Scalable Classification of Online Vaccine Concerns

Rickard Stureborg **Chloe Qinyu Zhu** **Christopher Li** **Bhuwan Dhingra**
Duke University
{rickard.stureborg, qz124, cl619}@duke.edu
bdhingra@cs.duke.edu

Abstract

Concerns regarding vaccines impact vaccine uptake, and these concerns can shift quickly as seen during the COVID-19 pandemic. Identifying longitudinal trends in vaccine concerns and misinformation might inform the healthcare space by helping strategically allocate resources or information campaigns. Large Language Models (LLMs) have been shown to perform well under zero-shot settings when labels are well defined. However, their use on large corpora remains prohibitive due to computational costs. Therefore, we explore using LLMs to label training datasets on top of which cheaper models (such as BERT-based models) can be finetuned. Given the naturally varying granularity of concerns expressed in online text there are several potential approaches for how to prompt LLMs to provide multi-label outputs. Our results indicate that classifying the concerns over multiple passes through the LLM, each consisting a boolean question whether the text mentions a vaccine concern or not, works the best. GPT-4 can strongly outperform crowdworker accuracy when compared to ground truth annotations provided by experts on the recently introduced VaxConcerns dataset, achieving an overall F1 score of 78.7%. We use this model in conjunction with a binary classifier to first retrieve texts that discuss vaccination and then classify which potential concerns are raised by the text.

1 Introduction

Much public health work has been focused on describing the landscape of misinformation and concerns surrounding vaccination. One such work introduces VaxConcerns, a disease-agnostic taxonomy of concerns that may drive people towards hesitancy (Stureborg et al., 2023). The VaxConcerns taxonomy organizes concerns into two levels, one of broad granularity with concern categories such as "Health Risks" and another of finer granularity with specific claim categories such as vaccines having "Harmful Ingredients" or "Specific Side-Effects". VaxConcerns is composed of 5 parent categories and 19 child classes each associated with one (and only one) parent category.

One of the proposed tasks for VaxConcerns is to classify text (e.g. blog articles, tweets) into the taxonomy. Since the taxonomy is hierarchical, and multiple concerns can be brought up by a single passage of text, this constitutes a hierarchical multi-label classification task. Hierarchical multi-label classification requires an independent binary ("present" or "not present" in the text) prediction for every label in the taxonomy. These predictions must be made separately for parent and child categories, since a text can invoke the broad category (e.g. Lack of Benefits) without a specific rationale (e.g. Existing Alternatives). For example, consider the YouTube comment: "`I don't need the vaccine! No reason to get it`", which clearly invokes the parent Lack of Benefits without invoking any of the child labels.

Using this dataset, we build automatic classifiers by prompting Large Language Models (LLMs), and investigate considerations such as (a) cost, (b) output consistency, (c) performance, and (d) prompting strategies. We ultimately build a system that strongly outperforms the best crowdsource annotation by using GPT-4-Turbo, and find four Pareto-optimal system designs to optimize cost and performance. We offer guidelines that may help public health efforts during design considerations of their automatic classifiers.

2 Methodology

2.1 Datasets

Relevant Passages To limit overall inference costs and potential false positives, we filter out any passages which are not on the topic of vaccination. We use GPT-4-Turbo and a simple prompt which

defines the topic of vaccination to label the data in our training set. Later, we build a keywords based classifier using the labels provided by GPT-4-Turbo to create a highly scalable binary classifier. The test set consists of 25 articles scraped from 2 known anti-vaccination blogs, annotated at the paragraph level for related/unrelated to vaccination. There are 504 total annotated passages in the test set.

Concerns To build automatic systems for classifying which concerns are raised by a passage of interest, we experiment on the dataset of anti-vaccination passages introduced by Stureborg et al. (2023). We use this dataset as our test set, and contains 200 fully labeled passages (each passages maps to one binary label for each of the 24 concerns in VaxConcerns). An example passage from the dataset is shown below:

> "The Minister of Fear (the CDC) was working overtime peddling doom and gloom, knowing that frightened people do not make rational decisions - nothing sells vaccines like panic."

2.2 Short-hand Notation for Prompting Strategies

A *fully labeled* example is a passage which has a binary ("present" or "not present") prediction for every label in the taxonomy. These predictions can be either batched together all at once (single-pass) or split into multiple passes (multi-pass) to the model. Further, the choice of which groups to batch together introduces more options. Labels can be shown on their own as a binary choice (binary), as a short flat list of options (multi) or as a tree-structure (hierarchical, or hrchl). For the purpose of notation, we will combine the notion of passes and label combinations by combining these short-hand names using the convention defined in Cartwright et al. (2019).

3 Results

GPT-4-Turbo correctly classifies whether passages are on the topic of vaccination 91.9% of the time, with a precision of 89.6%, recall of 97.7%, and F1 score of 93.5%. Note that this (fortunately) favors recall over precision, which ensures lower false negatives in our downstream system.

