LLM-Based Structurally Focused Topic Modeling

Sam Mucyo

May 13, 2025

Abstract

We investigate whether pre-extracting thematic elements via GPT-4 improves topic modeling by Latent Dirichlet Allocation (LDA) on Reddit mental-health posts. Using ∼10K documents from `r/depression` and `r/Anxiety`, we prompt GPT-4 to extract themes, emotional tone, cognitive patterns, and keywords, then compare LDA on structured text versus baseline processed text. We evaluate coherence (c_v), diversity, and perplexity; structured LDA shows a 25% coherence gain and a 113% perplexity reduction (4% diversity loss). Qualitative pyLDAvis and word-cloud analyses confirm more interpretable topics. Code and data are at `https://github.com/Sam-Mucyo/neuro_and_ai`.

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. 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, our project investigates 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 [2].

2 Related Work

Traditional methods like Latent Dirichlet Allocation (LDA) [1] analyze word co-occurrence patterns. While effective, they can produce topics that are collections of words not always intuitively meaningful. 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 another study on LLM-based topic modeling [4], 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 [2]. 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.

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 step *before* applying a traditional method like LDA, aiming to provide a more focused and potentially more interpretable input for the topic modeling algorithm. We then apply this specifically to mental health posts to see if similar trends are found.

3 Methods

The methodology for this approach involves several key steps, beginning with **Data Preprocessing**, where raw posts are cleaned and lemmatized using tools such as `spaCy` and `NLTK`. This process includes removing punctuation, numbers, and stopwords.

Structured Extraction follows, utilizing a large language model (LLM). Specifically, GPT-4 is prompted to systematically extract structured information—such as themes, emotional tone, cognitive patterns, and keywords—from each post, effectively transforming the unstructured text into a semi-structured data set (e.g., JSON or a list of fields). This LLM-derived representation then serves as the foundation for the subsequent **LDA Modeling**.

We then train two Gensim LDA models: one on the structured text derived from the LLM and a baseline model on the preprocessed raw text. Both are configured with ten topics, twenty passes, and α set to `auto` for easier qualitative analysis and interpretation.

The final stage is **Evaluation**, which combines quantitative metrics and qualitative analysis. The quantitative assessment includes computing **Topic Coherence** using the c_v metric, which measures the semantic similarity of the top words within a topic—a higher score indicating more meaningful groupings. **Topic Diversity** is calculated as the ratio of unique top words across all topics, reflecting the breadth of vocabulary covered; a higher value indicates greater diversity. **Perplexity** evaluates how well the model predicts the word distribution in the corpus, where a lower score signifies a better data fit and a more reliable explanation.

For qualitative analysis, we visualize topics with interactive `pyLDAvis` plots and word clouds. We also built a lightweight web app for human evaluation via *document-intrusion tests*: each screen shows four candidate documents—three from the same topic and one "intruder." Participants must spot the outlier, letting us gauge topic interpretability and coherence. Although we have not yet gathered enough responses

to report quantitative results, readers can experiment with the demo at our interactive document-intrusion site.

3.1 Implementation Challenges

Processing thousands of posts with GPT-4 presented significant technical challenges, including API rate limits and long processing times. We addressed these by developing a parallel processing pipeline that distributes work across multiple CPU cores while implementing exponential backoff for API rate limit handling. After benchmarking on sample documents, we found that our original, serial pipeline—which processes each post in about 3-5 seconds would require roughly 15–18 hours to handle all 10,000+ posts (and could stretch past 20 hours in the worst case), posing significant risks of lengthy runtimes and lost progress if interrupted. By contrast, our parallel pipeline shards LLM calls across multiple worker processes, employs exponential backoff to gracefully handle API rate limits, and checkpoints intermediate results to enable robust, resumable execution—bringing the end-to-end runtime down to under 90 minutes, a roughly ten-fold throughput improvement.

