NOT hallucinate methods, datasets, locations or findings not present in the abstracts.
- If information is missing, simply omit it and do not guess.
- Stay concise, factual and grounded.
- Maximum {max_words} words.

Figure 3: Shared Prompt for Mistral and Llama 3

Before summary generation, the combined token length of the prompt and input abstracts was evaluated using the Llama 3 chat template to determine whether single-pass summarization was feasible. If the total length fell below an approximate threshold of 6000 tokens, Llama 3 generated the topic summary in a single pass. When this threshold was exceeded, a hierarchical chunking strategy identical in structure to that used for BART was applied, ensuring that no input content was truncated even for topics containing multiple long abstracts.

Mistral Summarization

Mistral 7B Instruct v0.2 has a more limited context window than Llama 3 and therefore requires explicit handling of long topic files. To avoid unintended truncation, we applied the same token-based control logic used for Llama 3. The total prompt length was computed in advance and compared against a threshold of approximately 3500 tokens. When the input fell below this limit, the full topic was summarized in a single pass. For longer inputs, the topic text was divided into smaller chunks of abstracts, each constrained to approximately 2200 tokens. Each chunk was summarized using the shared academic prompt template shown in Figure 3, with conservative generation limits (max_new_tokens = 192). The intermediate summaries were then concatenated and summarized again using the same prompt template and a larger generation budget (max_new_tokens = 256) to produce the final topic-level summary.

To ensure consistency and comparability across the three summarization models, all models were provided with the same topic-level input files generated by BERTopic. Llama 3 and Mistral were prompted using an identical instruction template to ensure comparable summarization behavior. No external knowledge, chain of thought reasoning, or few-shot examples were used, so the models relied solely on the information contained in the input abstracts. Structural consistency was further maintained by enforcing a single paragraph format for all generated summaries.

Evaluation Metrics

Topic coherence was used to evaluate the quality of the topics produced by BERTopic. Coherence measures how closely related the most important words within a topic are and indicates whether a topic represents a clear and consistent theme. In this study, coherence served as a supporting check to confirm that the discovered topics were suitable for topic-level summarization and further evaluation.

Our evaluation strategy for the generated summaries combines traditional ROUGE scores with an LLM-based judge. This follows the approach used in the LimTopic study, which argues that word similarity on its own cannot capture the actual quality of multi-document summaries. ROUGE provides a quantitative measure of how much content from the human reference summary is preserved, but it does not evaluate coherence, factual grounding, or how well the generated summaries reflect the underlying abstracts. Because of the limitations, LimTopic supplemented ROUGE with LLM-based assessment, and we adopted the same rationale in this study [1].

We created human reference summaries for each topic, and these served as the ground truth for all evaluation steps. These human summaries were used both for the ROUGE comparisons and as the baseline text that the LLM judged compared against when scoring each model's output. We applied ROUGE 1, ROUGE 2, and ROUGE-L to compare each model's summaries against human references. ROUGE 1 measures single-word overlap (unigrams), ROUGE 2 evaluates two-word sequence overlap (bigrams), and ROUGE-L captures the longest matching sequence between the generated and human summaries. Each metric was reported using precision, recall, and F-measure to provide a balanced view and summarize similarity [7].

Since ROUGE only measures how similar they are in terms of shared words and phrases, we incorporated an LLM-based judge to evaluate context quality, which ROUGE cannot capture. The judge used our human summaries as the reference and scored each model's summary based on a set of criteria inspired by the LimTopic evaluation framework. These criteria included grammaticality, readability, cohesiveness, understandability, likability, coherence, relevance, fluency, and description quality. Each category was scored on a five-point scale, where one indicates poor quality and five indicates very good. We collected scores from Chat GPT 5.1 and

Claude Sonnet 4.5 and averaged their output across the three discovered topics. They were accessed externally. Figure 4 illustrates the prompt inspired by LimTopic used by the LLM judges [1].

You are a very helpful and respectful assistant. Please take time to fully read and understand the context.

