http://dx.doi.org/10.62762/TETAI.2025.871572

RESEARCH ARTICLE

GPT vs. Other Large Language Models for Topic Modeling: A Comprehensive Comparison

Muhammet Bora Meram[1], **Çağatay Kalkan**[1], **Tuğba Çelikten**[1,2,*] **and Aytuğ Onan**[2]

[1] Department of Software Engineering, Manisa Celal Bayar University, Manisa 45140, Turkey
[2] Department of Computer Engineering, Izmir Katip Celebi University, Izmir 35620, Turkey

Abstract

Topic modeling is a widely used unsupervised natural language processing (NLP) technique aimed at discovering latent themes within documents. Since traditional methods fall short in capturing contextual meaning, approaches based on large language models (LLMs)—such as BERTopic—hold the potential to generate more meaningful and diverse topics. However, systematic comparative studies of these models, especially in domains requiring high accuracy and interpretability such as healthcare, remain limited. This study compares ten different LLMs (GPT, Claude, Gemini, LLaMA, Qwen, Phi, Zephyr, DeepSeek, NVIDIA-LLaMA, Gemma) using a dataset of 9,320 medical article abstracts. Each model was tasked with generating five topics per article; the outputs were analyzed using metrics such as diversity, relevance, cosine similarity, and entropy. The highest diversity was achieved by Phi (0.717) and LLaMA (0.682), while the highest relevance was observed in Qwen (0.610) and GPT (0.536). The average cosine similarity was 0.5778, and the topic overlap score between GPT and LLaMA was 0.0024. The Zephyr model generated both the longest topics (98.5 words) and the greatest vocabulary variety (67,295 unique words). Overall, an inverse relationship was observed between diversity and relevance across models, suggesting that model selection should be carefully aligned with the intended application. This study offers a methodological foundation for future research by revealing the strengths and weaknesses of LLM-based topic modeling methods in critical domains such as healthcare, where precision and explainability are paramount.

Keywords: large language models, topic modeling, LLM-based topic generation, topic diversity.

1 Introduction

Topic modeling methods are unsupervised techniques used to uncover hidden topics within documents and are widely applied across various domains [1]. In other words, topic modeling is a powerful natural language processing (NLP) technique for discovering latent themes in large text collections [2].

Traditional techniques such as Latent Dirichlet Allocation (LDA) and Non-negative Matrix

Submitted: 28 May 2025
Accepted: 11 July 2025
Published: 27 July 2025

Vol. 2, **No.** 3, 2025.
10.62762/TETAI.2025.871572

***Corresponding author:**
✉ Tuğba Çelikten
tugba.celikten@cbu.edu.tr

Citation
Meram, M. B., Kalkan, C., Celikten, T., & Onan, A. (2025). GPT vs. Other Large Language Models for Topic Modeling: A Comprehensive Comparison. *ICCK Transactions on Emerging Topics in Artificial Intelligence*, 2(3), 116–130.

Factorization (NMF) often possess limited contextual understanding and fail to fully capture the semantic depth of language [3]. To overcome these limitations, Large Language Models (LLMs) have recently emerged as an innovative approach in topic modeling. Models like GPT, Claude, and Gemini possess the ability to produce semantically rich topics, thanks to their contextual comprehension capabilities, going beyond purely statistical methods [4].

However, there is still a lack of systematic and comprehensive comparative analyses of these models in the context of topic modeling in the current literature [6]. Especially in domains such as healthcare, where accuracy and explainability are critical, it is essential to analyze in detail how consistent, diverse, and interpretable the topics generated by these models truly are.

One of the main reasons for this gap is that LLM-based topic modeling remains a relatively new research area. Additionally, the lack of standardized evaluation metrics across different models further contributes to this issue [7].

Addressing this gap, the present study provides one of the first systematic analyses that compares the topic modeling performance of ten different LLMs in the healthcare domain. Unlike traditional embedding-based approaches, this study adopts a prompt-guided strategy tailored for each model.

Specifically, the study compares the topic modeling performance of ten LLMs (GPT, Gemini, LLaMA, Qwen, Phi, Zephyr, DeepSeek, NVIDIA-LLaMA, Gemma, Claude) on a corpus of 9,320 academic articles in the healthcare field using prompt-based methods. By employing metrics such as diversity, relevance, cosine similarity, and entropy, the strengths and weaknesses of each model are revealed from both qualitative and quantitative perspectives.

2 Related Work

Pham et al. [8] introduced a prompt-based framework called *TopicGPT*, aiming to enhance the topic modeling capabilities of LLMs. This framework generates more interpretable topics by incorporating natural language labels and descriptions, enabling users to exert greater control over model outputs.

Wang et al. [9] proposed an approach named *PromptTopic*, which seeks to improve the performance of LLMs in modeling topics from short texts. This method enables sentence-level topic inference, resulting in more coherent and meaningful topics.

Tan et al. [7] conducted an automatic evaluation of dynamically evolving topic taxonomies using LLMs. Their study demonstrated that LLMs could assess topic quality through metrics such as consistency, redundancy, diversity, and topic-document alignment, offering an alternative to traditional human-centric evaluation methods.

Another study utilizing LLMs, conducted by Çelikten et al. [34], performed topic modeling on scientific texts generated by both humans and AI. In this work, semantic representations of the texts were obtained using large language models, and prominent topics were identified through a rank-based aggregation method. The approach enables a comparative analysis of the thematic structure across texts from different sources.

Isonuma et al. [10] quantitatively evaluated the topic modeling performance of LLMs. Their findings showed that while LLMs are capable of generating consistent and diverse topics, they sometimes take "shortcuts" by focusing on specific sections of documents, thereby limiting controllability.

Arora et al. [11] developed an open-source benchmark called *HealthBench* to assess the performance and safety of LLMs in the healthcare domain. The benchmark comprises 5,000 multi-turn dialogues and evaluates LLM competencies in health-related tasks.

Asmussen et al. [6] emphasized the challenges of traditional LDA-based literature reviews, such as high topic redundancy and low topic distinctiveness. These issues underscore the need for more advanced modeling approaches.

In addition to LLM-based topic modeling approaches, Onan et al. [31] investigated the coherence and diversity of AI-generated and paraphrased scientific abstracts using a fuzzy topic modeling method. Their work allows for a more flexible and realistic representation of multi-topic scientific texts, providing a novel perspective for assessing the quality of AI-generated content.

Rosenfeld et al. [5] compared the generation styles of different LLMs to identify which models produce more consistent and original texts. Their assessment highlighted LLMs' sensitivity to linguistic structures and stylistic differences.Topic modeling has long been employed to uncover latent themes in text corpora. One of the most well-known techniques

in this domain is LDA [14], which infers hidden topic structures based on word distributions in documents. Other methods, such as NMF [12] and the Hierarchical Dirichlet Process (HDP) [13], have also been used, each incorporating different assumptions and structures. However, these techniques often rely on word co-occurrence frequencies and tend to produce semantically shallow representations due to their lack of contextual understanding.

