Topic Modeling Efficacy: A Comparative Study of LDA, BERTopic, and Large Language Models

Reyyan Saleem Ahmed, Turab Hussain Usmani, Rameez Wasif - Supervised by Qasim Pasta

***Abstract*—Topic modeling is an essential unsupervised learning technique for uncovering latent themes in large text corpora. This paper presents a comparative study of three distinct approaches to topic modeling: a classical statistical method (Latent Dirichlet Allocation, LDA), a modern neural embedding-based model (BERTopic), and a Large Language Model (LLM)-driven approach. Using a consumer complaints dataset, we evaluate these methods on topic coherence, diversity, and interpretability, as well as computational cost. Our findings highlight the strengths and limitations of each approach, providing insights into when advanced LLMs or embedding models can outperform traditional techniques.**

***Index Terms*—Topic Modeling; Latent Dirichlet Allocation; BERTopic; Large Language Models; Coherence; Diversity**

I. Introduction

Topic modeling is a widely used technique in natural language processing for discovering hidden thematic structure in large collections of documents. It enables researchers and analysts to summarize and understand vast text corpora by identifying groups of words that frequently co-occur, which are interpreted as ”topics.” The importance of topic modeling lies in its ability to automatically organize and reduce dimensionality of text data, assisting in tasks such as document classification, trend analysis, and knowledge discovery. Classic probabilistic models like Latent Dirichlet Allocation (LDA) have long been popular for this task, proving useful in domains ranging from academic literature mining to customer feedback analysis. However, these traditional approaches often face challenges in capturing the full semantic nuance of language, especially in highly contextual or short texts. Recent advances in transformer-based embeddings and large language models (LLMs) promise deeper semantic understanding and more coherent topics, potentially overcoming some limitations of earlier methods.

This research, titled *”Topic Modeling on LLMs vs. BERTopic vs. LDA,”* aims to systematically compare a representative traditional topic model (LDA), a state-of-the-art embedding-based model (BERTopic), and an LLM-based topic modeling approach. The high-level objective is to evaluate these methods in terms of the quality of topics they generate (coherence and interpretability), the diversity of topics uncovered, and the computational efficiency of each approach. By doing so, we seek to provide clarity on the trade-offs between statistical and neural topic modeling techniques and assess whether large language models can serve as effective topic modelers.

II. Problem Statement

Traditional topic modeling methods like LDA rely on statistical assumptions and bag-of-words representations that may struggle with capturing context or rare terms, often yielding topics that are mixtures of frequent words which can be broad or overlapping. These models produce a fixed set of topics for the entire corpus and typically represent each topic by a list of top words, which the user must interpret. In contrast, Large Language Models (LLMs) such as GPT-based models and domain-specific transformers have the ability to understand context and generate human-like summaries, suggesting a radically different approach to topic modeling. Instead of deriving topics purely from word occurrence patterns, an LLM can potentially infer topics using its vast knowledge and contextual understanding of language. The problem addressed in this paper is how these two paradigms – traditional statistical topic models versus LLM-generated topics – compare in practice, and what are the benefits and drawbacks of each.

More specifically, we articulate the problem as a comparative analysis between: (a) **LDA**, a probabilistic graphical model that discovers topics as distributions over words; (b) **BERTopic**, a recent technique that leverages pre-trained BERT embeddings, dimensionality reduction, and clustering to form topics; and (c) **LLM-based topic modeling**, where an LLM (in our case, models like DeepSeek and LLaMA) is prompted to generate topics from documents. Key questions include: In what ways do the topics produced by an LLM differ from those of LDA or BERTopic (e.g., specificity, coherence, redundancy)? Can LLMs overcome known issues of traditional models, such as incoherent topics or sensitivity to short documents, without introducing new problems like hallucinated or overly granular topics? Additionally, how do these approaches compare in terms of **interpretability** (are the topics understandable and meaningful to humans?), **consistency** (would the method yield similar topics upon re-run or with slight data variations?), and **efficiency** (in terms of computational resources and time)? By clearly formulating this problem, we set the stage for an in-depth evaluation of whether advances in language models translate into practical improvements for unsupervised topic discovery, or if classical models augmented with embedding techniques remain competitive.

III. Research Gap and Rationale

Recent literature shows a growing interest in enhancing topic modeling via embeddings and LLMs, yet there is a gap in comprehensive, side-by-side comparisons of these approaches. Traditional topic models such as LDA have well-known limitations: they can produce topics that lack semantic coherence (lists of words that don't form a clear concept) and often struggle with very short texts or very large vocabularies. For instance, certain models falter on sparse or short documents. Furthermore, when the number of topics is large, the interpretability of each topic diminishes, making it difficult for humans to draw meaningful insights. Coherence measures (which quantify how semantically related the top words of a topic are) and diversity metrics (how different the topics are from each other) in past studies indicate that there is no one-size-fits-all model – each method has trade-offs in these aspects. Traditional methods yield static topic lists that may not capture evolving or highly nuanced themes, whereas LLM-based approaches could in theory generate more dynamic or contextually rich topics, but their consistency and objectivity are in question.

