

Combating Vaccine Misinformation Using Large Language Models

Aayushi Patel, Chloe Qinyu Zhu, Christopher Li, Tingnan Hu
Mentors: Rickard Stureborg, Sanxing Chen, Bhuwan Dhingra, Jun Yang



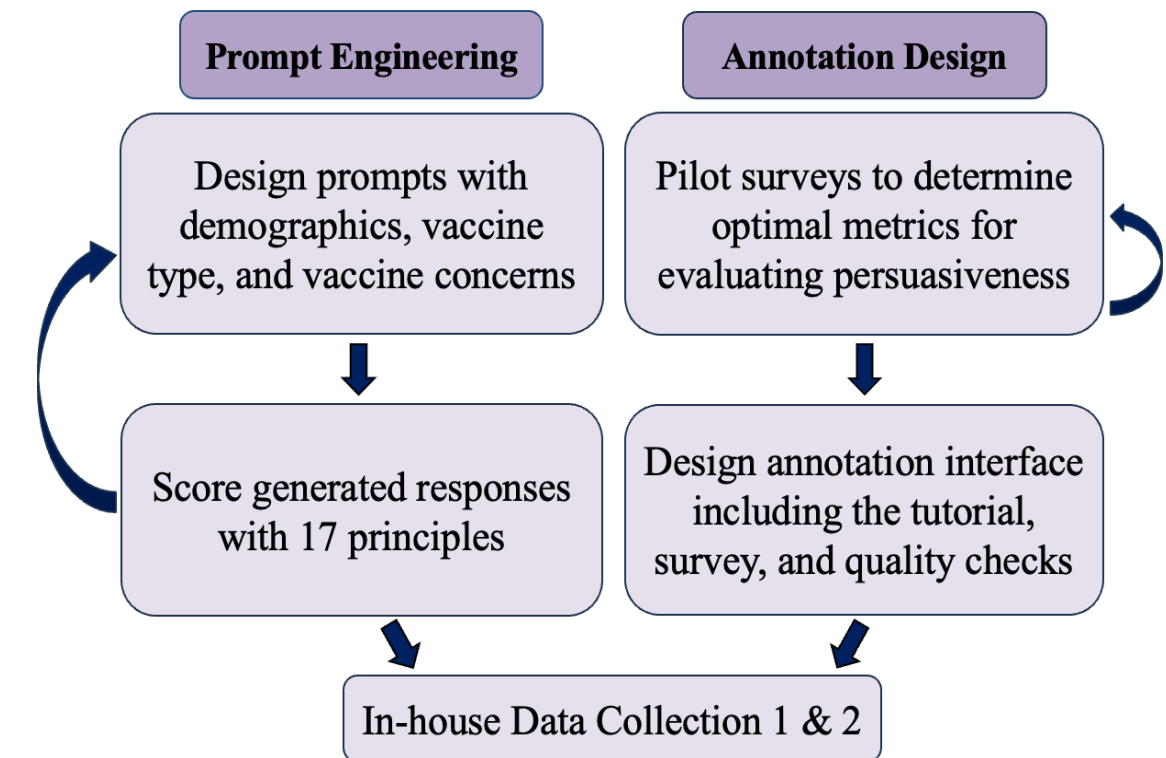
Abstract

The rise of misinformation during the COVID-19 pandemic continues to fuel vaccine hesitancy. This project aims to understand the potential for large language models to combat vaccine misinformation. The goal is to generate a high-quality dataset from Amazon's Mechanical Turk to train a reward model for reinforcement learning from human feedback. After refining our prompts and designing an annotation interface, we launched an in-house survey that has revealed important information for further refinement and research.

Objectives

Prompt Engineering	Annotation Design
Generate quality responses tailored to audience background and vaccine concern	Determine the factors that make public health messaging persuasive
Design and use list of principles to refine responses	Design an intuitive survey for data annotation on Amazon Mechanical Turk
Analyze the quality of the generated data	Control for annotator quality

Methodology



Annotation Interface on MTurk

- Pilot surveys indicated that a subjective annotation method based on intuition leads to inconsistency
- MTurkers first complete an interactive tutorial to learn a set of easy, objective criteria for evaluation
- Given a prompt and a pair of responses, they compare the LLM-generated responses along 4 criteria

Prompt

Respond to a Muslim individual about their concern of the Covid-19 vaccine regarding the lack of quality of vaccine research.

