DARTMOUTH
Guarini School of Graduate and Advanced Studies

Utilizing LLMs in Identifying Divergent Health Claims

Kishore Kumaran

Project Instructor: Prof. Sarah M. Preum

Motivation

- An increasing number of people now rely on online platforms to meet their health information needs.
- This makes it imperative to censor incorrect health advice or claims; the first step in this is identifying divergent health claims.

> Statement 1: Genetically modified crops can be harmful to the environment and ecosystems, threatening biodiversity.
> Statement 2: GM crops are often modified in ways that can make them better for the environment than their unmodified counterparts.

- These are the kind of content which are present in online platforms, and more and more people are consuming and contributing to such content.
- To prevent such content from causing real-world harm to unsuspecting people, an automated way to detect such diverging claims needs to be developed.

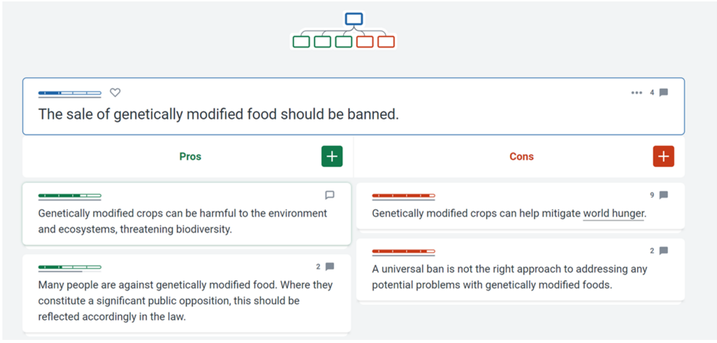
Challenges & Contributions

Lack of prior work in computational health | Absence of comprehensive medical datasets to work on | Requirement of high-computing GPUs to trains LLMs

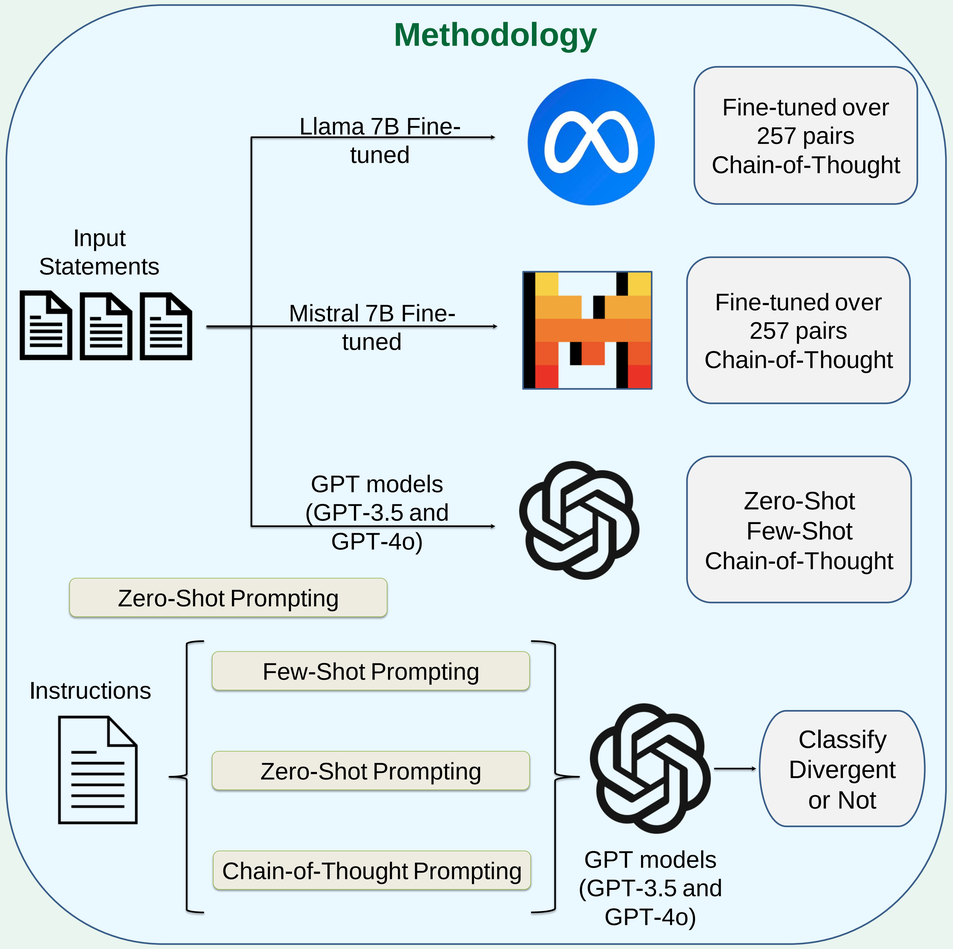
- **Resource**: develop a divergent health statements dataset comprising a variety of health topics.
- **Social impact:** Address some of the limitations of online health platforms by flagging inconsistent statements regarding health.
- **Benchmarking:** investigate the performance of state-of-the-art models including ChatGPT for such knowledge-intensive task.

