LLM-Based Structurally Focused Topic Modeling (Draft)

Sam Mucyo

May 13, 2025

Abstract

We investigate how large language models (LLMs) can enhance topic modeling by extracting structured information from unstructured text prior to applying Latent Dirichlet Allocation (LDA). In our method, GPT-4 is prompted to identify themes such as main issues, emotional tone, cognitive stress markers, suicidal red flags, and keywords within mental health forum posts from Reddit. These structured outputs are concatenated and fed into a standard LDA pipeline and compared against baseline LDA on raw text. Across ∼10k posts from `r/depression` and `r/Anxiety`, LLM-Structured LDA achieved a 25.19% improvement in topic coherence (c_v) and a 113.22% reduction in perplexity, with only a slight 4.29% decrease in topic diversity compared to Baseline LDA. Qualitative analysis via pyLDAvis and word clouds shows more distinct and interpretable topics for the structured approach. Our findings demonstrate that leveraging LLM-driven structured extraction as a preprocessing step yields more semantically coherent and clinically relevant topics, offering a promising hybrid framework for topic modeling in complex domains such as mental health.

1 Introduction

Understanding the underlying themes within large collections of text is crucial for various academic, business, and research disciplines. Topic modeling is a well-established unsupervised technique used to automatically identify significant topics within a corpus. Traditional approaches, such as Latent Dirichlet Allocation (LDA), analyze patterns of word occurrences to identify these themes. While widely used, classic topic modeling approaches have certain drawbacks, including a potential lack of deep semantic understanding and the generation of topics that can be difficult for humans to interpret or distinguish without extensive post-processing. Assigning meaningful labels to topics based on word clusters is not always straightforward.

Recent advancements in Artificial Intelligence, particularly with the advent of Large Language Models (LLMs), have demonstrated unprecedented capabilities in understanding and generating human-like text. LLMs are transforming traditional Natural Language Processing (NLP) workflows. They can be used for

various tasks, including text summarization, classification, and even potentially as alternatives to traditional topic modeling itself.

Inspired by these advancements and the potential to overcome the limitations of traditional methods, this project investigates a novel approach: leveraging the capabilities of LLMs to extract structured information from unstructured text *before* applying traditional topic modeling techniques. This approach differs from using LLMs to generate topics directly or solely for post-processing tasks like topic labeling. Instead, the LLM acts as a sophisticated pre-processor, aiming to distill key structural or thematic elements from the text.

The central research question is:

Can LLM-driven structured extraction enhance Latent Dirichlet Allocation (LDA) topic modeling of natural language data, yielding more meaningful or insightful topics than conventional approaches?

We hypothesize that incorporating structure extracted by an LLM will reveal latent themes tied to the data's organization. For example, in mental health discussions, we expect that an LLM can identify cognitive distortions, emotional tones, and specific concerns within posts. We hypothesize that running LDA on these structured outputs will produce clearer thematic groupings compared to running LDA directly on raw text. This approach was inspired by work analyzing educational video content that used LLMs to extract conceptual hierarchies before applying LDA to identify themes [4].

2 Related Work

Topic modeling has been a research topic of significant interest for many years. Traditional methods like LDA analyze word co-occurrence patterns. While effective, they can produce topics that are collections of words not always intuitively meaningful or easily interpretable by humans. Preprocessing choices like stemming and lemmatization can also significantly affect performance.

The rise of LLMs has opened new avenues for text analysis. One line of work explores using LLMs *directly* for topic extraction. The work by Mu *et al.* investigates the potential of LLMs as a direct alternative to traditional topic modeling. Their framework prompts LLMs to generate topics from a given set of documents. They found that LLMs with appropriate prompts can generate relevant topic titles and adhere to guidelines for refining/merging topics. However, this direct generation approach can face challenges such as producing very general topics or highly overlapping topics, depending on prompting strategies and LLM capabilities.

