Topic Modeling and Semantic Similarity-Based Evaluation of Biomedical Text Abstracts Generated by LLM

Araştırma Makalesi/Research Article

Burcu BAŞTÜRK[1], Aytuğ ONAN[2*]

[1]Bilgisayar Mühendisliği Bölümü, İzmir Katip Çelebi Üniversitesi, İzmir, Türkiye
[2]Bilgisayar Mühendisliği Bölümü, İzmir Yüksek Teknoloji Enstitüsü, İzmir, Türkiye

burcubasturk58@gmail.com, aytugonan@iyte.edu.tr
(Geliş/Received:21.12.2025; Kabul/Accepted:19.06.2026)
DOI: 10.17671/gazibtd.1846644

***Abstract*—** This study investigates how large language models (LLMs)—ChatGPT, Gemini, and DeepSeek—preserve thematic consistency in biomedical text summarization. A dataset of 56,000 cardiovascular texts was constructed, including 8,000 original PubMed abstracts and 48,000 LLM-generated summaries. After domain-specific preprocessing, topic modeling was performed using Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and Non-negative Matrix Factorization (NMF) across topic sizes (k = 10–50). Thematic consistency was evaluated using C_v coherence scores. In addition, semantic similarity between original and generated texts was assessed using SBERT-based cosine similarity to capture meaning preservation beyond lexical overlap. The results indicate that abstractive summaries, particularly those generated by Gemini and ChatGPT, achieve coherence levels comparable to or higher than original abstracts under LDA and LSA, while extractive summaries exhibit greater variability. For NMF, original abstracts show more stable trends, whereas ChatGPT's extractive summaries remain competitive at higher topic sizes. Overall, coherence tends to decrease with increasing topic numbers in LDA, whereas LSA and NMF benefit from larger topic sizes. These findings highlight the importance of combining coherence and semantic similarity for a more comprehensive evaluation of LLM-based biomedical summarization.

***Keywords*—** large language models (LLMs), biomedical text summarization, topic modeling, semantic similarity, SBERT, coherence analysis, PubMed abstracts

LLM Tarafından Oluşturulan Biyomedikal Metin Özetlerinin Konu Modelleme ve Semantik Benzerlik Tabanlı Değerlendirilmesi

***Özet*—** Bu çalışma, büyük dil modellerinin (LLM'ler) -ChatGPT, Gemini ve DeepSeek- biyomedikal metin özetlemesinde tematik tutarlılığı nasıl koruduğunu araştırmaktadır. 8.000 orijinal PubMed özeti ve 48.000 LLM tarafından oluşturulan özeti içeren 56.000 kardiyovasküler metinden oluşan bir veri seti oluşturulmuştur. Alana özgü ön işleme sonrasında, konu boyutları (k = 10–50) boyunca Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA) ve Non-negative Matrix Factorization (NMF) kullanılarak konu modellemesi gerçekleştirilmiştir. Tematik tutarlılık, C_v tutarlılık puanları kullanılarak değerlendirilmiştir. Ek olarak, sözcüksel örtüşmenin ötesinde anlam korumasını yakalamak için orijinal ve oluşturulan metinler arasındaki semantik benzerlik, SBERT tabanlı kosinüs benzerliği kullanılarak değerlendirilmiştir. Sonuçlar, özellikle Gemini ve ChatGPT tarafından oluşturulan soyut özetlerin, LDA ve LSA altında orijinal özetlere kıyasla benzer veya daha yüksek tutarlılık seviyelerine ulaştığını, çıkarımsal özetlerin ise daha fazla değişkenlik gösterdiğini ortaya koymaktadır. NMF için, orijinal özetler daha istikrarlı eğilimler gösterirken, ChatGPT'nin çıkarımsal özetleri daha yüksek konu boyutlarında rekabetçi kalmaktadır. Genel olarak, LDA'da tutarlılık artan konu sayılarıyla azalma eğilimindeyken, LSA ve NMF daha büyük konu boyutlarından fayda sağlamaktadır. Bu bulgular, LLM tabanlı biyomedikal özetlemenin daha kapsamlı bir değerlendirmesi için tutarlılık ve anlamsal benzerliğin birleştirilmesinin önemini vurgulamaktadır.

***Anahtar Kelimeler*—** büyük dil modelleri (LLM'ler), biyomedikal metin özetleme, konu modelleme, anlamsal benzerlik, SBERT, tutarlılık analizi, PubMed özetleri

1. INTRODUCTION

Today, the volume of biomedical research is rapidly increasing, reaching millions of publications in databases such as PubMed. Within this flood of information, it has become increasingly difficult for researchers to identify, access, and interpret relevant literature. In this context, summarization systems and natural language processing (NLP) techniques play a crucial role in enabling researchers to extract meaningful and structured information from scientific publications—saving time, facilitating hypothesis generation, and fostering interdisciplinary connections. Advances in biomedical language processing have radically transformed how scientists interact with the literature, and summarization now stands out as a strategic component at the core of this transformation [1, 2].

Moreover, since clinical data and patient information are often recorded in unstructured free text, traditional data processing methods face major limitations. NLP-based summarization can extract clinically relevant information from such text, supporting clinical decision-making systems and improving the efficiency and accuracy of healthcare delivery. Therefore, summarizing biomedical texts not only simplifies doctors' work but also helps improve patient care [3].

Recent advances in large language models (LLMs) such as ChatGPT, Gemini, and DeepSeek have revolutionized NLP by enabling machines to generate fluent, coherent, and contextually aware text across domains [4]–[6]. In the biomedical field, these models have shown promising performance in scientific text summarization, helping researchers and clinicians efficiently interpret large volumes of literature [7, 8]. LLMs are increasingly capable of capturing domain-specific terminology and reasoning patterns, as demonstrated in both clinical and academic benchmarks [9].Consequently, LLM-based summarization systems are emerging as powerful tools for knowledge extraction, hypothesis generation, and automated synthesis of scientific evidence.

Text summarization approaches are broadly categorized into extractive and abstractive methods [10]. Extractive summarization identifies the most informative sentences within the source text and concatenates them verbatim, which tends to preserve factual accuracy but may lead to redundancy or reduced fluency, particularly in biomedical documents. In contrast, abstractive summarization generates novel sentences that paraphrase the underlying meaning of the source material, enabling more natural and coherent summaries but also introducing risks such as semantic drift and factual inconsistencies [11].

Ensuring topic coherence is fundamental to maintaining the factual reliability of machine-generated summaries, particularly in the biomedical domain where even subtle deviations in thematic structure may result in misleading or clinically unsafe interpretations [12, 13]. LLMs, while highly fluent, are known to occasionally introduce semantic drift or hallucinated content when the underlying topic structure is not preserved. This makes coherence-based evaluation essential for assessing the trustworthiness of AI-generated scientific summaries.

