Project: LLM-Based Structurally Focused Topic Modeling

1. Research Question and Hypothesis

Inspired by **project 12** (“Use large language models for structurally focused topic modeling of natural language datasets”) in the list of project ideas, we propose to investigate whether large language models (LLMs) can improve topic modeling by extracting structured information from text before applying topic models. 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 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 in posts, and that running LDA on these structured outputs will produce clearer thematic groupings (e.g., types of negative thinking or coping themes) than running LDA on raw text.

Link to the Github Repo:

- https://github.com/Sam-Mucyo/neuro140/tree/main

2. Literature Review

Key Study: Talbot et al. (2023) demonstrated combining LLMs with topic modeling by analyzing educational YouTube videos. They used GPT to extract structured knowledge graphs of concepts from video transcripts, then applied LDA on these extracted concept graphs to identify themes. This approach revealed that popular videos tended to introduce core concepts early and reinforce them with supporting concepts.

Other Related Work:

- Wang et al. (2023) introduced PromptTopic, using LLMs to generate fine-grained topics at the sentence level
- Singh et al. (2023) developed a framework to detect cognitive distortions in patient-therapist conversations using LLMs.

Traditional topic modeling (LDA) has been used in mental health text analysis but often produces topics that are difficult to interpret without careful preprocessing

3. Dataset Selection

We will use mental health discussions data from forums like Reddit (r/depression, r/Anxiety) or therapy conversation transcripts. This domain naturally contains structural elements (cognitive distortion patterns, therapy dialogue structure) that align with our methodology. The dataset is publicly available from Reddit Mental Health Dataset.

4. LLM and Topic Modeling Methodology

Chosen LLM: We will use OpenAI GPT-4 for structured extraction due to its superior language understanding and ability to follow complex instructions. Alternatives include GPT-3.5 Turbo (cheaper) or open-source models like LLaMA-2 if budget constraints arise.

Topic Modeling Technique: We will use Latent Dirichlet Allocation (LDA) because:

1. It was used in the key paper, allowing comparative statements
2. It is well-known, interpretable, and has numerous existing implementations
3. Prior researchers have applied LDA to mental health forum posts, providing benchmarks

5. Project Plan and Experiments

5.1 Structured Data Extraction with LLM

In our initial exploratory analysis (EDA), we designed a prompt template for LLM to systematically extract structured information from each post. For mental health forum posts, the prompt will contain the main issue or concern described, the emotion or tone expressed, cognitive stress markers, etc. This process will transform unstructured text into a semi-structured dataset, with each document having an LLM-derived representation.