Topic modeling has long served as a fundamental unsupervised learning technique for uncovering latent semantic structures in large-scale text corpora. Classical statistical methods such as Latent Dirichlet Allocation (LDA) [14], Non-negative Matrix Factorization (NMF) [12], and Hierarchical Dirichlet Process (HDP) [13] have dominated early research, relying on word co-occurrence statistics and often yielding topics with limited contextual depth and interpretability.

To overcome these limitations, recent research has turned to leveraging the contextual understanding of large language models (LLMs) for topic modeling. One of the earliest frameworks in this direction, **TopicGPT** by Pham et al. [8], introduced a prompt-based architecture to steer LLMs toward more coherent and interpretable topic generation. Similarly, **LLM-TAKE** [3] focused on theme-aware keyword extraction using encoder-decoder LLMs, enabling finer semantic control in topic formulation.

Other innovations have explored architectural enhancements and evaluation perspectives. **DeTiME** [1] proposed a diffusion-based encoder-decoder approach to refine topic representations generated by LLMs. In parallel, Kapoor et al. [4] introduced **QualIT**, a framework that augments LLM-based topic modeling with qualitative insight metrics to enhance interpretability and user alignment.

Evaluation of topic quality remains a challenge. Tan et al. [7] highlighted the inconsistency of traditional metrics and proposed a taxonomy-aware framework for evaluating LLM-generated topics using coherence, redundancy, and alignment metrics.

A related study by Isonuma et al. [10] quantitatively benchmarked LLMs for topic modeling using metrics such as topic diversity, specificity, and controllability, echoing the need for systematic multi-model comparisons. Likewise, Çelikten et al. [34] presented a rank-based aggregation method for comparing human- and AI-generated scientific text structures using LLM embeddings.

While these works mark significant progress, most focus on individual models or specific tasks such as keyword extraction or label generation. To our knowledge, no prior study has conducted a domain-specific, head-to-head comparative evaluation of multiple LLMs in the context of full-topic modeling, particularly within critical domains such as healthcare where interpretability, precision, and semantic grounding are paramount.

This study addresses this gap by systematically comparing the topic modeling performance of ten state-of-the-art LLMs using both qualitative and quantitative metrics on a corpus of over 9,000 healthcare-related article abstracts. Our contribution lies not only in breadth—covering a diverse set of models (e.g., GPT, Claude, Zephyr, Phi, LLaMA)—but also in our use of application-aligned metrics and structured evaluation methodology to assess model strengths and limitations across multiple dimensions.

3 Methodology

In this study, the performance of LLMs in the context of topic modeling is systematically compared.

Unlike traditional topic modeling approaches, this study treats each article abstract as an *independent data unit* rather than processing the entire dataset at once. Accordingly, the experimental design is based on providing individual *prompts* to LLMs for each abstract.

The dataset consists of a total of 9,320 academic articles from the healthcare domain, each analyzed using ten different LLM models. Each model generated five topic models per article, and the resulting outputs were evaluated using both qualitative and quantitative criteria. Algorithm 1 outlines the sequential steps followed throughout the study.

This section introduces the topic modeling approach Section 3.1. The dataset used in the experiments and the data preparation process are described in Section 3.2. Experimental details and the use of large language models for topic modeling are presented in Section 3.3.

3.1 Topic Modeling

Topic modeling is a statistical method used to uncover hidden themes within large text corpora. It aims to classify documents into topics by analyzing word patterns. One of the most widely used approaches is the LDA model [32].

Algorithm 1: Topic Generation and Evaluation Process using LLMs

Data: A total of 9,320 academic articles from the healthcare domain
Result: Five topics generated by each of the ten LLMs per article and corresponding performance evaluation results

```
foreach Article ∈ ArticleSet do
    foreach LLM ∈ {GPT, Claude, Gemini, LLaMA, Qwen, Phi, Zephyr, DeepSeek, NVIDIA-LLaMA, Gemma} do
        Run the LLM with the specified prompt;
        Save the five generated topics;
    end
end
foreach Set of topics generated for each article do
    foreach LLM ∈ Models do
        Measure Diversity and Relevance scores;
        Perform Entropy calculation;
        Generate topic vectors and calculate Cosine Similarity;
    end
end
```

This technique is particularly common in the analysis and classification of semantic structures across documents. LDA, one of the best-known topic modeling methods, assumes that each document is a mixture of multiple topics and that each topic is represented by a distribution over specific words. This allows documents to be meaningfully grouped based on their content and enables the extraction of underlying themes [14].

Topic modeling facilitates the extraction of meaningful information clusters from large-scale text data sources such as digital archives, news websites, social media content, and academic publications, without the need for time-consuming manual labeling or classification. Consequently, it enhances user access to relevant information and enables deeper content analysis. In data-intensive domains such as healthcare, education, and law, topic modeling also contributes to the development of decision support systems [29].

3.2 Data Collection and Preparation

In this study, scientific article data were programmatically collected from the BioMed Central (BMC) [39] platform using the Python programming language and the Selenium library. During the data collection process, ten different medical themes were targeted, and approximately 1,000 articles were retrieved for each theme. The targeted themes are as follows: Human Health, Radiological Treatment, Breast Cancer, Treatment Protocols, Disease Prevention, Drug Development, Chronic Diseases, Medical Imaging, Clinical Decision Support Systems, and Personalized Medicine Approaches.

Unlike traditional data preprocessing approaches, this method performs independent evaluation for each text instance, thereby allowing the NLP capabilities of LLMs to be assessed on a more individual level. This structure enhances the comparability and explainability of the topic modeling performance across different LLM models.

To transform the data into a format that can be understood and processed by the models, a multi-step preprocessing pipeline was implemented, as summarized in Figure 1. Following the preprocessing stage, duplicate entries were removed, and corrupted or structurally inconsistent samples were filtered out during the cleaning phase. As a result, a total of 9,320 data instances were prepared for the experiments.

3.3 Large Language Models (LLMs) for Topic Modeling

The following ten LLMs were used on the same collection of articles: GPT (OpenAI), Claude (Anthropic), Gemini (Google), LLaMA (Meta), Qwen, Phi (Microsoft), Zephyr, DeepSeek, NVIDIA LLaMA, and Gemma.

Each model was provided with a similar prompt structure, and parameters such as output length and temperature were held constant to ensure comparability across model outputs.

3.3.1 GPT (OpenAI)

Developed by OpenAI, the GPT series has gained significant prominence in the field of NLP. GPT-4 is a multimodal model capable of processing both text and visual inputs, and it demonstrates human-level performance across various professional and academic benchmarks [15].

3.3.2 Claude (Anthropic)

Anthropic's Claude model is developed with a strong focus on safety and transparency. Claude 2.0 exhibits high performance in real-world tasks, and research on its inner workings has revealed how millions of concepts are represented within the model [16].

