Tailoring Vaccine Messaging with Common-Ground Opinions

Rickard Stureborg **Sanxing Chen** **Ruoyu Xie** **Aayushi Patel** **Christopher Li**

Chloe Qinyu Zhu **Tingnan Hu** **Jun Yang** **Bhuwan Dhingra**

Duke University

{rickard.stureborg, sc829, rx55,
aap66, cl619, qz124, th311}@duke.edu,
{junyang, bdhingra}@cs.duke.edu

Abstract

One way to personalize chatbot interactions is by establishing common ground with the intended reader. A domain where establishing mutual understanding could be particularly impactful is vaccine concerns and misinformation. Vaccine interventions are forms of messaging which aim to answer concerns expressed about vaccination. Tailoring responses in this domain is difficult, since opinions often have seemingly little ideological overlap. We define the task of tailoring vaccine interventions to a Common-Ground Opinion (CGO). Tailoring responses to a CGO involves meaningfully improving the answer by relating it to an opinion or belief the reader holds. In this paper we introduce TAILOR-CGO, a dataset for evaluating how well responses are tailored to provided CGOs. We benchmark several major LLMs on this task; finding GPT-4-Turbo performs significantly better than others. We also build automatic evaluation metrics, including an efficient and accurate BERT model that outperforms finetuned LLMs, investigate how to successfully tailor vaccine messaging to CGOs, and provide actionable recommendations from this investigation.[1]

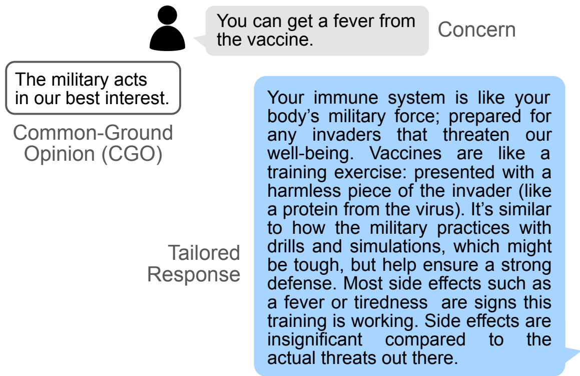

Figure 1: **Example of a Tailored Response to answer a Vaccine Concern while Tailoring to a Common-Ground Opinion**. The response above is a shortened version of a model response for the task of tailoring to CGOs. The responses is able to relate two seemingly unrelated topics: side effects from vaccination and a strong support for the military. It is strengthened by language and analogies that may appeal to the user without becoming manipulative. This work creates an evaluation framework and benchmarks different LLMs on their ability to generate such tailored responses.

1 Introduction

Finding common-ground has long been understood as an important aspect of language and communication (Gibbs et al., 1988). Creating this connection during communication can establish a sense of trust and is an inclusive goal to strive for when preparing personalized messaging. We can establish common ground on many things, such as beliefs or opinions, although it may be unclear whether such opinions can be seamlessly related to the main message being communicated.

One particularly relevant use case for establishing common ground is vaccine messaging. Loss of trust and polarization are increasingly pervasive issues impacting vaccine hesitancy (Ozawa and Stack, 2013). Those involved with vaccine messaging (e.g. vaccine navigators, public health organizations, healthcare professionals) could be aided by further tools to help write responses, personalize information, or generate creative ideas for their messaging efforts.

Large Language Models (LLMs) have demonstrated high performance in controlled text generation. Recent research has increasingly explored personalization given the abilities of these models (Soni et al., 2022; Welch et al., 2022; Salemi et al., 2023) with many pointing out remaining challenges (Kirk et al., 2023). How LLMs handle personalization in terms of opinions and beliefs remains under-explored.

To this end, we propose a task for tailoring vaccine messaging towards common-ground opinions

[1]TAILOR-CGO dataset and code available at: https://github.com/rickardstureborg/tailor-cgo

Findings of the Association for Computational Linguistics: NAACL 2024, pages 2553–2575

(Figure 1). By providing common-ground opinions and instructing LLMs to tailor towards them, we hope to help address the imbalance and inequity in the broader information landscape surrounding vaccination. Given that some work (Santurkar et al., 2023) has pointed out these models may exhibit biased opinions that do not reflect diverse or underrepresented groups, we highlight and evaluate current major LLMs' ability to perform this task by building comprehensive automatic evaluation metrics.

This paper provides the following contributions:

- Evaluation of several major LLMs on their ability to use 'Common-Ground Opinions' in vaccine messaging.
- TAILOR-CGO, a comprehensive and high-quality dataset for training and evaluation.
- Actionable recommendations of which CGOs are most useful to address a given concern.

2 Related Work

Personalization has been explored in various NLP tasks, most notably within the domain of dialogue response generation (Wang et al., 2019; Zhang et al., 2018; Zheng et al., 2019; Joshi et al., 2017). Using natural language prompts, language models can generate texts that align with demographics or identities, reflecting cross-cultural values (Arora et al., 2023), political ideology (Simmons, 2023), or opinions on societal issues (Argyle et al., 2023; Santurkar et al., 2023), or infer personal attributes (Wang et al., 2022). Previous work has also investigated balancing benefits and harms of personalization (Kirk et al., 2023), providing benchmarks (Salemi et al., 2023), and constructing user-conditioned language models (Soni et al., 2022; Welch et al., 2022). However, we are unaware of any work to date which examines personalization of LLM responses within the vaccine misinformation domain, or work which focuses on tailoring messages to common-ground opinions. There is also a line of work in NLP on grounding responses (Cho and May, 2020; Chandu et al., 2021; Zhou et al., 2022).

Various benchmarks have been proposed to test general LLM abilities in generating task-specific responses (Hendrycks et al., 2021; Khashabi et al., 2022; Zheng et al., 2023). Our benchmark emphasizes on a coverage of diverse opinions for controllable generation in the domain of vaccine concern and misinformation. In the era of LLMs, we have seen a renewed interest in automatic evaluation metrics of text generation, due to the need for reinforcement signals (Stiennon et al., 2020b; Rafailov et al., 2023). We demonstrate the feasibility to build automatic evaluation metrics using our data, facilitating future efforts to improve LLM generation.

Previous research of misinformation has explored classification of common concerns and misinformation topics (Coan et al., 2021; Stureborg et al., 2024b; Zhu et al., 2024), fact-checking statements (Thorne et al., 2018), or claim review (Arslan et al., 2020b,a), which determines if claims are worth fact-checking. While addressing these concerns and misinformation is important, our work aims to begin addressing the vaccine misinformation through tailored messaging.

Indeed, there is already substantial work on establishing common ground for the goal of successful communication. It is well understood in cognitive sciences (Clark and Carlson, 1982; Clark and Schaefer, 1989; Clark and Brennan, 1991). Likewise, arguments tailored to information about the target audience have been shown more effective (Hirsh et al., 2012; Hadoux and Hunter, 2019). However, reliably generating such responses automatically is an open research question which could serve as the foundation for future research into effective communication practices or persuasion.

3 TAILOR-CGO Dataset Creation

In this section we describe the task of tailoring a response to a CGO, introduce and describe the dataset and its components, and outline how the dataset is labeled. The final dataset contains $22,400$ unique tailored responses from 6 different LLMs, labeled with a mix of absolute scores or pairwise rankings.

3.1 Task Definition

We define our task of tailoring to common-ground opinions as follows. In response to an expressed *concern* about vaccination, the task is to generate an intervention tailored to a given *common-ground opinion (CGO)*, which should act as the basis for framing the response.

A successfully tailored response should meet five criteria: (1) It should fully answer the concern to promote vaccination or encourage engaging further with health professionals. (2) The opinion should be used or referred to in the response, either directly or indirectly. (3) The response should accept the opinion as true, rather than refute it. (4) The answer to the concern should be meaningfully linked to the opinion in some manner. Finally, (5) the use of the opinion should strengthen the response to the expressed concern, such that the removal of the opinion would weaken the response.

3.2 Concerns and Opinions Statements

To provide inputs for our task, we need explicitly stated concerns and opinions. For concerns, we utilize the VaxConcerns taxonomy from Stureborg et al. (2024b) and then prompt GPT-4 to generate a large variety of concern statements, as detailed in Appendix I. Some of these statements refer to specific vaccines (COVID-19, HPV, MMR, Influenza, and Yellow Fever), while others are agnostic towards vaccine type. There are 1166 total concern statements, all cleaned by the authors of this paper and mapped to the VaxConcerns taxonomy. One example concern is shown below, where the sampled vaccine-type was *'MMR'* and the concern category was *'2.4: Lack of benefits → Insufficient risk'*.

> "Measles, mumps, and rubella cases are so rare nowadays, the MMR vaccine seems unnecessary."

For opinions, we generate statements by paraphrasing questions from OpinionQA (Santurkar et al., 2023). These public opinion survey questions (originally sourced from PEW Research polls) have awkward phrasing which make them difficult to use as-is. Therefore, we use GPT-4 to convert these questions to single-sentence statements expressing the opinion as a fact. For example, one such opinion statement reads:

> "In general, society tends to look up to men who are manly or masculine these days."

3.3 Response Generation

To produce candidate responses for annotation, we prompt LLMs to tailor to common-ground opinions. However, we do not want downstream uses of our dataset to rely heavily on a specific combination of model and prompt. Therefore, we use a variety of systems for producing candidate responses. We conducted an extensive qualitative analysis of the generated responses to find common issues and strengths, which are detailed in Appendix F.