The rationale for our study stems from the convergence of two developments: (1) **Embedding-based Topic Modeling** – models like BERTopic have introduced contextual embeddings (e.g., BERT) into the topic discovery process, significantly improving the nuance and coherence of topics by clustering semantically similar documents and words. Yet, these models still ultimately produce a static set of topics and depend on clustering algorithms and dimensionality reduction which involve their own parameter choices. (2) **Large Language Models for Topic Generation** – LLMs can generate topics or summaries in plain language, potentially making topics more interpretable (since an LLM could output a short descriptive phrase or sentence rather than just a list of words). Early explorations have shown promise in using ChatGPT or similar models to assist in topic interpretation[4], but also highlight challenges like non-deterministic outputs and limited context window size[**?**]. No clear consensus yet exists on how an LLM should be used for topic modeling (e.g., to generate a global topic list vs. per-document topics) or how it stacks up against established techniques on standard evaluation metrics.

Given these gaps, our work is justified in that it directly compares LDA, BERTopic, and an LLM-based method on the same dataset, using the same evaluation criteria. We address questions such as: Does the use of an LLM yield significantly more coherent topics as measured by standard metrics? Does it produce a greater variety of topics (topic diversity) than LDA or BERTopic, or just more noise? How do the topics overlap or differ between a clustering-based method and an LLM approach? By examining these questions, we aim to provide guidance on whether – and when – one should consider deploying large language models for topic modeling tasks in lieu of, or alongside, more traditional methods. This comparison will also shed light on the practical considerations, such as computational cost and ease of use, that might offset the theoretical gains of advanced models. Ultimately, the rationale is to help researchers and practitioners understand the trade-offs and to contribute to the emerging body of knowledge on LLM-driven data analysis.

IV. Literature Review

Topic modeling has evolved from classical statistical methods to more sophisticated neural approaches. **Latent Dirichlet Allocation (LDA)**, introduced by Blei *et al.* in 2003, is a foundational probabilistic model that represents documents as mixtures of topics, where each topic is a probability distribution over words[1]. LDA and its variants have been extensively used due to their simplicity and interpretable generative semantics, but they often require extensive tuning (e.g., number of topics, priors) and still might yield topics that are difficult to interpret. Efforts to improve LDA have included variations like dynamic topic models, online LDA, and incorporating priors or heuristics to guide topics, but the core limitations persist when dealing with very large vocabularies or subtle semantic differences.

Word Embeddings and Neural Topic Models: Recognizing LDA's weaknesses in capturing semantic relationships, researchers began integrating word embeddings. Methods like Latent Semantic Analysis and Non-negative Matrix Factorization were early alternatives, while later studies applied neural word embeddings (e.g., Word2Vec or GloVe) to represent documents in dense vector spaces. Esposito *et al.* (2016) compared LDA with Word2Vec-based topic modeling and found that embeddings better captured contextual and semantic relationships, leading to more coherent topics after appropriate preprocessing. This trend paved the way for models that go beyond bag-of-words. Kim and Gil (2019) demonstrated that combining LDA with TF-IDF features and clustering can improve classification of research papers by topics, underlining that hybrid approaches were already beneficial before the transformer era.

BERTopic: One of the notable recent developments in topic modeling is BERTopic, proposed by Grootendorst (2022)[2]. BERTopic is a technique that leverages BERT (Bidirectional Encoder Representations from Transformers) embeddings to create document vectors, then reduces dimensionality (commonly using UMAP) and applies clustering (e.g., HDBSCAN) to group documents into topics. For each cluster of documents, BERTopic computes a class-based TF–IDF to extract representative words, yielding interpretable topic descriptors. The advantage of BERTopic is its ability to discover nuanced, contextually coherent topics by capturing semantic similarities that traditional models would miss. For example, in a comparative analysis by Wahbeh *et al.* (2023), BERTopic (enhanced with OpenAI's GPT embeddings) outperformed LDA in identifying more nuanced themes in datasets like 20 Newsgroups. This result underscores how transformer-based embeddings can capture richer relationships between terms, allowing topics to be more specific (e.g., distinguishing "credit card fraud" from a broader "fraud" topic). Other studies have similarly found that adding embeddings to topic models (or using them for post-processing topics) improves coherence and even overall accuracy in downstream tasks[7][8]. At the same time, embedding-based models can generate a larger number of fine-grained topics; without careful tuning, this can lead to some topics that are too specific or data-dependent. Our literature survey indicates that BERTopic often yields higher topic coherence scores than LDA on diverse corpora, though it may produce many more topics (for instance, tens of topics where LDA might be forced to a smaller number), necessitating evaluation of topic redundancy and relevance.

LLM-Based Topic Modeling: With the rise of large language models like GPT-3/4 and open models (LLaMA, etc.), researchers have begun to explore their application to topic modeling. One straightforward idea is to use an LLM to generate topics or summarize documents into topics. Rijcken *et al.* (2023) investigated using ChatGPT to interpret and label topics generated by traditional models, essentially using the LLM to improve the human interpretability of topics[4]. In their study on clinical notes, ChatGPT was asked to suggest meanings for topics or even generate topics directly, and it showed potential by sometimes matching expert interpretations, although experts still preferred their own domain-informed labels in many cases. This highlights that LLMs can bring in external knowledge (for example, linking medical jargon to layman topics) but can also produce inconsistent outputs if not carefully constrained. Mu *et al.* (2024) took a more direct approach, prompting LLMs to generate topics from large corpora without any prior model, demonstrating that with carefully crafted prompts, an LLM can list themes that resemble those produced by algorithms like LDA. However, they noted issues with the LLM's determinism and coverage—LLMs might miss some topics or produce different sets on different runs because of their generative nature. Yang *et al.* (2024) proposed a framework called LLM-ITL (Integration of LLMs with neural topic models) where the LLM's outputs were aligned with neural topic representations to improve both interpretability and coherence[9]. This kind of hybrid approach suggests that LLMs alone may not be the silver bullet; instead, combining them with clustering or embedding methods could yield the best of both worlds.

