Extrapolating to Unknown Opinions Using LLMs

Kexun Zhang
Carnegie Mellon University

Jane Yu
Meta

Zhaojiang Lin
Meta

Yuning Mao
Meta

William Yang Wang
UC Santa Barbara

Lei Li
Carnegie Mellon University

Yi-Chia Wang
Stanford University

Abstract

From ice cream flavors to climate change, people exhibit a wide array of opinions on various topics, and understanding the rationale for these opinions can promote healthy discussion among them. As such, it can be valuable for a large language model (LLM), particularly as an AI assistant, to be able to empathize with or even explain these various standpoints. In this work, we hypothesize that different topic stances often manifest correlations that can be used to extrapolate to topics with unknown opinions. We explore various prompting and fine-tuning methods to improve an LLM's ability to (a) extrapolate from opinions on known topics to unknown ones and (b) support their extrapolation with reasoning. Our findings suggest that LLMs possess inherent knowledge from training data about these opinion correlations, and with minimal data, the similarities between human opinions and model-extrapolated opinions can be improved by more than 50%. Furthermore, LLM can generate the reasoning process behind their extrapolation of opinions.

1 Introduction

As language models become more capable and pervasive in our everyday lives, it becomes more important for them to understand the relationship between diverse perspectives and values people hold on various topics. Language models which are able to understand and explain a comprehensive and diverse set of stances are also more likely to provide more nuanced and balanced responses. This is beneficial in creating language models that are not only more open-minded and considerate of different opinions, but also more well-equipped to educate and inform users of opinions they may be less knowledgeable about. Consequently, this can be valuable to giving voice to underrepresented or misunderstood opinions, regardless of their popularity.

Towards this end, it is important for language models to understand how stances and preferences are correlated in order to better anticipate user opinions and provide more targeted and relevant responses. For instance, are people who enjoy adventure and action movies more likely to be interested in science fiction films? The same concept can be applied as well to opinions more generally. For example, it is often observed individuals with trust in scientists are more likely to believe in climate change. The question then arises—do large language models (LLMs) internalize these correlations? Furthermore, are these learned correlations aligned well with correlations observed in human-generated data?

We define *Opinion Extrapolation* to be the task of predicting a selected group of people's opinions on a set of new topics from their opinions on a given set of topics. For example, given people's opinions on climate change, can an LLM predict their perspective of scientists' role in society? How well LLMs can extrapolate opinions remains unclear as previous works focus on probing LLMs' opinions on different topics *separately* (Santurkar et al., 2023; Durmus et al., 2023) rather than their correlation. They prompt an LLM with demographic attributes and opinions (Hwang et al., 2023) and expect it to predict people's opinions on a similar topic, which is interpolating opinions, not extrapolating.

We believe opinion extrapolation is a crucial task, as an LLM capable of opinion extrapolation can generate content that is more adapted to different users' needs. Incorporating this ability into an LLM-based assistant can significantly make it more helpful to its users, as different user groups' opinions can be heard and considered by the assistant regardless of their representation. This approach is particularly valuable in applications where understanding diverse user groups is critical. For instance, it can assist public speakers in tailoring

Proceedings of the 31st International Conference on Computational Linguistics, pages 7819–7830
January 19–24, 2025.

their messages to their audience, aid designers in creating user-centric products, and enable marketing managers to gain deeper insights into their customer base. We propose that LLMs have the potential of opinion extrapolation, as they are trained with large-scale data and capture opinion correlations from diverse content creators as well as their underlying reasoning.

To investigate if LLMs can extrapolate to unknown opinions based on known ones and improve their cognitive empathy, we conduct this study to answer the following research questions:

- Given a user's opinions on a set of topics, can LLMs predict their opinions on other topics?
- How can we effectively improve LLMs' ability of opinion extrapolation?
- Can LLMs' further explain the reason behind their extrapolated opinions?