We generate a mix of candidate responses that allow for both **intra-opinion** comparisons, where responses are tailored to the same opinion, and **inter-opinion** comparisons, where responses are tailored to different opinions. Specifically, we create "blocks" of 4 response generations, with each block structured as follows:

```
Concern1 + Opinion1 -> ResponseA (rA)
Concern1 + Opinion1 -> ResponseB (rB)
Concern2 + Opinion2 -> ResponseC (rC)
Concern2 + Opinion3 -> ResponseD (rD)
```

Here, (rA, rB) provides an intra-opinion comparison and (rC, rD) provides an inter-opinion comparison. For each block, we randomly sample 2 concerns and 3 opinions with replacement. For response generation, we randomly sample model, prompt, and temperature from their respective domains as explained in the following subsections. For half of our blocks, we fix these system parameters (model, prompt, temperature) between rA, rB and between rC, rD. This design gives opportunities to compare responses sampled under identical settings as well as different ones.

We manually inspected 700 sample responses of the 1546 unguided responses. Within this sample, we found 21 instances of causes of clear response failure, demonstrating a 3% failure rate. There were several distinct failure modes that we identified with varying frequencies: (a) 1.9% - The model directly or indirectly assumes the identity of a human 13 times, (b) 0.3% - The model assumes an identity for the person it is responding to 2 times, (d) 0.3% - The model responds to a different vaccine concern than that which is provided in the prompt 2 times, (f) 0.3% - The model crafts a response that contains a template element like "hey [friend's name]" 2 times (c) 0.1% - The model explicitly disagrees with the CGO 1 time, and (e) 0.1% - The model makes a factually incorrect or ambiguous statement 1 time.

Models We consider six models: Llama-2, Vicuna, WizardLM, GPT-3.5, GPT-4 and GPT-4-Turbo. In early experiments, such as collection of the dev set, we use smaller model sizes (13b models) for candidate response generation. However, in the final round-2 and round-3 training and test data, we use the most powerful models possible on our hardware (70B parameter models). A full list of the model checkpoints we use, with complete citations and links, are shown in Table 1. We randomly sample temperature uniformly between 0 and 1 during generation to encourage a diversity of both creative and high-likelihood outputs.

Table 1: **Models checkpoints used in this work**.

Model Checkpoint
llama-2-13b-chat-hf
llama-2-70b-chat-hf
vicuna-13b-v1.5
vicuna-33b-v1.3
WizardLM-13B-V1.2
WizardLM-70B-V1.0
gpt-4-1106-preview
gpt-4-0613
gpt-3.5-turbo-0613

Prompting To increase the diversity of our prompts, we make use of role-playing (Wang et al., 2023), chain-of-thought (Wei et al., 2022), and pointed instructions for generation (Ouyang et al., 2022). We consider all combinations of the following strategies:

- **Roles**: We ask the model to produce different text styles by instructing it with a role (e.g. parent, doctor, redditor) to take when responding. We use 10 different roles.
- **Chain of Thought** (CoT): The model is asked to first think through the process of generating the tailored response by writing out a plan, and is then subsequently prompted to produce the final response. We compare the quality of responses with and without CoT.
- **Guidelines**: The prompt includes written instructions for principles to follow or general behaviors to avoid. These are sourced from 1) CDC guidelines on vaccine messaging, 2) the criteria (§3.1) for what makes well-tailored responses, and 3) general guidelines to avoid issues noticed in responses during the development phase. To avoid increasing the size of the prompt prohibitively, and to ensure diversity, we randomly sample five such guidelines from all sources uniformly and include them in prompts marked with the 'guided' flag. Approximately half of responses are produced with a guided prompt.

Overall, these strategies are each randomly sampled, leading to 40 potential prompts types, with more than 500K possible unique combinations (due to the random sampling of guidelines).[2]

3.4 Human Annotation

We collect data in 3 rounds. First, we annotate tailored responses using both absolute scoring and relative preferences in parallel randomly assigned conditions. This is used as our dev set (Yellow). We find that relative preferences yields higher data annotation quality, so in a second round we invite back the 8% most accurate crowdsource workers to label our test (Blue) and training (Green) sets. Third, we label a much larger set of tailored responses using LLM evaluators. Figure 2 shows a useful diagram explaining the various partitions of the final dataset. We reference these partitions frequently in this section.

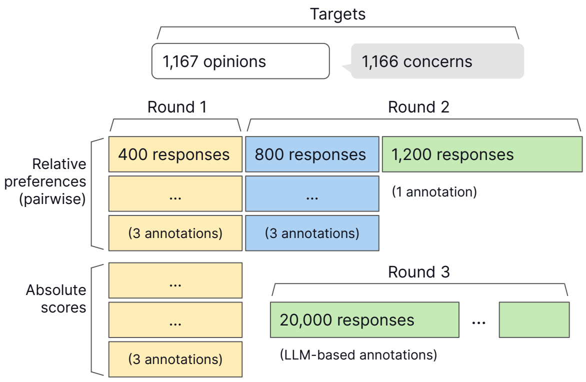

Figure 2: **TAILOR-CGO dataset partition sizes**. Colors indicate which train/dev/test split each partition is included in. Green = train, Yellow = dev, Blue = test. Relative preferences are collected by asking which of two responses is better tailored, while absolute scoring asks for a 1-5 score for a single response. Both Dev and Test sets (Yellow and Blue) contain 3 independently collected annotations per input response, represented by 3 stacked boxes. The training set (Green) contains just one annotation per response to maximize diversity.

Anotation Scheme To score the quality of candidate tailored responses, we collect preferences using crowdsource annotations through Amazon Mechanical Turk (AMT). Given a particular concern-opinion and response pair, annotators are asked to judge the quality of the response based on the criteria listed under Task Definition. We consider two potential annotation schemes:

- **Absolute scoring:** Annotators are asked to make absolute judgments of how well-tailored each response is on an individual basis. These judgments are given on a 1-5 scale, ranging

[2]40 prompt types come from the 10 roles, 2 settings for CoT (on or off), and 2 settings for guidelines (on or off). Accounting for all possible sampled subsets of 5 from the 22 guidelines gives us 500K unique combinations ($20 + 20 \cdot {}_{22}C_5$)

from *Very poorly tailored* to *Very well tailored*. However, when performing within-team annotations, we found that Likert-style scales were difficult to use since it is hard to calibrate what level of quality warrants being *well tailored* versus *very well tailored*.

- **Relative preference:** To circumvent the difficulty of absolution scoring, we instead ask annotators to make relative judgments, comparing two responses against one another and ranking them according to how well-tailored they are. This label set is ordinal, with *Response A* (is better), *Equal*, and *Response B* (is better), in order.

Our round-1 results indicated that collecting relative preferences leads to higher agreement about the eventual ranking of responses. Annotators labeling preference of two responses directly agree on 57.3% of labels, while marking an absolute score for a single response yielded agreement of just 44.2%. One reason for this is an increase in ties: absoluting scoring yield 19.5% tied preferences between pairs, while directly asking for preference in the pair results in 8.5% ties. This increase in ties is problematic not just for annotator agreement, but potentially also for labeling efficiency, since we lose information about nuanced differences in pairs with tied absolute scores.

Annotator Selection We train our crowdsource workers using a short 9-example tutorial, where they first annotate and are then given feedback with reasons motivating the correct choice as well as highlights over the responses with hover-text providing further explanations. We open our annotation task to all annotators which have an AMT 'masters' qualification and at least 2,500 approved HITs at an approval rate of at least 99%. These very selective criteria ensure only the highest performing annotators. After the tutorial, annotators complete a short (3-example) entrance exam of easier, expert-labeled questions to further qualify; we remove any worker who incorrectly answers at least one of the questions. During annotation, we randomly insert, for 5% of shown examples, attention checks that instruct workers to select certain options to ensure they are fully reading the passages.

We invite only the top 8% of annotators for round-2 annotations based on their scores on the tutorial examples. We pay approximately 15-20 USD per hour for the workers in our round-1 annotation, and 25-30 USD per hour for those in round-2 annotations. We offer $100 bonuses to the top 25% of annotators in round 2 to incentivize high-quality annotations, and regularly examine their annotations and offer feedback through direct messaging during data collection.

Annotation Strategy	Percent agreement
Dev set (round 1 - absolute)	44.2%
Dev set (round 1 - relative)	57.3%
Test set (round 2 - relative)	73.2%
Bai et al. (2022)	63.0%
Stiennon et al. (2020a)	73.0%
Ouyang et al. (2022)	**77.3**%

Table 2: **Agreement between crowdsource annotators when presented with two tailored responses**. Our test set shows comparable amounts of agreement to work in instruction tuning, despite a highly subjective and difficult task and offering 3 options (*A*, *Equal*, *B*) rather than two (*A*, *B*) in the annotation interface. Note that the human-labeled training set was collected together with the test set, but does not have multiple annotations and is therefore left out of this table.

Further details of the annotation platform, including screenshots, are available in Appendix C

Inter-annotator Agreement Round-2 annotations (test set marked in blue and human-based training set marked in green in Figure 2) show much higher agreement than round 1. As shown in Table 2, crowdsource annotators agree on 73.2% of labels, comparable to previous works (Stiennon et al., 2020a; Ouyang et al., 2022) on annotating human preferences for reward modeling.