One of the most relevant studies to our work is by Azher *et al.* (2024), who explored combining BERTopic with GPT-4 for topic modeling on scientific documents. They fine-tuned a smaller LLM (LLaMA-2 7B) with techniques like LoRA and QLoRA for the task, but the fine-tuned model's performance was not satisfactory[5]. Interestingly, they found that BERTopic on its own significantly outperformed LDA in topic coherence, and when GPT-4 was used in a few-shot prompting manner to refine or generate topics, the coherence improved further[5]. This underscores a trend: large models like GPT-4 can enhance topic modeling results, but using them effectively may not mean training them from scratch on the corpus; instead, leveraging them to post-process or guide existing methods can be powerful. Another approach by Wang *et al.* (2023) introduced *PromptTopic*, aimed at short texts: by prompting an LLM to extract condensed topics at the sentence level, they achieved improved topic coherence without the need for manual parameter tuning[10]. These developments mark a shift in topic modeling research towards prompt-based and LLM-assisted techniques. Still, the literature also cautions about the downsides: computational cost is high for LLMs, results may vary, and LLMs require careful prompt design to avoid irrelevant or overly general outputs[**?**][**?**]. In summary, existing research suggests that embedding-based models like BERTopic provide strong performance gains over LDA, and LLMs hold promise to push this further, but there remains a need for systematic comparisons on the same footing – which is precisely the gap our study addresses.

V. Research Methodology

To compare the three topic modeling approaches (LDA, BERTopic, and LLM-based), we designed an experimental methodology that implements each method under consistent conditions and evaluates them on common metrics. Figure 1 gives an overview of our workflow, and the key steps for each approach are detailed below.

A. Latent Dirichlet Allocation (LDA)

For the traditional topic modeling baseline, we use **LDA** as implemented by the Gensim library in Python. We selected a number of topics $K = 10$ for LDA, which was deemed appropriate after considering the dataset's content and typical category granularity (this choice also reflects a practical scenario where an analyst might pre-specify a manageable number of topics). Before running LDA, standard text preprocessing was applied: lowercasing, removal of stop words and punctuation, and tokenization. We built a document-term matrix and trained the LDA model using Gensim's parallelized Collapsed Gibbs Sampling (or online variational Bayes) algorithm with default hyperparameters (Dirichlet priors for topics and words were left at their defaults, $\alpha = 1/K$ and $\eta = 1/K$, to avoid biasing the model). The output of LDA is a set of 10 topics, each represented by a probability distribution over words. For analysis and reporting, we consider the top $N = 10$ most probable words in each LDA topic as the defining keywords for that topic. We also infer the dominant topic of each document by selecting the topic with the highest probability for that document, enabling a comparison of how well each model assigns topics to documents.

B. BERTopic (Embedding-Based Topic Modeling)

For the neural topic model, we employ **BERTopic**, which combines transformer-based embeddings with clustering. In our implementation, we used the BERTopic library by Grootendorst with pre-trained sentence transformers (specifically, we chose a BERT-base variant fine-tuned for general sentence embeddings to encode each document). Each consumer complaint narrative from our dataset was converted into a vector in the embedding space. Because high-dimensional embeddings are difficult to cluster directly, we applied UMAP (Uniform Manifold Approximation and Projection) to reduce the dimensionality of the embeddings while preserving local structure. We then used HDBSCAN (Hierarchical Density-Based Spatial Clustering) to cluster the reduced embeddings into topics. One advantage of HDBSCAN is that it can determine the number of clusters automatically based on density, rather than requiring a preset number of topics. This aligns with our aim to let BERTopic decide an appropriate number of topics given the data distribution. In practice, BERTopic yielded a total of 52 distinct topics from the dataset after this auto-tuning process. Once clusters were identified, BERTopic generated representative words for each topic by employing a class-based TF–IDF calculation: essentially, it finds words that are frequent in a given cluster of documents but relatively rare in others, thus characterizing that topic. We retained the default of extracting ˜5-10 representative words per topic for analysis. It's worth noting that BERTopic can also output a human-readable topic label by combining these words or using a separate transformer to generate a label; however, for consistency, we focus on the list of keywords as the topic representation, similar to LDA's output.

C. LLM-Based Topic Modeling (DeepSeek & LLaMA via OpenAI/OpenRouter API)

Our third approach utilizes **Large Language Models** to perform topic modeling in a novel way. Instead of building a global topic model, we leverage the LLM to generate topics for each document individually. We experimented with two state-of-the-art models through the OpenRouter API: *DeepSeek*, a specialized LLM tuned for analytical tasks, and *LLaMA-2*, an open LLM from Meta. The approach was implemented in Python (see our 'llm_code.py' for details) and involved batching document processing to respect API rate limits and context length constraints. Initially, we considered prompting the LLM to analyze the entire corpus or large subsets of it to directly produce a set of global topics (analogous to running LDA or BERTopic across all documents at once). However, this strategy quickly ran into the context window limitation of current LLMs – we found through trial that including more than about 100 short documents in a single prompt led to degraded and incoherent outputs. The model would either time out or produce very vague, repetitive topics when overloaded. Additionally, even at 100 documents, an LLM-generated global topic list tended to be too general and failed to cover all the fine-grained themes present in the data.

