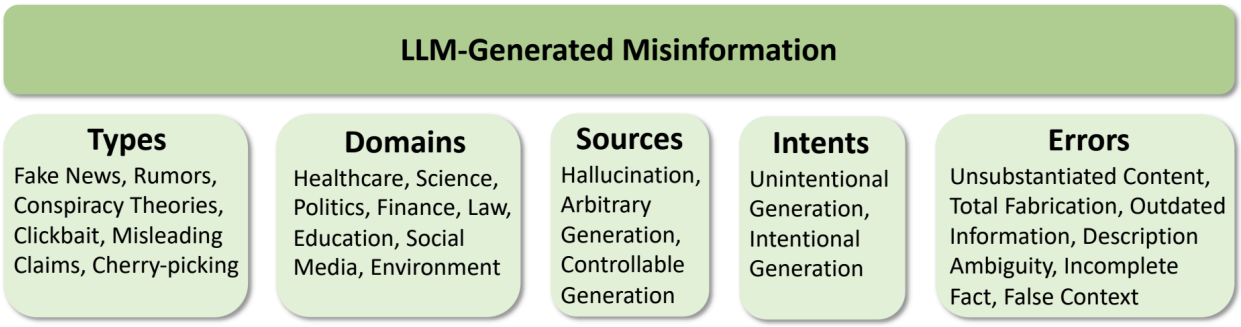
Intelligent Document Question Answering System

This project aims to build a Q&A system that takes long documents (e.g. academic papers, manuals, legal docs) and allows users to ask questions in natural language. The system retrieves relevant content and uses an LLM to generate accurate, grounded answers.

Context

LLMs play a significant role in spreading misinformation, which can lead to serious real-world consequences. One reason is that they "exploit weaknesses in how we monitor our world to determine what is real and what is not" (1). This misinformation can be both unintentional (via hallucinations) and intentional (due to user manipulation or malicious prompts) (2).

It is also claimed that LLM-generated misinformation is harder to detect than misinformation produced by humans (2). The following figure (3) categorizes LLM-generated misinformation into four main types:



Taxonomy of LLM-Generated Misinformation (3)

Problem statement

The goal is to create an end-to-end, LLM-powered document QA system that provides fast and reliable access to scientific knowledge within 3 months, with the following performance goals:

≥ 80% Recall@5

≥ 0.5 ROUGE-L score

These metrics are used to evaluate the retrieval and faithfulness of answers, respectively, with an emphasis on minimizing hallucinations and maximizing factual accuracy.

Criteria for success

To consider the system successful, it should:

1. Ingest and parse long documents
2. Embed and interpret user questions
3. Conduct vector-based semantic search over the document corpus
4. Generate clear, grounded answers using LLMs
5. Achieve ≥ 80% Recall@5 on a test set of scientific QA pairs

Scope of solution space

The initial scope of the system will be limited to one scientific domain: **neuroscience or neurology**. This allows for more focused retrieval and tuning, and makes evaluation more manageable.

Constraints

- **Domain knowledge requirement:** Working with specialized content (e.g., neuroscience) may require expert input or domain-specific embeddings.
- **Evaluation difficult**y: Selecting appropriate automated metrics to assess factual accuracy without human feedback is challenging.
- **Computational cost**: Vector search, embedding generation, and LLM inference require significant compute resources.

Stakeholders

Potential audiences and stakeholders for this project are medical doctors and researchers, hospitals and medical institutions, and academic institutions/students/researhcers.

Data sources

I will be using scientific articles published on arXiv.org, accessing them using their API.

References

(1) Garry, M., Chan, W. M., Foster, J., & Henkel, L. A. (2024). Large language models (LLMs) and the institutionalization of misinformation. *Trends in cognitive sciences*.
(2) Chen, C., & Shu, K. (2024). Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*, *45*(3), 354-368.
(3) Chen, C., & Shu, K. (2023). Can llm-generated misinformation be detected?. *arXiv preprint arXiv:2309.13788*.