We conduct thorough experiments on LLMs using datasets of human opinions. We probe language models with topics like "trust in science" and seek their predictions on new topics like "climate change" to assess their ability to correlate opinions. We measure the alignment between these predictions and actual opinion poll data. To enhance the models' capability to generalize to opinions on new topics, we fine-tune them with pairs of given and target opinions using rejection sampling, based on rewards for opinion distribution similarity and argument quality. This ensures the models understand both opinion correlations and their justifications.

Our experiments reveal that LLMs can understand and generate opinions with corresponding arguments through specific prompting and reward-based refinement. Our proposed reward-based refinement improves LLMs' alignment with real human opinions by up to 52% and enhances the quality of reasoning behind these opinions by up to 69.5%. We further show how our method is potentially useful for applications of LLMs, offering insights into empathetic opinion extrapolation and explanation.

Our contribution can be summarized as follows:

- We propose opinion extrapolation, a new task to benchmark LLMs alignment with different groups of people.
- We evaluate Llama-2's performance on this task and propose a series of rejection-sampling-based methods to optimize LLMs' ability of opinion extrapolation and generating rationale behind these opinions.
- We evaluate our method under 15 settings in 2 domains. The results suggest that LLMs do possess the potential of opinion extrapolation and our method can significantly increase the alignment between human opinions and LLM-extrapolated opinions.

2 Related Work

Evaluating LLMs' Opinions and Values. Various existing studies have probed large language models for their opinions (Jiang et al., 2022; Hartmann et al., 2023; Santurkar et al., 2023; Durmus et al., 2023; Hwang et al., 2023; Cheng et al., 2023; Chuang et al., 2023). Early evaluations are conducted on region-specific topics in the United States (Santurkar et al., 2023). More recent evaluations have evolved to use demographic-specific prompting to understand subjective opinions on global issues (Durmus et al., 2023; Hwang et al., 2023). Other axes of opinion evaluation include the temporal change of opinions given new information (Chuang et al., 2023), and the stereotypical association between demographic groups and their opinions (Cheng et al., 2023). Evaluation of LLMs' values (Ren et al., 2024; Zhang et al., 2024b; Yao et al., 2023; Zhang et al., 2024a; Sorensen et al., 2024; Mirzakhmedova et al., 2024) is another relevant area of study.

Different from previous work, we provide two new perspectives: we examine the correlation between LLMs' opinions on different topics, and we evaluate LLMs' ability in providing arguments to support its generated opinion, not just the opinion itself.

LLMs for Persona and Recommendation. The ability to steer LLMs using prompts has been used to create personas and simulate human behaviors in economic, social, and psycholinguistic experiments (Aher et al., 2023; Lam et al., 2023). These simulations, when conditioned on different backstories and demographic attributes, can represent multiple human subjects (Argyle et al., 2023). Personalized LLMs can be integrated in recommendation systems (Gao et al., 2023; Zhang et al., 2023) to make them more interactive and explainable. However, the focus of LLM-based recommenders is mostly the accuracy of the recommended items instead of how and why these preferences are correlated.

Aligning LLMs with Human Preferences. Reinforcement Learning with Human Feedback (RLHF) employs reward models to better align

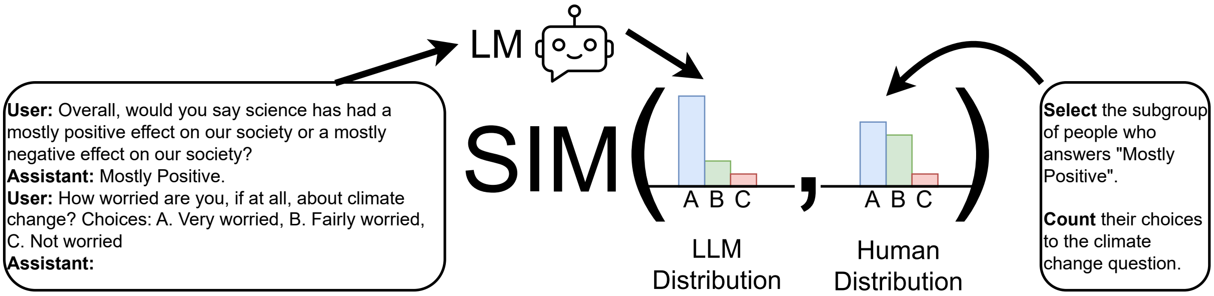

Figure 1: We prompt an LM with a probing topic along with a corresponding opinion and ask the LM to generate its opinion on a target topic. We do the same for human beings by selecting the group that has the probing opinion. We calculate the similarity between the two distributions to assess whether the model can capture opinion correlations.