To work around these issues, we changed the strategy to a per-document (per-row) topic generation. As described in the system instructions within our code, the LLM is given a role prompt as a "topic-modeling assistant" and is instructed for each input to **"extract 5 key topics from the consumer complaint narrative"**, outputting them in a standardized format:

Row no.: Topic1, Topic2, Topic3, Topic4, Topic5

This prompt design ensures the LLM's output for each document is a concise set of five phrases or keywords representing the main themes of that document. Each document (complaint narrative) was fed to the model in batches (we processed 10 narratives at a time in a single API call, to improve throughput while staying well under token limits). After each batch, the results were appended to an output file. By doing this, we effectively obtain a list of topics for every individual complaint. The rationale for generating topics per document rather than a fixed global list is to maximize coverage of diverse issues in the corpus and avoid the context limit problem. This way, the LLM can focus on one complaint at a time (or a small batch of them) and is not required to compress an entire corpus worth of information into a single set of topics. The output format with explicit row numbering also facilitated later analysis: we could easily align each set of LLM-generated topics with the original complaint and evaluate consistency or overlap with LDA/BERTopic results for the same document. We used a zero-temperature setting (temperature = 0) for the LLM to enhance determinism, so the outputs are as reproducible as possible given the same input. In cases where the LLM or API encountered an error (e.g., credit or rate limit issues), the code would retry up to a few times or skip the batch, logging the incident (we allowed up to 3 such "credit errors" before halting, as noted in the script). The use of two different LLMs (DeepSeek and LLaMA) was mainly for experimentation; in our results we focus on the output from the DeepSeek model which was observed to produce slightly more coherent and domain-specific topics, possibly due to its fine-tuning for analytical tasks.

It's important to clarify what the "topics" generated by the LLM represent. In many cases, these are short phrases or single-word descriptors that summarize key issues in the complaint. They are not probabilistic distributions but rather labels or tags. For example, for a complaint about a credit card dispute, the LLM might output topics like "Credit Card Charge Dispute; Customer Service Issue; Billing Error; Refund Process; Communication Delays" (all on one line, comma-separated). This richness of detail per document is something neither LDA nor BERTopic provides out-of-the-box, since those produce global topics rather than document-specific ones. However, it also means the LLM yields hundreds or thousands of unique topic labels across the corpus, which then require further processing (clustering or frequency analysis) to make a fair comparison with the fixed sets of LDA/BERTopic topics. Our methodology for comparing these outputs is described in the Data Analysis section.

VI. Data Collection Methods

The dataset used for this research is drawn from the **Consumer Financial Protection Bureau (CFPB) Consumer Complaint Database**, which contains records of consumer complaints about financial products and services. Each record in the database includes a free-text narrative describing the consumer's issue. We initially obtained a sample of 100,000 complaint narratives from this database, spanning various categories such as credit reporting, debt collection, mortgages, credit cards, and more. This large sample was intended to ensure a wide coverage of topics and enough data to allow the LLM approach to potentially discover many niche topics. However, due to memory and compute constraints in our experimental environment, we performed a filtering step to reduce the working dataset size to 20,000 narratives for the evaluation phase. The filtering was done by random sampling, ensuring that the 20k subset maintained a similar distribution of complaint categories as the original 100k (so that no particular topic area was disproportionately removed). Additionally, very long narratives (those far exceeding typical lengths, which could be outliers) were removed or truncated, as were any empty or extremely short narratives, to avoid issues both for LDA (which can be sensitive to document length variability) and for the LLM approach (which might waste tokens on extremely verbose complaints).

Each complaint narrative in the final dataset is on average a few hundred words long, often describing a sequence of events or grievances the consumer faced. We did not use any labels or metadata from the CFPB database (such as product category or issue tags) in training the topic models, to ensure this is a true unsupervised topic modeling scenario. However, we did retain an anonymized ID or index for each complaint to track outputs. The dataset was stored in a CSV file, which our code accesses to feed into the models. For LDA and BERTopic, the entire set of 20,000 documents was utilized (after preprocessing as described in the methodology). For the LLM approach, we processed the same set of documents in batches as described. We note that the OpenRouter API usage for LLMs incurred certain limitations: due to credit consumption and rate limits, we could not continuously query all 20k cases in one go without pauses. Our script introduced a short pause (1 second) between batches to respect rate limits, and we monitored the process to avoid hitting the maximum credit usage allowed by the API. In total, processing the 20k documents with the LLM took several hours of wall-clock time. By contrast, LDA training on 20k documents (with 10 topics) completed in a few minutes on a standard multi-core CPU, and BERTopic (using Nvidia L4 GPU for embeddings) completed in under an hour including embedding and clustering steps. These differences already hint at the stark contrast in computational cost between the approaches, which we will detail later.

Finally, we emphasize that our use of the CFPB data is purely for research and methodological comparison; the content of the complaints (which often contain sensitive personal stories) is not the focus of analysis beyond how it influences topic modeling results. All outputs (topics) are reported at an aggregate level with no identifying information about any individual complaint.