Kialo: Online Structured Debate Platform

- Kialo is an online platform specifically designed for structured debates and discussions on a wide range of topics, from political issues to health topics.
- Discussions on Kialo are organized into argument trees. These arguments are then categorized into pro and con sections, allowing for a clear visualization of different perspectives on the topic.
- For our project, we have scraped the debates with the 'Health' tag and are matching statements with their immediate cons to obtain potential divergent statement pairs.
- I manually went through 500 of these pairs and annotated them as either divergent or non divergent pairs.
- This forms the data for fine-tuning Llama and Mistal LLMs



Methodology

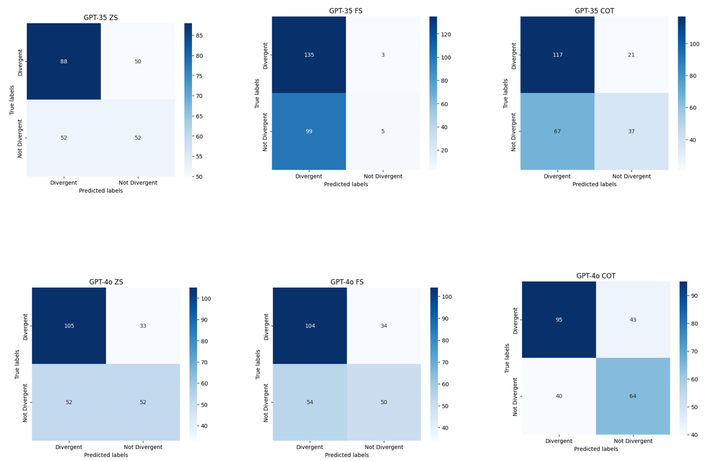


Results

Method	Classifier	Precision	Recall	F1-Score
Base Models	ChatGPT (ZS)	0.57	0.57	0.57
	ChatGPT (FS)	0.60	0.51	0.41
	ChatGPT (COT)	0.64	0.60	0.59
	GPT-4o (ZS)	0.64	0.63	0.63
	GPT-4o (FS)	0.63	0.62	0.62
	GPT-4o (CoT)	**0.65**	**0.65**	**0.65**
Fine-Tuned Models	Llama Fine-tuned (CoT)	0.58	0.56	0.54
	Mistral Fine-tuned (CoT)	0.59	0.58	0.56

Table. Performance comparison for classification of divergent statements.The shorthand indicates ZS: Zero-shot,, FS: Few-shot and CoT: Chain-of-Thought prompting.

Error Analysis of GPT models



- **ChatGPT tends to over-predict every sample as divergent. ChatGPT with Zero-Shot prompting produces better results compared to Few-shot and Chain-of-Thoughts prompting.**
- **GPT-4o is the best performing model with Chain-of-Thought prompting giving the best result.**
- **Llama 7B and Mistral 7B models even with fine-tuning do not perform as good as the base GPT models.**
- **This can be due to ChatGPT and GPT-4o having 175B and 1.8T parameters respectively. With more parameters, the model is able to understand complex data patterns, which is pertinent to our problem, as subtle distinctions between the statements are what causes divergence in most pairs.**

Conclusion & Future Work

- We develop a dataset of Divergent pairs in the realm of Health.
- We fine-tune Llama 7 billion and Mistral 7 billion and see how models compare with GPT models (GPT 3.5 and GPT-4o)
- The dataset can to be extended. Out of the scraped data from Kialo site, a total of 12,127 potential divergent pairs were obtained but only 500 pairs were annotated for this project. More pairs can be annotated and experimented on.
- More Language models needs to be analyzed on, especially LLMs fine-tuned for medical domain applications like Med-Gemini.