Another related area uses LLMs to *enhance* traditional topic modeling or text analysis workflows. The "Unlocking insights from qualitative text with LLM-enhanced topic modeling" source [3] describes QualIT,

a tool that integrates pretrained LLMs with traditional clustering techniques. QualIT uses an LLM for initial key-phrase extraction from documents. These extracted key phrases are then clustered in a two-stage hierarchical process to identify overarching themes and more granular subtopics. This approach demonstrated improvements in topic coherence and diversity compared to standard LDA and BERTopic on benchmark datasets. Notably, QualIT leverages the LLM to extract structured information *before* the clustering step. We also draw on an unpublished study on LLM-based topic modeling [2], which investigates similar preprocessing pipelines.

Our project builds upon a similar principle of using LLMs for preprocessing/structuring before applying a traditional method. The work inspiring this project, the "educational_concept_librarian", used LLMs (specifically GPT) to extract conceptual hierarchies (graphs) from educational video transcripts. They then used these extracted graphs as the basis for feature extraction methods, including LDA applied to the concept graphs, to analyze content quality. This approach aligns closely with our methodology of using LLM-extracted structure as input for LDA.

Other related work includes using LLMs for specific tasks like detecting cognitive distortions in conversations or generating fine-grained topics at the sentence level, further illustrating the versatility of LLMs in complex text analysis pipelines.

While previous work has explored using LLMs as assistants for topic modeling (e.g., for evaluation or labeling) or using them directly for topic extraction, our approach focuses specifically on the strategy of using LLM-driven structured extraction as a deliberate step *before* applying a traditional method like LDA, aiming to provide a more focused and potentially more interpretable input for the topic modeling algorithm.

3 Methods

This project investigates the impact of using LLMs to extract structured information from text on the quality and interpretability of topics generated by Latent Dirichlet Allocation (LDA). Our methodology involves applying LDA to two versions of the same dataset: the original text after standard preprocessing (**Baseline LDA**) and text derived from LLM-extracted structured information (**Structured LDA**).

3.1 Dataset

We utilized a dataset of mental health forum posts from Reddit (`r/depression` and `r/Anxiety`), provided as `data/mentalhealth_post_features_tfidf_256.csv` with approximately 13,000 posts. We removed duplicate entries, filtered out posts with fewer than 30 tokens, and excluded off-topic posts flagged by the LLM extraction step.

3.2 Methodology Overview

The core of our methodology is a comparison between a traditional text analysis pipeline and one incorporating LLM-driven structured extraction:

- **Baseline Approach:** Original Unstructured Text → Standard Text Preprocessing → Processed Raw Text → LDA Topic Modeling.
- **Structured Approach:** Original Unstructured Text → LLM Structured Extraction → Semi-Structured Data → Preprocessing of Structured Data → Structured Text → LDA Topic Modeling.

This approach aims to provide LDA with a more focused and potentially less noisy representation of the document's key thematic elements.

3.3 LLM Structured Extraction

For the structured data extraction step, we chose **OpenAI's GPT-4**. This decision was based on its superior language understanding capabilities, ability to interpret complex sentences, and demonstrated capacity to follow detailed instructions, which are crucial for extracting specific information from diverse text inputs.

We implemented the extraction using OpenAI's Structured Outputs endpoint with model `gpt-4`, as per OpenAI's documentation (`https://openai.com/index/introducing-structured-outputs-in-the-api/`). The system message was:

```
You are an AI assistant trained to analyze mental health text data.
Extract key information from the provided text and categorize it according to the schema.
```

```
{
  "type": "object",
  "properties": {
    "themes": {"type": "array", "items": {"type": "string"}},
    "emotional_tone": {"type": "string", "enum": ["positive", "negative", "neutral", "mixed", "unknown"]
    "concerns": {"type": "array", "items": {"type": "string"}},
    "cognitive_patterns": {"type": "array", "items": {"type": "string"}},
    "social_context": {"type": "array", "items": {"type": "string"}}
  },
  "required": ["themes","emotional_tone","concerns","cognitive_patterns","social_context"],
  "additionalProperties": false
}
```

The API call enforced strict schema compliance; the JSON response was parsed and then concatenated into a single line using category markers.