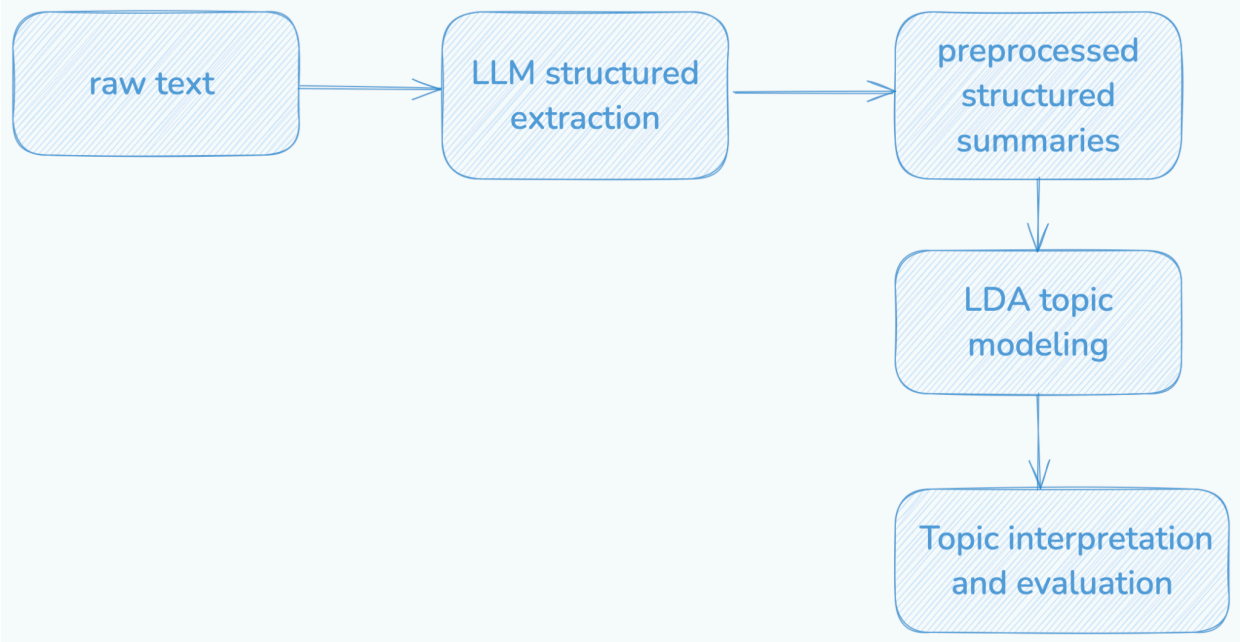
5.2 Topic Modeling with LDA

Once we have structured outputs, we will preprocess them for LDA by concatenating key fields into a single text per post, performing standard text preprocessing (lowercasing, removing stopwords), then finding optimal number of topics (K) through coherence analysis. The resulting topics might correspond to themes like "social anxiety and fear of judgment" or "health anxiety and physical symptoms."

5.3 Baseline and Comparative Experiments

To quantify benefits, we'll run baseline topic modeling on the raw dataset and compare with theLLM-structured approach. We will also conduct ablation studies using only certain structured fields to determine which contributes most to topic quality.

5.4 Flowchart of the Process



5.5 Implementation Details

We will implement these steps incrementally:

1. Test the LLM prompt on a small subset and refine as needed
2. Scale extraction to the full dataset
3. Run LDA with various topic counts
4. Analyze resulting topics in context of our hypothesis

6. Evaluation Metrics and Benchmarks

We will use a combination of quantitative metrics and qualitative analysis. (1) topic coherence by measuring how semantically related the top words of a topic are (using measures like Cv or UCI coherence) and/or (2) topic diversity, by examining the fraction of unique words in top-N words across all topics. Then we'll compare metrics between LDA on raw text (baseline) and LDA on LLM-structured text (our method). We also plan to do human evaluation by doing a qualitative assessment of topic interpretability through manual examination

7. Computational Feasibility

Our plan is designed for available resources: an academic compute node (8 CPU cores, 32GB RAM, 24GB GPU VRAM) and a MacBook M3 Max. Using GPT API for prompting shifts computation to external servers. We can parallelize API calls using the 8-core node. If using local models or finetunig, the 24GB GPU can handle moderately sized models with quantization techniques. For LDA, the algorithm is CPU-bound and memory-bound. We think 32GB RAM is sufficient for our dataset size, and 8 CPU cores can run parallel LDA sampling efficiently. For runtime, GPT-4 processing might take a few seconds per post. For 1000 posts, extraction would complete in minutes plus overhead. LDA training for moderate K should converge within minutes.

8. Potential Challenges and Mitigation Strategies

- **LLM Output Consistency and Accuracy:** The quality of our structured data hinges on GPT's ability to follow the prompt and not hallucinate or omit important info. In mental health text, the language can be idiosyncratic (typos, slang like "IDK" for "I don't know", etc.) or very emotional, which might throw off the model or cause variability in output format. We plan to test and refine the prompt with a variety of examples, including edge cases (very short posts, very long posts, use of sarcasm or figurative language, etc.). We'll enforce a structured format by asking for JSON or a list of fields. We will also look into Introducing Structured Outputs in the API | OpenAI
- **Determining Number of Topics and Topic Interpretation:** choosing *K* for LDA can be tricky – too low and topics are overly broad, too high and they splinter. We plan to use the coherence score cross-validation approach as mentioned pmc.ncbi.nlm.nih.gov. That is, run LDA for a range of K (e.g., 5 to 20) and pick the K that yields the best coherence. This data-driven method should give a reasonable choice. We will also manually verify that the chosen K is sensible by looking at the topics; if the top coherence K still yields some redundant topics, we might slightly adjust.
- **Dataset Quality and Noise.** User-generated text can be noisy (irrelevant content, very short posts that don't have enough info, or multi-topic posts that confuse LDA). If our dataset includes unrelated tangents or trivial posts like "Anyone here? Hello.", these could create spurious topics or waste LLM quota. We plan to apply filtering criteria to the dataset. For Reddit, we might remove posts below a certain length or those that moderators marked as off-topic. We can also have the LLM perform a relevance check: e.g., the prompt could instruct "If the post is not actually about a personal mental health experience, just label it as off-topic." and we'd exclude those from LDA. This way, we ensure we're modeling relevant data.