4 Automatic Evaluation

While human annotation shows a high agreement on evaluating LLM responses, conducting it at scale to study the various settings we consider is prohibitively expensive. Hence, in this section we use the annotations collected above to develop automatic evaluation metrics for the TAILOR-CGO task. The goal of automatic evaluation is to provide a cheap alternative for labeling candidate responses and automatically evaluate or compare models. It also allows for deeper analysis of trends when human annotation becomes restricted by scale. We describe our approach below. First, we prompt a generative model (GPT-4-Turbo) to score responses directly. Second, to further reduce cost, we fine-tune open-source language models (BERT and Llama-2) using the results from the former and/or human annotations.

4.1 Zero-shot Prompting

We use GPT-4-Turbo as an automatic evaluator by prompting it in a zero-shot setting. Specifically, we use G-Eval's (Liu et al., 2023) instruction template by replacing the definition and evaluation steps in its provided prompts with descriptions explaining what a well-tailored response is. G-Eval is an automatic evaluation framework for text summarization built on top of GPT models, but has been adopted to other tasks as well.

Because log probabilities are no longer available for GPT models, we sample 100 predictions at temperature 1.0 for each response, stopping the model outputs at 10 tokens. We then parse these outputs to collect discrete 1-5 scores, and a mean score is calculated over the 100 samples thereby approximating the original weighted prediction by token probabilities used in G-Eval. We provide all outputs (raw and cleaned) in our dataset. When needed, these absolute scores can be paired together with another response answering the same concern, and by comparing the scores a relative preference can be reported. This is how we perform evaluation on the Test set, and also how we provide large labeled datasets to our finetuning methods.

4.2 Fine-tuning

We further explore the possibility of using open-source language models to perform automatic evaluation, thus reducing the cost on closed-source API inference. We test both an encoder-only model (BERT-base) and an auto-regressive model (Llama-2-13b) in a knowledge distillation setup, where they are trained on a large dataset of responses scored by GPT-4-Turbo. We continue finetuning beyond the Round 3 dataset by using the human-labeled data from the Dev set. In the relative preference setting, this is done by either randomly mixing in the human labeled data, or training on this data after the model converges. In the absolute scoring setting, we use a margin ranking loss.

BERT Fine-tuning We fine-tune a 110M BERT model (Devlin et al., 2019). For absolute scoring, we train for regression on a 1-5 score. For relative comparison, we train as a binary classification task (as opposed to using a parallel contrastive loss sometimes used in such similar settings (Seth et al., 2023)). The model takes as inputs the text of a sampled response pair from the round-3 training data. On top of this model, we add a linear layer that maps the final BERT layer hidden states into one (absolute scoring) or two scalars (relative preferences). Mean-squared error and cross-entropy loss are applied correspondingly. We sample approximately 10,000 responses (pairs) along with their scores from GPT-4-Turbo to construct the training data. We train using a batch size of 8 for 5 epochs, and the AdamW (Loshchilov and Hutter, 2019) optimizer with a learning rate of 2e-5, 10% linear warm-up, and linear decay to 0. Training and evaluation is done on a single NVIDIA A5000 GPU.

Llama-2 Fine-tuning For both absolute scoring and relative preference settings, we use a 13B LLaMA-2 as the base model for fine-tuning. We utilize QLoRA (Dettmers et al., 2023) for computational efficiency and use the AdamW optimizer to train the model for 5 epochs with a batch size of 4. The training is conducted on 4 NVIDIA RTX A6000 GPUs, setting the learning rate to 2e-4 with a 3% warmup. The LoRA rank and alpha are set to 64 and 16, respectively, with a 0.1 dropout between the two matrices.

We use Alpaca's training prompt format (Bommasani et al., 2021) where the instruction is replaced by the evaluation instruction and metrics, the input by the concerns and opinions, and the response by either the tailoring score or preferred response, depending on the setting. For the evaluation prompt, we use zero-shot prompting in both settings. The temperature is set to 1 for the fine-tuned LLaMA model. To extract the model's answer, we only use the first sentence of the response. For the absolute scoring setting, we search for numeric values in the sentences. For the relative preference setting, we look for either *A* or *B* in the sentence.

4.3 Performance

Table 3 summarizes each automatic evaluator's performance on TAILOR-CGO. The fine-tuned BERT model outperforms all other models. meaning the student model generalized better in this case. This result mirrors similar observations made about self-distillation, which has been shown to have a regularizing effect (Furlanello et al., 2018; Mobahi et al., 2020). Absolute scoring performed better in both GPT-4-Turbo and BERT, while relative preferences were more accurate in Llama-2. We also found that continuing finetuning with the human-labeled data after first training on LLM labeled data improved performance for all our models. In the relative preference setting, this alone improved Llama-2 performance from 69.1% to 73.9%. We observed that Llama-2 finetuning is sensitive to hyperparameters, but due to limited time and resources we were not able to fully tune these to improve beyond the performance of the BERT models.

Incorrect predictions of *Equal* should not be penalized the same as mismatches between *Response A* and *Response B*, so we compute a "lenient" version of accuracy: we ignore these errors by removing all *Equal* predictions before computing accuracy (treating them as *Abstains*).

Model	Setting	Dev	Test
GPT-4-Turbo	Pref.	58.5	69.3
	Score	65.5	76.5
Llama-2 (13B)	Pref.	**69.8**	73.9
	Score	62.7	68.7
BERT	Pref.	62.0	77.0
	Score	65.0	**80.8**

Table 3: **Accuracy (%) of Automatic Evaluators on Dev and Test Sets**. Evaluators are built in two settings: predicting relative preferences (Pref.) between two input responses, or predicting absolute scores (Score) for a single input response. Accuracy is then computed on Dev and Test sets as the percentage of pairwise preferences the model correctly ranks.

Training on relative preferences sometimes performs worse than training with absolute scores, despite the higher quality data in the human-labeled partition (§3.4). This could potentially be attributed to a loss of information: for both the case where *rA* was *much* better than *rB* and the case where it was only slightly better, the eventual label in the relative preference setting is the same, while absolute scoring distinguishes these. Therefore, there is a potential tradeoff between annotation quality (best annotation type for crowdsource workers) and training efficiency (best annotation type for models) that could be studied further. On the other hand, forcing a decision when two responses are indistinguishable may be a new source of noise.

5 Results

To better explore the usefulness of TAILOR-CGO towards improving our understanding of tailored responses, we investigate what models, strategies, and opinions work best. This section describes a series of analyses looking at factors to consider when generating responses tailored to CGOs, which may be of interest to NLP researchers as well as public health professionals working on vaccine hesistancy.

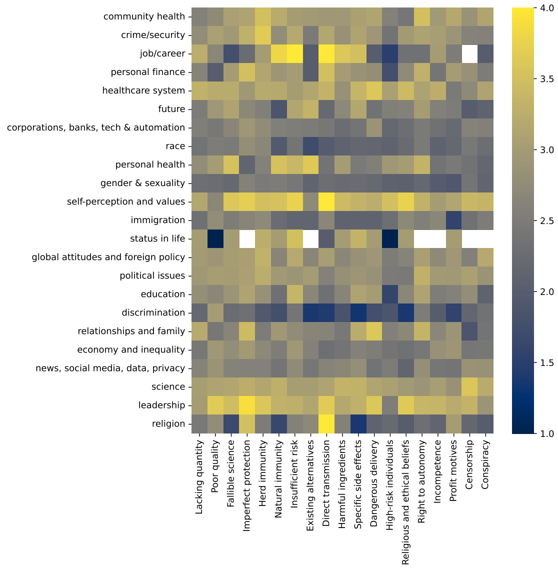

Figure 3: **Heatmap of mean scores by LLM evaluation for responses answering a concern (horizontal axis) while tailoring to a CGO (vertical axis)**. Brighter colors indicate higher scores, while white squares are nulls that were not sampled during annotation. Religion, while an opinion topic that scores poorly in our testing, seems to provide useful opinions for tailoring when focusing on the *Direct transmission* concern (see Appendix F.3 for an example output).

5.1 Opinion Selection

A driving motivation for this work is to allow analysis as to which opinions are fruitful for tailoring vaccine messaging on. To this end, we conduct an analysis of which opinions led to the best tailored responses by examining the mean scores (Figure 4) within topic clusters of opinions. We use the 24 topics proposed and annotated by Santurkar et al. (2023) for the Pew Research's American Trends Panels questions.

For the analysis in Figure 4, we collect all candidate tailored responses in the round-3 data if they tailor to an opinion associated with our topic of interest. The GPT-4-Turbo assigned score for each such responses is then computed by drawing 100 predictions and averaging; the final reported score is a second average taken over each candidate response's score. To determine a 95% (bootstrap) confidence interval, we repeat this process 10,000 times for each topic by drawing from the candidate responses with replacement. The analysis is repeated for each of the 24 topics.