You will evaluate the quality of the following text based on the reference text.
Rate it on these metrics: Grammaticality, Readability, Cohesiveness, Understandability, Likability, Coherence, Relevance, Fluency, and Description Quality. Each metric should be scored from 1 to 5, where 5 is best and 1 is worst.

Carefully read both the reference text and the summary before rating it.

Reference text:
[REFERENCE HERE]

Text to evaluate:
[SUMMARY HERE]

Figure 4: LLM judge prompt used to score generated summaries against human references.

This combined approach gives a more complete view of summarization quality. ROUGE quantifies summary similarity, while the LLM judges evaluate meaning, quality, and faithfulness. Together, they allow us to assess both the content preservation and the quality of the summaries produced by BART, Llama 3, and Mistral.

V. RESULTS

Topic Clustering

BERTopic identified three coherent topics across the 100. The distribution was relatively balanced, with Topic 0 containing 37, Topic 1 containing 35, and Topic 2 containing 28. The coherence scores support the validity of these clusters, with Topic 0 achieving the highest coherence (*0.578*), while Topic 1 (*0.402*) and Topic 2 (*0.415*) were moderately coherent. The overall coherence score was *0.465*, which is reasonable given the small dataset size and the diverse research topics represented in the dataset. The top words for each topic illustrate three semantically distinct research areas. Topic 0 is characterized by keywords such as *news*, *learning*, *detection*, *machine*, *social*, and *media*. This cluster appears to capture work related to news-

oriented machine learning tasks, including classification, social media analysis, and detection-focused applications.

Topic 1 contains words such as *analysis*, *sentiment*, *data*, *news*, *text*, *classification*, and *machine* learning. These terms point to research-centered methodological sentiment analysis, often focusing on model development, comparative evaluation, or text-based analytical pipelines. Topic 2 includes *market*, *sentiment*, *financial*, *news*, *analysis*, and *learning*, forming a distinct theme around financial sentiment analysis, particularly studies connecting news sentiment to market movement or financial prediction models.

Figure 5 illustrates the top ten words for each topic produced by BERTopic, along with their corresponding c-TF-IDF scores. The bar lengths represent the relative importance of each word within its topic, where higher values indicate terms that are more characteristic of that topic compared to others. These scores are computed using class-based TF-IDF, which aggregates all abstracts assigned to a topic into a single representation and weights words based on both their frequency within the topic and their distinctiveness across topics. The values are relative rather than absolute and should be interpreted within each topic individually. As a result, semantically related terms such as *news* or *sentiment* may appear across multiple topics with different scores, reflecting their varying importance in distinguishing each topic.

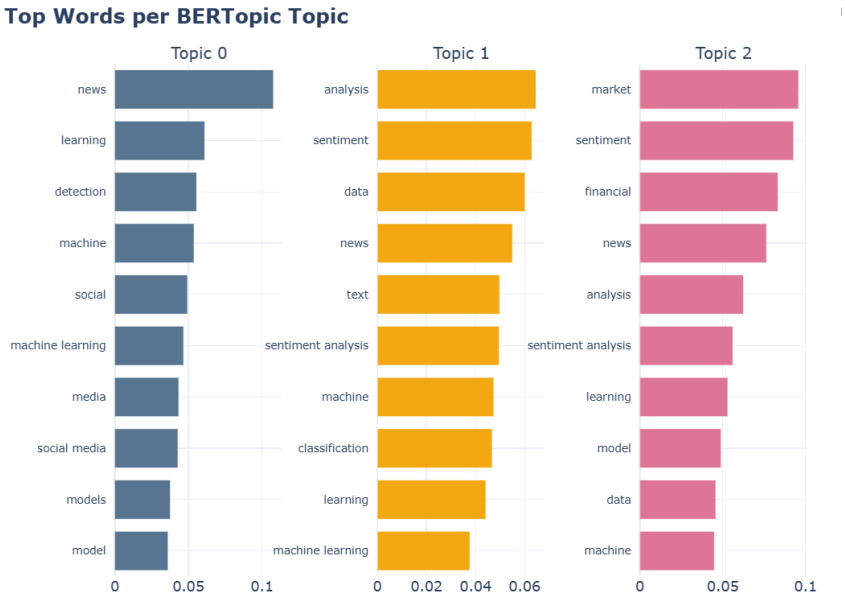


Figure 5: Top ten words for each BERTopic topic, where bar values indicate the relative importance of each word within its topic.