4 Results

4.1 Quantitative Evaluation Metrics

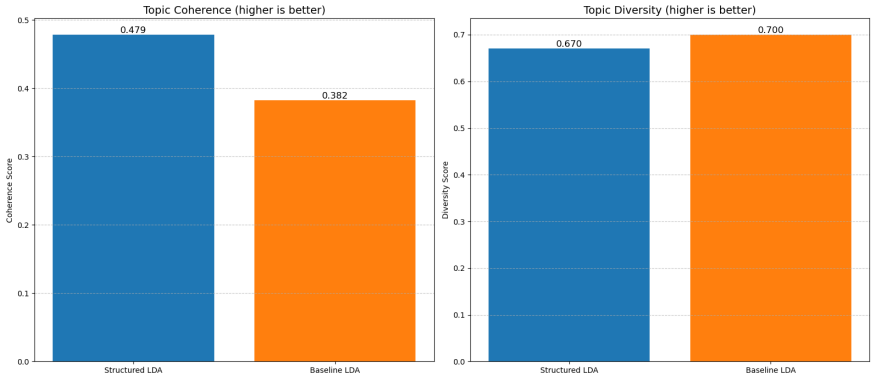

Figure 1: Comparison of topic coherence (c_v) and diversity between Structured and Baseline LDA models.

Figures 1 and 2 (see Appendix B) show that the structured approach achieves higher topic coherence with a 25% gain (0.479 vs. 0.382) and a 113% reduction in perplexity (91.2 vs. 194.5) than the baseline, while incurring only a slight drop in topic diversity by 4% (0.67 vs. 0.70). Taken together, these results indicate that the structured representation yields more focused, semantically consistent topics that fit the data better, outweighing the marginal loss in vocabulary breadth.

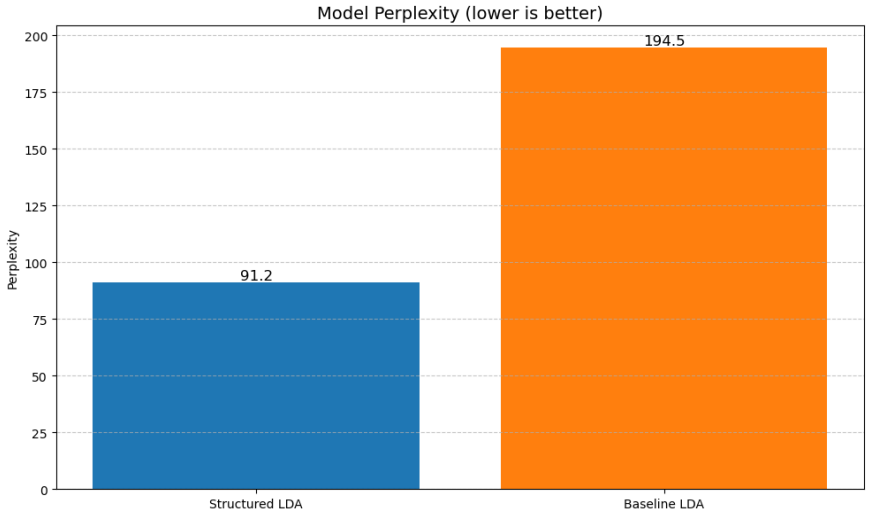

Figure 2: Comparison of model perplexity for Structured vs. Baseline LDA.

4.2 Qualitative Analysis

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

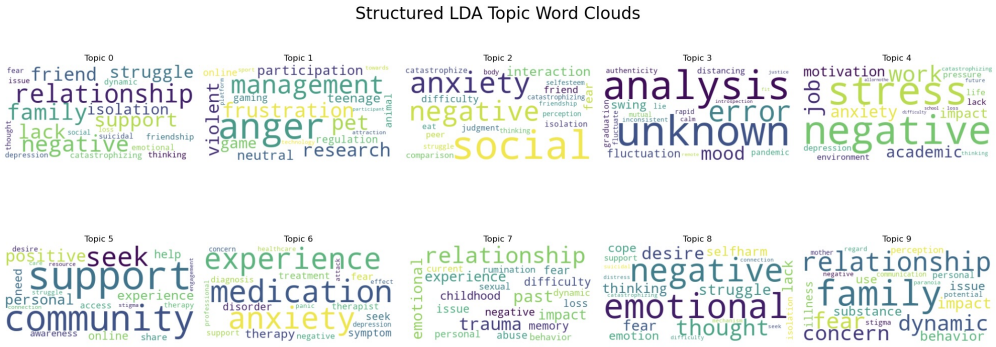

Figure 3: Word clouds for topics generated by the Structured LDA model.