However, topic coherence alone cannot capture semantic equivalence between the original and generated texts. Therefore, recent studies increasingly emphasize embedding-based similarity measures, such as Sentence-BERT (SBERT), which enable the evaluation of semantic preservation beyond lexical overlap.

Motivated by this need, the present study aims to evaluate how effectively abstractive and extractive summaries generated by advanced LLMs -ChatGPT, Gemini, and DeepSeek- preserve the thematic consistency of original PubMed abstracts. Using topic modeling techniques (LDA, LSA, NMF) and coherence metrics (C_v), systematically compare models across multiple topic coherences. The results provide insights into the strengths and limitations of current LLMs in producing domain-faithful, reliable biomedical summaries, and highlight the importance of coherence-aware evaluation for future AI-driven scientific workflows.

The main contributions of this study can be summarized as follows:

(1) A large-scale and systematically constructed dataset consisting of 56,000 biomedical texts, including both original PubMed abstracts and LLM-generated summaries.

(2) A comprehensive comparative evaluation of three widely used topic modeling techniques (LDA, LSA, and NMF) under multiple topic configurations (k = 10–50).

(3) The development of a multi-dimensional evaluation framework that integrates topic coherence with semantic similarity (SBERT-based cosine similarity) to assess the quality and meaning preservation of LLM-generated summaries.

(4) A systematic comparison of multiple LLMs (ChatGPT, Gemini, DeepSeek) and summarization styles (abstractive vs. extractive), providing insights into their impact on thematic consistency in biomedical texts.

Unlike studies that focus solely on model development, this work aims to provide a systematic and scalable evaluation framework to better understand the strengths and limitations of current LLM-based summarization systems.

Although topic coherence provides valuable insights into the thematic structure of texts, it does not directly measure semantic correctness, factual accuracy, or faithfulness to the original content. Therefore, in this study, coherence is treated as a complementary metric rather than a standalone indicator of summary quality.

2. LITERATURE REVIEW

Biomedical literature continues to grow at an unprecedented pace, making it increasingly difficult for clinicians and researchers to efficiently access and interpret relevant information. Text summarization has therefore become an essential tool for reducing cognitive load by condensing lengthy scientific documents while preserving their core findings. Early biomedical summarization systems relied on extractive selection or domain-specific neural models, but these approaches often struggled with long-range dependencies and semantic abstraction. With the emergence of LLMs, the ability to generate fluent and human-like summaries has dramatically improved, yet concerns remain regarding factual accuracy and thematic fidelity in high-stakes biomedical contexts. These limitations highlight the need for complementary evaluation methods that assess not only linguistic quality but also the preservation of the underlying topics and semantic structure of the original texts.

Recent advances in LLMs have significantly influenced biomedical text summarization, enabling more fluent and human-like outputs compared to earlier neural approaches. Early systematic reviews, such as Wang et al. , showed that most biomedical summarization systems were extractive and targeted research literature rather than clinical data, with evaluations relying heavily on surface-level metrics like ROUGE. With the emergence of models such as GPT-4, BioGPT and domain-adapted transformer architectures, researchers have begun exploring abstractive summarization for both literature and electronic health records [14].

Chaves et al. and Huang et al. reviewed biomedical text summarization methods with a particular focus on neural and pre-trained models, highlighting a shift from feature-engineering and extractive pipelines to encoder–decoder architectures and transformer-based models. These surveys emphasize that pre-trained models significantly improve fluency and readability, but also raise concerns about domain adaptation and factual consistency in high-stakes biomedical applications [15, 16].

With the emergence of general-purpose LLMs such as GPT-4 and related foundation models, a growing body of work evaluates their use for clinical and biomedical summarization. Van Veen et al. adapted LLMs to clinical text summarization and reported that LLM-generated summaries can match or even surpass expert-written summaries in readability, but they remain vulnerable to factual errors and hallucinations [17]. Other recent studies benchmark LLMs on multi-document EHR summarization and clinical note summarization consistently finding that while LLMs substantially reduce cognitive load and produce linguistically high-quality summaries, ensuring factual accuracy and reliable evaluation remains a major challenge [18] - [20].

LLM-based abstractive summaries can be highly readable yet may omit crucial clinical details, introduce incorrect relations, or over-generalize findings. In contrast, extractive summaries typically maintain better content fidelity, because they remain anchored to source spans. As a result, several recent studies advocate hybrid strategies—for example, using extractive steps to identify key evidence and then applying controlled abstractive generation—to balance fluency and faithfulness in medical settings [19].

Topic modeling remains a fundamental tool for analyzing large biomedical corpora, especially as a complementary method to evaluate the semantic fidelity of LLM-generated summaries. Although classical models such as LDA, LSA, and NMF were introduced earlier, they continue to be widely applied and extended in recent biomedical and NLP research.

Latent Dirichlet Allocation (LDA) remains the most widely used probabilistic approach because it models word–topic and topic–document distributions, offering stable performance on sparse datasets such as PubMed abstracts [21]. In contrast, Latent Semantic Analysis (LSA) relies on singular value decomposition to capture global semantic structure; while computationally efficient, it often performs inconsistently on highly specialized biomedical terminology due to its linear assumptions [22].

Non-negative Matrix Factorization (NMF) provides another effective alternative by decomposing term–document matrices into additive components that lead to more interpretable topic-word groupings, which has been shown useful in biomedical text mining tasks [23].

Topic coherence metrics are widely used to quantify how semantically meaningful and interpretable the discovered topics are. Among these metrics, C_v coherence has become one of the most commonly adopted measures due to its ability to combine multiple sources of semantic information into a single score. The C_v measure is based on a sliding window, uses a one-segmented probability distribution, and leverages normalized pointwise mutual information (NPMI) together with cosine similarity to evaluate the degree of semantic relatedness among the top words within each topic [24]. This hybrid design allows C_v to capture both global co-occurrence patterns and local contextual dependencies, which makes it particularly suitable for scientific and biomedical text collections, where terminology is highly domain-specific.

For this reason, C_v is frequently used in research that evaluates topic structure across LLM-generated and human-written texts. In the context of the present study—where topic models are applied to PubMed abstracts and LLM-generated abstracts—the C_v metric provides a robust way to assess whether LLM outputs preserve the underlying thematic consistency of the original biomedical documents.

In addition to topic coherence, semantic similarity metrics have gained increasing attention for evaluating the faithfulness of generated summaries. Embedding-based approaches, particularly Sentence-BERT (SBERT), enable the comparison of sentence-level semantic representations using cosine similarity. Unlike traditional overlap-based metrics such as ROUGE, SBERT captures contextual meaning and semantic equivalence, making it especially suitable for biomedical text evaluation where paraphrased expressions are common. Recent studies emphasize that combining topic-level coherence with embedding-based similarity measures provides a more comprehensive evaluation of summary quality [33].