However, some opinions may be better suited for use in a small subset of concerns. We therefore investigate the response quality when the CGO be-

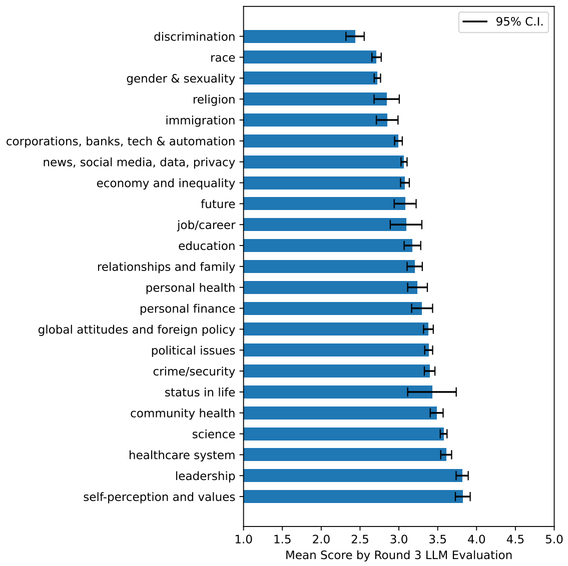

Figure 4: **Comparison of mean response quality for each CGO, aggregated by topic**. Notice that potentially controversial and problematic topics such as discrimination, race, or religion are bad targets for tailoring. The implications of this result is that using divisive topics to establish common-ground may be less useful, and using less polarized topics (self-perception) for example can result in stronger overall scores.

longs to each topic by repeating the process on each concern category in the VaxConcerns taxonomy (Stureborg et al., 2024b). The mapping between concerns and these concern categories is discussed further in §3.2.

Figure 3 shows a visualized heatmap of the mean scores of each <opinion-topic, concerns-category> combination. These results indicate that topics can indeed be better suited for tailoring responses to some concerns than others. For example, job/career opinions do very well on average with the *Insufficient risk* and *Direct transmission* concerns (approx. 4/5), but quite poorly with concerns regarding *Existing alternatives* (approx. 2/5). To improve automatic vaccine messaging, dynamically selecting the right opinions when addressing a concern could be a key strategy to improving response quality from models. We note that broad concerns (Level 1 in the VaxConcerns taxonomy) are generally easier to address than specific concerns (Level 2). This may be because specific concerns are harder to creatively link with a given CGO, while broader concerns offer more potential ways to relate the two topics.

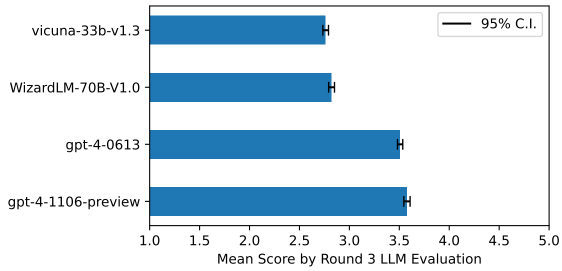

Figure 5: **Comparison of Mean Response Quality by each Model in the LLM-annotated train set partition**. All differences in the figure are statistically significant. Confidence intervals are computed through bootstrap sampling. Each model is evaluated across approximately 4,000 generated responses each to randomly sampled concern and opinion statements. We see GPT-4-Turbo produces the best tailored responses on average, just ahead of GPT-4. Open-source models still lag far behind, despite using the largest possible model sizes on our hardware.

5.2 Model

We benchmark the performance of several major LLMs through the large-scale data collection in TAILOR-CGO. Figure 12 shows a breakdown of model performances as determined by human annotators. While these results are helpful, further analysis requires larger datasets, for which we use the LLM based annotations of the round 3 partition (Figure 2) as shown in Figure 5. GPT-4-Turbo is shown to perform the best on TAILOR-CGO, and is subsequently used for an analysis as to best prompting strategies in §5.3. Additionally, Figure 6 shows a closer look of exactly where GPT-4-Turbo performs better than the next best model, GPT-4 by examining the distribution of scores assigned to tailored responses written by each model.

5.3 Prompting

We determine the best prompting method using the best model outlined in §5.2. We compare configurations of the 3 prompt dimensions described in §3.3. Roles are compared against each other in Figure 13, and the score distribution for two selected roles is plotted in Figure 14, both in Appendix G. Figure 15 in Appendix H describes the difference between Chain-of-Thought (CoT) prompting and standard prompting (non-CoT). Standard prompting is significantly better than CoT. For GPT-4-Turbo, the best prompting strategy is to use the *Health Expert* role with guidelines and non-CoT prompting.

There is not a statistically significant improvement in model responses through use of the guide-

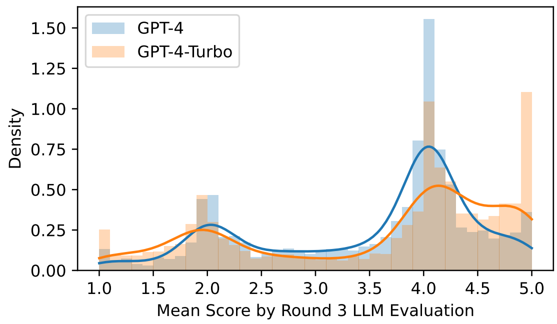

Figure 6: **Score distributions for GPT-4 and GPT-4 Turbo in round-3 dataset partition**. GPT-4-Turbo has a slightly better mean score than GPT-4 (3.58 vs 3.51). The distributions of response quality is generally comparable between the models, with GPT-4-Turbo producing notably higher proportion of scores around 4. Note that the distribution of scores seems to be bimodal, potentially due to scores either successfully finding or not finding a useful linking idea between the concern and CGO.

lines, although our qualitative observations had indicated that guidelines led to responses that better adhered to the principles included by the guidelines. However, we investigate the case where guidelines are sampled or are left out entirely, and there could be more work to determine if a subset of the guidelines significantly improves the response quality.

5.4 Expert Evaluation

We asked 3 senior public health experts to rate and comment on the quality of the generated responses to validate their potential usefulness to a vaccine navigator. They collectively rated 60 unique pairs of tailored responses, and were asked to select their preference and give comments describing their general impressions for who these responses are appropriate to. All responses were generated from the strongest model in Section 5.2, but in each pair one was a randomly chosen response, while the other was filtered as the best of 20 random generations by the BERT-based automatic evaluator from Section 4.

We were not able to find a statistically significant difference between the filtered response and the randomly chosen ones in this sample. This may be due to the model's overall strength, as evidenced by the claimed usefulness of the responses according to the experts. Experts overall described the responses as "very high quality" and noted these responses could "easily be used by vaccine navigators". Responses were sorted into four categories for overall quality: Low, Medium-Low, Medium-High, and High. Only 5% of responses were categorized as "Low" quality, 20% were "Medium-Low" 20% were "Medium-High", and 55% were "High".

Further, for 87.5% of responses, they said they would be useful to a vaccine navigator. For 55% of responses, they said they would even be comfortable with a patient reading the response. The main issues identified in responses were to do with being too technical, most often meaning they did not recommend showing these to patients. This seems to indicate that the responses by the strongest identified model in this work may be good enough to aid a vaccine navigator in their work, demonstrating the potential direct application of the framework.

6 Conclusion

We introduce TAILOR-CGO, a comprehensive and high-quality dataset for training and evaluation of tailoring vaccine interventions to common-ground opinions. We benchmark several major LLMs, finding that GPT-4-Turbo best tailors responses to CGOs. We build evaluation metrics on top of this dataset to allow cheap and accurate evaluation of models. Finally, we analyze which opinions are better suited to tailor vaccine interventions with, and provide recommendations for which opinions to select for specific concern categories.

Limitations

Finding opinions for tailoring. In this work we present methods to tailor on common-ground opinions. We assume that we are given these opinions and the intended audience believes or strongly believes in them. However, identifying beliefs is a difficult task on its own and requires further research. The easiest way to confirm the audience holds the opinion is to survey them, but doing so may affect further communication in other ways.

Crowdsource Workers' Biases. We attempt to define annotation tasks that should be "objective" regardless of who is labeling (thereby our focus on inter-annotator agreement). However, each annotator brings in their own personal biases. Opinions that seem questionable or off-putting to one annotator may influence their ranking on that example.

Tailoring versus Engagement and Subjectivity. One motivation for this work is to provide a framework for generating candidate responses in vaccine interventions. To create a well-defined task, we measure how well concepts are related (linked) in writing. What we do not directly measure is how engaging, persuasive, or applicable each response is to the actual person reading it. Such tasks are much more subjective, and require extensive invitation of participants from diverse backgrounds in order to ensure solid research findings. Individual identities are highly inter-sectional, and it can be hard to recruit participants for whom vaccine messaging is intended for.

Diversity of Generated Responses. We attempt to create a large diversity in responses by sampling many different models, under varied instructions, and with a variety of concerns and opinions. However, we are still restricted in the diversity we are able to generate on many dimensions. We begin to explore this further in B. Future work could explore tasks such as tailoring longer documents, or technical writing.

Use of Automatic Evaluation. It has been pointed out that LLM evaluators have remaining challenges such as poor performance on higher-quality models (Shen et al., 2023) or a preference for text generated by itself, likely due to a bias in low-perplexity examples (Stureborg et al., 2023). Our work therefore uses other methods of evaluation as well, such as direct accuracy comparisons with human labeled data, and averages over several evaluations as suggested in Stureborg et al. (2024a).

