Comparative Analysis of LLMs in Identifying Divergent

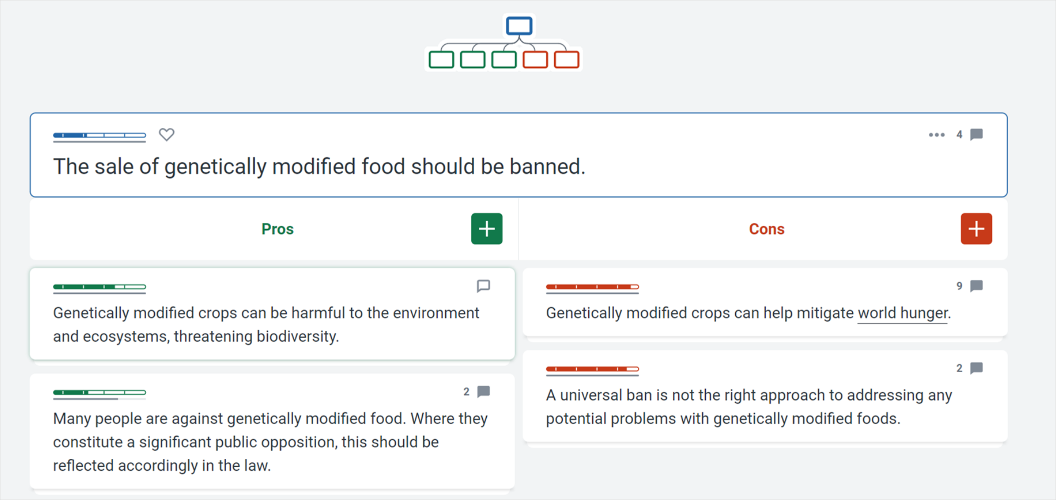
Kishore Kumaran
Dartmouth College

1 Introduction

An increasing number of people now rely on online platforms to meet their health information needs. Gatto et al. (2023) Thus identifying incorrect or invalid advice or claims has become an imperative task as the intake of wrong medication or treatments can have serious consequences to the patient's health and in worst case scenario it can lead to fatalities. A step forward in this direction is to identify conflicting or divergent statements or claims. This is the problem statement for this project. Given two statements in the health domain, our aim is to predict if these statements are divergent or not divergent.

In this project we create and make available a dataset for training machine learning models to tackle this problem and we are evaluating different large-language models' capabilities to distinguish divergent statements from non-divergent statements.

2 Dataset



The sale of genetically modified food should be banned.

Genetically modified crops can be harmful to the environment and ecosystems, threatening biodiversity.

Many people are against genetically modified food. Where they constitute a significant public opposition, this should be reflected accordingly in the law.

Genetically modified crops can help mitigate world hunger.

A universal ban is not the right approach to addressing any potential problems with genetically modified foods.

Figure 1: Example Debate

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. As part of previous work done under Prof. Sarah, Madhusudan Basak and in collaboration with Zhouxin Ma, I have scrapped all the health debates from Kialo.com. Out of a total of 159 debates with the health tag, 10 random debates were chosed for manual annotation and the rest 149 debates were chosen for weak labeling. From each debate, we are matching statements with their immediate 'con' to form a potential divergent statement pair. We use this logic to generate potential divergent pairs from all 159 debates from the kialo website. In the 10 debates, a total of 499 pairs of statements were manually annotated by me. I annotated them as either divergent or non-divergent pairs. These 499 annotated pairs were split into train and test datasets containing 257 pairs and 242 pairs respectively. The remaining 149 debates produced a total of 10155 pairs of statements which were weak-labelled using gpt-4o.

This forms the data for fine-tuning Llama and Mistal LLMs.

3 Methodology

3.1 Prompt Analysis

I've utilized chatGPT to analyze which prompting technique is best suited for this task. I considered zero-shot prompt, few shot prompts (varying from 1 shot to 9 shot prompts) and chain of thought prompting.

I took 10 pairs from the manually annotated sample and tried all the different prompts for all these pairs and compared their predictions with the (manually annotated) ground-truth; based on this analysis, Chain-of-Thought prompting gave the best performance and Zero-Shot prompting produced the second best performance. Based on this, Chain-of-Though prompting was chosen to be the best prompting style for this specific problem and was used as the prompt method for all further analysis.