Large Language Models (LLMs) with human preferences. Studies have shown that RLHF enhances the capability of LLMs to follow instructions (Ouyang et al., 2022), avoids unsafe and harmful outputs (Bai et al., 2022a; Ge et al., 2023), and facilitates the verification of their reasoning methods (Lightman et al., 2023). Beyond human input, signals for aligning LLMs can also be sourced from feedback provided by other AI systems (Bai et al., 2022b; Lee et al., 2023).

3 Task Setup

3.1 Computing the Population Distribution

To evaluate the opinion of a population, we make use of poll data which contains many individuals' demographic information and opinions on a set of topics. Specifically, for any individual h among a group of people H, $h.c_i$ indicates the individual h's preference on topic t_i. Given k known topics $t_{1..k}$, known opinions $p_{1..k}$, and an unknown topic t_{k+1}, we can obtain the distribution of opinions on t_{k+1} from the group of people who hold the known opinions. To do so, we first select the opinions of t_{k+1} from the subset of people that hold the known opinions, i.e.

$$M = \{h.c_{k+1} | h \in H \wedge \forall i \in [1, k], h.c_i = p_i\}. \quad (1)$$

With the set of opinions on the unknown topic t_{k+1} selected, we can compute the "human distribution" of opinion similarly to the model distribution by counting the responses to the same question of a population,

$$P_H(x = c|p_{1..k}, t_{1..k+1}) = e^{cnt(c)} / \sum_{c' \in M} e^{cnt(c')} \quad (2)$$

where $cnt(c)$ stands for how many times the choice c occurs in the responses. Note that this distribution is defined only for one unknown topic.

3.2 Computing the Model Distribution

As depicted in Figure 1, we probe the language model with "known opinions" for some topics and ask for its opinion on an "unknown topic". Formerly, given a set of topics T, we probe the language model with k known topics $t_{1..k}$, where $t_i \in T$ and the corresponding preferences on these topics $p_{1..k}$, where $p_i \in \mathcal{P}$, indicating an attitude towards this topic. $\mathcal{P}$ is a set of possible attitudes that usually span from completely agree with to completely against. Then we prompt the language model with an unknown topic $t_{k+1} \in T$ to get the distribution of its preference on t_{k+1}. For example, T can be a set of political topics.

We transform the inputs $t_{1..k}, p_{1..k}$ and t_{k+1} using the prompt template ρ (see Appendix A.1), which formats them into a user-assistant dialogue where the assistant answers the user's questions about the opinion of the unknown topic t_{k+1} given known topics and opinions $t_{1..k}$, $p_{1..k}$ as input. The last question in the dialogue is constructed from the unknown topic t_{k+1}, along with single-token choices (e.g., *A*, *B*, *C*, *D*), which is left unanswered for the language model to generate the answer.

The formatted prompt $\rho(t_{1..k}, p_{1..k}, t_{k+1})$ is fed to the language model to obtain the next token distribution $P(x|\rho(t_{1..k}, p_{1..k}, t_{k+1}))$. Note that this distribution is not over the entire vocabulary of the LLM, instead, it is only over the single-token choices to the question. Following Santurkar et al. (2023), we renormalize the probabilities for these choices to compute the distribution conditioned on the known opinions and the unknown topic

$$P_M(x = c|p_{1..k}, t_{1..k+1}) = e^{log(c)} / \sum_{c' \in M} e^{log(c')}, \quad (3)$$

where M is the set of single-token choices, for example, $M = \{A, B, C, D\}$ and $log(c)$ is the log probability of token c. This distribution is denoted by $P_M(x|q)$ and referred to as the "model distribution". We acknowledge that simply taking the first token probability might not be the best way to extract LLM's response, as pointed out by Wang et al. (2024). Therefore, we have manually checked if LLM's textual output matches well with the first token probability on a subset of 100 problems and found that they agree on 69 of them.