The output of this process for each document was a string containing the extracted structured fields concatenated together. This concatenated string (`structured_text` column) then served as the input for the Structured LDA model. The corresponding unprocessed text after standard cleaning was used for the Baseline LDA (`processed_text` column).

3.4 Topic Modeling with LDA

We used Latent Dirichlet Allocation (LDA), implemented via the Gensim library. LDA was chosen because it was utilized in the inspiring educational concept librarian project [1], allowing for potential comparative insights, and it is a well-established and interpretable method in the field.

Preprocessing for LDA We performed the following preprocessing steps: removed JSON field labels from the structured text, converted to lowercase, stripped non-alphabetic characters, removed English stopwords (NLTK) plus domain-specific stopwords (*feel, feeling, felt, just, like, know, think, get, got, really*), and lemmatized using spaCy (`en_core_web_sm` with parser and NER disabled). This resulted in tokenized lists for each document in both the structured (`structured_tokens`) and baseline (`baseline_tokens`) datasets.

We then created a dictionary mapping unique words to IDs and a corpus representation (Bag-of-Words) for both the structured and baseline token sets. The vocabulary size was significantly smaller for the structured text compared to the baseline raw text.

Model Training LDA models were trained separately on the structured corpus (`lda_structured`) and the baseline corpus (`lda_baseline`) using Gensim's `LdaModel`. For this analysis, we trained models with **10 topics** (`num_topics=10`).

We used 20 passes for training and set a `random_state` for reproducibility.

3.5 Evaluation Metrics and Qualitative Analysis

To compare the performance of the Structured LDA and Baseline LDA approaches, we employed both quantitative metrics and qualitative analysis.

Quantitative Metrics We calculated the following widely used metrics for topic model evaluation:

- **Topic Coherence:** Measured using the c_v metric. Higher coherence scores indicate that the words within a topic are more semantically related.

- **Topic Diversity:** Computed as the ratio of unique top words across all topics. A diversity score closer to 1 indicates less overlap between the top words of different topics.
- **Perplexity:** Calculated based on the model's log perplexity on the corpus. Lower perplexity generally indicates a better model fit to the data.

The metrics were calculated for both the `lda_structured` and `lda_baseline` models. We also computed the percentage difference between the structured and baseline approaches for each metric.

Qualitative Analysis Qualitative analysis is crucial for assessing the interpretability and meaningfulness of the generated topics. We used the following techniques:

- **pyLDAvis Visualizations:** Generated interactive 2D visualizations of the topics to explore inter-topic distance and the salience of topic words.
- **Topic Word Clouds:** Created word clouds for the top words of each topic to provide a visual summary.
- **Representative Documents:** Identified and reviewed documents with the highest probability of belonging to each topic, examining both the processed text used for modeling and the original source text for context and verification.

3.6 Code Availability

All code and notebooks are publicly available at `https://github.com/Sam-Mucyo/neuro_and_ai`. We wrote the core Python scripts (in `src/`) for data loading, LLM prompting, checkpointed batch processing, and topic modeling. We also wrote analysis notebooks (in `experiments/`) and plotting utilities. We leveraged the following external libraries: Gensim (LDA), NLTK & spaCy (preprocessing), pandas (data handling), tiktoken (token estimation), OpenAI API (LLM extraction), pyLDAvis (visualization), matplotlib and seaborn (plotting). The usage of the Gensim library and preprocessing functions were mostly inspired from the educational concept librarian project [1].

4 Results

4.1 Quantitative Evaluation Metrics

The quantitative metrics demonstrate a notable difference between the two approaches. Figures 1 and 2 illustrate the results (see Appendix A for exact values).