3.2 GPT model Analysis

The latest model GPT-4o is chosen as the baseline and the performance of this model is tested against all the other model including fine-tuned smaller models like LLama 7B and Mistral 7B.

For GPT models, GPT 3.5 along with GPT-4o is used, with three different prompting strategies: Zero-shot, few-show (9 shot) and Chain of Thoughts promptings. The task from the test dataset of matched pairs, classify whether the statements are divergent or not.

3.3 Llama and Mistral Fine-Tuned model Analysis

Llama 7B and Mistral 7B model are used for this project for finetuning, taking into consideration the computing resources required for generating predictions and finetuning. Fine-tuning with the train dataset of 257 pairs is performed and the model is tested using Chain of Thoughts prompt to classify whether pairs in the test dataset are divergent or not.

3.4 Llama and Mistral Fine-Tuned model (with Manually Annotated and Weak Labels) Analysis

The Llama 7B and Mistral 7B are again taken for fine-tuning using both the manually annotated data and weakly annotated data. With the manually annotated data, the weight for fine-tuning is set to 1, and for the weakly labeled data, the weight for fine-tuning is set to 0.5. The models are fine-tuned with the train dataset of 257 pairs and another 3000 pairs sampled randomly from the weakly labeled dataset; the models are fine-tuned using a total of 3257 pairs of statements. These models are again tested using Chain of Thoughts prompt to classify whether pairs in the test dataset are divergent or not.

4 Results and Discussion

Method	**Classifier**	**Precision**	**Recall**	**F1-Score**
Base Models	GPT 3.5 (ZS)	0.57	0.57	0.57
	GPT 3.5 (FS)	0.60	0.51	0.41
	GPT 3.5 (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 (COT)	0.58	0.56	0.54
	Mistral (COT)	0.59	0.58	0.56
	Llama with weak labels (COT)	0.53	0.52	0.50
	Mistral with weak labels (COT)	0.59	0.60	0.59

Table 1: Precision, Recall and F1-Score for all models

As evident from Table 1, GPT-4o model with Chain-of-Thought prompting outperforms all other models.

Chain-of-Thought prompting is the most suitable prompting method for this use-case, followed by Zero-shot and lastly Few-shot prompting.

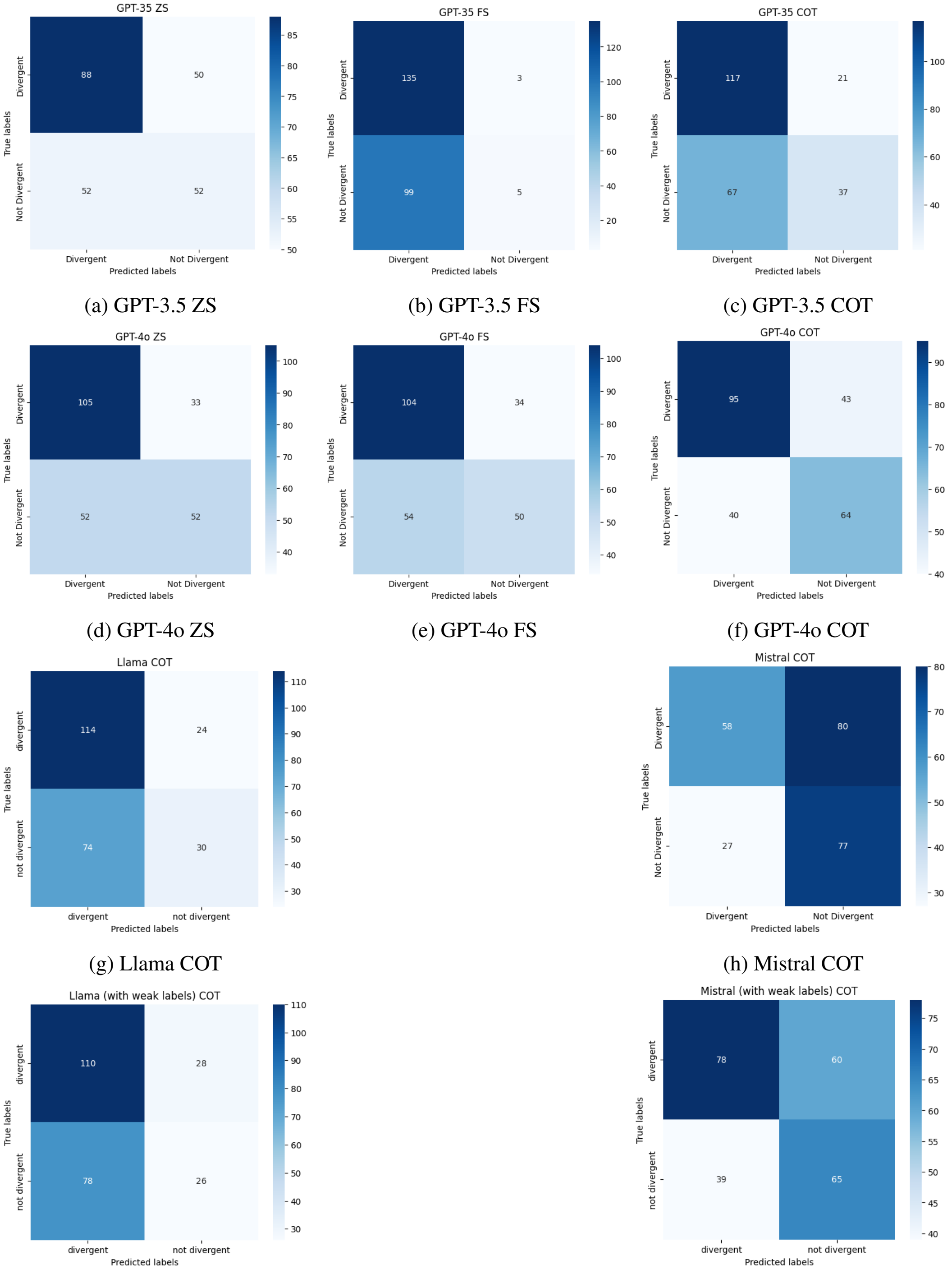
Llama and Mistral model even though containing only 7B parameters give comparable performance to GPT-3.5 ZS and GPT-3.5 FS. Although GPT-3.5 COT outperforms Llama and Mistral fine-tuned models.