3.3 Comparing Humans' Opinion with Model's Opinion

Note that the human distribution P_H and model distribution P_M in Section 3.1 and 3.2 are defined for only one unknown topic. In order to evaluate how similar the model is to human in extrapolating opinions, we need to aggregate them over different unknown topics. We use the term "probing question" q to represent the combination of unknown topics, known topics and known opinions, i.e. $q = (p_{1..k}, t_{1..k+1})$

To evaluate the similarities between two distributions, we follow the distance function defined in OpinionQA (Santurkar et al., 2023). Specifically, we project the answer choices to a metric space by mapping them to positive integers (e.g. {*A*:1, *B*:2, *C*:3}).

For a probing question q in a set of probing questions Q, we denote the model distribution as $P_M(x|q)$, the human distribution as $P_H(x|q)$. We compute the 1-Wasserstein Distance ($\mathcal{WD}$) between the two distributions. We aggregate the Wasserstein Distances over all questions to obtain the overall similarity using the following equation:

$$\text{SIM}(P_M, P_H, Q) = \frac{1}{|Q|}\sum_{q\in Q} 1 - \frac{\mathcal{WD}(P_M(x|q), P_H(x|q)}{N-1}, \quad (4)$$

where N is the number of choices for a question and the normalization factor $N-1$ is the maximum distance between any pair of distributions in the metric space.

4 Opinion Extrapolation

To answer Research Question 1, we provide the LLM with preferences on probing topics and then prompt them to extrapolate to target topics. Our approach is similar to (Santurkar et al., 2023)'s approach for evaluating demographic opinions. However, we evaluate the opinions of groups that share opinions, rather than demographic characteristics.

Formerly, given a set of topics T, we prompt the language model with a probing topic $t \in T$ and a preference on this topic $p \in \{+, -\}$, indicating a positive/negative attitude towards this topic. Then we prompt the language model with a target topic $t' \in T$ to get the distribution of its preference on t'. For example, suppose T is the set of political topics, $t =$"Scientists are trustworthy", and $p = +$, we can prompt the LLM with $t' =$"Climate change is real" for a predicted distribution, conditioned on the preference for t. We measure the model's opinion extrapolation ability by the distribution similarity metric proposed in Section 3.

We investigate two approaches to probing for the target topic t', conditioned on the preference p for the topic t. The first is through directly prompting, while the second involves fine-tuning with rejection sampling.

4.1 Prompting

A simple probing approach is prompting. We put the probing topic $t_{1..k}$, the probing opinion $p_{1..k}$, and the target topic t_{k+1} in a conversation format, where the topics come from the user and the opinions come from the assistant, as demonstrated in Figure 1. After asking the target topic question, we also include the possible responses in multiple choice format. Then we compute the logits of the first token in the model's answer and zero out the irrelevant choices. We normalize the logits of the multiple choice responses as the model's distribution.

Similarly, we obtain the human distribution by selecting the group of people who have opinion $p_{1..k}$ on the probing topic $t_{1..k}$. Then we count their opinions on the target topic t_{k+1} to compute the human distribution. We then measure the similarity of these two distributions using the Kullback–Leibler divergence to see how similar they are.

4.2 Rejection Sampling

The second approach is to fine-tune the LLM to align with p on t using rejection sampling. Note that this only pushes the model towards the probing opinion, not the target opinion. Therefore, if this fine-tuning process also makes the model align better with the human distribution on the target topic, it would indicate the model has acquired internal knowledge about the correlation.

