Project Proposal

Proposal 1: Distilling Verifiable Medical Knowledge: A Teacher-Student Approach for Accurate LLM Citations

Motivation:

- LLMs are prone to hallucination, inventing facts and sources. A system that can provide more accurate answers with verifiable citations is a critical step toward trustworthy medical AI assistants.
- While RAG can improves factuality of LLMs by providing external context, it introduces inference-time and complexity. For many real-world applications, a self-contained, efficient model that produces reliable, cited information is highly desirable.

Statement: The project aims to develop and evaluate a teacher-student knowledge distillation pipeline to create compact model for generating accurate medical information with verifiable citations.

The key question: to what extent can teacher model be distilled into a smaller student model, enabling it to produce trustworthy without real-time retrieval?

Proposed Approach:

1. Scape `PubMed` open access dataset as medical knowledge base
2. Prompt teacher model (such as `GPT-5`) to retrieve relevant medical text chunks
3. Synthesize and generate high-quality datasets to teach student
4. Fine-tune a open-source model (such as `Qwen2.5-7B-Instruct-1M`)
5. Evaluate the held-out test set on `factual accuracy`, `citation recall` and `hallucination rate`, benchmarked against zero-shot student, teacher model and student + RAG

Related Literature Review:

1. Closing the gap between open source and commercial large language models for medical evidence summarization
2. How well do LLMs cite relevant medical references? An evaluation framework and analyses

Proposal 2: A Hybrid Stylometric and Semantic Fluctuation Approach for Detecting AI-Generated Media News

Motivation:

- LLMs poses a direct threat to integrity of public information and vast quantities of convincing but false media news are generated.
- AI text dectors often rely on content-based classifiers, creating need for more robust detection methods.

Statement: The project aims to develop and evaluate a hybrid detection model that combines statistical stylometry with semantic fluctuation analysis to detect AI-generated media news.

The key question: can a hybrid model can achieve greater accuracy of detecting AI-generated news compared to standard content-based classifiers?

Proposed Approach:

1. Scape human-written news sources as human baseline
2. Generate a synthetic counterpart (eg. GPT-4o) on same topics
3. Create a 3rd adversarial dataset by having humans slightly paraphase AI-generated articles to simulate evasion tactics
4. Develop a hybrid model: stylometric analysis, semantic flucutation analysis
5. Train a robust classifier and evaluate and benchmark

Related Literature Review:

1. Detecting AI-Generated Text: Factors Influencing Detectability with Current Methods
2. Defending Against Neural Fake News
3. Automatic Detection of Machine Generated Text: A Critical Survey