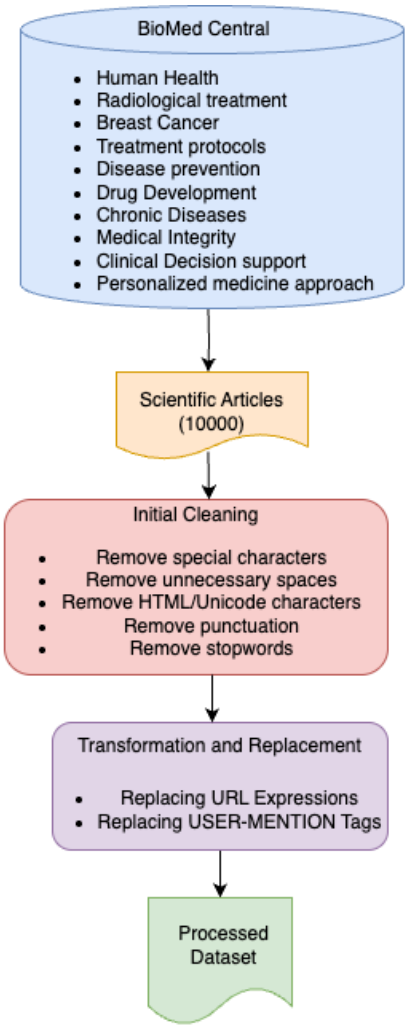

Figure 1. Flowchart of the data preprocessing pipeline.

3.3.3 *Gemini (Google DeepMind)*

Gemini, developed by Google DeepMind, is a family of multimodal models capable of processing text, images, audio, and video. The Gemini models are offered in various configurations (Ultra, Pro, Nano) and demonstrate superior performance across a range of tasks [17].

3.3.4 *LLaMA (Meta)*

Meta's LLaMA series offers open-source and efficient LLMs. LLaMA 3 is equipped with capabilities such as multilingual support, coding, and logical reasoning, and it performs well across various benchmarks [18].

3.3.5 *Qwen (Alibaba)*

Qwen, developed by Alibaba, is a comprehensive language model series available in different parameter sizes. Qwen models have demonstrated effective performance across a variety of NLP tasks [19].

3.3.6 *Phi (Microsoft)*

Microsoft's Phi series aims to deliver high performance with relatively small model sizes. The Phi-3-mini model, with 3.8 billion parameters, achieves results comparable to much larger models [20].

3.3.7 *Zephyr (Hugging Face)*

Zephyr is a project aimed at developing compact language models aligned with user intent. Trained using the Distilled Supervised Fine-Tuning (dSFT) method, Zephyr models perform effectively across a range of tasks [21].

3.3.8 *DeepSeek (DeepSeek-AI)*

Developed by DeepSeek-AI, DeepSeek provides open-source LLMs. The DeepSeek LLM 67B model demonstrates strong performance particularly in tasks involving coding, mathematics, and logical reasoning [22].

3.3.9 *NVIDIA-LLaMA*

NVIDIA has extended the capabilities of LLaMA models by enabling the construction of 3D networks, allowing the models to understand and generate 3D objects. This approach bridges the gap between language processing and 3D modeling [23].

3.3.10 *Gemma (Google)*

Gemma is an open-source model family developed by Google, based on the research and technology behind Gemini. Gemma models show strong performance in tasks related to language understanding, logical reasoning, and safety [24].

It is important to note that the ten LLMs included in this study vary in terms of their exposure to domain-specific training data. Models such as GPT-4 (OpenAI) and Claude 2.0 (Anthropic) have been evaluated on biomedical benchmarks and are known to exhibit strong reasoning capabilities in health-related contexts. In contrast, models like Zephyr and Phi have not been explicitly fine-tuned on medical corpora but were included to assess how general-purpose LLMs perform on domain-specific tasks. This distinction allows us to explore whether domain alignment significantly influences topic modeling outcomes in the healthcare setting.

4 Experiments and Results

A prompt-based approach was adopted to leverage the semantic capacity of each model. The methodology followed in this study is illustrated in Figure 2. Accordingly, each model was presented with explicit

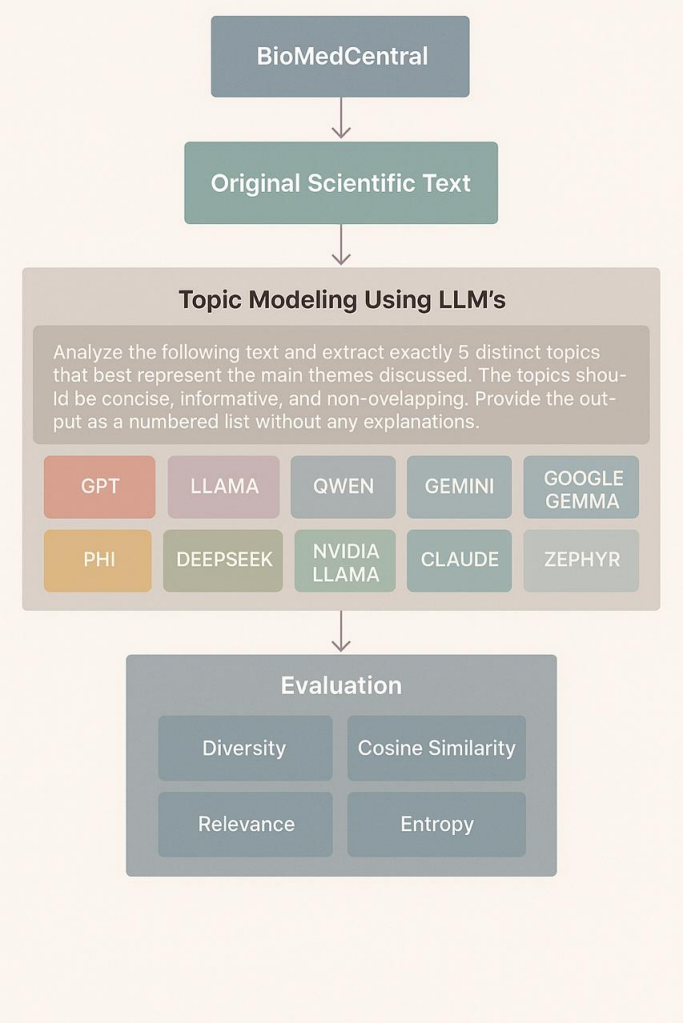

Figure 2. The pipeline of the study.

GPT:

prompt = ''

Analyze the following text and extract exactly 5 distinct topics that best represent the main themes discussed. The topies should be concise, informative, and non-overlapping. Provide the output as a numbered list without any explanations.

Data:

The knee joint is the predilection site of osteoarthritis [1].
Knee osteoaerthitis (KOA) is a chronic disease characterized by progressive damage to the cartilage, subchondral bone, synovial tissues, and other intra-articular structures of the knee joint [2, 3]. The increasing incidence of knee osteoarthritis reduces peoples quality of life, increases disability and imposes a huge...

Figure 3. Sample of Prompt–Input pair.

prompts instructing it to generate "five short, meaningful, and original topics" based on the given abstract.

Figure 3 presents the common prompt employed in the experiments along with a sample input, while Figure 4 shows the corresponding output generated based on that input.

The results were obtained individually for each LLM, and these outputs were evaluated using the following metrics: Diversity, Relevance, Cosine Similarity, and Entropy.

Diversity: In topic modeling, *diversity* is a key metric that measures how distinct the generated topics are from one another [25, 30].