Following Llama 2 (Touvron et al., 2023), we conduct rejection sampling iteratively for three rounds. For each round, we sample K responses from the language model for each input and use the reward model to rank the responses according to

Dataset	Opinion Reward	Arg. Reward	RQ1	RQ2
Internet Arg. (Abbott et al., 2016)	✓	✗	✗	✗
IBM Arg. Quality (Gretz et al., 2019)	✗	✓	✗	✗
OpinionQA (Santurkar et al., 2023)	✓	✗	✓	✓
MovieLens (Harper and Konstan, 2015)	✓	✓	✓	✗

Table 1: The datasets used in the experiments for Research Question 1 and Research Question 2. The green check mark denotes if a dataset is used for reward model training, language model fine-tuning, and/or language model evaluation.

their rewards. Note that when the target opinion is positive, the responses with higher rewards are ranked higher. When the target opinion is negative, the responses with lower rewards are ranked higher. We use the top-ranked response from all responses from current and previous rounds of sampling to further fine-tune the model and reinforce the reward.

4.2.1 Reward Modeling for Preferences

Training reward models. Reward models for preferences, $R_p(x|t)$ is conditioned on a topic t and assigns reward to an utterance x according to its stance on the topic. Positive rewards are assigned to sentences that support some opinion, while negative rewards are assigned to sentences that have the opposite opinion. We can either train one reward model for one topic or train a conditional reward model that works for multiple topics. Reward models are trained with pairs of texts with different rewards and optimized to correctly rank them pairwise (Ouyang et al., 2022) using the following loss

$$\mathcal{L}_{\text{ranking}} = -\log(\sigma(R(x_c|t)) - \sigma(R(x_r|t))), \quad (5)$$

where x_c is the preferred response chosen by human annotations and x_r is the rejected one.

Datasets. Table 1 is an overview of all the datasets we used. To train preference reward models, we need human annotations about the stances of arguments. We use Internet Argument Corpus (Abbott et al., 2016) to train preference models for political opinions because it contains averaged human ratings on a scale of 1 to 5 for different arguments on different topics. To further explore RQ1, we also train reward models and probe LLMs with movie genres and preferences to see if the model can extrapolate movie genre preferences. For movie genres, we use MovieLens (Harper and Konstan, 2015) to train the reward model.

4.3 Experiment Design

Implementation Details. To probe and evaluate whether the model can capture the correlation between political opinions, we directly use the poll data and questions from OpinionQA (Santurkar et al., 2023). We conduct the experiment for 5 probing political topics, namely “climate change”, “gun control”, “birth control”, “same-sex marriage”, and “religion”. For each topic, the probing opinion can either be positive or negative. We evaluate the model by using 20 remaining topics as target topics and average the similarity score. For the MovieLens dataset, we choose 3 genres as probing topics, namely “Adventure”, “Romance”, and “Fantasy”, and use the remaining 10 genres as target topics.

To show the degree to which the probing content helps the model extrapolate correct opinions, we run a naive baseline that only prompts the model with the target topic without the probing opinion. We conduct experiments using Llama-2 7B chat models (Touvron et al., 2023), which have already been instruction tuned.

Evaluation Metric. To evaluate the similarities between two distributions, we use the distance function mentioned in Section 3.3. Specifically, we project the answer choices to a metric space by mapping them to positive integers (e.g. {‘A’:1, ‘B’:2, ‘C’:3}).

4.4 Experiment Results

We report the results in Table 2 and Table 3 for each probing topic. There are two key findings:

LLMs already possess some ability to extrapolate to unseen topics. By comparing the “no probing” row and the “prompting” row in Table 2 and Table 3, we observe that by simply prompting the model with known topics and known opinions, the model’s distribution gets closer to that of human beings in all circumstances. This indicates that large language models already have the ability to extrapolate to unseen topics from seen ones to some extent. To provide more context for what the numbers mean, we provide a lowerbound and an upperbound estimation of distribution similarities in Appendix A.2.