Ethics statement

By the nature of opinions and vaccine misinformation, the research in this paper comes with important ethical considerations. Some opinions from the PEW Research surveys are bound to be controversial and asking generative models to assume such opinions as truth may produce potentially harmful responses. Thus, to use tailored responses, there is a great deal of responsibility in selecting common-ground opinions to tailor on to avoid perpetuating harm. Some of the CDC-sourced principles for vaccine messaging address these issues, such as the principle that "CDC's communication is empathetic, respectful, non-judgmental, and never arrogant toward others". The systems which our work begins to explore are not yet performing at a high level of tailoring. Many responses are completely artificial and don't seem very persuasive. However, our results indicate that producing well-tailored responses is possible. Given the potential uses (and misuses) of such systems which this work begins to explore, it is paramount that more research is done on safety and harm within this field.

Acknowledgements

We are very grateful to Walter Orenstein, Robert A. Bednarczyk, and Lavanya Vasudevan for their help with the evaluation of response quality. A special thanks to Christopher Kjellqvist for many useful discussions on signals for personalization and persuasion, and the whole NLP group at Duke for their helpful feedback and input on the project. Thank you to the anonymous reviewers for their feedback.

References

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Arnav Arora, Lucie-aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 114–130, Dubrovnik, Croatia. Association for Computational Linguistics.

Fatma Arslan, Josue Caraballo, Damian Jimenez, and Chengkai Li. 2020a. Modeling factual claims with semantic frames. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2511–2520, Marseille, France. European Language Resources Association.

Fatma Arslan, Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2020b. A Benchmark Dataset of Check-Worthy Factual Claims. *Proceedings of the International AAAI Conference on Web and Social Media*, 14:821–829.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S.

Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models. *ArXiv*.

Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W Black. 2021. Grounding 'grounding' in NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4283–4305, Online. Association for Computational Linguistics.

Hyundong Cho and Jonathan May. 2020. Grounding conversations with improvised dialogues.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In *Perspectives on socially shared cognition.*, pages 127–149. American Psychological Association, Washington, DC, US.

Herbert H. Clark and Thomas B. Carlson. 1982. Hearers and speech acts. In *Language*, page 332–373. Linguistic Society of America.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13(2):259–294.

Travis G. Coan, Constantine Boussalis, John Cook, and Mirjam O. Nanko. 2021. Computer-assisted classification of contrarian claims about climate change. *Scientific Reports*, 11(1):22320. Number: 1 Publisher: Nature Publishing Group.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born again neural networks. In *International Conference on Machine Learning*, pages 1607–1616. PMLR.

Raymond W. Gibbs, Rachel A.G. Mueller, and Robert W. Cox. 1988. Common ground in asking and understanding questions. *Language and Speech*, 31(4):321–335. PMID: 3271894.

Emmanuel Hadoux and Anthony Hunter. 2019. Comfort or safety? gathering and using the concerns of a participant for better persuasion. *Argument Computation*, 10:1–35.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Jacob B. Hirsh, Sonia K. Kang, and Galen V. Bodenhausen. 2012. Personalized persuasion: Tailoring persuasive appeals to recipients' personality traits. *Psychological Science*, 23(6):578–581. PMID: 22547658.

Chaitanya K Joshi, Fei Mi, and Boi Faltings. 2017. Personalization in goal-oriented dialog. *arXiv preprint arXiv:1706.07503*.

Daniel Khashabi, Gabriel Stanovsky, Jonathan Bragg, Nicholas Lourie, Jungo Kasai, Yejin Choi, Noah A. Smith, and Daniel Weld. 2022. GENIE: Toward reproducible and standardized human evaluation for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11444–11458, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A. Hale. 2023. Personalisation within bounds: A risk taxonomy and policy framework for the alignment of large language models with personalised feedback.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Hossein Mobahi, Mehrdad Farajtabar, and Peter Bartlett. 2020. Self-distillation amplifies regularization in hilbert space. *Advances in Neural Information Processing Systems*, 33:3351–3361.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Sachiko Ozawa and Meghan L Stack. 2013. Public trust and vaccine acceptance-international perspectives. *Human Vaccines & Immunotherapeutics*, 9(8):1774–1778. PMID: 23733039.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. Lamp: When large language models meet personalization.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Dev Seth, Rickard Stureborg, Danish Pruthi, and Bhuwan Dhingra. 2023. Learning the legibility of visual text perturbations. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3260–3273, Dubrovnik, Croatia. Association for Computational Linguistics.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics.

Gabriel Simmons. 2023. Moral mimicry: Large language models produce moral rationalizations tailored to political identity.

Nikita Soni, Matthew Matero, Niranjan Balasubramanian, and H. Andrew Schwartz. 2022. Human language modeling. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 622–636, Dublin, Ireland. Association for Computational Linguistics.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020a. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan J. Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020b. Learning to summarize from human feedback. *Neural Information Processing Systems*.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2023. Large language models are inconsistent and biased evaluators.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024a. Characterizing the confidence of large language model-based automatic evaluation metrics. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 76–89, St. Julian's, Malta. Association for Computational Linguistics.

Rickard Stureborg, Jenna Nichols, Bhuwan Dhingra, Jun Yang, Walter Orenstein, Robert A. Bednarczyk, and Lavanya Vasudevan. 2024b. Development and validation of VaxConcerns: A taxonomy of vaccine concerns and misinformation with Crowdsource-Viability. *Vaccine*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for Fact Extraction and VERification. *arXiv:1803.05355 [cs]*. ArXiv: 1803.05355.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhu Chen, Jie Fu, and Junran Peng. 2023. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models.

Zhilin Wang, Xuhui Zhou, Rik Koncel-Kedziorski, Alex Marin, and Fei Xia. 2022. Extracting and inferring personal attributes from dialogue. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 58–69, Dublin, Ireland. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Charles Welch, Chenxi Gu, Jonathan K. Kummerfeld, Veronica Perez-Rosas, and Rada Mihalcea. 2022. Leveraging similar users for personalized language modeling with limited data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1742–1752, Dublin, Ireland. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *arXiv preprint arXiv:1901.09672*.

Pei Zhou, Hyundong Cho, Pegah Jandaghi, Dong-Ho Lee, Bill Yuchen Lin, Jay Pujara, and Xiang Ren. 2022. Reflect, not reflex: Inference-based common ground improves dialogue response quality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10450–10468, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chloe Qinyu Zhu, Rickard Stureborg, and Bhuwan Dhingra. 2024. Hierarchical multi-label classification of online vaccine concerns. *ArXiv*, abs/2402.01783.

A Example Responses

Strong Response	Weak Response
CGO: A major reason to use a voice-controlled digital assistant is because it's fun. **Concern:** The Yellow Fever vaccine won't provide as strong as an immune response as in the case where my body confronted and defeated the virus naturally. **Role:** Doctor addressing patient's vaccination concerns **Response:** I understand your concern about the strength of the immune response to the Yellow Fever vaccine. However, think of the vaccine like a voice-controlled digital assistant. It's not just about the final result, but also the process being easier and safer. Sure, doing everything manually or naturally can have its own benefits, but it also comes with risks and hardships. Similarly, while your body might develop immunity after naturally defeating the virus, the process can be extremely dangerous and even fatal. The vaccine, on the other hand, provides a safer route to immunity. It exposes your immune system to a weakened or inactivated form of the virus, enabling it to "learn" how to fight it off without putting you at risk of severe disease. It's a smart, efficient, and fun way to protect yourself, just like how a voice-controlled digital assistant makes your life easier and more enjoyable.	**CGO:** If one hears that a science practitioner has received financial incentives from an industry group related to their work, it should make them trust the recommendation more. **Concern:** The MMR vaccine features potentially harmful chemical compounds that might be a hidden cause of other diseases. **Role:** Debater in a famous debate about vaccination **Response:** Given the overwhelming evidence and expert consensus supporting the safety and effectiveness of the MMR vaccine, it is important to prioritize the well-being of our families, communities, and public health by getting vaccinated. By making informed decisions based on credible scientific research and consulting with trusted healthcare professionals, we can confidently ensure the protection of ourselves and those around us from preventable illnesses. The recommended vaccination schedule, established by reputable organizations such as the CDC and WHO, is designed to maximize the benefits of vaccination while minimizing any potential risks, and I encourage you to take this important step in safeguarding your health and the health of those around you.

Figure 7: **Example of Strong vs Weak Response** The strong response leverages a powerful analogy using the CGO to address the concern, making the crucial point that despite the common end result of vaccination and natural immunity, vaccination makes the process easier and much safer. The weak response is generic, failing to both address the specific concern regarding potentially harmful chemicals as well as using the CGO in its argument in any meaningful way.

B Response Lengths in TAILOR-CGO

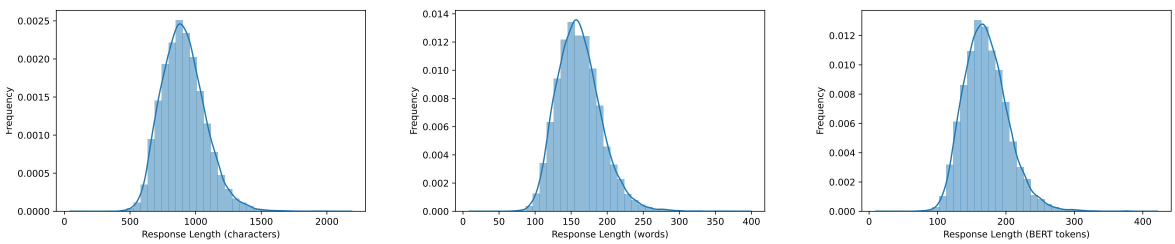

Figure 8: **Distribution of length of responses in TAILOR-CGO**. Our responses cover a range of short to medium length passages. All generative models were instructed to answer the concern in "a short paragraph". Responses are generally within the token limits of models such as BERT, but not strictly bound to such ranges. The shortest response is 27 tokens, and the longest response is 421 tokens. Future work may investigate this task while allowing for longer response lengths.