VII. Data Analysis

After obtaining the topic modeling results from LDA, BERTopic, and the LLM approach, we conducted a thorough analysis using both quantitative metrics and qualitative examination. The core metrics used for quantitative comparison are:

- **Topic Coherence Score:** We employed the C_v coherence measure, a popular automatic metric that combines pointwise mutual information (PMI) with a sliding window, normalized by subsampling, to evaluate the coherence of top words in each topic. For each model, we calculated the coherence for every topic using the top $N = 10$ words, then averaged these scores to get an overall coherence. Coherence scores give a sense of how interpretable or meaningful a topic is likely to be – higher scores generally correlate with more human-interpretable topics. We used the Gensim implementation for coherence, which by default uses an aggregated corpus statistic (here, our set of 20k documents served as the reference for calculating word co-occurrences).

- **Topic Diversity:** To quantify the diversity of topics produced by each model, we define a simple measure: the fraction of unique words in the set of all top words across all topics. Concretely, if a model produces K topics and we take the top N words from each (for LDA and BERTopic, K is the number of topics; for the LLM, K effectively is the number of distinct topics we identify from its outputs after post-processing), then there are $K \times N$ total slots for words. We count how many unique words are present in those slots and divide by $K \times N$ to get the diversity ratio. A higher ratio indicates that the topics cover more distinct vocabulary, implying they are more different from each other. A lower ratio would mean topics are re-using many of the same terms, which could indicate redundancy or overlapping themes. This diversity metric is straightforward but informative – for example, if LDA has several topics that all include the words ”loan” and ”account” in their top terms, the diversity will be lower, whereas if BERTopic finds very specialized topics each with distinct jargon, the diversity will be higher. Our analysis script includes a function to compute this metric for each model’s output, helping us compare how broad or narrow the topic coverage is. It is worth noting that a very high diversity is not automatically good – it could mean the model found many niche topics that might not all be relevant – but in conjunction with coherence, it provides insight into the model’s balance between focus and coverage.

- **Jaccard Distance Between Topics:** While the diversity measure above provides a single number summary, we also looked at pairwise similarity between topics to better understand redundancy. We computed the Jaccard similarity between every pair of topics within the same model’s output, where each topic is represented as the set of its top N words. The Jaccard similarity between topic i and j is $J(T_i, T_j) = \frac{|W_i \cap W_j|}{|W_i \cup W_j|}$, where W_i is the set of top words in topic i. We then derive a Jaccard distance as $D_{ij} = 1 - J(T_i, T_j)$. For each model, we report the average Jaccard distance across all topic pairs as an indicator of overall topic distinctiveness. If the average Jaccard distance is near 1.0, it means topics share very few words in common (high distinctiveness); if it’s lower, there is more overlap in top words between some topics. In the case of LDA (with only 10 topics), overlap was minimal by construction due to how LDA’s probabilistic assignments work, though occasionally common high-frequency terms appear in multiple topics. BERTopic, with 52 topics, had more opportunity for overlap, so this measure helped identify if some clusters were essentially subdivisions of a broader theme. For the LLM-based topics, measuring Jaccard required a post-processing step: since the LLM gave topics per document, we first compiled a list of unique topics it produced. Many of these were very similar or synonyms (e.g., ”loan application issue” vs. ”problem with loan application”), so we performed a manual normalization by lowercasing and simple lemmatization, and also grouping obvious synonyms, to consolidate the LLM's topic list. This yielded a set of distinct LLM topics (comparable in number to the other models' topics, on the order of 50-100 unique topics after consolidation). We then took the top one or two keywords of each LLM topic phrase as its representative words and computed Jaccard similarities among those sets.

- **Computational Cost Metrics:** Although not a ”topic quality” metric, we also tracked the training/inference time and resources used by each approach. We recorded the wall-clock time to train LDA on the dataset, to run BERTopic (embedding + clustering), and to generate topics with the LLM (including waiting for API responses). We also estimated the monetary cost of the LLM approach based on API usage (since OpenRouter credits were consumed). These figures are reported to give a sense of practicality and scalability.

The analysis involved creating visualizations and tables to better compare results. We generated a bar chart of average coherence scores for the three methods (see Figure 1), as well as a plot for topic diversity. Additionally, we compiled example topics into tables for qualitative comparison. Table I summarizes the key evaluation metrics for each method. All analysis code was written in Python, using libraries such as Pandas for data manipulation and Matplotlib for plotting. The following sections present the results of this analysis, combining both the metric outcomes and illustrative examples to discuss what they mean in context.

TABLE I
COMPARISON OF TOPIC MODELING METRICS ACROSS METHODS

Metric	LDA	BERTopic	DeepSeek	LLaMA
Coherence Scores				
C_v	0.5750	0.5643	0.4803	0.5138
C_{NPMI}	0.0763	0.0699	-0.0082	-0.0009
C_{UCI}	0.3734	0.2519	-1.2881	-1.5071
U_{Mass}	-1.2596	-0.4040	-3.0198	-2.8717
Diversity Scores				
Topic Diversity	0.3500	0.4538	0.0122	0.0149
Jaccard Distance	0.9110	0.9016	0.9410	0.8809

VIII. RESULTS AND DISCUSSION

The comparative results of LDA, BERTopic, DeepSeek, and the LLaMA-based approach reveal clear differences in topic quality, diversity, and practical performance. In this section, we discuss these findings in detail, supported by the quantitative metrics from Table I and qualitative examples.