We look at various prompting strategies for hierarchical multi-label classification of vaccine concerns. Table 1 shows that for GPT-4, the best strategy is to run inference several times, asking for a single, binary label on every iteration.

Table 1: **F1 Score (%) by each model under various prompting strategies**. We use the notation defined by Cartwright et al. (2019). Values are the mean score from two runs with temperature 0 and 1.

Prompting	GPT-3.5-Turbo	GPT-4	GPT-4-Turbo
single-pass multi	65.63	76.41	77.03
single-pass hrchl	63.95	77.05	**78.33**
multi-pass multi	**67.51**	76.99	75.83
multi-pass hrchl	64.25	75.02	72.87
hrchl-pass multi	59.73	78.56	74.70
hrchl-pass binary	57.59	74.19	74.22
multi-pass binary	61.81	**78.65**	77.45

As multi-pass binary is the most expensive prompting strategy, we also look at the relationship between inference cost and model performance, as shown in Figure 1.

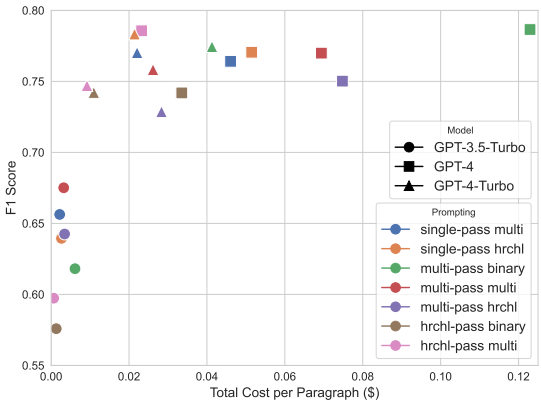

Figure 1: **Total cost versus performance by model and prompting strategy**. The approximate curve above describes the current abilities to trade off performance for cost in LLMs. Running the best model on $10,000$ passages would cost approximately 220 USD.

4 Conclusion

We explore use of LLM classifiers to monitor vaccine concerns online. In order to scale these systems, we develop training datasets labeled by LLMs and defined a target curve of cost versus performance tradeoffs which must be exceeded in order to claim success by future models. We find that LLMs have higher accuracy than crowdsource workers on this task, indicating the potential upside of using these datasets for downstream training.

Limitations

Limited Model Selection. We make use of both instruction tuned chatGPT models, as well as two Llama-2 variants. Due to constraints in the project scope, time, and performance of Llama-2 (especially on format errors), we performed most of the experiments using the GPT models exclusively. This is problematic for a few reasons. These models are opaque systems, with little insight into the training procedures and training data, pre- and post-processing that is carried out, and other manipulations by OpenAI. However, this is a difficult limitation to entirely get rid of due to the strength of this class of models. GPT models are SOTA across many tasks in NLP, and ignoring them would be to ignore the currently most relevant models in the field.

Dataset Size and Focus. We use the dataset provided by Stureborg et al. (2023). This dataset only contains 200 passages and 4,800 passage-label pairs for evaluation. This reduces the statistical power of any findings our results show. Our experiments are therefore preliminary and we believe helpful, but further experiments have to be carried out to verify the trends we have uncovered. Further, the dataset focuses exclusively on anti-vaccination blog text. This is very different from the domain of social media text both in style and content, and we can therefore not know whether or the performance will generalize out of this domain.

Limited to a Single Taxonomy. In our experiments, we only investigate a single taxonomy as the target for hierarchical multi-label classification. It is possible that results will vary if introducing taxonomies of different sizes or even domains. That being stated, the core motivation for this work is to allow for the detection of vaccine concerns in online text. For this purpose, VaxConcerns is the best option due to its crowdsource viability, high quality evaluation set, and the fact that it is disease-agnostic and therefore robust to new viruses/diseases that may arise in a future global health crisis. We therefore believe furthering this resource is of value to the healthcare community.

Ethics Statement

This research investigates text claiming misinformation, falsehoods, and general propaganda from anti-vaccination sources. Sometimes, sources we examine are engaged in other seemingly conflicting interests such as selling herbal supplements or alternative medicines. To limit potential misconstrued or misquotations of our work, we therefore limit quoting such passages unless absolutely necessary for understanding our work. We also do not link to or mention directly these websites from our work, though the sources are shared through our public datasets.

Further, given that tools or recommendations coming out of this research may be (directly or indirectly) used in future public health efforts we recognize the potential implications on individual's health and healthcare choices. To ensure we trust the conclusions of our research, we examine by hand the ground-truth labels provided in each dataset.

Acknowledgements

We thank Jun Yang for his advise on research directions and feedback. This work was supported by NSF award IIS-2211526 and an award from Google.

References

Mark Cartwright, Graham Dove, Ana Elisa Méndez Méndez, Juan P. Bello, and Oded Nov. 2019. Crowdsourcing multi-label audio annotation tasks with citizen scientists.

Rickard Stureborg, Bhuwan Dhingra, and Jun Yang. 2023. Interface design for crowdsourcing hierarchical multi-label text annotations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17.