According to this definition, the top-ranked words from all topics are aggregated, and the proportion of non-repeating (unique) words is calculated. The diversity metric is mathematically formulated as shown in Eq 1:

$$\text{Diversity} = \frac{\left|\bigcup_{k=1}^{K} T_k\right|}{K \times N} \tag{1}$$

- K = total number of topics,
- N = number of top words selected from each topic,
- T_k = set of the top N words in the k-th topic.

To compute the diversity metric in practice, we extracted the top $N = 10$ representative words from each of the $K = 5$ topics generated for a given article by a particular model. We then calculated the proportion of unique words across these $K \times N = 50$ words. This yields a normalized score between 0 and 1, where a higher score indicates greater lexical variation and thematic diversity among the generated topics. For example, if 42 out of the 50 words are unique across five topics, the diversity score would be 0.84. This metric captures how well a model differentiates the topics it generates, thereby reflecting its capacity for semantic breadth.

A higher diversity value indicates that the topics are more unique and varied. In contrast, a low diversity score suggests significant overlap and repetition among topics, implying that the model fails to sufficiently distinguish between different themes. Therefore, the diversity metric plays a crucial role in evaluating the quality of topic modeling.

Relevance: In topic modeling, *relevance* is a metric used to determine how specific a word is to a given

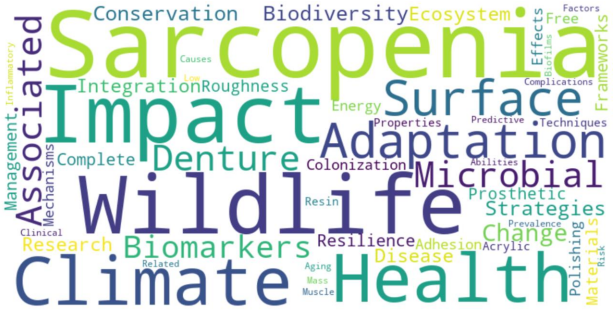

Figure 4. Output generated from the sample input and prompt.

topic and to what extent it represents that topic. This metric reduces the influence of frequently occurring but non-topic-specific words, thereby enhancing the interpretability and meaningfulness of the generated topics. A vector-based similarity analysis was performed between topic sets generated by different models for the same article. This analysis employed embedding models based on Sentence-BERT [26]. According to this definition, the relevance of a word to a topic is calculated using the formula shown in Eq 2:

$$\text{Relevance}(w, k \mid \lambda) = \lambda \cdot P(w \mid k) + (1 - \lambda) \cdot \frac{P(w \mid k)}{P(w)} \tag{2}$$

- $P(w \mid k)$: The probability of word w within topic k.
- $P(w)$: The overall probability of word w in the entire corpus.
- λ: A balancing parameter between 0 and 1. As λ approaches 1, more weight is given to the word's probability within the topic; as it approaches 0, more emphasis is placed on the word's topic specificity (lift).

This relevance metric was originally introduced by Sievert and Shirley [26] in the LDAvis framework to enhance topic interpretability by balancing term probability and distinctiveness (lift).

The relevance metric is particularly useful in visualization tools such as *pyLDAvis*, as it contributes to making topics more meaningful and interpretable. A high relevance score indicates that a word is specific to a given topic and represents it well. In contrast, a low relevance score suggests that the word is commonly used across the corpus but is not specific to any particular topic, making it less effective in representing that topic.

Therefore, the relevance metric is a valuable tool for enhancing the quality of topic modeling and ensuring the interpretability of the resulting topics.

Cosine Similarity: In topic modeling, *cosine similarity* is a widely used metric for measuring the similarity between two topics. This metric calculates the cosine of the angle between two vectors, which indicates how similar their directions are. In the context of topic modeling, each topic is represented as a distribution over words, and these distributions can be expressed as vectors [27]. Mathematically, the cosine similarity between two vectors A and B is computed using Eq 3:

$$\begin{aligned}\text{Cosine Similarity} = \cos(\theta) &= \frac{A \cdot B}{\|A\|\|B\|} \\ &= \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}}\end{aligned} \tag{3}$$

- A and B: The two vectors being compared.
- A_i and B_i: The i-th components of vectors A and B, respectively.
- $\cdot$: The dot product of the two vectors.
- $\|A\|$ and $\|B\|$: The norms (magnitudes) of vectors A and B.

Cosine similarity is a standard metric in NLP and information retrieval for comparing vectorized text representations [27]. In the context of topic modeling, it serves as an effective means of quantifying semantic overlap between topic-word distributions.

Cosine similarity ranges from -1 to 1. However, since the vectors used in topic modeling typically contain non-negative values, the effective range becomes 0 to 1. A value close to 1 indicates that the vectors (topics) are highly similar, while a value close to 0 implies that the vectors are significantly different from one another.

In topic modeling, cosine similarity plays a crucial role in comparing topics generated by different models, tracking the evolution of topics over time, and clustering similar topics. A high cosine similarity suggests that two topics share similar themes, whereas a low similarity indicates that the topics are associated with distinct thematic content.

Entropy: In topic modeling, *entropy* is a metric used to measure the uncertainty or dispersion of a topic's word distribution. This metric is particularly useful for evaluating how distinct or specific a topic is. The

entropy value is calculated using the formula as Eq [28]:

$$H(k) = -\sum_{w \in V} P(w \mid k) \log_b P(w \mid k) \qquad (4)$$

- $H(k)$: The entropy of topic k.
- V: The entire vocabulary.
- $P(w \mid k)$: The probability of word w within topic k.
- b: The base of the logarithm (commonly 2, e, or 10).

In topic modeling, the entropy metric plays a significant role in assessing the distinctiveness and interpretability of topics. A low entropy value indicates that the topic is more specific and well-defined, while a high entropy value suggests that the topic is more general and ambiguous.

The evaluation process was conducted using the Python programming language. The analysis utilized several libraries, including `transformers`, `scikit-learn`, `numpy`, `pandas`, and `sentence-transformers` [33, 35–38].

Access to the LLM models was provided through Hugging Face and OpenAI APIs [15, 33].

The results of the experiments are presented in Table 1.

To statistically validate the observed inverse relationship between diversity and relevance, we computed the Pearson correlation coefficient across the average scores of the ten evaluated LLMs. The analysis revealed a strong negative correlation ($r = -0.76$, $p = 0.011$), suggesting that models with higher diversity scores tend to produce less relevant topics, and vice versa. This statistically supports our qualitative observation of a trade-off between topic diversity and contextual relevance.

The analysis of model outputs revealed notable differences in terms of *diversity* and *relevance*. As shown in Table 1, the models with the highest diversity scores were Phi (0.717) and DeepSeek (0.695). The topics generated by these models are highly distinct from one another, indicating their ability to produce creative content that spans a broad thematic range. However, this diversity may come at the cost of reduced alignment with the original content; for instance, Phi's relevance score was relatively low (0.395). This suggests that high diversity can sometimes negatively impact content coherence.

Table 1. Overall performance evaluation of LLMs in topic modeling.