3. METHODOLOGY

This study investigates the extent to which large language models (LLMs)—ChatGPT, Gemini, and DeepSeek—preserve the thematic structure of biomedical abstracts when generating both abstractive and extractive summaries. To enable a systematic and reproducible comparison, a structured methodological framework is developed, incorporating topic modeling and semantic similarity–based evaluation to analyze variations across models and summarization strategies. The overall workflow of the proposed methodology is illustrated in Figure 1.

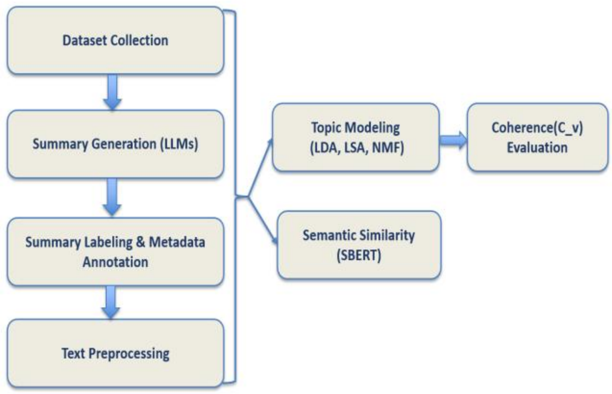

Figure 1. Flow Chart for Methodology

3.1. Dataset Collection

PubMed is a freely available bibliographic platform managed by the U.S. National Library of Medicine (NLM), offering extensive coverage of research across biomedical sciences, clinical medicine, nursing, dentistry, veterinary fields, public health, and life sciences. Owing to its comprehensive and authoritative content, it is one of the most widely utilized resources for obtaining reliable scientific and medical literature in academic research [25].

To construct the dataset, 20,000 biomedical abstracts were retrieved from PubMed using the NCBI Entrez Programming Utilities (E-utilities) through the Bio.Entrez module in Biopython. The query “heart disease[Title/Abstract]” was used to ensure cardiovascular relevance. Data was downloaded in batches of 500 records, and detailed article information was obtained via efetch in MEDLINE format, from which only the abstract field was extracted. A one-second delay between requests was implemented to comply with NCBI rate limits.

After collection, a simple filtering process was applied to ensure content quality: abstracts between 100 and 6,000 characters were retained, removing extremely short or excessively long entries. The resulting dataset contained clean, domain-specific biomedical abstracts suitable for summarization and subsequent topic coherence analysis.

3.2. Summary Generation

To generate model-based summaries, three widely used LLMs—ChatGPT (OpenAI), Gemini (Google DeepMind), and DeepSeek (DeepSeek AI)—were employed. The illustration is shown in Figure 2.

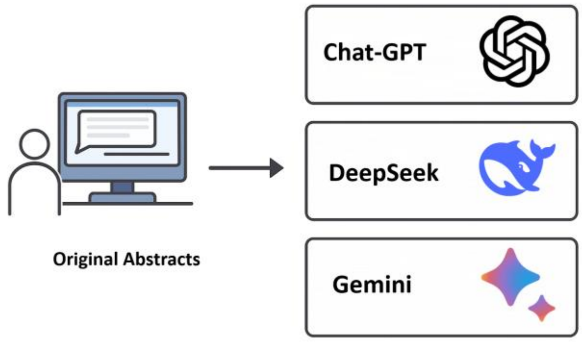

Figure 2. Illustration of Abstractive and Extractive Summarization of Original Abstracts Using ChatGPT, DeepSeek and Gemini.

ChatGPT is a transformer-based generative model capable of producing coherent, context-aware text and is frequently used for tasks such as question answering, content summarization, and instruction following [26]. Gemini, developed by Google DeepMind, is a multimodal architecture capable of processing text, images, audio, and code, offering high accuracy and advanced reasoning capabilities. DeepSeek focuses on efficient information retrieval and structured generation, emphasizing factual consistency and domain reliability, making it particularly suitable for scientific and technical contexts [27].

All summaries were generated between November 25, 2024 and June 1, 2025 using the publicly available web-based chat interfaces of ChatGPT (OpenAI), Gemini (Google), and DeepSeek. During this period, the default production models deployed on each respective platform were used. As generation was conducted via the official web interfaces rather than API-based access, specific API version identifiers were not directly accessible to the user. The generation time frame is reported to account for potential model updates and version changes.

Two distinct prompting strategies were employed. For extractive summarization, models were instructed with the prompt: “Extract the most informative sentences from the following abstract.” No explicit constraints on sentence count or compression ratio were applied. For abstractive summarization, the prompt “Summarize the following abstract.” was used, allowing free reformulation without length restriction. All summaries were generated using the official web-based chat interfaces of the respective models under default system configurations, and no manual post-editing was performed.

For extractive summarization, the models were instructed to select the most informative sentences directly from the original abstract without paraphrasing. No strict constraint was imposed on the number of sentences; however, the models were encouraged to produce concise summaries comparable in length to the original abstract. All generated outputs were manually collected and stored in CSV format for further analysis. This prompting strategy specifications are shown in Table 1.

Table 1. Prompting Strategy Specifications

Strategy	Prompt Text	Constraints	Generation Settings
Extractive	“Extract the most informative sentences from the following abstract.”	No explicit sentence or compression limit. No rephrasing instruction provided.	Generated via official web-based interface using default system parameters.
Abstractive	“Summarize the following abstract.”	No explicit length or sentence constraint. Free reformulation allowed.	Generated via official web-based interface using default system parameters.

This procedure resulted in a total of 48,000 summaries (3 models × 2 summary types × 8,000 abstracts).

3.3. Summary Labeling and Metadata Annotation

After generating abstractive and extractive summaries from each LLM, a structured metadata annotation procedure was applied to organize all outputs into a unified and analysis-ready dataset. Each summary file was imported into Python using pandas, and the original summary column was standardized by renaming it to a common field, summary_text. To ensure consistency across all summary files, a standardized metadata schema with four columns —id, summary_text, model, and style— was created.

The model column identifies the source system (ChatGPT, Gemini, DeepSeek, or PubMed), while the summary_text field contains the summary produced by each model. The style field encodes whether the summary is abstractive, extractive, or an original PubMed abstract. For example, ChatGPT summaries were labeled with model="chatgpt" and style="abstractive" or style="extractive", and the same schema was applied to Gemini and DeepSeek.