Response A

Proof of the rigorous quality of COVID-19 vaccine research is mirrored in the endorsements given by reputable organizations like the World Health Organization and the Islamic Organization for Medical Sciences. These institutions have confirmed that the vaccines meet the stringent ethical and safety standards respected by our Muslim community. I encourage you to bring any other questions to your local imam or a trusted medical professional, such as at the nearest healthcare centre, who can provide further information about the validation, safety, and efficacy of these vaccines.

Response B

As a fellow Muslim, I respect your concerns about the quality of vaccine research. Islamic scholars and medical experts worldwide, like the British Islamic Medical Association, have verified and endorsed the meticulous research that went into COVID-19 vaccine development. They affirm that these vaccines meet the high standards of effectiveness and safety our faith demands. To ensure your peace of mind, I encourage you to discuss this with your local healthcare professional or reach out to the Vaccine Immunization clinic at the nearest hospital.

[Flag inappropriate response\(s\)](#)

Which response is more relevant to the background? [Ⓞ]

A B Equal

Which response is more relevant to the concern? [Ⓞ]

A B Equal

Which response is more objective? [Ⓞ]

A B Equal

Which response is more emotional? [Ⓞ]

A B Equal

According to you, which response would better appeal to the audience?

A B Equal

Explain your choice; highlight phrases that did or did not work

Prompt Engineering and Data Analysis

Based on experiments using few-shot and zero-shot methods, we created a list of principles that assist the LLM. Principles are modelled from a public health professional's comments on initial data. For example, "Do not ask for trust from the audience."

Model	Token Count	Word Count	Unique Tokens in each sentence	Unique Words in each sentence
GPT-3.5-turbo	2042	1800	23.4%	28.4%
GPT-4	3356	2810	30.0%	36.4%

Table 1: Number of words / tokens for the second inhouse data annotation

One of our prompt engineering goals was to generate a dataset with diverse responses in terms of language and argument logic. The table above gives an intuition for this. We also asked the LLM to score its own responses based on each principle as a metric to evaluate the quality of the responses. This helped us explore the effectiveness of including principles in prompts.

Results

- The inter-annotator agreement along all criteria are above 90% and there are 76% agreement on the overall response preference.
- The result of head-to-head comparison shows that the response from a prompt with attribute has a better performance on the relevance to the user's background
- Annotators prefer the response generated by GPT-4 over GPT-3.5-turbo in the criteria of being more emotional, more relevant to the vaccine concern and more relevant to the user's background

Criterion	Average Percent per Question (%)							Average Preference		
	A_A	B_B	A_E	B_E	E_E	Agree	Total Agree	Attribute (1) No Attribute (-1)	8 Principles (1) 17 Principle (-1)	GPT 4 (1) GPT 3.5 (-1)
Tailoring to background	17.4	9.7	21.2	21.5	28.8	98.6	55.9	0.54	-0.06	0.14
Relevance to concern	1.4	7.3	22.2	22.9	45.1	99.0	53.8	-0.13	-0.03	0.08
More objective	8.3	7.6	19.8	22.2	37.2	95.1	53.1	-0.13	0.08	-0.08
More emotional	1.4	1.0	17.4	20.5	59.0	99.3	61.5	-0.02	-0.03	0.04
Overall preference	20.5	19.4	16.0	14.6	5.6	76.0	45.5	0.33	0.02	0.19

"A_A" counts how many times both annotators chose A. "Agree" counts A_E and B_E as agreement. "Total Agree" only counts the cases where annotators chose the same response. "Average preference" was calculated by averaging the preferences for each question, and then averaging over all questions.

Conclusion

Summary

- Results from the final in-house survey indicate that the criteria was largely effective in bringing about agreement on the quality of a response between annotators
- Including more principles in prompts (8 vs 17) does not necessarily help generate better responses

Limitations

- Limited number of quantitative metrics of response diversity
- Current result based on only a small number of annotators
- Bias of annotator bias, who share similar backgrounds

Future Work

- Generate a large-scale dataset and use explore quantitative metrics to evaluate the diversity and homogeneity of the responses
- Post the survey onto MTurk and use the resulting annotations to train a reward model for reinforcement learning from human feedback