ROUGE

As shown in Table II, Mistral achieves the highest ROUGE 1 and ROUGE 2 scores, whereas Llama 3 attains the highest ROUGE-L score, indicating different patterns of alignment with the human-written reference summaries.

Metrics	BART	Mistral	Llama 3
ROUGE 1	30.67	**41**	38
ROUGE 2	5.67	**10.33**	8.33
ROUGE-L	14.33	18.67	**20**

TABLE II: Average ROUGE F-measure scores (percent) for BART, Mistral, and Llama 3 across all topics.

At the word level, Mistral's ROUGE 1 score of (*41.00*) exceeds that of Llama 3 (*38.00*) and BART (30.67), indicating a higher degree of unigram overlap with the human reference summaries. Given the fixed maximum length of 180 words, this three-point gap between Mistral and Llama 3 suggests that Mistral tends to retain a greater proportion of important terms, rather than replacing them through paraphrasing, than Llama 3. In contrast, BART's substantially lower ROUGE 1 score reflects weaker alignment with the reference at the word level.

A similar pattern is observed at the phrase level. Mistral's ROUGE 2 score of (*10.33*), compared to (*8.33*) for Llama 3 and (*5.67*) for BART, indicates stronger preservation of short word sequences. This difference implies that Mistral not only retains individual key terms but also maintains local phrase continuity more closely aligned with the human references. The relatively low ROUGE 2 scores across all models, however, also reflect the abstractive nature of the task, where paraphrasing and sentence restructuring are expected.

In contrast to the word- and phrase-level results, ROUGE-L highlights a different strength. Llama 3 achieves the highest ROUGE-L score (*20.00*), outperforming Mistral (*18.67*) and BART (*14.33*). Since ROUGE-L measures the longest common subsequence between summaries, this result suggests that Llama 3 more closely mirrors the global structure and ordering of information found in human-written references. This indicates stronger alignment in narrative flow and sentence-level organization, even when exact wording differs.

Taken together, these results point to a clear trade-off between word fidelity and structural alignment. Mistral demonstrates stronger performance on ROUGE 1 and ROUGE 2, indicating higher word- and phrase-level overlap, whereas Llama 3 excels on ROUGE-L, reflecting greater preservation of overall structure. BART performs consistently lower across all ROUGE variants, suggesting limitations in both content retention and structural similarity under the same constraints. Importantly, while these ROUGE scores quantify alignment with human references,

they do not fully capture perceived coherence or readability, motivating the qualitative evaluation discussed in the following section.

LLM as a Judge

The qualitative evaluation results obtained using the LLM-as-a-judge approach are summarized in Table III (GPT-5.1) and Table IV (Claude 4.5 Sonnet). The tables report average scores across nine qualitative dimensions for summaries generated by BART, Mistral, and Llama 3, aggregated across the three topics.

GPT 5.1			
Metrics	**BART**	**Mistral**	**Llama 3**
Grammaticality	3.17	**5**	**5**
Readability	3.33	4.17	**4.33**
Cohesiveness	2.83	**4.33**	4.17
Understandability	3.67	4.17	**4.33**
Likability	3.0	**4**	**4**
Coherence	2.83	**4.67**	4.5
Relevance	3.5	**4.83**	3.83
Fluency	3.17	**5**	**5**
Description quality	3	**4.17**	**4.17**

Table III: Average qualitative scores assigned by GPT-5.1.

