# Scalable Classification of Online Vaccine Concerns

Rickard Stureborg, Chloe Q. Zhu, Christopher Li, Bhuwan Dhingra
Duke University

## 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.

## Introduction

**Taxonomies for Vaccine Hesitancy**
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., 2024).

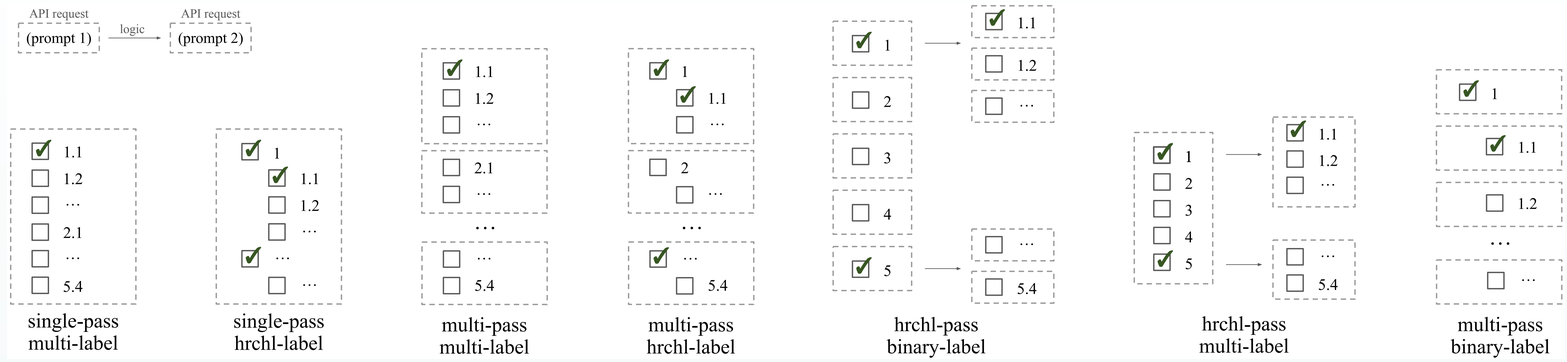
The VaxConcerns taxonomy organizes concerns into two levels, one of broad granularity with concern categories such as “Health Risks” and another of finer granularity with specific claim categories such as vaccines having “Harmful Ingredients” or “Specific Side-Effects”.

**Hierarchical multi-label text annotation**
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 multilabel 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”

It clearly invokes the parent Lack of Benefits without invoking any of the child labels.




## Methodology

**Example target output**

```
{"VaxConcerns 1": "Yes",
 "VaxConcerns 1.1": "Yes",
 ...
 "VaxConcerns 2": "No",
 ...
 "VaxConcerns 5.3": "Yes",
 "VaxConcerns 5.4": "No"
}
```

**Classification Models**
To build automatic classification systems, we use Large Language Models (LLMs). LLMs have recently shown a high degree of accuracy across many tasks, even under zero-shot settings.

We investigate five models in this work:

- GPT-3.5-Turbo, checkpoint `gpt-3.5-turbo-0613`
- GPT-4, checkpoint `gpt-4-0613`
- GPT-4-Turbo, checkpoint `gpt-4-1106-preview`
- Llama-2-7b, checkpoint `llama-2-7b-chat-hf`
- Llama-2-13b, checkpoint `llama-2-13b-chat-hf`

We investigated larger models (such as Llama-2-70b) but did not have enough compute at the time of the experiments to run this size model.

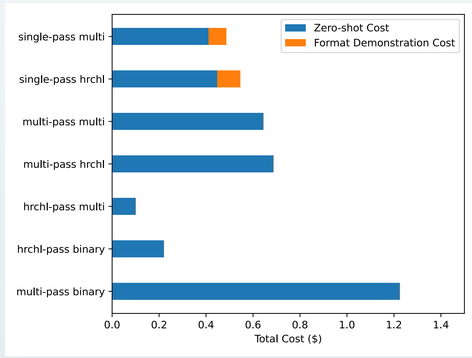
To prompt the models to give us accurate predictions, we use the following system and user prompts:

```
You are a healthcare expert. Your job is to find out
vaccine related concerns from a given PARAGRAPH. You
will be given a map of all concerns, where the key is
as VaxConcerns 1.1 and values are the actual
concerns.

Here’s the CONCERN MAP: {concerns}

Please read the PARAGRAPH and tell me whether the
concern is mentioned in the PARAGRAPH one by one. In
your response, only return a map, same length as the
CONCERN MAP, with the key exactly the one in the map,
values as Yes or No.
```

To produce a fully labeled prediction for a passage of interest (that is, a binary decision for every label in the taxonomy) we consider 7 potential prompting strategies, shown above.

## Results

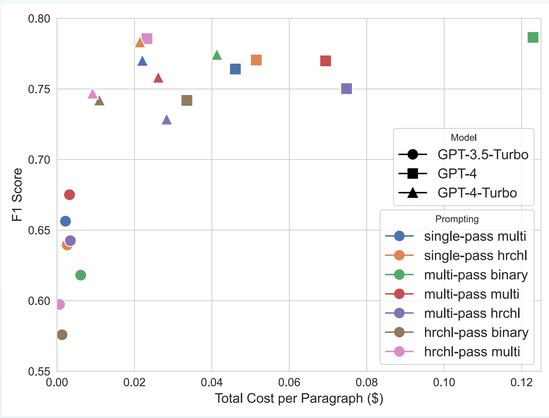
**Inference Costs for each prompting strategy, incl. format demonstrations**
Despite using the same model, costs can vary massively (multi-pass binary is 9.4x more expensive than hrchl-pass multi). Costs of single-pass strategies are higher due to needing format demonstrations to reduce the failure rate to a reasonable level, with GPT-3.5-Turbo for example (Figure 2). Overall, performing hierarchical passes in small groups of labels is the cheapest prompting strategy by far, while binary labeling is the most expensive. Cost is given for the whole dataset of 200 examples.




**F1 Score (%) by each model under various prompting strategies**
Values are the mean score from two runs with temperature 0 and 1. GPT-4 scores better than GPT-4-Turbo on average with a mean 76.7% F1 score compared with GPT-4-Turbo’s 75.8% across all prompting strategies.

| 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 |

**Total cost vs performance by model and prompting strategy**.
Throughout our experiments, we show a positive relationship between cost of prediction and performance. However, this relationship is largely driven by model cost differences. Yet, the relationship between the cost of prompting strategies and their performance is positive. This could potentially hint that models perform better when focusing on fewer labels per generation.




## 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.

## Acknowledgements

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