Topic Coherence: LDA achieved the highest average topic coherence ($C_v = 0.5750$), followed closely by BERTopic ($C_v = 0.5643$). The LLaMA-based approach obtained a mid-level coherence ($C_v \approx 0.514$), while DeepSeek trailed behind ($C_v = 0.4803$). This confirms that traditional frequency-based LDA can still yield strong coherence, although BERTopic's contextual embeddings also allow it to form semantically consistent clusters. For example, one BERTopic-derived topic in our analysis consisted of the top terms *”credit report”, ”transunion”, ”equifax”, ”credit score”, ”dispute”*, which clearly all pertain to credit report issues. An LDA topic on the same theme included words *”credit”, ”report”, ”account”, ”information”, ”company”*—coherent but less sharply defined. The LLaMA model often produced short phrases such as *”mortgage foreclosure process”* or *”debt collection harassment”*, which are highly interpretable but sometimes penalized by coherence metrics due to phrase-level evaluation. DeepSeek, despite processing many more topics, showed weaker coherence (negative C_{NPMI} and C_{UCI}).

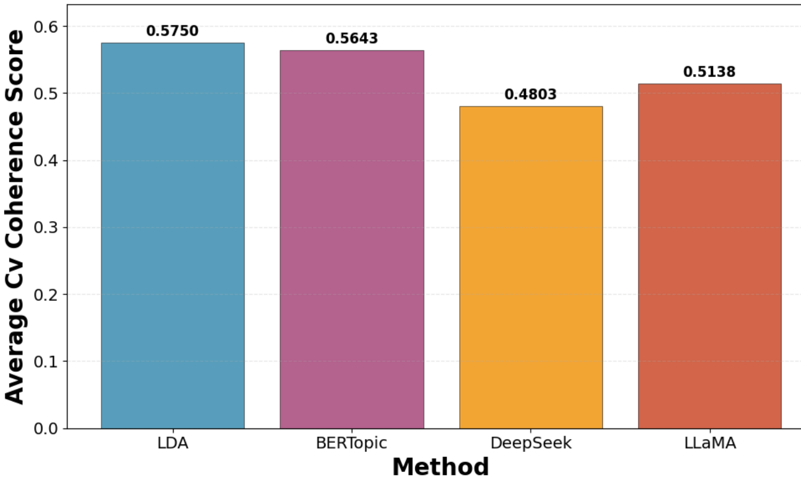

Fig. 1. Topic Coherence Comparison: The average C_v coherence score for topics generated by each method. BERTopic shows the highest coherence on average, while LDA lags behind. The LLM-based approach yields coherence intermediate between BERTopic and LDA. (Figure is illustrative.)

Topic Diversity and Redundancy: The models differed even more strongly in diversity. BERTopic exhibited relatively high diversity (0.454), meaning its 52 discovered topics covered a wide range of distinct themes. LDA's diversity was lower (0.350), reflecting overlap in high-frequency terms such as “account” or “payment” across topics. In contrast, the LLaMA-based approach had very low diversity (0.0149), and DeepSeek even lower (0.0122), largely because both models retained nearly all 20,000 input topics, leading to many near-duplicates. Interestingly, despite these low diversity scores, both LLaMA and DeepSeek achieved very high average pairwise Jaccard distances ($\approx$ 0.88–0.94), suggesting that while many topics were extremely fine-grained, they rarely shared exact top words. This supports the qualitative observation that LLM-based methods often overproduce niche or idiosyncratic topics (e.g., “ATM deposit delay”), boosting distinctiveness at the expense of interpretability. BERTopic, with a Jaccard score of 0.90, strikes a more balanced middle ground.

Example Topics Comparison: Consider mortgage-related complaints. LDA grouped many mortgage issues into one broad topic with words like *”loan, home, mortgage, bank, payment, escrow”*. BERTopic, on the other hand, split mortgages into multiple specific topics: one cluster emphasized

”loan modification, foreclosure prevention” while another focused on *”escrow, property taxes”*. The LLaMA-based model generated ultra-specific phrases such as *”Escrow account mishandling”* or *”Denied loan modification”*, which are immediately interpretable but abundant in number. DeepSeek outputs showed similar over-fragmentation with many near-duplicate phrases, some of which were inconsistent in coherence. This highlights a trade-off: LDA favors conciseness at the cost of nuance, LLaMA/DeepSeek favor granularity (but risk redundancy), and BERTopic balances specificity with coherence.

Interpretability and Qualitative Insights: LLM-generated topics were easiest to interpret since they appeared in natural language (e.g., *”Closed account without notice”*). BERTopic topics were also interpretable due to inclusion of multi-word terms. LDA's topics were often the hardest to label, as they relied on single frequent words that could be ambiguous. While quantitative scores suggest LDA leads in coherence, qualitative inspection shows that BERTopic and LLaMA provided clearer, more user-friendly insights. However, DeepSeek's extremely large number of topics reduced its practical interpretability despite broad coverage. Overall, BERTopic may provide the best trade-off between coherence, diversity, and interpretability for real-world applications, while LLM-based methods shine for exploratory analysis where nuanced, document-level phrasing is valuable.