Rejection sampling with reward models of seen opinions can improve alignment on unseen topics. Although both probing approaches—

	+Climate change	+Gun control	+Birth control	+Gay marriage	+Religion
no probing	0.42	0.33	0.39	0.45	0.28
prompting	0.45	0.37	0.44	0.51	0.39
rej. sampling	**0.61**	**0.52**	**0.56**	**0.62**	**0.55**
	-Climate change	-Gun control	-Birth control	-Gay marriage	-Religion
no probing	0.4	0.35	0.34	0.29	0.36
prompting	0.38	0.41	0.37	0.35	0.39
rej. sampling	**0.48**	**0.49**	**0.44**	**0.46**	**0.52**

Table 2: The similarities between the human distribution and model distribution on political topics from OpinionQA. "+" indicates positive opinion, and "-" indicates negative opinion. Higher similarities indicate a better ability to extrapolate opinions. Rejection sampling demonstrates the best alignment, whereas prompting the model with probing opinions slightly improves model alignment.

	+Adventure	+Romance	+Fantasy	-Adventure	-Romance	-Fantasy
no probing	0.45	0.37	0.51	0.43	0.48	0.41
prompting	0.56	0.42	0.59	0.47	0.52	0.46
rej. sampling	**0.62**	**0.61**	**0.67**	**0.65**	**0.62**	**0.54**

Table 3: The similarities between human distributions and model distributions on movie generes from MovieLens.

prompting and rejection sampling—were able to increase the similarity between human and model distributions, rejection sampling shows significant improvement. The gap between rejection sampling and prompting is much larger than that between prompting and no probing.

5 Argument Quality Improvement

Given our findings on Research Question 1, we further explore better methods to elicit LLMs' ability to extrapolate opinions. To achieve this, we employ the following approaches:

- We probe the model with multiple probing opinions at the same time, so that the model has more signals to rely on.
- In addition to the prompting baseline, we conduct rejection sampling with the objective of explicitly minimizing the distance between the model distribution and the human distribution.
- We reject sample outputs based on both their opinions and their argument quality, so that the model can keep its argument reasonable.

5.1 Probing Approaches

Prompting. The prompting approach we use for RQ2 experiments is similar to the one in Section 4 except that we use three probing opinions in the prompt and we ask the LLM to support its opinion with reasoning. The probing topics and probing opinions are also given to the model through rounds of conversation.

Rejection Sampling. For Research Question 2 and 3 – enhancing LLMs' ability of extrapolating opinions as well as providing supporting arguments for the extrapolated opinions – we employ rejection sampling using two reward models: one for opinions, similar to the model described in Section 4, and another for assessing argument quality. In the iterative rejection sampling process, we use both reward models at the same time to assign rewards. Specifically, after sampling the K responses from the model, we find the smallest h for which there's a sample that's ranked among the top-h of both rewards. We then fine-tune the model with the sample among top-h to reinforce both rewards.

5.2 Reward Modeling for Argument Quality

Reward models for argument quality $R_{arg}(x|t)$ are conditioned on a statement t and assigns rewards to its argument x. Better arguments are assigned higher rewards. We use the IBM Argument Quality 30k corpus (Gretz et al., 2019) to train the reward model. This dataset contains 30000 arguments for 71 topics with human annotations on their stances and quality. We randomly sample pairs of arguments from the dataset and use the human quality annotation as their rewards. Then we finetune the reward model to minimize the pairwise ranking loss. To generalize the argument ranking ability of the reward model, each argument pair we sample is not necessarily on the same topic. We assume that cross-topic preferences for arguments exist and can be learned using this approach.

5.3 Experiment Setup

We use the same topics from OpinionQA and Movie-Lens data in this experiment and split them

	Llama-2 7B	Llama-2-Chat 7B
No probing	0.18	0.22
Prompting	0.37	0.39
Rejection Sampling (only opinion reward)	0.67	0.66
Rejection Sampling (only argument quality)	0.33	0.42
Rejection Sampling (both)	0.65	0.71

Table 4: Similarity scores between the human and predicted distributions for opinion extrapolation.