GPT-5.1 Evaluation

As shown in Table III, the GPT-5.1 evaluation highlights clear qualitative differences between the models. Mistral receives the highest scores across several key dimensions, including grammaticality (*5.00*), fluency (*5.00*), relevance (*4.83*), and coherence (*4.67*). These values indicate summaries that are perceived as polished in terms of language quality and strongly aligned with the content of the human reference summaries. Llama 3 achieves closely comparable results, with maximum scores for grammaticality (*5.00*) and fluency (*5.00*), along with high ratings for coherence (*4.5*) and readability (*4.33*), suggesting well-structured and easy-to-read summaries.

BART consistently receives lower scores, particularly for cohesiveness (*2.83*) and coherence (*2.83*). These lower values suggest difficulties in maintaining clear structure and logical flow, even when summaries remain concise.

Claude Sonnet 4.5			
Metrics	**BART**	**Mistral**	**Llama 3**
Grammaticality	2.67	4	**5**
Readability	2.33	3.33	**4**
Cohesiveness	1.33	2.67	**3**
Understandability	2.33	3	**4.33**
Likability	1.67	3	**3.33**
Coherence	1.33	3	**4.33**
Relevance	1.67	**3.67**	2.67
Fluency	2.33	2.67	**4.67**
Description quality	2	**3.67**	**3.67**

Table IV: Average qualitative scores assigned by Claude 4.5 Sonnet.

Claude 4.5 Sonnet Evaluation

Claude 4.5 Sonnet (Table IV) reveals a slightly different emphasis. Here, Llama 3 achieves the highest scores for grammaticality (*5.00*), coherence (*4.33*), fluency (*4.67*), and understandability (*4.33*), indicating summaries with strong structural clarity and smooth information flow. Mistral performs strongest on relevance (*3.67*) and shares the highest description quality score (*3.67*) with Llama 3, suggesting effective coverage of key topic information.

Similar to the GPT-5.1 evaluation, BART performs substantially worse across most dimensions, with particularly low scores for cohesiveness (*1.33*) and coherence (*1.33*). These results further reinforce the observation that BART's summaries struggle with overall structure compared to those produced by LLMs.

Taken together, the results reported in Tables III and IV reveal a consistent qualitative pattern. Both evaluators clearly distinguish BART from the two LLM-based approaches, with Mistral and Llama 3 consistently achieving higher scores across key qualitative dimensions. While GPT-5.1 tends to favor Mistral, particularly in terms of relevance and language quality, Claude 4.5 Sonnet assigns higher scores to Llama 3 for coherence and understandability.

Despite these differences, both LLM judges show strong agreement in their overall model rankings, indicating that the observed qualitative patterns are robust across judging models. At the same time, the variation in absolute scores highlights the inherent variability of LLM-based qualitative assessment. As a result, such evaluations should be interpreted as supportive validation rather than definitive measures of summary quality.

VI. DISCUSSION

RQ1. How well can topics be identified in the research paper abstracts using BERTopic and sentence-level semantic embeddings?

Based on the results, RQ1 demonstrates that BERTopic, combined with sentence-level semantic embeddings, was able to identify meaningful and interpretable topics within the research abstracts. The extracted topics aligned well with recognizable research areas, suggesting that this approach is suitable for uncovering thematic structure in academic text. While the dataset was relatively small, the topics were coherent enough to support downstream summarization and evaluation, indicating that the method works effectively.

RQ2: How well do Large Language Models preserve key information from each topic when evaluated using ROUGE similarity against human-written reference summaries?

Regarding RQ2, the results suggest that modern instruction-tuned LLMs are generally better at preserving key information from topic-level abstract collections than a traditional summarization model. However, the observed differences between models indicate that information preservation can take different forms. Some models appear to prioritize retaining key terms and phrases, while others better preserve overall structure and flow. The weaker performance observed for BART may be related to both model characteristics and the experimental setup. In particular, BART is trained primarily for news summarization rather than academic text. It has a more limited context window and requires more aggressive chunking of the input. This likely reduced the contextual information available during summarization. Additionally, because abstracts are already condensed representations of full papers, the task effectively involved summarizing summaries, which may further amplify information loss, especially when combined with chunking. Overall, LLMs are effective in preserving the key information from each topic when evaluated using ROUGE similarity against human-written references.