All annotated files were exported as standardized CSVs and combined, ensuring full traceability and compatibility across the dataset. This metadata structure enabled systematic grouping, filtering, and comparison of summaries, allowing topic modeling and coherence evaluation to be performed reliably for each model–style combination.

3.4. Text Preprocessing

Text preprocessing is a fundamental step in NLP pipelines and aims to transform raw text into a clean, structured, and machine-interpretable format. It typically includes normalization, removal of noise such as URLs and punctuation, stopword filtering, and lemmatization to reduce lexical variability. These operations improve the quality of downstream tasks by ensuring that topic modeling and semantic analyses operate on linguistically consistent input. In biomedical text mining, domain-aware preprocessing is especially important, as specialized terminology, acronyms, and structured sections must be handled carefully to preserve semantic content [28].

After constructing the unified dataset of 56,000 summaries, a preprocessing pipeline was applied to prepare the text for topic modeling and coherence analysis. The combined dataset was merged into a single CSV file and processed using Python. These operations were implemented using standard NLP libraries, enabling clean and comparable textual inputs for downstream modeling.

For linguistic processing, a spaCy-based NLP model was used. The pipeline first attempted to load the biomedical model en_core_sci_sm (scispaCy); if unavailable, it fell back to en_core_web_sm, with named entity recognition, parsing, and text classification disabled to keep the pipeline lightweight. A customized stopword list was then constructed by combining the standard NLTK English stopwords with additional “structural” biomedical terms. To avoid losing important semantic information, common negation words (no, not, without, never, etc.) and domain-specific cardiovascular terms (e.g., cardiovascular, myocardial, atrial, hypertension, infarction, thrombosis) were explicitly removed from the stopword list.

Before token-level processing, each summary was normalized by replacing non-breaking spaces, standardizing dash characters, and removing URLs and email addresses via regular expressions. The text was then passed through the spaCy pipeline, and each token was processed according to the following rules:

- Spaces, punctuation, pure numbers, and symbol-only tokens were discarded.
- Tokens were lowercased and removed if they appeared in the stopword list.
- For proper nouns and all-uppercase tokens, the lowercased surface form was preserved instead of the lemmatized form.

- For all other tokens, the spaCy lemma was taken, stripped of leading/trailing punctuation, and discarded if empty or non-informative.

This standardized, cleaned corpus formed the basis for all subsequent topic modeling and coherence evaluations.

3.5. Topic Modeling

Topic modeling was applied to uncover latent thematic structures in the original PubMed abstracts and the summaries generated by ChatGPT, Gemini, and DeepSeek. Three commonly used topic modeling approaches—Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and Non-negative Matrix Factorization (NMF)—were using the cleaned corpus to examine how well each summary type preserved the thematic distribution of the source abstracts. Models were evaluated across five topic numbers (k = 10, 20, 30, 40, 50) to assess the stability of thematic representation at different coherences.

3.5.1. Latent Dirichlet Allocation (LDA)

LDA is a probabilistic topic modeling method that represents each document as a mixture of latent topics and each topic as a distribution over words. LDA estimates these hidden structures by assuming a generative process in which documents arise from underlying thematic proportions, making it one of the most widely used approaches for uncovering semantic patterns in large text corpora [29].

LDA was used to estimate topic coherence across the original PubMed abstracts and the summaries generated by LLMs. The analysis was performed on the preprocessed dataset. The data were grouped by model (PubMed, ChatGPT, Gemini, DeepSeek) and summary style (original, abstractive, extractive), and a separate LDA modeling pipeline was applied to each (style, model) subset.

For each group, the cleaned texts were split into token lists and used to build a Gensim Dictionary and bag-of-words corpus. To reduce noise, extremely rare and overly frequent terms were removed using filter_extremes. LDA models were then trained with k = 10, 20, 30, 40, 50 topics using Gensim's LdaModel (passes = 8, iterations = 200, random_state = 42, eval_every = None).

For each trained model, topic quality was quantified using the C_v coherence metric, computed via Gensim's CoherenceModel with the learned topics, the tokenized texts, and the corresponding dictionary. The resulting coherence scores were stored together with their (model, style, k) configuration and summarized in a table. Finally, the scores were visualized as grouped bar charts, with models and styles on the x-axis and different topic numbers (k) represented as separate bars, enabling direct comparison of topic coherence across summarization models and styles.

3.5.2. Latent Semantic Analysis (LSA)

LSA is a matrix factorization–based topic modeling technique that represents documents and terms in a reduced-dimensional semantic space. By applying Singular Value Decomposition (SVD) to a TF–IDF–weighted document–term matrix, LSA captures latent associations between words and documents, enabling the extraction of underlying semantic themes. As a linear and interpretable model, LSA serves as an effective baseline for comparing the thematic structure of original and LLM-generated summaries [30].

LSA was used as a linear-algebraic baseline to complement the probabilistic LDA models. The analysis was performed on the same preprocessed dataset, grouped by model and summary style. For each subset, the text_clean field was split into tokens, and a Gensim Dictionary and bag-of-words corpus were constructed. Extremely rare and overly frequent terms were removed via filter_extremes to obtain a cleaner vocabulary. Groups with fewer than 50 documents were skipped, and very large groups were capped at 2,000 documents to maintain computational efficiency.

For LSA, the corpus was first transformed into a TF–IDF–weighted representation using Gensim's TfidfModel, and then a LsiModel was trained with k = 10, 20, 30, 40, 50 topics for each group. The model was configured with onepass=True and chunksize=2000, corresponding to a classical LSA setup over the TF–IDF space. Topic quality was quantified using the C_v coherence metric computed by Gensim's CoherenceModel with the learned LSA topics, the tokenized texts, and the associated dictionary. Coherence scores for each configuration were stored and visualized using grouped bar charts, enabling direct comparison of LSA-based topic coherence across different models and summary types.

3.5.3. Non-negative Matrix Factorization (NMF)

NMF factorizes the term–document matrix into two non-negative components that capture document–topic and topic–word relationships. Because the factorized components remain non-negative, the resulting topics tend to be intuitive and easy to interpret, making NMF a valuable method for uncovering meaningful structure in textual data. The approach is widely used in text mining and pattern recognition, often yielding well-separated and semantically coherent topics [31, 32].

NMF was employed as a matrix factorization-based topic modeling approach to complement the LDA and LSA models. The analysis was conducted on the same preprocessed dataset, grouped by model and summary style. For each subset, the text_clean field was split into token lists, and a Gensim Dictionary and bag-of-words corpus were constructed. Extremely rare and overly frequent terms were filtered out using filter_extremes to obtain a stable vocabulary. Groups with fewer than 50 documents were excluded, and very large groups were limited to 2,000 documents to control memory usage and runtime.