Model	Diversity	Relevance
GPT	0.612	0.509
LLaMA	0.684	0.470
Gemini	0.526	0.536
Qwen	0.407	0.480
Phi	0.717	0.395
Zephyr	0.684	0.425
DeepSeek	0.695	0.456
NVIDIA-LLaMA	0.580	0.407
Gemma	0.643	0.466
Claude	0.634	0.486

On the other hand, the models with the highest relevance scores were Gemini (0.536) and GPT (0.509). The topics produced by these models showed stronger alignment with the original texts, yielding more coherent and contextually appropriate outputs. Notably, despite Gemini's high relevance score, its diversity score was relatively low (0.526), indicating a tendency to produce repetitive or pattern-bound topics. Similarly, although Phi demonstrated high diversity, its low relevance score (0.395) suggests that some models may generate original yet contextually detached topics.

In conclusion, an inverse relationship between diversity and relevance was generally observed across models. This finding highlights the importance of carefully selecting the appropriate model based on the specific objective of the task. Depending on whether the goal is to maximize diversity or relevance, the performance and suitability of the chosen model may vary significantly.

Diversity. As shown in Figure 5, the highest diversity scores were achieved by the Phi (0.717) and DeepSeek (0.695) models. This indicates that these models are capable of generating a broader range of topics with less internal repetition. However, an increase in diversity may also lead to reduced topical relevance. For example, the Phi model has a notably low relevance score (0.395), suggesting that the generated topics, while varied, may stray from the content of the original text. The lowest diversity score was observed for the Qwen model (0.407), indicating a tendency to produce similar topics for the same article.

Relevance. As shown in Figure 6, the highest average relevance scores were observed in the Gemini (0.536) and GPT (0.509) models. The topics generated by these models are more consistent with and better

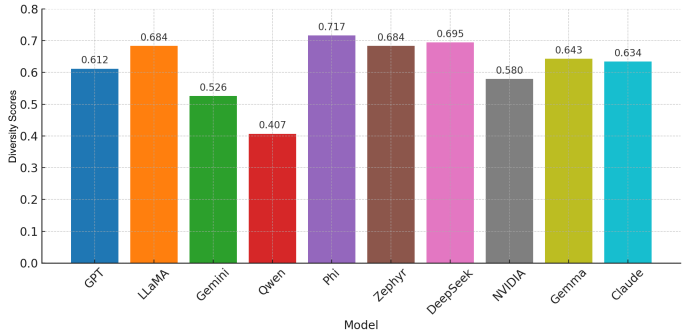

Figure 5. Diversity scores of the LLM models.

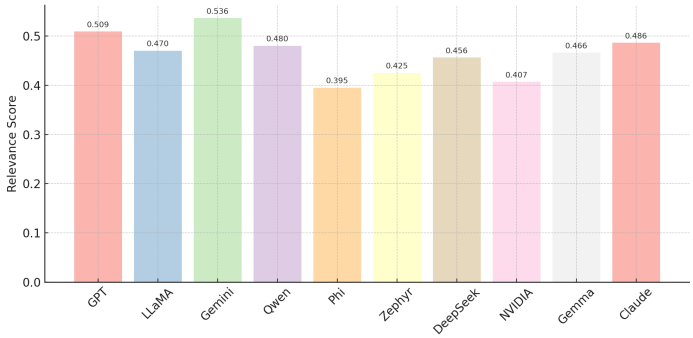

Figure 6. Relevance scores of the models.

aligned to the content of the original article abstracts. In contrast, the Phi (0.395) and NVIDIA-LLaMA (0.407) models yielded weaker results in terms of content relevance.

Cosine Similarity. The cosine similarity matrix presented in Figure 7 illustrates the semantic closeness between topic modeling outputs generated by different LLMs. This analysis was conducted by vectorizing the topic representations produced by each model and comparing them with those from the other models. High similarity scores indicate that the corresponding models tend to group texts around similar themes, while low scores suggest that the topic distributions differ either structurally or semantically.

Notably, high similarity scores were observed among the **DeepSeek**, **NVIDIA-LLaMA**, **Claude**, and **Gemma** models (e.g., DeepSeek–Gemma: 0.66; NVIDIA-LLaMA–Claude: 0.62), suggesting that these models generated closely aligned thematic outputs during the topic modeling process. This clustering effect may be attributed to factors such as shared training datasets, recent model architectures, or similar attention mechanisms applied during processing.

GPT, **LLaMA**, and **Gemini** also demonstrated relatively high similarity among themselves (e.g., GPT–Gemini: 0.64), indicating that the topic structures generated by these models reflect a somewhat distinct thematic framework compared to the aforementioned group.

On the other hand, the **Zephyr** model demonstrated consistently low similarity levels with all other models (approximately in the 0.27–0.28 range), suggesting that it follows a different representational strategy in topic modeling. This implies that the model tends to generate more unique topic distributions or produces themes that overlap less with those of other models.

The **Qwen** and **Phi** models, in contrast, showed moderate levels of similarity. They are not strongly integrated into any specific cluster, nor do they exhibit fully isolated behavior. This indicates that these models generate somewhat distinct topics while still maintaining partial thematic overlap with others.

In conclusion, this similarity analysis reveals that the topic generation tendencies of LLMs can be modeled and that latent thematic relationships among models can be uncovered. Such analyses are particularly valuable for determining which models are suitable for collaborative use in tasks such as LLM-based text classification, content summarization, or topic-driven information retrieval systems.

Entropy. Entropy is an important metric that reflects how concentrated or dispersed the word distribution of a topic is. A low entropy value indicates more focused and interpretable topics, whereas a high entropy value implies broader and more ambiguous topic distributions. Below, the entropy scores of each model are interpreted individually:

- **Phi** ($H = 4.02$): This model has the highest entropy score, indicating that the generated topics are highly dispersed and ambiguous, which may reduce their interpretability.
- **DeepSeek** ($H = 3.88$): Another model with high entropy. While the topics cover broader themes, this may come at the cost of reduced semantic clarity.
- **Zephyr** ($H = 3.75$): Produced relatively general topics but still maintained a certain degree of semantic focus.
- **Claude** ($H = 3.60$): With moderate entropy, this model strikes a balance between originality and breadth.
- **NVIDIA-LLaMA** ($H = 3.55$): Similar to Claude, it generated thematically balanced and relatively