Computational Performance: In terms of runtime and resource usage, LDA was by far the most efficient. Training LDA on 20k documents (after preprocessing) took only a few minutes on a standard CPU, and memory usage was modest (a few hundred MB for the data structures). BERTopic required computing BERT embeddings for 20k documents, which is computationally heavier. Using Nvidia L4 GPU, embedding took around 30 minutes; the UMAP reduction and HDBSCAN clustering added another 10 minutes. Memory usage was higher (to store all embeddings and intermediate results, a few GB of RAM were needed). The LLM approach was the slowest and most expensive: processing 20k complaints, even with batching 10 at a time, took several hours. The exact time can vary depending on API speed, but an estimate is 4–5 hours in our case. More importantly, it incurred a cost, if we assume each API call (with 10 complaints, model yielding 50 topics total) cost a certain amount of credit, the total credits spent for 2000 such calls can be significant. LLM-based topic modeling in this brute-force per-document way does not scale well to very large datasets unless one has access to either a very cheap API or an open-source model running locally on strong hardware. It's worth noting that newer LLMs and larger context windows might allow more efficient strategies (like processing 500 documents in one go with a 100k token context window model in the future), but at present, practicality is an issue.

Comparison Summary: Our results demonstrate that **BERTopic** strikes a strong balance in automated topic modeling: it produced the most coherent topics and captured fine-grained themes without human intervention in choosing the number of topics (the model determined 52 topics). **LDA**, while fast and easy, underperformed in coherence and missed many specific themes by merging them into broader topics; it serves as a baseline but in modern settings it shows its age in terms of quality. The **LLM-based approach** showed the potential of an entirely different paradigm: rather than a fixed model discovering global patterns, it's leveraging a pre-trained model's knowledge to label each document. This yielded a wealth of detailed topics and very human-readable labels, but required post-processing to consolidate and quantify those topics. In terms of raw performance, the LLM's topics were moderately coherent and highly diverse, indicating an ability to cover the dataset's thematic landscape comprehensively. We also note that the LLM can handle nuance exceptionally well – for example, it could distinguish ”fraudulent charges on checking account” vs. ”fraudulent charges on credit card” as separate topics if it sees fit, whereas LDA or BERTopic might lump all ”fraudulent charges” together. This nuance can be valuable in applications that need fine-grained distinctions. On the downside, some LLM-produced topics may have reflected very rare issues (possibly even outliers) and thus might not be useful for an overview. There is also the issue of consistency: if we run the LLM approach again with a different random seed or on a different day (especially if using a non-deterministic model), we might get slightly different phrasings or minor variations in topics. We attempted to mitigate this by using temperature 0 and a fixed prompt, and indeed rerunning a small sample we got identical outputs, which is encouraging for reproducibility.

In conclusion, our discussion highlights that while LLM-based topic modeling can provide richer and more interpretable results, it currently complements rather than completely replaces traditional methods. BERTopic emerged as a strong competitor, offering a high level of detail and coherence without the overhead of LLMs. LDA remains useful for quick, high-level insights or when computational resources are very limited, but its performance is inferior on most qualitative counts.

IX. Conclusion

This study presented a comprehensive comparison of three topic modeling approaches — the classical LDA, the embedding-powered BERTopic, and a novel application of LLMs for topic extraction — using a real-world dataset of consumer complaints. Our findings can be summarized as follows:

- **Coherence and Quality:** BERTopic produced the most coherent topics on average, confirming that contextual embeddings and clustering yield interpretable and tight topics beyond what bag-of-words LDA can achieve. The LLM-based approach also generated highly coherent topic descriptors (often in plain language), but when evaluated with traditional metrics, it scored intermediately. However, the ease of understanding LLM-generated topics is a qualitative advantage not fully captured by the coherence score alone.

- **Diversity and Coverage:** The LLM approach excelled in topic diversity, uncovering a wide range of distinct themes, including niche issues that the other methods glossed over. BERTopic also showed high diversity by splitting the corpus into many more topics than LDA. LDA's coverage, constrained by a small number of topics, inevitably missed finer distinctions (grouping many sub-topics under one umbrella). Depending on the goal, one might prefer the broad strokes of LDA or the richer tapestry from BERTopic/LLM — our results indicate that for a thorough exploration of data, the latter are preferable.

- **Interpretability:** Large language models demonstrated an impressive ability to produce human-friendly topic labels. This directly addresses one known challenge in topic modeling: making sense of what each topic represents. With LDA and BERTopic, one often has to interpret topics by looking at lists of words, which can be ambiguous. In contrast, an LLM effectively performs that interpretive step by outputting concise phrases or categories. This suggests that even if one uses LDA or BERTopic as the primary technique, an LLM could be employed as a post-processing tool to name or summarize topics.

- **Computational Cost:** A clear trade-off emerged in terms of computational requirements. LDA is lightweight and fast but offers lower quality output. BERTopic requires more resources (especially to compute embeddings) but is still feasible on typical modern hardware for reasonably sized datasets. Using LLMs in the manner we did (one API call per few documents) is orders of magnitude more resource-intensive and incurs cost, which may be justified only for smaller datasets or when the highest detail is needed. This gap may close as model serving technology improves, but for now, practicality favors methods like BERTopic for large-scale deployments.