Since NMF operates on non-negative, real-valued inputs, the bag-of-words corpus was transformed into a TF–IDF–weighted representation using Gensim's TfidfModel. For each group, an NMF model (gensim.models.Nmf) was then trained with k = 10, 20, 30, 40, 50 topics, using chunksize = 2000, w_max_iter = 200, h_max_iter = 200, and a fixed random_state = 42 to enhance reproducibility. Topic quality was evaluated using the C_v coherence metric, computed via Gensim's CoherenceModel with the learned NMF topics, the tokenized texts, and the corresponding dictionary. The resulting coherence scores were aggregated for each (model, style, k) configuration and visualized as grouped bar charts, enabling a direct comparison of NMF-based topic coherence across the original abstracts and the different LLM-generated summary types.

3.6. Topic Visualization via Word Clouds

To qualitatively interpret the thematic structures produced by LDA, LSA, and NMF, word clouds were generated for each combination. For every topic model, the most representative terms were extracted based on term weights and visualized using a frequency-based word cloud. This allowed intuitive comparison of dominant biomedical concepts across original abstracts and LLM-generated summaries. Word clouds were used as a complementary qualitative tool to support coherence-based quantitative evaluation. Word clouds were used only as supplementary qualitative visualizations to support the interpretation of topic modeling results and were not considered as primary analytical tools.

3.7. Multi-Dimensional Evaluation Metrics

Sentence-BERT (SBERT) is an extension of the BERT architecture specifically designed to generate semantically meaningful sentence-level embeddings. Unlike the original BERT model, which focuses on token-level representations, SBERT produces fixed-length vector representations for entire sentences by applying a pooling operation over token embeddings. This approach enables efficient and reliable semantic similarity comparisons between texts using standard distance metrics [33].

In this part, semantic similarity between original abstracts and LLM-generated summaries was measured using Sentence-BERT (SBERT) embeddings. Cosine similarity scores were computed between the embedding representations of the original and generated texts to quantify the degree of semantic preservation.

Unlike traditional surface-level metrics that rely on lexical overlap, SBERT captures contextual and semantic relationships between sentences, allowing for a more robust evaluation of meaning consistency. This is particularly important in biomedical text summarization, where semantically equivalent expressions may differ significantly at the lexical level.

By combining topic coherence with SBERT-based semantic similarity, the proposed evaluation framework enables a more comprehensive assessment of summary quality, addressing both thematic structure and semantic fidelity.

The all-MiniLM-L6-v2 model is a compact and efficient sentence embedding model provided by the Sentence-Transformers framework. It is built upon the MiniLM architecture, a distilled variant of BERT that reduces model complexity while preserving strong performance in semantic similarity and sentence representation tasks. Due to its optimized structure, the model offers a favorable trade-off between computational efficiency and accuracy, making it particularly suitable for large-scale applications that require fast and reliable text embedding generation[34].

To assess the semantic similarity between the original abstracts and the generated summaries, the Sentence-BERT (SBERT) model all-MiniLM-L6-v2 was utilized. The model generated embedding vectors for each original abstract and its corresponding summary. Subsequently, cosine similarity was computed in a pairwise (row-wise) manner between each abstract–summary embedding pair.
This procedure was applied independently for each model and for both summarization types as well as for the original PubMed abstracts. As a result, 8,000 similarity scores were obtained for each configuration, with values ranging between 0 and 1. The overall semantic similarity performance for each model and summary type was then calculated by averaging these scores.

By combining topic coherence with SBERT-based semantic similarity, the evaluation framework provides a more comprehensive assessment of summary quality by capturing both thematic structure and meaning preservation.

To assess whether the observed differences between models are statistically significant, statistical analysis was performed on SBERT-based cosine similarity scores. Since similarity scores were computed for each abstract–summary pair, they provide sample-level observations suitable for hypothesis testing.

Pairwise comparisons between models were conducted separately for abstractive and extractive summarization settings. Given the large sample size and potential non-normality of similarity distributions, non-parametric statistical tests were preferred. In addition to p-values, effect sizes were reported using Cohen's d to quantify the magnitude of differences.

It should be noted that statistical significance testing was applied only to SBERT-based metrics. Topic coherence scores represent aggregated corpus-level values and therefore are not suitable for direct statistical hypothesis testing.

4. RESULTS

This section summarizes the findings from the topic modeling and coherence analyses conducted on the original PubMed abstracts and the summaries produced by LLMs. Using LDA, LSA, and NMF across multiple topic sizes, the thematic consistency of each summary type was evaluated using the C_v coherence metric. Word clouds were additionally generated to provide qualitative insight into dominant terms and thematic patterns. The following subsections present the coherence results and comparative observations. To provide a more comprehensive evaluation, semantic similarity analysis was incorporated alongside topic coherence, allowing assessment of both thematic structure and meaning preservation.

4.1. LDA

LDA was used to assess how well each model–style combination preserved the thematic structure of the original PubMed abstracts. Coherence scores (C_v) were computed for all groups across topic sizes (k = 10–50), allowing comparison between abstractive and extractive summaries as well as across the three LLMs. This results shown in Table 2.

Table 2. LDA Coherence Scores (C_v) Across Models, Summary Styles, and Topic Numbers

Model	Topic Numbers				
	k=10	*k=20*	*k=30*	*k=40*	*k=50*
Original	0.393	0.401	0.383	0.380	0.362
ChatGPT(A)	0.420	0.415	0.420	0.403	0.382
ChatGPT(E)	0.401	0.348	0.373	0.363	0.341
DeepSeek(A)	0.417	0.385	0.356	0.362	0.365
DeepSeek(E)	0.389	0.405	0.376	0.363	0.357
Gemini(A)	0.433	0.436	0.387	0.397	0.383
Gemini(E)	0.389	0.374	0.353	0.370	0.341

According to the results, abstractive summaries generated by LLMs achieved coherence values comparable to—or in some cases higher than—the original PubMed abstracts. However, this observation should be interpreted as an indication of more compact and lexically consistent topic structures, rather than superior summary quality in terms of factual accuracy or semantic faithfulness.

Among all systems, Gemini (A) produced the highest coherence scores across most topic sizes, peaking at 0.436 for k = 20, followed by ChatGPT (A), which demonstrated relatively stable coherence values across different topic configurations.

Extractive summaries generally exhibited lower or more variable coherence patterns compared to their abstractive counterparts. For instance, ChatGPT (E) showed a decrease at k = 20, suggesting that extractive summarization may lead to less stable topic distributions depending on the selected topic configuration.

DeepSeek demonstrated moderate and relatively balanced performance, with extractive summaries occasionally yielding slightly higher coherence values than abstractive ones at certain topic sizes.

The original PubMed abstracts showed relatively stable coherence values across topic sizes. In comparison, some LLM-generated summaries exhibited more compact topic structures, which may reflect differences in lexical organization rather than improvements in semantic or factual quality.