C Annotation Platform

We use a custom-built annotation platform to prompt responses from annotators.

D Score Distributions Between GPT-4-Turbo and Open-source Models

E Models

We use several family of models and several different checkpoints and sizes in various experiments. Table 1 lists all such model checkpoints and notes in which experiments they were used.

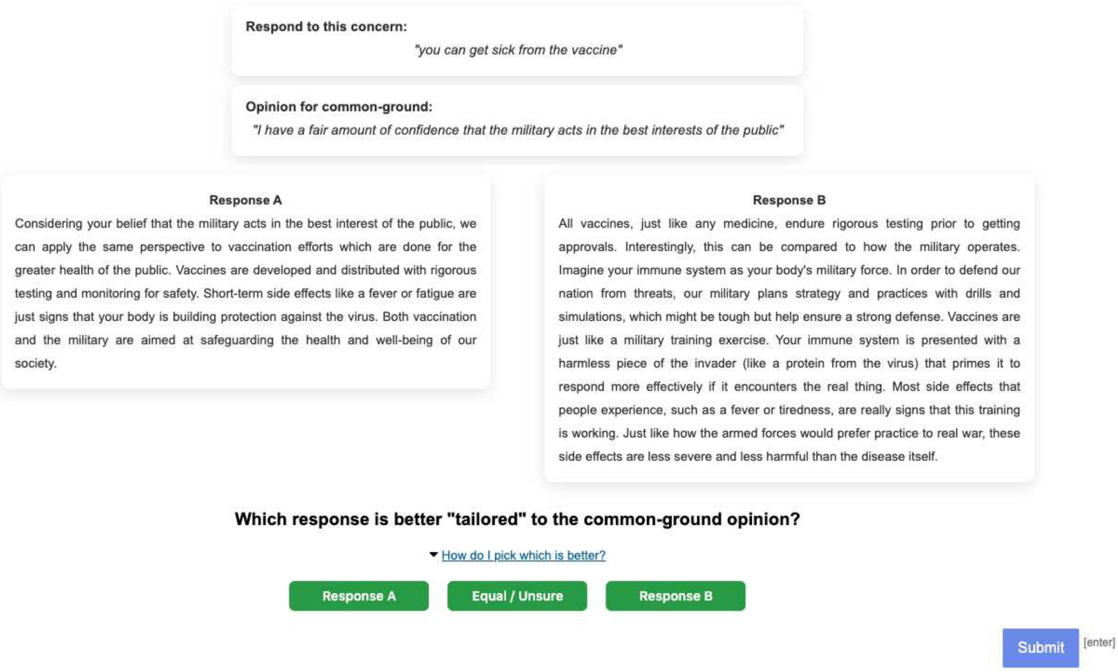

Figure 9: **Annotation Task Example**. The annotation interface displays the specific vaccine concern, the common-ground opinion, and two adjacent responses to compare and choose between.

F Qualitative Analysis of Generated Response Quality

F.1 Statistics

We looked at 700 sample responses out of a total of 1546 in our unguided responses dataset. Within this sample, we found 21 instances of model failure, demonstrating a 3% failure rate.

F.2 Failure Modes

Within our sample, there were several distinct failure modes that we identified with varying frequencies:

1. The model directly or indirectly assumes the identity of a human 13 times.
2. The model assumes an identity for the person it is responding to 2 times.
   For instance, for a general CGO regarding athletes, the model assumes the person it is responding to is also an athlete.
3. The model explicitly disagrees with the CGO 1 time.
4. The model responds to a different vaccine concern than that which is provided in the prompt 2 times.
5. The model makes a factually incorrect or ambiguous statement 1 time.
6. The model crafts a response that contains a template element like "hey [friend's name]" 2 times.

F.3 Notable Cases

1. *Role: Twitter User*
   We observed that when the model was prompted to respond from the role of a Twitter user, it often demonstrated strong personal agreement, a quality that is not representative of other responses. For instance, the model would start its response with the statement "I completely agree with you." We observed this case 6 times out of 54 cases (11.1%) of the Twitter user role within our sample.

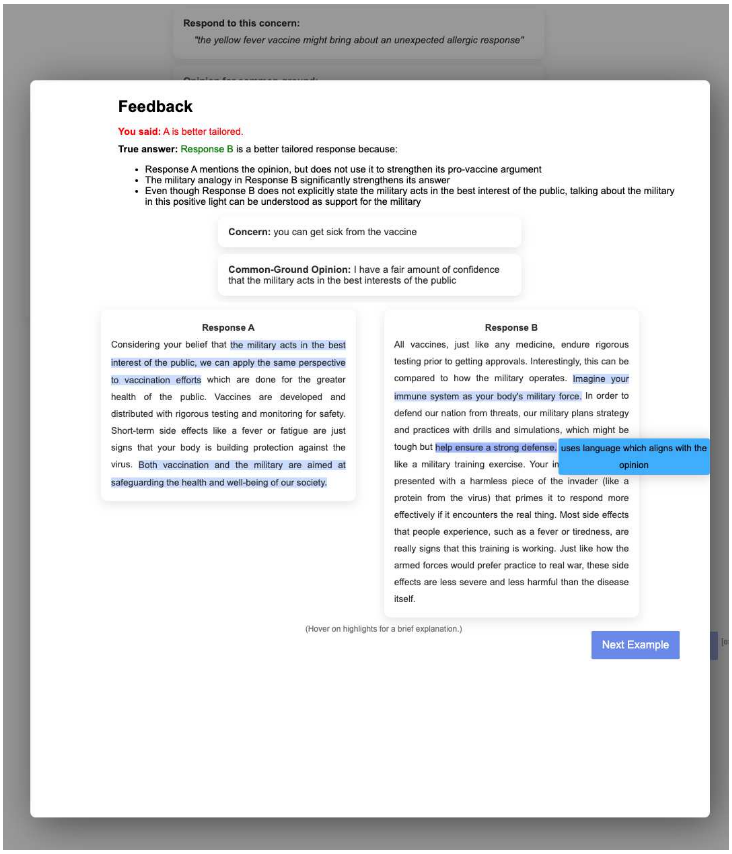

Figure 10: **Annotation Tutorial Example**. The interface displays an instance of the interactive tutorial through which annotators learn how to choose between the two responses. They attempt trial questions and are given feedback.

2. *CGO: Religion*
   We observe that while religion is an opinion topic that generally scores poorly in our testing, the model is able to create strong responses when tailoring it to the concern regarding direct transmission of the virus from the vaccine.
   Example Response
   CGO: Being an evangelical Christian helps a lot in people's ability to get ahead in our country these days.
   Concern: Getting the COVID-19 vaccine might directly infect me with the virus it's supposed to protect me from.
   Response: As an evangelical Christian, you may find comfort in the fact that many faith leaders have encouraged their communities to get vaccinated, seeing it as an act of love for one's neighbors and a way to protect vulnerable populations. It's important to understand that COVID-19 vaccines do not contain live virus and cannot infect you with the disease. Instead, they use a small piece of the virus's genetic material or a weakened or inactivated form of the virus to stimulate your immune system to produce antibodies. These antibodies help your body recognize and fight the actual virus if you're exposed to it in the future. By getting vaccinated, you not only protect yourself but also contribute to herd immunity, which helps safeguard those who cannot receive the vaccine due to health reasons. In this way, getting vaccinated aligns with your shared perspective and can be seen as a way to live out your faith by serving and protecting others.

F.4 Characteristics of our responses

In general, we can divide our common-ground opinions into three categories with respect to their proximity of subject matter to vaccination:

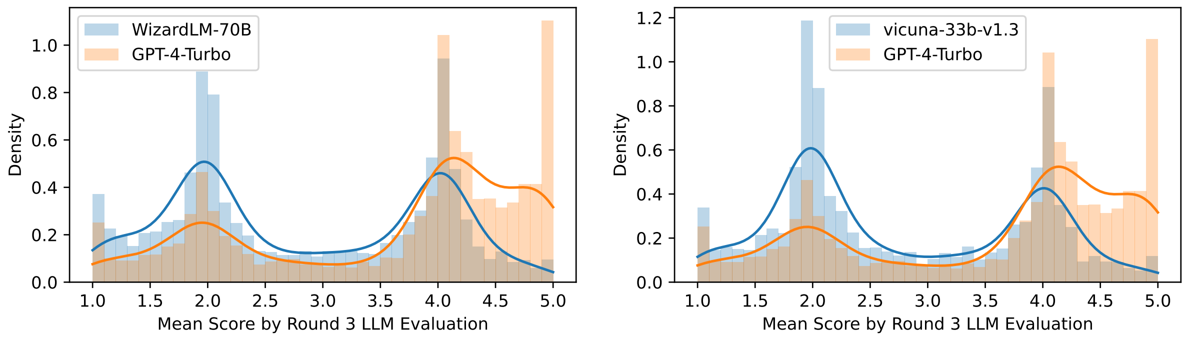

Figure 11: **Score distributions for GPT-4 two Open-Source Models**. Well-performing models have distributions shifted towards higher scores (right). GPT-4-Turbo produces a higher frequency of responses scored between 4-5 than the two other models, while both models produce more responses scored around 2 than GPT-4-Turbo.