RQ3. How effectively can a Large Language Model-based judge evaluate the coherence, quality, and factual grounding of the generated summaries, compared with human-written references?

For RQ3, the use of an LLM as a judge provided useful complementary insights into summary quality, but its role should be interpreted with caution. LLM-based evaluation is not yet theoretically grounded as a validated replacement for human judgment. In this study, it was

primarily explored as a scalable evaluation mechanism, motivated by the fact that manual evaluation becomes impractical as the dataset size increases. While the data set here consisted of only 100 abstracts, the approach is intended to scale to a larger collection. Notably, the qualitative patterns produced by the LLM judges broadly aligned with the ROUGE-based results, suggesting that LLM-based evaluation can reflect similar performance trends, even if it cannot serve as an objective ground truth.

Limitation

The human reference summaries were written by a single individual, which introduces subjectivity and may influence both ROUGE scores and the LLMs' judgments. Furthermore, the need to chunk topic-level abstract files due to context limitations affected model performance, as each chunk reduced the amount of cross-abstract context available. Finally, the multistage summarization process compounded compression effects, as the models summarized text that was already highly condensed. These factors may have constrained the models' ability to fully preserve information and should be considered when interpreting the results. Closed LLMs would most likely yield greater performance, but were not used in this study because of the cost.

Learning Reflections

This project provided hands-on experience with core natural language processing techniques, including text preprocessing, sentence-level embeddings, topic modeling, and multi-document summarization. It highlighted the importance of preprocessing and introduced new challenges related to token limits, chunking strategies, and evaluation. In particular, exploring an LLM as a judge was informative for understanding how qualitative evaluation might scale to larger datasets, while also emphasizing the need for cautious interpretation.

Future Work

Future work could apply this framework to larger datasets and to full research papers rather than abstracts, allowing models to operate on richer context. Using larger or API-based language models with extended context windows may reduce the need for aggressive chunking and improve context preservation. Further exploration of LLM-based judges could also help address RQ3 more thoroughly, for example, by studying evaluation consistency, agreement with human judgment, and sensitivity to different prompt designs. In addition, retrieval-based approaches such as retrieval-augmented generation or graph-based methods could be explored to support more structured summarization and evaluation at scale.

VII. CONCLUSION

This study combined topic modeling with large language model-based summarization to analyze and summarize a collection of scientific abstracts. BERTopic successfully identified coherent topics that supported topic-level summarization, and the results showed that modern, prompted large language models outperform a traditional summarization model in this setting. The evaluation further demonstrated that ROUGE metrics and LLM-based judging provide complementary perspectives on summarization performance. Overall, the proposed pipeline offers a practical approach for automated literature analysis, while highlighting the importance of context handling and careful evaluation.

References

[1] I. A. Azher, V. D. R. Seethi, A. P. Akella, and H. Alhoori, 'LimTopic: LLM-based Topic Modeling and Text Summarization for Analyzing Scientific Articles limitations', in *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*, Hong Kong China: ACM, Dec. 2024, pp. 1–12. doi: 10.1145/3677389.3702605.

[2] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, 'MPNet: Masked and Permuted Pre-training for Language Understanding', Nov. 02, 2020, *arXiv*: arXiv:2004.09297. doi: 10.48550/arXiv.2004.09297.

[3] M. Grootendorst, 'BERTopic: Neural topic modeling with a class-based TF-IDF procedure', Mar. 11, 2022, *arXiv*: arXiv:2203.05794. doi: 10.48550/arXiv.2203.05794.

[4] M. Lewis *et al.*, 'BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension', Oct. 29, 2019, *arXiv*: arXiv:1910.13461. doi: 10.48550/arXiv.1910.13461.