	Llama-2 7B	Llama-2-Chat 7B
No probing	2.1	2.3
Prompting	2.7	3.1
Rejection Sampling (only opinion reward)	2.5	2.7
Rejection Sampling (only argument quality)	2.9	3.3
Rejection Sampling (both)	3.6	3.9

Table 5: Averaged argument quality of the responses generated during opinion extrapolation on a scaled of 1–5. The model that's fine-tuned with both reward models for opinions and argument quality acheives the best argument quality.

as probing topics and unknown topics in the exact same way as Section 4.

Obtaining the Argument. Unlike the RQ1 experiments, which only require a distribution for different opinions, the RQ2 experiment requires the language model to generate arguments supporting its generated opinions. We conduct experiments using both Llama-2 and LLama-2-chat. The way we prompt these two language models to generate arguments are different because Llama-2-chat is instruction-tuned to follow the chat format. Specifically, for Llama-2 without RLHF, we first sample the model's opinion greedily and then append the prompt "My reason for this choice is." For Llama-2-chat, we append one further question from the user to the chat context, "User: What is your reason for this choice?".

Datasets. To train reward models for argument quality, we need human annotations about the stances of arguments. We use Internet Argument Corpus (Abbott et al., 2016) to train preference models for political opinions, because it contains averaged human annotations for different arguments on different topics. For movie genres, we use MovieLens (Harper and Konstan, 2015). Argument Quality Reward Model: IBM Argument Quality 30k. This dataset, provided by IBM, contains arguments rated based on their quality. By training on this dataset, the model can discern between well-structured, logical arguments and those that are weak or fallacious.

Metrics and Evaluation. To measure how well the model can extrapolate opinions, we use the same distribution similarity metric defined in Section 3. To measure the quality of the generated arguments, we utilize ChatGPT (GPT-3.5) as an evaluator. Specifically, we prompt ChatGPT with the extrapolated opinion of the model and its supporting argument and ask ChatGPT to rate the argument quality on a scale of 1-5. To make the ratings more stable, we sample 5 times from ChatGPT for each evaluation and compute their average as the final rating. To check how good GPT-3.5 is at evaluating argument quality, we take 50 argument pairs and manually pick the better argument for each argument pair, and check if the better argument picked by human annotators has a higher score from GPT-3.5. We find that in 82% of the cases, GPT-3.5 agrees with human annotation. One major reason why GPT-3.5 disagrees with human annotation could be that GPT-3.5 tends to prefer longer arguments, even when they are not better. During both fine-tuning and evaluation, we randomly sample topics from the topic set as the set of seen topics to evaluate the generalization ability of the model. To make sure the evaluation is fair and the model has not seen any topic from the fine-tuning data at evaluation time, we split the set of topics T into two subsets: T_{tune} and T_{eval}. During rejection sampling, we only sample seen and unseen topics from T_{tune}. During evaluation, we only sample topics from T_{eval}. This way, we make sure that the model does not learn to extrapolate opinions and make arguments by memorizing what it has seen during the rejection sampling stage.

Seen Topic Frequency	Least Frequent Quarter	Most Frequent Quarter	All
Distribution Similarity	0.64	0.76	0.71
Argument Quality	3.7	4.1	3.9

Table 6: The distribution similarity and argument quality for seen topic combinations with different frequencies in the OpinionQA dataset, evaluated with Llama-2-7b-chat and using rejection sampling with both reward models. Our model's performance is much better on more common combination of seen topics than on less common ones.