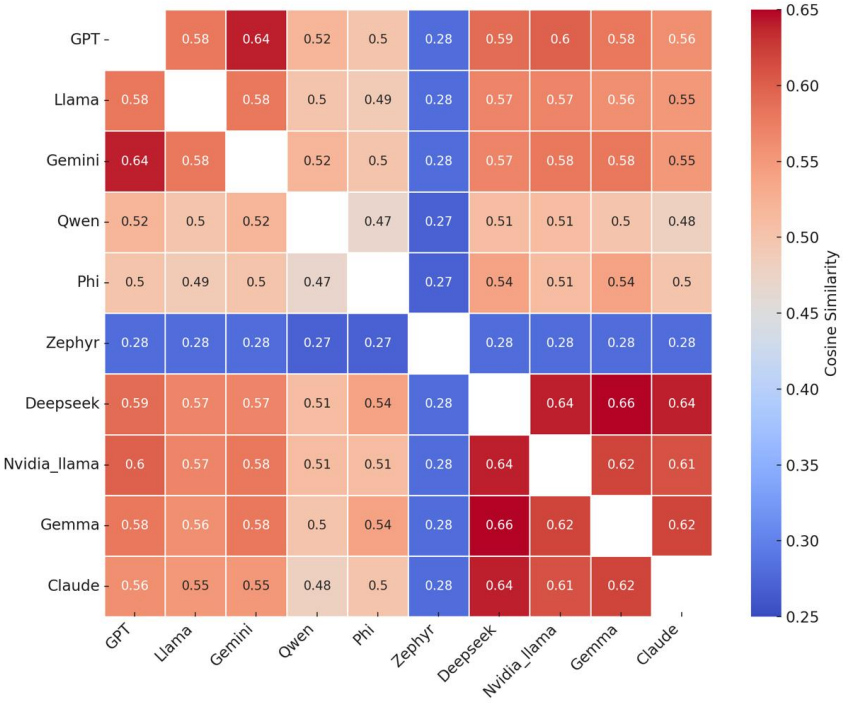

Figure 7. Cosine similarity scores of the models.

coherent topics.

- **LLaMA** ($H = 3.40$): With a lower entropy score, the model produced more focused and clearer thematic content.
- **Gemma** ($H = 3.32$): The topics it generated were reasonably dense and focused, offering satisfactory interpretability.
- **GPT** ($H = 3.21$): Produced more compact and structured topics, which facilitated easier interpretation.
- **Gemini** ($H = 3.10$): Generated more distinct topics with well-defined themes and lower variance.
- **Qwen** ($H = 2.98$): This model had the lowest entropy score, indicating that it produced very clear and narrowly scoped topics. However, such content may sometimes suffer from excessive repetition or drift from broader context.

These findings suggest that models such as Gemini and Qwen may be more suitable for applications that require high interpretability, whereas models like Phi and DeepSeek are better suited for scenarios where thematic diversity is prioritized. The distribution of entropy scores across models is visually summarized in Figure 8.

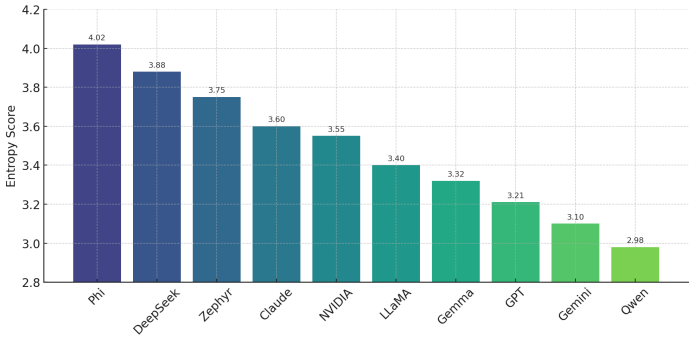

Figure 8. Entropi scores of the models.

In this section, the performance of LLMs in the topic modeling task is comprehensively evaluated. The analyses show that each model excels in certain metrics while exhibiting weaker performance in others. Based on the experimental results, an overall evaluation of the LLMs is presented in Table 2.

The **DeepSeek** model achieved the second-highest diversity score (*diversity* = *0.695*), indicating strong performance in generating original and varied topics. However, its relevance score (*relevance* = *0.456*) is comparatively lower. This suggests that while the model is capable of producing creative content, the generated topics may have weaker alignment with the source text.

Similarly, the **Qwen** model produced coherent and meaningful topics that aligned well with the input text, as reflected in its high relevance score (*relevance*

$= 0.480$). However, its relatively low diversity score ($diversity = 0.407$) indicates a tendency to generate similar and repetitive themes.

The **Phi** model achieved the highest diversity score among all LLMs ($diversity = 0.717$), demonstrating its ability to generate a wide range of diverse topics. However, its relevance score ($relevance = 0.395$) is among the lowest, suggesting that the model may have limited capability in establishing contextual coherence with the source text.

GPT, overall, exhibited a balanced performance profile. Both its diversity score (0.612) and relevance score (0.509) were above average, indicating that the model is capable of generating topics that are both creative and contextually coherent.

Among the other models, **Gemini** stood out with a high relevance score (0.536), suggesting its ability to establish meaningful relationships with the input text. However, its diversity score (0.526) remained at a moderate level, indicating that the topics it produced were somewhat limited in thematic range.

The **Zephyr** model demonstrated high diversity (0.684), but its lower relevance score (0.425) suggests that the contextual coherence of the generated topics may be limited.

Similarly, the **NVIDIA-LLaMA** model achieved a good diversity score (0.580); however, its low relevance score (0.407) indicates weaker performance in producing meaningful and consistent content.

The **Claude** model delivered a balanced performance in terms of both diversity (0.634) and relevance (0.486). While it did not dominate in any single metric, it offered a stable structure suitable for general-purpose use.

The **Gemma** model also exhibited a similarly balanced profile, with a diversity score of (0.643) and a moderate relevance score (0.466).

The **LLaMA** model attained a high diversity score (0.684), but its relevance score (0.470) was slightly below average. This suggests that while the model can generate content across a wide range of themes, the connection to the source text may be relatively weak.

This evaluation provides valuable insights into how the topic modeling performance of LLMs may vary under different scenarios and use cases.

To empirically demonstrate the task-aligned behavior of different LLMs, we compared the outputs of Phi and GPT for a representative abstract on "breast cancer detection technologies." Phi produced the following five topics: (1) Emerging diagnostic modalities, (2) Biomarker-driven early detection, (3) Multimodal treatment frameworks, (4) Genetic profiling in oncology, (5) Patient-centric innovation pathways.

Conversely, GPT generated the following: (1) Techniques in breast cancer screening, (2) Utilization of MRI and mammography, (3) Role of BRCA mutations in diagnosis, (4) Stratification of patient risk profiles, (5) Limitations of current diagnostic workflows.

While the topics generated by Phi exhibit high lexical and thematic dispersion—characteristics desirable in exploratory or hypothesis-generating contexts—GPT's topics exhibit greater contextual anchoring and semantic specificity. This empirical contrast supports our claim that LLMs exhibit differential strengths, and thus, the selection of a suitable model should be guided by the intended purpose of the task.

In this context, scenarios in which a single model achieves both high diversity and high relevance are quite limited. This highlights the importance of multi-model comparisons in LLM-based topic modeling applications.

According to the data, in terms of topic diversity, the **Zephyr** model produces the highest average number of topics with 7.32. In contrast, the **Claude** and **GPT** models generate the fewest topics, with averages of 4.98 and 4.99, respectively.

On the other hand, the topics generated by LLMs are analyzed in terms of word count, lexical diversity, and shared topic generation. This analysis enables a comparative evaluation not only through quantitative metrics but also based on the structural and content-related characteristics of the topics.