[5] A. Grattafiori *et al.*, 'The Llama 3 Herd of Models', Nov. 23, 2024, *arXiv*: arXiv:2407.21783. doi: 10.48550/arXiv.2407.21783.

[6] S. Kukreja, T. Kumar, A. Purohit, A. Dasgupta, and D. Guha, 'A Literature Survey on Open Source Large Language Models', in *Proceedings of the 2024 7th International Conference on Computers in Management and Business*, Singapore Singapore: ACM, Jan. 2024, pp. 133–143. doi: 10.1145/3647782.3647803.

[7] C.-Y. Lin, ‘ROUGE: A Package for Automatic Evaluation of Summaries’, *University of Southern California*, 2004.

[8] W. Janssens, M. Bogaert, and D. Van Den Poel, ‘A Comparative Analysis of Topic Reduction Techniques for BERTopic’, *IEEE Access*, vol. 13, pp. 204087–204103, 2025, doi: 10.1109/ACCESS.2025.3638956.

[9] ‘Xplore Resources and Help’. Accessed: Dec. 15, 2025. [Online]. Available: https://ieeexplore.ieee.org/Xplorehelp/overview-of-ieee-xplore

[10] *bibtexparser: Bibtex parser for python 3*. (Dec. 19, 2024). Accessed: Dec. 15, 2025. [Online]. Available: https://github.com/sciunto-org/python-bibtexparser

[11] N. Reimers and I. Gurevych, ‘Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks’, Aug. 27, 2019, *arXiv*: arXiv:1908.10084. doi: 10.48550/arXiv.1908.10084.

[12] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, ‘MTEB: Massive Text Embedding Benchmark’, Mar. 19, 2023, *arXiv*: arXiv:2210.07316. doi: 10.48550/arXiv.2210.07316.

[13] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, *spaCy: Industrial-strength Natural Language Processing in Python*. (2020). Python. doi: 10.5281/zenodo.1212303.

[14] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed., Draft version. 2025. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[15] Chaumond, ‘README.md · facebook/bart-large-cnn at main’. Accessed: Dec. 17, 2025. [Online]. Available: https://huggingface.co/facebook/bart-large-cnn/blob/main/README.md

[16] ‘NousResearch/Meta-Llama-3-8B-Instruct · Hugging Face’. Accessed: Dec. 17, 2025. [Online]. Available: https://huggingface.co/NousResearch/Meta-Llama-3-8B-Instruct

[17] mistralai, ‘mistralai/Mistral-7B-Instruct-v0.2 · Hugging Face’. Accessed: Dec. 17, 2025. [Online]. Available: https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