Across all models, coherence tended to decrease as the number of topics increased, which is consistent with the expected difficulty of maintaining coherent topic distributions at finer granularity levels.

The corresponding bar chart visualization for LDA is provided in Figure 3.

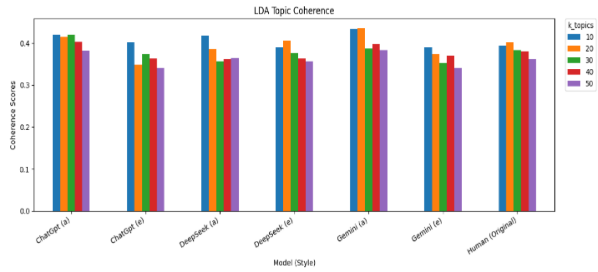

Figure 3. Bar chart visualization of LDA coherence scores

Also, LDA word clouds showing the dominant terms for each topic (k = 5) generated from the combined dataset are shown in Figure 4 as an example. Larger words represent higher term importance within each topic.

Figure 4. LDA word clouds illustrating the dominant terms for each topic (k = 5) generated from the combined dataset.

4.2. LSA

LSA was applied to evaluate the semantic structure of the original abstracts and LLM-generated summaries. Coherence scores were computed for multiple topic sizes (k = 10–50) to assess how effectively each model preserved latent semantic relationships within the text. The results are presented in Table 3.

Table 3. LSA Coherence Scores (C_v) Across Models, Summary Styles, and Topic Numbers

Model	Topic Numbers				
	k=10	*k=20*	*k=30*	*k=40*	*k=50*
Original	0.339	0.393	0.422	0.440	0.460
ChatGPT(A)	0.395	0.388	0.405	0.396	0.407
ChatGPT(E)	0.391	0.372	0.396	0.434	0.444
DeepSeek(A)	0.389	0.380	0.368	0.379	0.416
DeepSeek(E)	0.359	0.362	0.370	0.354	0.363
Gemini(A)	0.373	0.392	0.396	0.384	0.414
Gemini(E)	0.393	0.341	0.367	0.376	0.377

The original PubMed abstracts achieved the highest coherence values, particularly at larger topic sizes, indicating that the original biomedical texts maintain strong latent semantic organization.

Among the LLMs, ChatGPT (A) produced relatively consistent coherence values across topic sizes, suggesting stable semantic representations. ChatGPT (E) also demonstrated competitive performance at higher topic numbers, indicating that extractive summaries may preserve structural patterns under certain configurations.

DeepSeek showed moderate performance, with abstractive summaries gradually improving as the number of topics increased. In contrast, extractive summaries exhibited lower and less stable coherence values.

Gemini demonstrated balanced but slightly lower coherence overall, with some variability across topic sizes.

As expected for LSA, coherence values generally increased with the number of topics, reflecting the model's ability to capture richer latent structures in higher-dimensional semantic space. The strong performance of the original abstracts suggests that naturally written biomedical texts contain dense and well-organized semantic relationships, which are only partially preserved in LLM-generated summaries.

The corresponding bar chart visualization for LDA is provided in Figure 5.

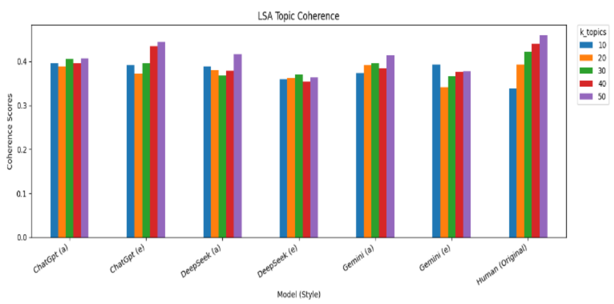

Figure 5. Bar chart visualization of LSA coherence scores

LSA word clouds showing the dominant terms for each topic (k = 5) generated from the combined dataset are shown in Figure 6 as an example.

Figure 6. LSA word clouds illustrating the dominant terms for each topic (k = 5) generated from the combined dataset.

4.3. NMF

NMF was applied to assess the topic structure of the original abstracts and LLM-generated summaries. Unlike LDA and LSA, NMF decomposes the term–document matrix into two non-negative components, enabling additive and parts-based topic representations. The results are presented in Table 4.

Table 4. NMF Coherence Scores (C_v) Across Models, Summary Styles, and Topic Numbers

Model	Topic Numbers				
	k=10	*k=20*	*k=30*	*k=40*	*k=50*
Original	0.327	0.328	0.377	0.380	0.386
ChatGPT(A)	0.313	0.334	0.347	0.381	0.386
ChatGPT(E)	0.324	0.354	0.374	0.405	0.407
DeepSeek(A)	0.315	0.345	0.344	0.380	0.395
DeepSeek(E)	0.308	0.312	0.343	0.340	0.358
Gemini(A)	0.290	0.324	0.362	0.371	0.401
Gemini(E)	0.307	0.325	0.352	0.366	0.397

The original PubMed abstracts demonstrated consistently increasing coherence values across topic sizes, suggesting that the additive structure captured by NMF aligns well with the inherent semantic organization of biomedical texts.

Among the LLMs, ChatGPT (E) showed the most consistent improvement across topic sizes, reaching the highest coherence values among generated summaries at larger topic numbers. ChatGPT (A) also exhibited a similar upward trend, indicating relatively stable topic representations.

DeepSeek demonstrated moderate performance, with abstractive summaries generally outperforming extractive ones in terms of coherence stability.

Gemini showed comparatively lower coherence values at smaller topic sizes, although both abstractive and extractive summaries improved as the number of topics increased.

Overall, coherence values increased with larger topic sizes, reflecting the ability of NMF to capture more detailed semantic components. However, these results should be interpreted as differences in structural topic representation rather than direct indicators of summary quality or factual correctness.

The corresponding bar chart visualization for NMF is provided in Figure 7.

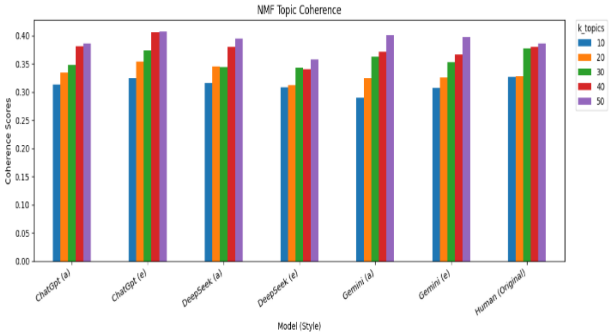

Figure 7. Bar chart visualization of NMF coherence scores

NMF word clouds showing the dominant terms for each topic (k = 5) generated from the combined dataset are shown in Figure 8 as an example.