As shown in Figure 9, there are notable differences among the models in terms of topic length. The **Zephyr** model exhibits the highest average word count (98.5 *words*), indicating that it generates topics with more detailed and extensive expressions. In contrast, models such as **Gemma** (18.5) and **Claude** (20.2) tend to produce shorter and more concise topics. This may reflect each model's stylistic preferences and summarization capabilities.

The graph in Figure 10 shows the total number of unique words used by each model across all generated topics. The highest lexical diversity was observed in

Table 2. Overall evaluation of LLMs based on their strengths and weaknesses.

Model	Strength	Weakness
GPT	Balanced relevance level	Moderate diversity
LLaMA	High diversity	Low relevance score
Gemini	Highest relevance	Moderate diversity
Qwen	Strong in relevance	Lowest diversity
Phi	Highest diversity	Lowest relevance
Zephyr	High diversity	Low relevance
DeepSeek	Very high diversity	Moderate relevance
NVIDIA-LLaMA	Balanced diversity	Low relevance
Gemma	Moderately balanced	No clearly dominant strength
Claude	High relevance	Less diverse compared to others

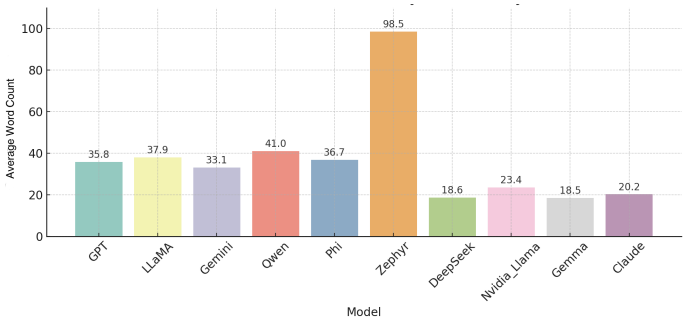

Figure 9. Average topic word count across LLM models.

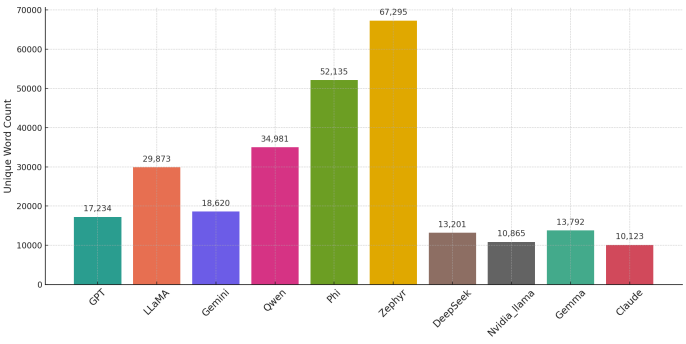

Figure 10. Total lexical diversity (number of unique words) across LLM models.

the **Zephyr** model (*67,295 unique words*), indicating that it utilizes a broader vocabulary to enhance diversity. This is followed by the **Phi** and **Qwen** models. On the other hand, models such as **Claude**, **NVIDIA-LLaMA**, and **DeepSeek**, which exhibited relatively lower lexical diversity, appear to rely on a more limited vocabulary when generating topics.

As shown in Figure 11, the analysis revealed that a total of **4,414 exact topic matches** were generated across different LLMs. This indicates that certain topics are more common and "predictable" for specific types of content, and also suggests that some models may exhibit similar tendencies in topic generation.

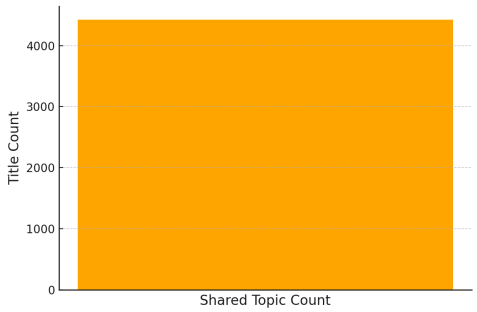

Figure 11. Number of exact topic matches across different LLMs.

Shared topic generation is considered a useful metric for identifying weaknesses in diversity and assessing the originality of generated content.

We computed the number of exact topic matches by comparing all topic strings generated by different models for each article. A match was considered valid only if two topic strings were fully identical in wording. A total of 4,414 such exact matches were observed across the dataset, indicating a moderate level of convergence among different LLMs. For example, both GPT and Claude generated the topic 'Advances in breast cancer treatment and diagnosis' for the same abstract in multiple instances.

In conclusion, it is clearly observed that LLMs differ not only in terms of numerical metrics but also in the structural characteristics of the topics they generate. Such textual analyses provide valuable insights into the content generation capabilities of models from a more qualitative perspective.

5 Conclusion

In this study, the topic modeling performance of ten different LLMs was comparatively evaluated on a large-scale collection of academic texts in the

healthcare domain. The analyses were based on both quantitative and qualitative metrics, including *diversity*, *redundancy*, *cosine similarity*, and manual evaluation. The findings revealed that models such as GPT and Claude produced meaningful and coherent topics, while models like Qwen and Phi yielded more narrowly scoped outputs due to higher redundancy rates.

An important consideration that emerged during our evaluation is the degree to which models are aligned with domain-specific content. While certain LLMs such as GPT-4 and Claude have demonstrated effectiveness in healthcare-related benchmarks due to their exposure to biomedical corpora and clinical QA datasets, others—such as Zephyr and Phi—have not been explicitly trained or fine-tuned for medical applications. Nevertheless, our results show that even these domain-agnostic models can generate diverse and occasionally insightful topics in a healthcare setting. This finding underscores the value of including both domain-informed and general-purpose LLMs in comparative evaluations, especially when assessing their generalization capabilities in high-stakes fields such as medicine. Future research may further explore how fine-tuning on specialized corpora affects both the semantic coherence and factual alignment of topics generated by LLMs in clinical domains.

In domains such as healthcare, where accuracy and explainability are of critical importance, the potential for employing LLM-based approaches is substantial. However, in assessing the generative performance of these models, it is essential to consider not only automated metrics but also expert judgment and application-specific context. This study highlights the strengths and weaknesses of current models and provides a valuable reference point for the development of more reliable and interpretable topic modeling systems in the future.