Appendix

Appendix A. Eda and Data Characteristics

Figure A.1. Distribution of Abstract Length

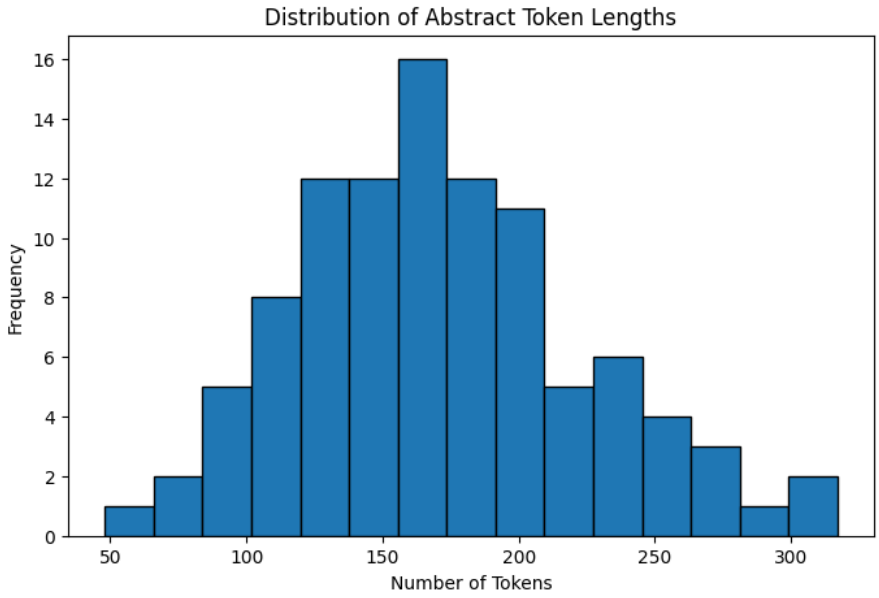


Figure A.2. Boxplot of Abstract Length (Tokens)

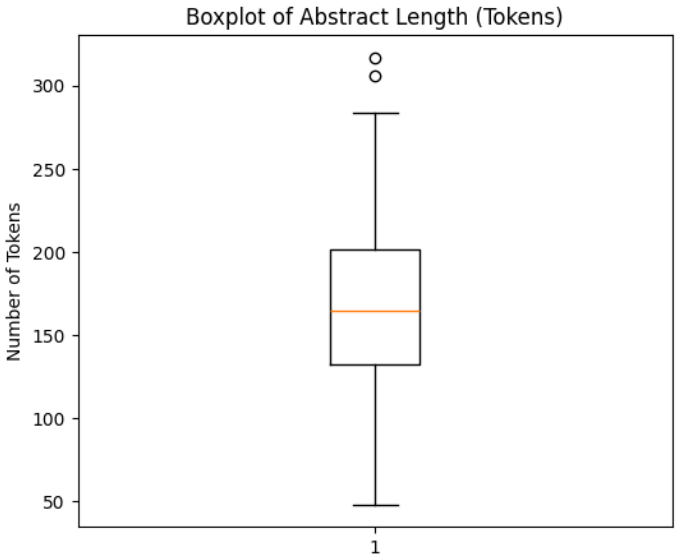


Figure A.3. Number of papers per Year

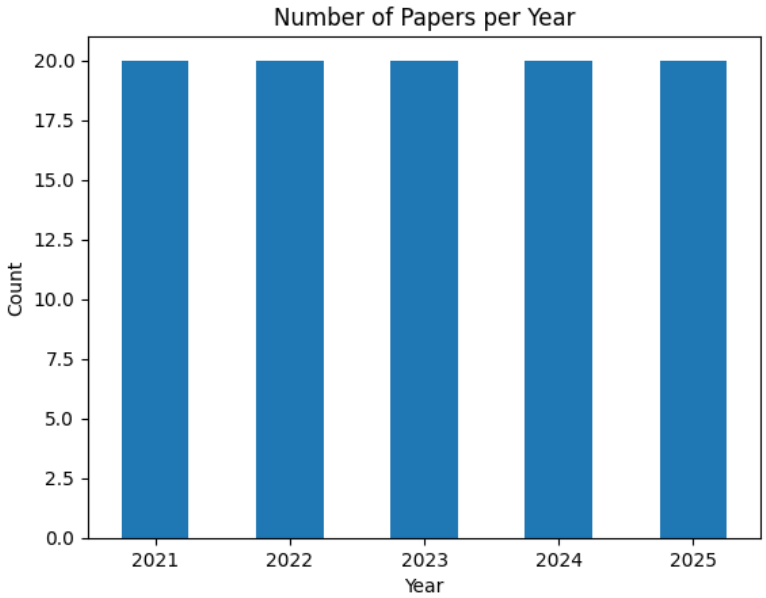
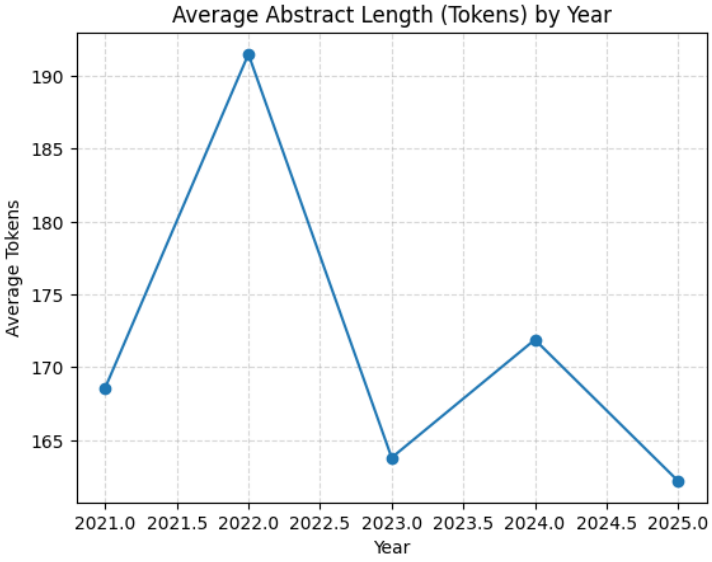