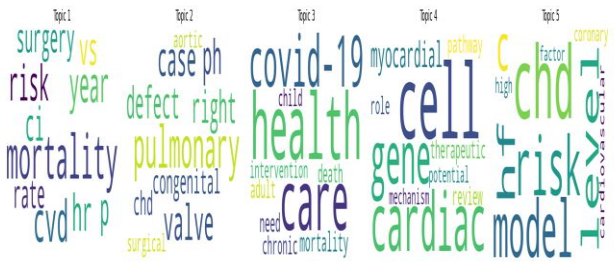

Figure 8. NMF word clouds illustrating the dominant terms for each topic (k = 5) generated from the combined dataset.

4.3. SBERT Results

To further evaluate semantic consistency, SBERT-based cosine similarity scores were computed between the original abstracts and the generated summaries. The results are presented in Table 4 for both abstractive and extractive summarization settings.

Table 4. SBERT-Based Semantic Similarity Scores Across Models and Summary Types

Summary Type	ChatGPT	DeepSeek	Gemini
Abstractive	0.8782	0.8263	0.8536
Extractive	0.8961	0.8158	0.7964

The results demonstrate clear differences in how effectively each model preserves the semantic content of the original biomedical abstracts.

In the abstractive setting, ChatGPT achieved the highest similarity score (0.8782), indicating its strong capability to retain the underlying meaning while generating paraphrased summaries. Gemini followed with a slightly lower score (0.8536), suggesting a relatively good level of semantic preservation. DeepSeek showed the lowest performance (0.8263) among the three models in this setting.

In the extractive setting, all models generally produced higher similarity scores compared to their abstractive counterparts. ChatGPT again outperformed the other models with the highest score (0.8961), reflecting its effectiveness in selecting semantically relevant content from the original texts. DeepSeek achieved a moderate score (0.8158), while Gemini exhibited the lowest similarity (0.7964), indicating comparatively weaker semantic alignment in extractive summaries.

Overall, the findings suggest that ChatGPT consistently maintains the highest level of semantic similarity across both summarization types. Additionally, the higher scores observed in extractive summaries confirm that preserving original sentence structures leads to stronger semantic alignment. These results complement the coherence-based analysis by showing that models with higher thematic consistency also tend to preserve semantic meaning more effectively.

4.4. Statistical Significance Analysis

To validate the reliability of the observed differences in semantic similarity, statistical significance testing was conducted on SBERT-based cosine similarity scores. Pairwise comparisons between models were performed separately for abstractive and extractive summarization settings. The results are presented in Table 5.

Table 5. Statistical Significance Results for SBERT-Based Semantic Similarity

Metric	**Condition**	**Comparison**	**p-value**	**Cohen's d**
SBERT	Abstractive	ChatGPT vs Gemini	< 0.001	0.46
SBERT	Abstractive	ChatGPT vs DeepSeek	< 0.001	0.73
SBERT	Abstractive	Gemini vs DeepSeek	< 0.001	0.37
SBERT	Extractive	ChatGPT vs Gemini	< 0.001	1.01
SBERT	Extractive	ChatGPT vs DeepSeek	< 0.001	1.06
SBERT	Extractive	Gemini vs DeepSeek	< 0.001	-0.18

All pairwise comparisons yielded statistically significant differences ($p < 0.001$).

In the abstractive setting, ChatGPT showed higher semantic similarity compared to both Gemini and DeepSeek, with moderate to large effect sizes. Gemini also outperformed DeepSeek with a smaller effect size.

In the extractive setting, ChatGPT demonstrated substantially higher performance, with large effect sizes when compared to both Gemini and DeepSeek. In contrast, the difference between Gemini and DeepSeek was negligible, indicating similar semantic performance despite statistical significance.

Overall, the results confirm that the observed differences are statistically significant, while the effect sizes highlight the practical magnitude of these differences.

5. CONCLUSION and FUTURE WORK

This study presented a large-scale and systematic evaluation of how different LLMs preserve thematic integrity when summarizing biomedical literature. By comparing original PubMed abstracts with 48,000 LLM-generated summaries using three complementary topic modeling techniques, the findings demonstrate that coherence preservation depends on both the summarization style and the underlying model architecture. Abstractive summaries produced by ChatGPT and Gemini generally achieved higher coherence levels, while extractive summaries exhibited more variable performance across topic configurations.

The consistency of these observations across LDA, LSA, and NMF indicates that the identified coherence differences reflect meaningful semantic patterns rather than artifacts of a specific modeling approach. In addition to topic coherence, semantic similarity analysis using SBERT provided further evidence on how effectively the generated summaries preserve the underlying meaning of the original texts. Together, these findings suggest that modern LLMs are capable of approximating—and in some cases exceeding—the thematic organization of human-written biomedical abstracts, although their performance remains sensitive to both model characteristics and summarization strategies.

Although this study does not propose a novel model architecture, its primary contribution lies in establishing a large-scale, systematic, and multi-perspective evaluation framework for analyzing LLM-generated biomedical summaries. By jointly considering topic coherence and semantic similarity, the proposed framework enables a more comprehensive understanding of both thematic structure and meaning preservation across different models and summary types.

It is important to note that higher coherence scores do not necessarily indicate better summary quality in terms of factual correctness or clinical reliability. Coherence should therefore be interpreted as a structural property rather than a direct measure of overall summary quality.

Despite its contributions, this study has several limitations. First, the evaluation relies on automatic metrics such as topic coherence and semantic similarity, which do not fully capture factual correctness or clinical validity. High coherence or similarity scores do not necessarily guarantee that generated summaries are factually accurate or medically reliable. Second, human expert evaluation was not included, which limits the ability to assess the practical usefulness of the summaries in real-world biomedical contexts. Third, the analysis was restricted to a single biomedical domain (cardiovascular abstracts), which may limit the generalizability of the findings to other medical areas. Future studies will expand the analysis to multiple biomedical domains beyond cardiovascular literature (e.g., oncology, neurology, and infectious diseases) to evaluate the generalizability and robustness of the proposed framework across diverse clinical contexts.

Future work will extend this framework by incorporating human-in-the-loop evaluation and factual consistency metrics to provide a more robust and clinically meaningful assessment of LLM-generated summaries. In addition, integrating linguistic feature-based classification and stylometric analysis may offer deeper insights into distinguishing human-written and machine-generated texts.

Further studies may also explore dynamic and neural topic modeling approaches to evaluate semantic consistency under evolving topic structures. Expanding the analysis to diverse biomedical domains will further enhance the generalizability and practical relevance of the proposed framework.