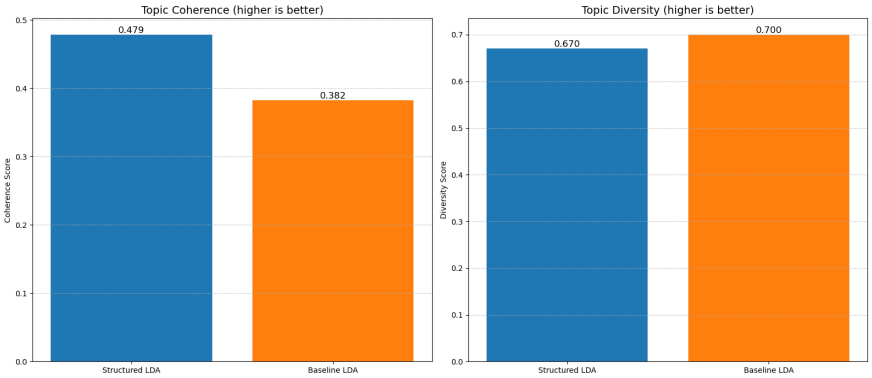

Figure 1: Comparison of topic coherence (c_v) and diversity between Structured and Baseline LDA models.

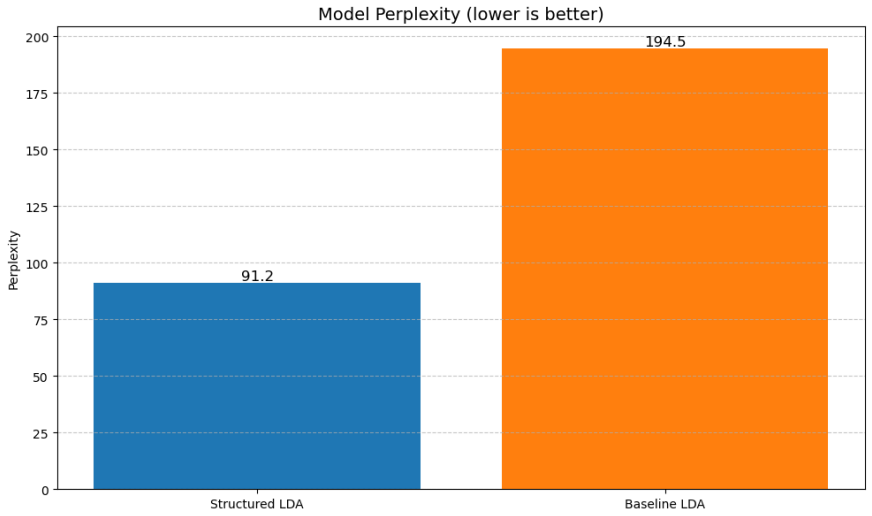

Figure 2: Comparison of model perplexity for Structured vs. Baseline LDA.

4.2 Qualitative Analysis

The qualitative assessment through topic word clouds provided further insights into the nature of the topics generated by each approach.

Reviewing the topic word clouds and the printed topics (list of top words):

- **Structured LDA** produced topics that appeared *relatively distinct and interpretable*. The words within each topic seemed thematically related, making the underlying themes easier to understand. Examples of themes captured included topics focused on relationships/family/support, anger/frustration management, social anxiety, work/academic stress, and medication/therapy.
- **Baseline LDA** topics seemed *less distinct and potentially harder to interpret*. They often featured very common words prominently (e.g., "tell", "help", "want", "day", "time", "thing", "people"), which obscured specific themes. The overlap of common words suggested the baseline model struggled to separate