Future research can explore similar comparative analyses across different languages and disciplines to better understand the generalization capabilities of LLMs. Additionally, the evaluation of hybrid approaches or interactive, user-guided dynamic modeling techniques may further enhance the applicability of topic modeling in sensitive domains such as healthcare.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Xu, W., Hu, W., Wu, F., & Sengamedu, S. (2023). DeTiME: Diffusion-enhanced topic modeling using encoder-decoder based LLM. *arXiv preprint arXiv:2310.15296*. [Crossref]

[2] Mu, Y., Dong, C., Bontcheva, K., & Song, X. (2024). Large language models offer an alternative to the traditional approach of topic modelling. *arXiv preprint arXiv:2403.16248*. [Crossref]

[3] Maragheh, R. Y., Fang, C., Irugu, C. C., Parikh, P., Cho, J., Xu, J., ... & Achan, K. (2023, December). LLM-TAKE: Theme-aware keyword extraction using large language models. In *2023 IEEE International Conference on Big Data (BigData)* (pp. 4318-4324). IEEE. [Crossref]

[4] Kapoor, S., Gil, A., Bhaduri, S., Mittal, A., & Mulkar, R. (2024). Qualitative insights tool (qualit): Llm enhanced topic modeling. *arXiv preprint arXiv:2409.15626*. [Crossref]

[5] Rosenfeld, A., & Lazebnik, T. (2024). Whose llm is it anyway? linguistic comparison and llm attribution for gpt-3.5, gpt-4 and bard. *arXiv preprint arXiv:2402.14533*. [Crossref]

[6] Asmussen, C. B., & Møller, C. (2019). Smart literature review: a practical topic modelling approach to exploratory literature review. *Journal of Big Data, 6*(1), 1-18. [Crossref]

[7] Tan, Z., & D'Souza, J. (2025). Bridging the Evaluation Gap: Leveraging Large Language Models for Topic Model Evaluation. *arXiv preprint arXiv:2502.07352*. [Crossref]

[8] Pham, C. M., Hoyle, A., Sun, S., Resnik, P., & Iyyer, M. (2023). Topicgpt: A prompt-based topic modeling framework. *arXiv preprint arXiv:2311.01449*. [Crossref]

[9] Wang, H., Prakash, N., Hoang, N. K., Hee, M. S., Naseem, U., & Lee, R. K. W. (2023, December). Prompting large language models for topic modeling. In *2023 IEEE International Conference on Big Data (BigData)* (pp. 1236-1241). IEEE. [Crossref]

[10] Isonuma, M., & Yanaka, H. (2024). Comprehensive Evaluation of Large Language Models for Topic Modeling. *arXiv preprint arXiv:2406.00697*. [Crossref]

[11] Arora, R. K., Wei, J., Hicks, R. S., Bowman, P., Quiñonero-Candela, J., Tsimpourlas, F., ... & Singhal,

K. (2025). Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*. [Crossref]

[12] Purpura, A. (2018, August). Non-negative Matrix Factorization for Topic Modeling. In *DESIRES* (p. 102).

[13] Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the american statistical association*, *101*(476), 1566-1581.

[14] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

[15] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*. [Crossref]

[16] Anthropic. (2024, May 21). *Mapping the mind of a large language model*. Anthropic. Retrieved from https://www.anthropic.com/research/mapping-mind-language-model

[17] Team, G., Anil, R., Borgeaud, S., Alayrac, J. B., Yu, J., Soricut, R., ... & Blanco, L. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*. [Crossref]

[18] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... & Scialom, T. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*. [Crossref]

[19] Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., ... & Zhu, T. (2023). Qwen technical report. *arXiv preprint arXiv:2309.16609*. [Crossref]

[20] Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., ... & Zhang, Y. (2024). Phi-4 technical report. *arXiv preprint arXiv:2412.08905*. [Crossref]

[21] Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., ... & Wolf, T. (2023). Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*. [Crossref]

[22] Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., ... & Zou, Y. (2024). Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*. [Crossref]

[23] Wang, Z., Lorraine, J., Wang, Y., Su, H., Zhu, J., Fidler, S., & Zeng, X. (2024). Llama-mesh: Unifying 3d mesh generation with language models. *arXiv preprint arXiv:2411.09595*. [Crossref]

[24] Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., ... & Kenealy, K. (2024). Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*. [Crossref]

[25] Dieng, A. B., Ruiz, F. J., & Blei, D. M. (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, *8*, 439-453. [Crossref]

[26] Sievert, C., & Shirley, K. (2014, June). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63-70).

[27] Aletras, N., & Stevenson, M. (2014, April). Measuring the similarity between automatically generated topics. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers* (pp. 22-27).

[28] Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379-423. [Crossref]

[29] Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. In *Handbook of latent semantic analysis* (pp. 439-460). Psychology Press.

[30] Azher, I. A., Seethi, V. D. R., Akella, A. P., & Alhoori, H. (2024, December). 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* (pp. 1-12). [Crossref]

[31] Onan, A., & Çelikten, T. (2024, July). Evaluating the Coherence and Diversity in AI-Generated and Paraphrased Scientific Abstracts: A Fuzzy Topic Modeling Approach. In *International Conference on Intelligent and Fuzzy Systems* (pp. 149-157). Cham: Springer Nature Switzerland. [Crossref]

[32] Kherwa, P., & Bansal, P. (2020). Topic modeling: a comprehensive review. *EAI Endorsed Trans. Scalable Inf. Syst.*, *7*(24), e2.

[33] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38-45). [Crossref]

[34] Çelikten, T., & Onan, A. (2025). Topic modeling through rank-based aggregation and LLMs: An approach for AI and human-generated scientific texts. *Knowledge-Based Systems*, *314*, 113219. [Crossref]

[35] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

[36] Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... & Oliphant, T. E. (2020). Array programming with NumPy. *nature*, *585*(7825), 357-362. [Crossref]

[37] McKinney, W. (2010). Data structures for statistical computing in Python. *scipy*, *445*(1), 51-56.

[38] Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*. [Crossref]

[39] BioMed Central, *BioMed Central - Open Access Publisher*, Springer Nature. Available at: https://www.biomedcentral.com

Muhammet Bora Meram received his B.Sc. degree in Software Engineering from Manisa Celal Bayar University, Turkey, in 2025. He is currently an IT Intern at GUESS, where he is involved in internal network monitoring, ERP systems, and iOS mobile development. He has also completed internships in cybersecurity and database management, both in Turkey and the Netherlands. His research interests include machine learning, natural language processing, large language models, and mobile application development. He has worked on several academic and applied projects and aims to further pursue his career in software development and AI technologies. (Email: borameram411@hotmail.com)

Çağatay Kalkan received his B.Sc. degree in Software Engineering from Manisa Celal Bayar University, Manisa, Turkey, in 2025. He completed an internship at Hayat Finans, where he worked on iOS mobile application development. His research interests include mobile application development, iOS ecosystem, and user-centric interface design. He is currently working on publishing his own iOS applications on the App Store and continues to improve his skills in Swift and modern iOS development frameworks. He aims to pursue a professional career in iOS development and mobile technologies. (Email: kalkancag@gmail.com)

Tuğba Çelikten received the M.Sc. degree in software engineering from Manisa Celal Bayar University, Manisa, Turkey, in 2023. She is currently pursuing her Ph.D. degree in computer engineering at İzmir Katip Çelebi University, İzmir, Turkey. Her research interests include artificial intelligence, natural language processing, text mining and language models. She has contributed to national research projects and serves as a reviewer for scientific conferences in the field of artificial intelligence. (Email: tugba.celikten@cbu.edu.tr)

Aytuğ Onan (Editor-in-Chief, ICCK) received the Ph.D. degree in computer engineering from Ege University, İzmir, Turkey, in 2016. Since 2024, he has been serving as a professor at the Department of Computer Engineering, İzmir Katip Çelebi University, İzmir, Turkey. His research interests include artificial intelligence, data science, natural language processing, text mining and large language models. He has published numerous articles in international journals and conferences and actively participates in national and international research projects. (Email: aytug.onan@ikcu.edu.tr)