REFERENCES

[1] L. Hunter and K. B. Cohen, “Biomedical language processing: what's beyond PubMed?”, *Molecular Cell*, 21(5), 589–594, 2006.

[2] Z. Lu, “PubMed and beyond: a survey of web tools for searching biomedical literature”, Database, 2011, baq036.

[3] D. Demner-Fushman, W. W. Chapman and C. J. McDonald, “What can natural language processing do for clinical decision support?”, *Journal of Biomedical Informatics*, 42(5), 760–772, 2009.

[4] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman et al., “GPT-4 Technical Report”, arXiv preprint, arXiv:2303.08774, 2023.

[5] G. Team, R. Anil, S. Borgeaud, J. B. Alayrac, J. Yu, R. Soricut et al., “Gemini: A family of highly capable multimodal models”, arXiv preprint, arXiv:2312.11805, 2023.

[6] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu et al., DeepSeek-V3 Technical Report, arXiv, 2024.

[7] I. Jahan, M. T. R. Laskar, C. Peng, J. Huang et al., “Evaluation of ChatGPT on biomedical tasks: A zero-shot comparison with fine-tuned generative transformers”, arXiv preprint, arXiv:2306.04504, 2023.

[8] H. Nori, N. King, S. M. McKinney, D. Carignan, E. Horvitz, “Capabilities of GPT-4 on medical challenge problems”, arXiv preprint, arXiv:2303.13375, 2023.

[9] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung et al., “Large language models encode clinical knowledge”, *Nature*, 620(7972), 172–180, 2023.

[10] R. Nallapati, B. Zhou, C. Gulcehre, et al., “Abstractive text summarization using sequence-to-sequence RNNs and beyond”, Proceedings of ACL, 2016.

[11] R. Mishra, J. Bian, M. Fiszman, C. R. Weir, S. Jonnalagadda, J. Mostafa et al., “Text summarization in the biomedical domain: A systematic review of recent research”, *Journal of Biomedical Informatics*, 52, 457–467, 2014.

[12] J. Maynez, S. Narayan, B. Bohnet, R. McDonald, “On faithfulness and factuality in abstractive summarization”, arXiv preprint, arXiv:2005.00661, 2020.

[13] S. Syed, M. Spruit, “Full-text or abstract? Examining topic coherence scores using Latent Dirichlet Allocation”, **2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)**, Tokyo, Japan, 165–174, October 2017.

[14] M. Wang, M. Wang, F. Yu, Y. Yang, J. Walker, and J. Mostafa, “A systematic review of automatic text summarization for biomedical literature and EHRs,” *J. Am. Med. Inform. Assoc*., vol. 28, no. 10, pp. 2287–2297, 2021.

[15] A. Chavs, C. Kesiku, and B. Garcia-Zapirain, “Automatic text summarization of biomedical text data: a systematic review,” *Information*, vol. 13, no. 8, p. 393, 2022.

[16] Z. Huan, X. Chen, Y. Wang, J. Huang, and X. Zhao, “A survey on biomedical auomatic text summarization with large language models,” *Inf. Process. anag.*, vol. 62, no. 5, p. 104216, 2025.

[17] D. Van Veen, C. Van Uden, L. Blankemeier, J. B. Delbrouck, A. Aali, C. Bluethgen, et al., “Clinical text summarization: adapting large language models can outperform human experts,” Research Square, rs-3, 2023.

[18] E. Croxford, Y. Gao, E. First, N. Pellegrino, M. Schnier, J. Caskey, et al., “Evaluating clinical AI summaries with large language models as judges,” *npj Digital Medicine*, vol. 8, no. 1, p. 640, 2025.

[19] L. Bednarczyk, D. Reichenpfader, C. Gaudet-Blavignac, A. K. Ette, J. Zaghir, Y. Zheng, et al., “Scientific evidence for clinical text summarization using large language models: scoping review,” ., vol. 27, *Journal of Medical Internet Research*, p. e68998, 2025.

[20] A. Naemi and A. Sahafi, “Benchmarking large language models for MIMIC-IV clinical note summarization,” *J. Healthc. Inform. Res*., pp. 1–21, 2025.

[21] D. Maier, A. Waldherr, P. Miltner, G. Wiedemann, A. Niekler, A. Keinert, … S. Adam, “Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology,” **in Computational Methods for Communication Science**, (Eds.: D. Nguyen, C. Trilling), Routledge, New York, A.B.D., 2021, pp. 13–38.

[22] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, “Indexing by Latent Semantic Analysis,” *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[23] Y. Chen, H. Zhang, R. Liu, Z. Ye, J. Lin, “Experimental explorations on short text topic mining between LDA and NMF based schemes”, Knowledge-Based Systems, 163, 1–13, 2019.

[24] M. Röder, A. Both, A. Hinneburg, “Exploring the Space of Topic Coherence Measures”, **8th ACM International Conference on Web Search and Data Mining (WSDM)**, Shanghai, China, 399–408, 2–6 February, 2015.

[25] U.S. National Library of Medicine. “PubMed Overview”. https://pubmed.ncbi.nlm.nih.gov/about/ (01.12.2025).

[26] OpenAI.“GPT-4 Technical Report”.https://openai.com/research/gpt-4 (01.12.2025).

[27] M. E. de Carvalho Souza, L. Weigang, “Grok, Gemini, ChatGPT and DeepSeek: Comparison and Applications in Conversational Artificial Intelligence”, *Inteligencia Artificial*, 2(1), 2025.

[28] D. Jurafsky, J. H. Martin, Speech and Language Processing, 3rd ed., Draft Version, Stanford University, Stanford, CA, U.S.A., 2021.

[29] D. M. Blei, A. Y. Ng, M. I. Jordan, “Latent Dirichlet Allocation”, *Journal of Machine Learning Research*, 3(1), 993–1022, 2003.

[30] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman, “Indexing by Latent Semantic Analysis”, *Journal of the American Society for Information Science*, 41(6), 391–407, 1990.

[31] D. D. Lee, H. S. Seung, “Algorithms for Non-negative Matrix Factorization”, *Advances in Neural Information Processing Systems (NIPS)*, vol. 13, pp. 556–562, 2001.

[32] Y. X. Wang, Y. J. Zhang, “Nonnegative matrix factorization: A comprehensive review”, *IEEE Trans. Knowl. Data Eng*., 25(6), 1336–1353, 2012.

[33] N. Reimers, I. Gurevych, “Sentence-BERT: Sentence embeddings using siamese BERT-networks”, **Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)**, Hong Kong, China, 3982–3992, November, 2019.

[34] N. Reimers, I. Gurevych, “Making monolingual sentence embeddings multilingual using knowledge distillation”, **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**, Online, 4512–4525, November, 2020.