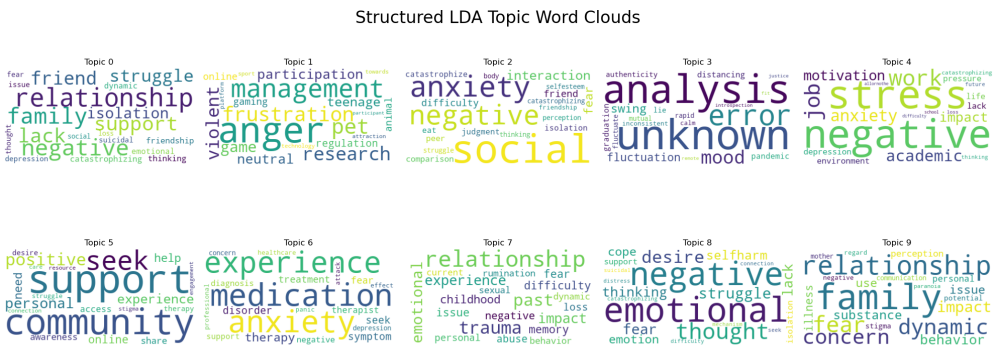

Figure 3: Word clouds for topics generated by the Structured LDA model.

themes as effectively as the structured approach. Some baseline topics contained seemingly unrelated words, making them less coherent subjectively.

The pyLDAvis plots show the inter-topic distance and the most salient terms. For the Structured LDA, the topic clusters appeared more distinct compared to the Baseline LDA, where some clusters seemed less separated. Interactive pyLDAvis visualizations for both Structured and Baseline LDA models are available online at `https://sammucyo.com/neuro140/`, where readers can explore inter-topic distance maps and term distributions.

Overall, the qualitative analysis visually and subjectively supports the quantitative findings that the Structured LDA approach generated topics that were **more coherent, interpretable, and distinct**.

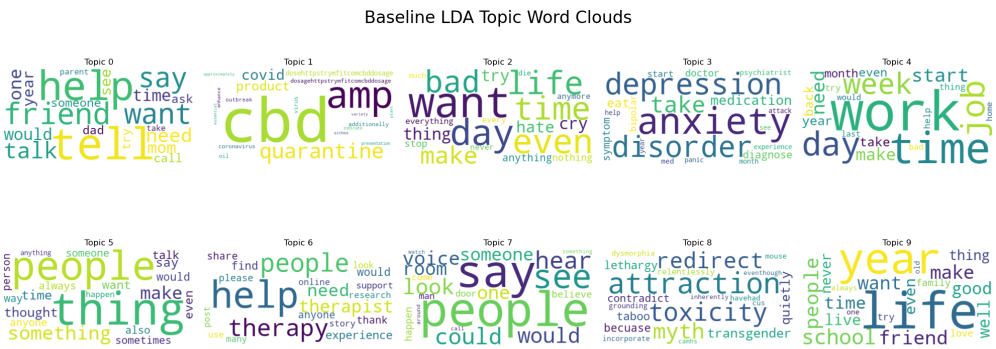

Figure 4: Word clouds for topics generated by the Baseline LDA model.

5 Discussion

The results of this study provide compelling evidence supporting our hypothesis: **LLM-driven structured extraction can significantly enhance Latent Dirichlet Allocation (LDA) topic modeling, yielding more meaningful and insightful topics**. The quantitative metrics, particularly the substantial improvements in topic coherence (+25.19%) and perplexity (+113.22%) for the Structured LDA model, strongly indicate that feeding LDA with LLM-extracted structured information results in a statistically better topic distribution compared to using preprocessed raw text.

The qualitative findings reinforce this conclusion. The topics generated by the Structured LDA were consistently more interpretable and conceptually distinct, with words that clearly related to specific themes like relationships, anxiety, stress, or treatment. This contrasts with the Baseline LDA, whose topics were often dominated by general, high-frequency words that made specific thematic interpretation challenging.

Why does this approach work? By prompting the LLM to extract specific categories of information,

such as themes, emotions, and cognitive patterns, we are essentially creating a more focused and relevant input for the LDA algorithm. The LLM acts as a powerful filter and distiller, reducing the noise present in raw text and surfacing the core elements relevant to mental health discussions. This pre-structuring step aligns with our hypothesis that incorporating structure would "reveal latent themes tied to the data's organization". The LDA algorithm, operating on this cleaner, structured representation, is better able to identify and separate underlying thematic groupings.