Upon further analyzing the confusion matrices of each model for both the classes, one interesting point is GPT-3.5 model with FS and COT prompting is biased to classify the majority of statement pairs as Divergent.

This bias is not present in GPT-4o model, highlighting the improvements from GPT-3.5 to GPT-4o.

LLama and Mistral models fine-tuned with only manually annotated data provides comparable performance to GPT-3.5 FS and GPT-3.5 ZS models.

When fine-tuning with weak-labels, we see both Mistral and Llama are getting biased to classify more statement pairs as divergent. This may be due to the unbalanced nature of the weak-labels dataset. From the random sampling, more divergent samples might have been chosen. This is an oversight on my part for not ensuring the dataset is balanced.



(a) GPT-3.5 ZS

(b) GPT-3.5 FS

(c) GPT-3.5 COT

(d) GPT-4o ZS

(e) GPT-4o FS

(f) GPT-4o COT

(g) Llama COT

(h) Mistral COT

(i) Llama (weak labels) COT

(j) Mistral (weak labels) COT

Figure 2: Confusion Matrix of all models over divergent and non-divergent classes

5 Conclusion

In this project, we develop a dataset of Divergent pairs in the domain of Health.

We fine-tune Llama 7 billion and Mistral 7 billion with manually annotated data and also utilizing weakly annotated data and perform a comparative study on how these models perform against GPT models (GPT 3.5 and GPT-4o).

6 Limitations and Future Work

This project is severely constrained by the computing resources and lack of availability of annotated data.

As for future direction, I hope to fine-tune models with higher number of parameters. Llama and Mistral with 7 billion parameters even with fine-tuning do not have the complex reasoning skills of GPT-4o model with 1.8 trillion parameters. Since subtle distinctions can change the implications of a statement, higher reasoning power would theoretically provide better performance.

For future scope, Llama 70B models can be fine-tuned as this might have a better chance at giving comparable performance to GPT-4o model.

7 Acknowledgements

We thank Professor Preum, Omar and Madhusudan for their help and insightful suggestions during the final presentation. It helped shape the direction of this project and significantly helped me to improve the quality of our final report.

References

Joseph Gatto, Madhusudan Basak, and Sarah Masud Preum. Scope of pre-trained language models for detecting conflicting health information. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pp. 221–232, 2023.