Figure A.4. Average Abstract Length (Tokens) by year



Appendix B. Coherence and ROUGE Visualization

Figure B.1. Topic Coherence per Topic

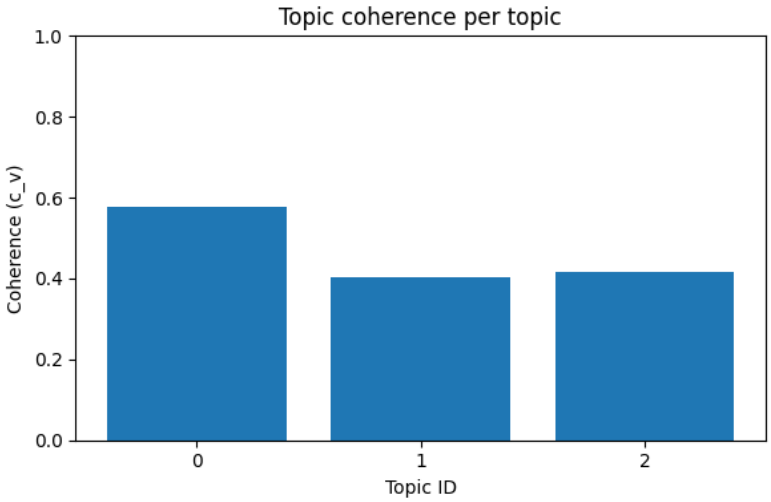


Figure B.2. Number of Abstracts per Topic

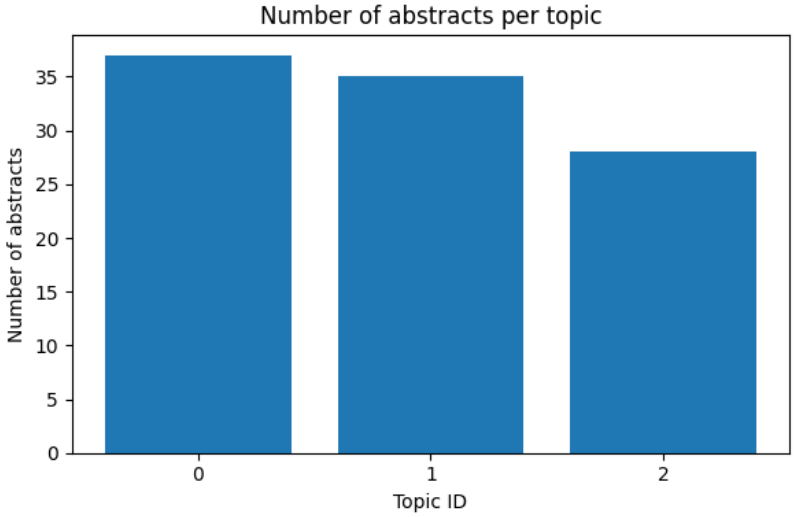


Figure B.3. Average ROUGE F measure per model in Percent