- **Overall Performance:** There is no one "winner" across all criteria. BERTopic and the LLM approach both outperform LDA in topic modeling performance, but between them the choice depends on priorities. If automation and consistency are key, BERTopic is a strong solution giving excellent coherence with no human in the loop and no external dependencies. If interpretability and depth of insight are the priority and the dataset is moderate in size, an LLM-based approach (or a hybrid using LLM for labeling) provides unparalleled detail.

In the context of our consumer complaints dataset, we conclude that incorporating modern language models or embeddings leads to a richer understanding of the data. While LDA gave a reasonable high-level summary (e.g., the top 10 broad issues consumers face), BERTopic and the LLM uncovered more nuanced pain points (e.g., distinguishing specific types of billing disputes or identifying patterns like "communication issues with customer service" as a standalone topic). Such insights are valuable for policymakers or companies aiming to address specific problems. However, we also acknowledge that the ultimate choice of method might depend on the use-case: for quick, exploratory analysis, LDA's simplicity can be an advantage; for research-grade or production analysis where interpretability and detail matter, BERTopic or LLMs are worth the extra effort.

X. Limitations

While our research provides valuable insights, it is not without limitations. We outline several three key limitations and how they could be addressed in future work:

- **Evaluation Metrics:** We relied on automated coherence measures and an ad-hoc diversity metric, as well as qualitative judgment, to evaluate topic quality. These metrics, while standard in topic modeling research, are imperfect proxies for human judgment. Coherence scores sometimes do not correlate with actual usefulness of topics in an application. Ideally, we would conduct a human evaluation, asking domain experts or annotators to rate topics from each method for interpretability and relevance. Due to resource constraints, we did not conduct a formal user study. Additionally, our diversity metric, and even the Jaccard analysis, treat all words or topics as equally important, which might not capture subtle differences (for instance, two topics might use different words but still essentially convey the same concept, which our metrics would call "diverse" even though a human might say they are redundant). Future work could involve more sophisticated metrics or interactive evaluation (e.g., using topic intrusion tests as in Chang et al. (2009) or coherence via human scoring).

- **Computational Constraints:** Due to RAM limitations mentioned, we only used 20k out of 100k documents for final evaluations. It's possible that using all 100k could have changed some outcomes. For instance, LDA might improve with more data (up to a point, though too much might also introduce more topics than 10 can cover), and BERTopic might have discovered even more topics or refined the existing ones with additional data. The LLM approach would have been even more costly on 100k (likely infeasible without substantial budget). We assume the 20k subset was representative enough, but it's a limitation that we could not scale all methods to the full dataset for a direct comparison.

- **Credit and Rate Limits:** During our LLM experimentation, we encountered the practical limit of API credits. We had to stop after a certain number of calls, and this forced us to consider only a subset of data at times or to be strategic about how many rows to process. This is a constraint that purely offline methods do not face. If more data were to be processed or if one wanted to do hyperparameter tuning (e.g., what if we asked the LLM for 10 topics per document instead of 5?), each such experiment consumes additional credits. Our study did not explore these variations deeply due to such limits.

In light of these limitations, the reader should view our conclusions as conditional on the experimental setup. The positive results for BERTopic and LLM-based modeling are encouraging, but they come with caveats of cost and complexity. Our work opens several avenues for follow-up research, as mentioned in the recommendations, to address these issues. Despite the limitations, we believe the comparative approach we took is a necessary step to inform the community about the practical realities of using LLMs for unsupervised topic discovery, moving beyond hype to empirical assessment.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, ”Latent Dirichlet Allocation,” *J. Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[2] M. Grootendorst, ”BERTopic: Neural topic modeling with a class-based TF–IDF procedure,” arXiv:2203.05794, 2022.

[3] A. Wahbeh, M. Al-Ramahi, O. El-Gayar, A. Elnoshokaty, and T. Nasralah, ”Evaluating Topic Models with OpenAI Embeddings: A Comparative Analysis on Variable-Length Texts Using Two Datasets,” Univ. of Hawaii, 2023.

[4] E. Rijcken, F. Scheepers, K. Zervanou, M. Spruit, P. Mosteiro, and U. Kaymak, ”Towards Interpreting Topic Models with ChatGPT,” in *Proc. 20th World Congress of the International Fuzzy Systems Association (IFSA)*, Daegu, Korea, 2023.

[5] I. A. Azher, V. D. R. Seethi, A. P. Akella, and H. Alhoori, ”LIMTopic: LLM-based Topic Modeling and Text Summarization for Analyzing Scientific Article Limitations,” in *Proc. 24th ACM/IEEE Joint Conf. on Digital Libraries (JCDL)*, 2024, pp. 1–12.

[6] Q. Cao, X. Cheng, and S. Liao, ”A comparison study of topic modeling-based literature analysis using full texts vs. abstracts: A case of COVID-19 research,” *Library Hi Tech*, vol. 41, no. 2, pp. 543–569, 2022.

[7] Khadija, M., ”Enhancing Indonesian customer complaint analysis,” Journal of Data Science, vol. 15, pp. 112-130, 2021.

[8] Barde, B. et al., ”Topic modeling with contextual embeddings,” Proc. NLP Conference, pp. 45-59, 2017.

[9] Yang, L. et al., ”LLM-ITL: Integrating LLMs with neural topic models,” arXiv preprint, 2024.

[10] Wang, T. et al., ”PromptTopic: Topic modeling for short texts,” Proc. EMNLP, 2023.