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 seem 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 seem *less distinct and potentially harder to interpret*. They often featured very common words prominently (e.g., "tell", "help", "want", "day", "time", "thing", "people"), which ob-

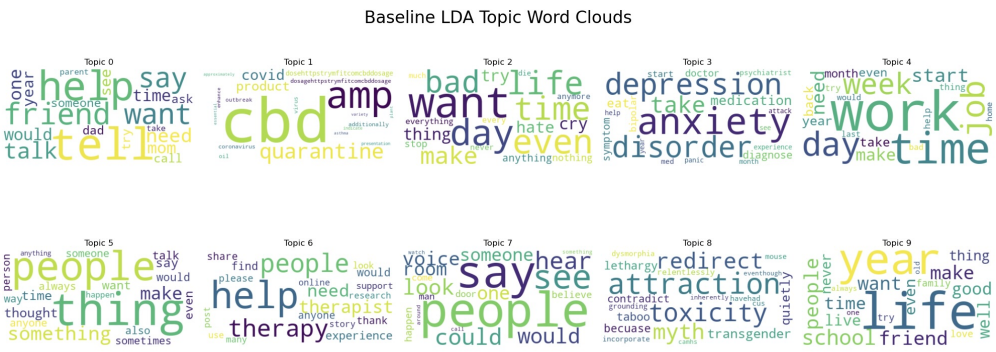

Figure 4: Word clouds for topics generated by the Baseline LDA model.

scured specific themes. The overlap of common words suggested the baseline model struggled to separate 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**.

5 Discussion

Our findings confirm that integrating LLM-extracted structure into the input corpus enables LDA to discover clearer, more semantically coherent themes than a conventional preprocessing pipeline. Topics produced with this hybrid workflow were qualitatively easier to interpret, consistently surfacing domain-relevant concepts (e.g., relationships, anxiety, treatment) instead of generic high-frequency terms.

The improvement arises because the LLM acts as a high-precision filter, distilling posts into salient thematic, emotional, and cognitive cues before probabilistic grouping. This "best-of-both-worlds" design combines the language understanding of GPT-4 with the controllable clustering of LDA, echoing but extending earlier artifact-extraction approaches such as QualIT [3]. Although evaluated on a single mental-health dataset, the method should transfer to any domain where targeted structural cues matter, provided the LLM prompt is robust and the usual LDA hyper-parameters (e.g., K) are tuned.

6 Future Work

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 may 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 assessment.

Code and Data Availability

The complete code, processed data (`results/processed_data.csv`), and analyses notebooks are available at:

- Main repository: `https://github.com/Sam-Mucyo/neuro_and_ai`
- Doc-intrusion test repository: `https://github.com/Sam-Mucyo/doc-intrusion-test`.

References

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] Hramir. educational_concept_librarian: A tool for extracting and organizing educational concepts. GitHub repository, `https://github.com/Hramir/educational_concept_librarian`, 2025. [Online; accessed 13-May-2025].

[3] Satya Kapoor, Alex Gil, Sreyoshi Bhaduri, Anshul Mittal, and Rutu Mulkar. Qualitative insights tool (qualit): Llm enhanced topic modeling. In *LREC-COLING 2024*, 2024. arXiv:2409.15626 [cs.IR]; Submitted on 24 Sep 2024; `https://doi.org/10.48550/arXiv.2409.15626`.

[4] Yida Mu, Chun Dong, Kalina Bontcheva, and Xingyi Song. Large language models offer an alternative to the traditional approach of topic modelling. In *Accepted at LREC-COLING 2024*, 2024. arXiv:2403.16248 [cs.CL]; https://doi.org/10.48550/arXiv.2403.16248.

A Detailed JSON Prompt 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
}
```

B 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.