Group 13 Week 9 Proposal:

Subash Ramanathan, Summer Habarneh, Zack Lichtenberg

Comparative Attention Analysis of Generative Language Models on Medical vs. Legal Question Answering

1. Final Topic Area

Field: Natural Language Processing (NLP)

Rationale:
This project investigates interpretability in generative language models applied to question answering (QA) in medical and legal domains. Given its emphasis on transformer-based models and attention mechanisms, it directly aligns with the course focus on NLP. The motivation stems from understanding how models process and differentiate domain-specific information, especially in high-stakes areas like medicine and law.

2. Dataset Description

Medical Domain:
- Dataset: MedQA (USMLE-style questions)
- Source: AI2 / Hugging Face
- Size: ~12,000 multiple-choice clinical questions
- Format: JSON (text-based)
- Characteristics: Contains complex, high-level clinical scenarios requiring domain-specific knowledge.

Legal Domain:
- Dataset: dzunggg/legal-qa-v1
- Source: Hugging Face
- Size: ~3,700 rows
- Format: Text-based QA pairs
- Characteristics: Involves formal language, legal citations, and logical reasoning based on statutes or precedents.

3. Model Selection

- Model: GPT-2 (small) and/or T5-small
- Architecture: Transformer-based decoder (GPT-2) or encoder-decoder (T5)

Justification:
These models are open-source, pretrained, and support access to attention weights. GPT-2 enables decoder-only attention analysis, while T5 allows cross-attention inspection. Their interpretability and support for text generation tasks make them ideal for this analysis. We will use pre trained versions via Hugging Face Transformers.

4. Research Questions

Our investigation will address the following:
1. How do generative language models distribute attention across tokens when answering medical vs. legal questions?
2. Do attention focal points vary by domain-specific terms (e.g., 'diagnosis' vs. 'plaintiff')?
3. Can entropy and focus span be quantified to distinguish domain reasoning patterns?
4. What insights can attention visualization provide for model trust in sensitive domains?

5. Plan of Action

Data Preprocessing:
- Sample 20–50 QA pairs per domain
- Annotate domain keywords using spaCy or SciSpacy
- Standardize QA format for model input

Model Implementation:
- Load GPT-2 and T5-small from Hugging Face with `output_attentions=True`
- Generate answers and extract attention weights

Experimental Design:
- Compare attention heatmaps across correct/incorrect predictions
- Measure entropy, keyword focus, and token span

Analysis Techniques:
- Heatmap visualization (matplotlib, BertViz)
- Dimensionality reduction (PCA) on attention vectors
- Statistical comparisons between domains

Timeline:
Week 1: Data sampling and annotation
Week 2: Model loading and testing
Week 3: Inference and attention capture
Week 4: Analysis and visualization
Week 5: Evaluation and interpretation
Week 6: Final report and presentation

6. Team Contribution

Member 1:
- Role: Dataset preparation and keyword tagging
- Assist with QA formatting

Member 2:
- Role: Model loading and inference pipeline
- Support attention extraction and debugging

Member 3:
- Role: Statistical analysis and visualization (entropy, span, heatmaps)
- Report writing and formatting

Shared Responsibilities:
- Weekly updates
- Final report and results presentation
- Peer review of analysis findings

Sources:

GBaker. (2023). *MedQA-USMLE-4-options* [Data set]. Hugging Face. https://huggingface.co/datasets/GBaker/MedQA-USMLE-4-options

dzunggg. (2023). *Legal QA v1* [Data set]. Hugging Face. https://huggingface.co/datasets/dzunggg/legal-qa-v1