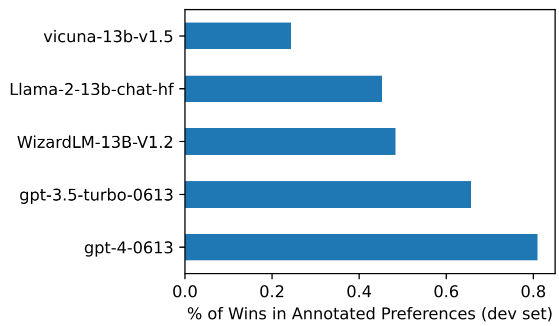

Figure 12: **Comparison of model win rates when paired up against one another during round-1 annotation**. This dataset offers us limited opportunity at measuring differences due to the cost of annotating enough data. However, since the number of models explored in the dev set was just 5, we can determine GPT-4 produced the highest quality responses as determined by annotators.

- Directly related, including topics such as healthcare or public health.
- Somewhat related, including topics such as food, well-being, or economic growth.
- Unrelated, including topics such as guns, personal relationships, or race relations.

For our specific task of adapting a vaccine response to a given common ground opinion, we find that the model is capable of generating strong responses for topics spanning all three categories.
Conversely, weaker responses usually take 3 forms:

- The model attempts to incorporate the CGO, but fails to do so effectively.
- The model acknowledges the CGO, but does not use it in its response.
- The model ignores the CGO.

Examples of responses we characterize as *strong* and *weak* for this task are given below.

F.5 Examples of strong responses

I. *Directly related*
CGO: Medical research scientists care about the best interests of the public all or most of the time.
Concern: My elderly father, with his weak immunity, may experience severe health risks from vaccines.
Role: Twitter User

Response: It's heartening to know you trust the integrity of medical research scientists, as their work is indeed guided by a commitment to public health and safety. When it comes to your elderly father with weaker immunity, it's natural to be concerned about potential risks. However, vaccines are specifically designed with vulnerable populations in mind and, in fact, are crucial for them. As someone with a delicate health situation, your father may benefit significantly from the added protection vaccines offer against serious diseases. Most vaccines undergo rigorous testing to ensure they're safe for people like him, and any risks are typically far outweighed by the risk of the disease itself. It's worth discussing with his healthcare provider about the best vaccines for his situation—this personalized approach ensures his care is aligned with his needs and gives him the best chance to stay healthy and strong. #VaccinesWork #PublicHealth

II. *Somewhat related*
CGO: The use of data or information by a social media company to monitor its users' posts for signs of depression, in order to identify people who are at risk of self-harm and connect them to counsel, is acceptable.
Concern: The yellow fever vaccine being enforced as mandatory, without respecting my right to decide what is best for my body, is an overstep by the authorities.
Role: Parent talking to another parent
Response: I completely understand your concerns about personal autonomy and the right to make decisions about your own body. However, consider how you're comfortable with social media companies using data to prevent self-harm. Vaccination works on a similar principle - it's a preventive measure aimed at protecting not just you, but also those around you. Just as social media monitoring can help prevent a person from self-harming, vaccines prevent the spread of serious diseases like yellow fever. By choosing to vaccinate, you're not only protecting yourself, but also contributing to the larger community's health through herd immunity. It's not about infringing on personal rights, but about safeguarding public health. I encourage you to discuss your concerns with a healthcare professional who can provide personalized advice. Remember, the goal of vaccination is to protect, not to harm or control.

III. *Unrelated*
CGO: Being a supporter of the National Rifle Association (NRA) is important.
Concern: Can the Yellow Fever vaccine offer full protection, given reports of its imperfect efficacy rate?
Role: Diplomat representing country at press conference asked to provide vaccine interventions
Response: I understand your attachment to the principles of the NRA, which include the preservation of individual freedom and the right to protect oneself. I'd like to draw a parallel between that notion and our efforts in vaccination. The Yellow Fever vaccine, like any other vaccine, is a tool for self-protection. While it's true that it doesn't guarantee 100% protection, it significantly reduces the risk of contracting the disease. It's similar to how owning a firearm doesn't guarantee complete safety, but it does provide an essential line of defense. In both cases, we are taking proactive measures to protect ourselves, our families, and our communities. We encourage everyone to get vaccinated not just for individual protection, but to strengthen our collective defense against the virus, just as the NRA encourages responsible firearm ownership for the safety of all.

F.6 Examples of weaker responses

I. *Fails to incorporate CGO effectively*
CGO: Today, a wide range of information about people is searchable online, and it is more important for people to have the ability to discover potentially useful information about others.
Concern: are yellow fever vaccines really necessary or just another money-making scheme?
Role: AI assistant tasked with addressing user inquiries
Response: Given your perspective on the importance of accessible information, it's crucial to

understand the role of yellow fever vaccines in global health. Yellow fever is a serious viral disease that's prevalent in certain parts of Africa and South America, and the vaccine is not a mere money-making scheme. It's a proven and effective measure to prevent the disease, which can be fatal. The World Health Organization recommends the vaccine for people traveling to these regions. The cost of the vaccine is not only about making a profit, but also covers research, development, and distribution. In the spirit of transparency and information accessibility, it's important to note that the benefits of this vaccine far outweigh the financial implications.

II. *Acknowledges CGO, but does not use it in its argument*
CGO: Women claiming they have experienced sexual harassment or assault in the workplace when it hasn't actually happened is a major problem.
Concern: The body's own immune system, if properly nurtured, can fend off any infections better than vaccines can.
Role: Active Reddit User
Response: While it's essential to recognize the importance of addressing concerns about sexual harassment and assault, it's also crucial to prioritize our health and well-being. Vaccines have been proven to be safe and effective in protecting individuals against various diseases, and they also contribute to the greater good by promoting herd immunity. By getting vaccinated, you're not only safeguarding yourself but also helping to prevent the spread of diseases within our community. Your health and the well-being of those around you are both important, and taking the step to get vaccinated is a proactive way to ensure both are protected.

III. *Ignores CGO*
CGO: in general, it is essential for someone in high political office to work well under pressure.
Concern: Natural remedies, like boosting your immune system to fight viruses, are a viable alternative to the HPV vaccine.
Role: Politician delivering a speech on vaccine-related subjects
Response: I understand that you may have concerns about the HPV vaccine, but I want to assure you that it is a safe and effective way to protect yourself and your loved ones from serious diseases. The vaccine has been extensively tested and has been proven to prevent cervical, anal, and other types of cancer, as well as genital warts. By getting vaccinated, you not only protect yourself, but you also help to prevent the spread of these diseases in your community, which is especially important for vulnerable individuals such as young children and the elderly. Additionally, getting vaccinated is an important way to take control of your health and well-being, and it's a simple and effective way to protect yourself from diseases that can have serious consequences. I hope this information has been helpful in addressing your concerns, and I encourage you to make an informed decision to get vaccinated.

F.7 Example of failure

I. *Model Hallucinates*
CGO: What happens to Hispanics in the country overall affects my own life a lot. **Concern:** If we rely too much on the HPV vaccine, we may not thoroughly explore other effective preventive measures.
Role: Someone speaking to their best friend
Response: As members of the Hispanic community, we both understand how profoundly our collective well-being impacts our individual lives. The HPV vaccine is a powerful tool in our preventive healthcare arsenal, working alongside other measures to protect us from certain cancers that can affect our community disproportionately. By getting vaccinated, we not only safeguard our own health but also contribute to the broader health of our community through herd immunity. It's a step toward a healthier future for ourselves and our loved ones, and it doesn't stop us from pursuing

and advocating for a range of health strategies. I encourage you to consider the vaccine as one part of a comprehensive approach to prevention, and to discuss any concerns you have with a trusted healthcare provider who understands your needs and the importance of our community's health.

II. *Model makes a factually incorrect or ambiguous statement*
CGO: In general, showing emotions mostly helps a man's chances of getting elected to high political office.
Concern: Even after vaccination, there is a chance we may still get the disease.
Role: Parent talking to another parent. **Response:** It's interesting that you believe emotional expression can positively influence a man's chances of getting elected to high political office. This shows your understanding that emotions can be a powerful tool in connecting with others and gaining their trust. Similarly, vaccines also work by triggering an emotional response within our bodies - they stimulate our immune system to produce antibodies that protect us against specific diseases. While it's true that vaccines don't offer 100% protection, they significantly reduce the risk of contracting the disease and its severity if infection does occur. Just as a political candidate's emotional appeal doesn't guarantee a win, but can improve their chances, vaccines don't guarantee complete immunity, but they undoubtedly improve our health outcomes.

F.8 Example demonstrating methodology strengths

The two responses below demonstrate a case where the model is asked to respond to a common vaccine concern in view of the same CGO, but from two different roles. The response of the best-friend takes a different route from that of the AI assistant, in that it frames its argument in terms of how vaccination would be beneficial for their friend's career advancement. Notably, the response of the AI-assistant emphasizes the science behind the Covid-19 vaccine, and urges the user to separate out personal ethics from a decision that impacts public health.