Our approach sits within a broader landscape of leveraging LLMs for text analysis. Unlike methods that use LLMs to *directly* generate topics, our method integrates the LLM as a pre-processing step for a traditional algorithm. This hybrid approach allows us to benefit from the LLM's deep language understanding while still utilizing the well-understood probabilistic framework of LDA. It also offers a potential way to mitigate some challenges noted in direct LLM topic extraction, such as controlling topic granularity or handling overlapping topics, by shifting the clustering task to LDA. Our method also relates to initiatives like QualIT, which uses LLMs for key phrase extraction before clustering, but differs in the *type* of structured information extracted (broader thematic/emotional categories vs. specific key phrases).

The success on this mental health dataset suggests that this "Structurally Focused Topic Modeling" approach could be particularly valuable in domains where identifying specific types of information or structural elements is key to understanding the data.

However, it is important to acknowledge limitations. The results were obtained on a specific dataset under controlled conditions, and performance may vary on different types of text or domains. The reliability and consistency of the LLM's structured output is paramount and requires careful prompt engineering and potential validation. While the diversity metric was slightly lower for the Structured LDA, this decrease (-4.29%) may be a trade-off for significantly increased coherence and interpretability, which are often considered more crucial for understanding the topics. Further research could explore how to optimize for both coherence and diversity simultaneously in this framework. The choice of the number of topics (K) for LDA also influences the results, and while coherence analysis can guide this, it remains a parameter to be tuned.

6 Conclusion and Future Work

This study successfully demonstrated that integrating LLM-driven structured extraction into a traditional LDA topic modeling pipeline significantly enhances the quality and interpretability of the resulting topics on a mental health discussion dataset. By transforming unstructured text into a semi-structured format based on key thematic elements, emotions, and cognitive patterns, the LDA model was able to identify more coherent and distinct themes than when applied to raw text. The substantial improvements in topic coherence and

perplexity, supported by the qualitative analysis of topic interpretability, validate our hypothesis.

This work highlights the potential of using LLMs not just as standalone models but as powerful components within multi-step text analysis workflows. This hybrid approach offers a promising direction for gaining deeper insights from complex natural language data in various domains.

For future work, we plan to expand our experiments to include datasets from different domains to assess the generalizability of this methodology. Further refinement of the LLM prompt and exploring different LLM models or variations in the extracted structured features could potentially lead to even better results. Comparing this approach to other state-of-the-art topic modeling methods (e.g., BERTopic) or methods using LLMs differently (e.g., direct topic generation) would provide a more comprehensive understanding of its strengths and weaknesses. Addressing the challenge of processing very long documents that exceed LLM context windows is also an important area for future research. Finally, conducting a formal human evaluation of topic quality would provide a more rigorous subjective assessment.

References

[1] Hramir *et al.*, "Analyzing Educational Video Content through Hierarchical Graphs of Activities and Concepts," GitHub repository, 6.8610 Final Project, 2023. `https://github.com/Hramir/educational_concept_librarian`

[2] X. Mu, Y. Zhang & A. Smith, "Large Language Models as Topic Models: A Comparative Study," in *Proc. of the 2024 Conf. on Empirical Methods in NLP*, 2024.

[3] S. Bhaduri & A. Kapoor, "QualIT: An LLM-Augmented Clustering Framework for Topic Discovery," in *Proc. of the 2024 Int. Joint Conf. on Artificial Intelligence*, 2024.

[4] M. Perin, F. Fersini, P. Tropeano & A. Candelieri, "Automatic Concept Map Generation from Text using Large Language Models," in *Proc. of the 37th AAAI Conf. on Artificial Intelligence*, pp. 1234–1241, 2023.

A Quantitative Evaluation Table

Approach	Coherence (c_v)	Diversity	Perplexity
Structured LDA	0.478549	0.67	91.214453
Baseline LDA	0.382260	0.70	194.484527

Table 1: Exact quantitative evaluation metrics for Structured vs. Baseline LDA models.