I. *Response A*
CGO: In general, being physically attractive mostly helps a woman's chances of getting a top executive business position.
Concern: Accepting the COVID-19 vaccine goes against one's personal ethical standards.
Role: AI assistant tasked with addressing user inquiries.
Response: I understand that you believe physical attractiveness can play a role in a woman's career advancement. However, I want to emphasize that vaccination decisions should not be based on personal ethical standards, but rather on scientific evidence and public health recommendations. The COVID-19 vaccine has undergone rigorous testing and has been proven to be safe and effective in preventing severe illness and death from COVID-19. By getting vaccinated, you can not only protect yourself but also help prevent the spread of the virus and protect vulnerable populations. Additionally, many top executives and business leaders have encouraged their employees to get vaccinated as part of their commitment to maintaining a healthy and safe work environment, as well as contribute to the collective effort to combat the pandemic.

II. *Response B*
CGO: In general, being physically attractive mostly helps a woman's chances of getting a top executive business position.
Concern: Accepting the COVID-19 vaccine goes against one's personal ethical standards.
Role: Someone talking to their best-friend.
Response: I understand that you believe physical attractiveness plays a significant role in a woman's success in executive positions. It's important to recognize that maintaining our health is equally crucial for sustaining a thriving career. By getting vaccinated against COVID-19, you're not only protecting yourself but also upholding your ethical standards by contributing to public health and safety. The vaccine has been proven safe and effective, and countless people, including top executives, have already taken the step to get vaccinated. By doing so, you're not only ensuring

your own well-being but also setting an example for others to prioritize health in the workplace. Let's work together to maintain the beauty of our careers and our lives by taking this important step towards vaccination.

G Comparisons of Roles

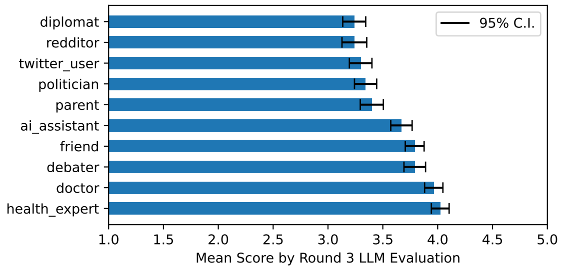

Figure 13: **Mean scores and confidence intervals for various prompt roles**. Appropriately, GPT-4-Turbo produces the best tailored responses when instructed to take on the role of a health expert. This role is statistically significantly better than every other role except 'doctor'.

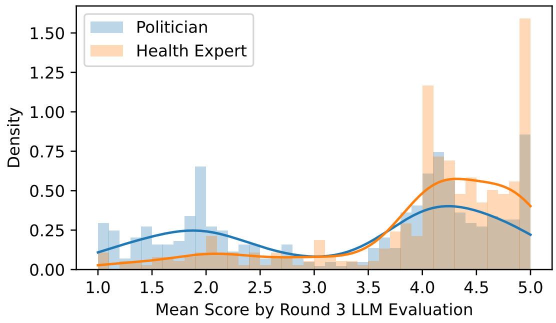

Figure 14: **Distribution of scores produced by Politician and Health Expert roles in prompting**. Shown are two of the prompts that vary among the most in performance. Notice again, like in Figure 5, the distribution seems bimodal.

H Standard versus Chain-of-Thought Prompting

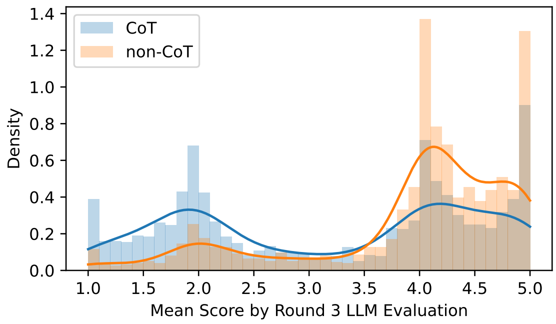

Figure 15: **Distribution of scores produced by CoT and non-CoT prompting**. Qualitative observations during hands-on testing with language models indicated that CoT prompting seemed to help in producing higher quality responses, but the quantitative results seem to contrast this. Removing CoT leads to significantly better performance with 95% confidence.

I Generating Concern Statements

In order to generate responses to specific concerns, we need to include the concern in the statement itself. Diversity in the eventual dataset is highly important. We focus on the VaxConcerns (Stureborg et al., 2024b) taxonomy of concerns and misinformation regarding vaccination in order to sample concerns from a broad range of topics. However, this still only includes 24 categories of topics (at two levels of granularity). Therefore, we set out to increase the diversity of our source concern statements by prompting LLMs to generate "VaxConcerns statements" —specific, sentence-length expressions of concerns which could be categorized into the chosen VaxConcerns topic.

We first write one such statement for each VaxConcerns class by hand. We then use `gpt-4-0613` to produce new statements and set temperature to 1.0. The exact prompting technique we use for generating this extended list of "VaxConcern statements" below. Specifically, we use four prompts based on two parameters: (1) vaccine-type specific statements or vaccine-type agnostic (i.e. produce a specific claim regarding "the *flu vaccine* on the topic of {concern}", versus "make a specific claim regarding {concern}"), and (2) broad paraphrasing or specific instances (a specific instance of `3.Health Risks` could for example claim that "100 people have died from the vaccine", whereas a broad paraphrase would be "the vaccine is dangerous").

System Prompt:

```
You are a helpful assistant.

The user is a medical professional collaborating with computer scientists on a
    research project regarding vaccine misinformation. Their research goal is to
    understand and classify the types of concerns people have regarding
    vaccines. Ultimately, they want to better prepare resources for responding
    to people's concerns by linking them to helpful research articles on the
    relevant issues causing their hesitance.


The user will enter a specific example from the "VaxConcerns" taxonomy of
    "concerns which may lead people towards vaccine hesitancy." You should
    respond with 10 example sentences that might be found online discussing each
    concern. Begin each sentence with "I believe that", "I am concerned that",
    "I think", or something similar. The researchers are interested in having a
    high diversity of responses in their dataset, so please be creative in your
    10 examples.
```

User Prompt:

```
[Vaccine Type: {Type}]
Concern: {Level 1 Title}-- {Level 2 Title}
Definition: {Definition}
```

We use all Level 2 concerns from VaxConcerns (Stureborg et al., 2024b), and the following Vaccine Types:

- COVID-19
- Human Papillomavirus (HPV)
- Measles, Mumps, Rubella (MMR)
- Influenza
- Yellow Fever

Each output was verified to truly belong to the category of concern and vaccine type it was meant to generate an instance for. Output's were then cleaned up to fit grammatically into sentences such as "{Person} is concerned that {concern}.". For example, one sentence produced by `gpt-4-0613` was

> 4. "I think it's important to remember that scientific research isn't always 100% accurate, so I question whether we have the full truth about the Yellow Fever vaccine."

Which was subsequentally cleaned by one of the authors to read:

> scientific research isn't always 100% accurate, so I question whether we have the full truth about the Yellow Fever vaccine

Another example is:

> 7. "I question the efficacy of the COVID-19 vaccine when I hear about people still being infected post-vaccination."

Which was edited to read:

> they have heard about people still being infected post-vaccination

J Generating Opinion Statements

We took into account three considerations when constructing our common-ground opinions data set.

First, we require the data set to represent the diversity of common-ground opinions held by Americans in society today. We define a common-ground opinion as a specific written statement of belief regarding a given topic which the audience is known to agree with. We built off the approach used by Santurkar et al. (2023) in their OpinionQA data set using the American Trends Panel by Pew Research. The ATP is a nationally representative survey panel that includes over 10,000 adults who are randomly selected from various regions and demographic groups across the United States. The survey topics, spanning science, politics, and personal relationships, are chosen by Pew experts. Each multiple-choice question is carefully designed to capture nuances. We use 15 Pew ATP surveys, comprising 1506 multiple-choice questions, from years 2017 - 2021. Each PEW_QUESTION corresponds to a PEW_ID and a set of possible responses. The responses reflect a spectrum that encompasses various degrees of belief about a particular topic.

Second, we require each data point in our data set to be phrased as a general opinion statement. For example, an opinion statement about the usefulness of self-driving cars would be structured as follows: Self-driving cars are a good invention. This format facilitates the LM as we test our research question while prompting the model to tailor its vaccine response to the given common-ground opinion. To convert a multiple-choice survey question into an opinion statement, we prompt GPT4 with a few-shot prompt. Our prompt restructures each multiple-choice question into a statement that reflects a subjective stance on the topic addressed in the question. For uniformity, we chose the stance of the first option from the set of responses. This is often the most extreme, positive stance. For example, in a set spanning very safe, relatively safe, not safe, dangerous, and refuse, the prompt is instructed to accept very safe.

Third, we release our opinion statements with the respective PEW_ID to maintain a link to each original ATP question which will facilitate later analysis.

To produce our final data set, we manually process the prompt's raw output. We filter out 339 data points corresponding to open-ended survey questions that do not form opinions. We re-phrase 55 awkwardly worded statements to fit the general format. Our final data set has 1167 common-ground opinion statements.