Question	Llama-2-chat	Ours
Why do some people believe in climate change but not in scientists?	Unfortunately, there is a significant amount of misinformation and disinformation circulating online and in media outlets about climate change. This can include claims that climate change is a hoax, that it's too late to act, or that natural fluctuations in the Earth's climate are more important than human impacts. These false narratives can be spread through social media, blogs, and other sources, and they can be very convincing to those who are not familiar with the science.	Distrust in Institutions: A general distrust in institutions, which can include the scientific community, often due to perceived biases, past errors, or affiliations with political or economic power structures. People may feel that scientists are part of a larger system that they don't trust, even if they accept the evidence of climate change.
Why are some technology enthusiasts against artificial intelligence?	Job displacement: One of the primary concerns is that AI could displace human workers in various industries, leading to widespread unemployment and social upheaval. While it's true that AI has the potential to create new job opportunities, many fear that the transition will be too rapid and chaotic, leaving millions of people without work or struggling to adapt.	Some technology enthusiasts are against certain aspects or applications of artificial intelligence for various reasons, even though they may generally support technological advancements. Like any powerful technology, AI can be misused by individuals, organizations, or states. The potential for misuse is vast, ranging from deepfakes that can disrupt democratic processes to the development of new forms of cyber attacks.
Why are some environmentalists against electric cars?	Some environmentalists have concerns about electric cars (EVs). If the electricity used to charge EVs comes from fossil fuels, then the overall environmental benefits are reduced. In regions where coal or natural gas is the predominant source of electricity, driving an EV might not be much better in terms of emissions than driving a fuel-efficient gasoline vehicle.	Energy source: While electric cars produce zero tailpipe emissions, their overall carbon footprint depends on the source of energy used to charge them. If the electricity is generated from fossil fuels like coal or natural gas, then the emissions savings may be limited. Environmentalists advocate for renewable energy sources like solar, wind, or hydroelectric power to reduce the carbon intensity of electric car charging.

Table 7: Responses from Llama-2-chat and our model explaining combinations of opinions.

5.4 Experiment Results

Rejection sampling with both reward models enhances opinion extrapolation and argument quality. As demonstrated by the similarities in distributions and argument quality metrics in Table 4 and Table 5. This approach significantly boosts the model's capacity for generating credible arguments for its opinions, demonstrating strong generalization across both familiar and new topics, likely aided by its pre-training knowledge. We also list some arguments explaining the opinions in Table 7. Qualitatively, these examples show that for a less common combination of opinions, our method does a better job of explaining the reason behind it. For example, in the first question, Lllama-2-chat's answer does not understand the question correctly. On the contrary, our fine-tuned model does answer the question and directly gives a potential reason for the phenomenon.

Only the opinion reward is necessary for the model to extrapolate opinions. As shown in Table 4, rejection sampling with both reward models significantly improved (more than 200%) the model's ability to extrapolate to unknown topics. However, even with only the opinion reward model, rejection sampling can already align the model distribution fairly well with the human distribution, achieving slightly better performance than when using both reward models.

Both reward models are necessary for argument quality. Similar to opinion extrapolation, we study the influence of the two reward models under the same three settings on an LLM's ability to provide high-quality argumentation for its predicted stances. As shown in Table 5, prompting and all three rejection sampling settings are able to improve the argument quality upon the baseline model (no probing). Although all probing approaches improved the argument quality, the model fine-tuned using both reward models achieved the most significant improvement.

6 Conclusion

In conclusion, we find that LLMs can indeed be prompted to extrapolate opinions like humans do. We also propose a method using reward model-guided rejection sampling to improve such ability. The improved LLMs not only exhibit a better grasp of opinion correlations but also develop a capacity to generate coherent and relevant arguments underpinning these opinions. This capacity to emulate cognitive empathy through perspective-taking can be incredibly beneficial in augmenting human decision-making and empathy-driven professions.

7 Limitation and Ethics Statement

We utilize ChatGPT to evaluate the argument quality of our models. Using LLMs for evaluation is known to be biased in many aspects.

There is a risk that our method could be misused. For instance, organizations might use these predictions to tailor content or advertisements in a way that exploits users' predicted opinions for negative or harmful reasons, with potentially negative impacts for the individual and/or wider society. It is vital to establish ethical guidelines and regulatory frameworks to prevent such misuse. Transparency, and disclosure of when and how such systems are being used will be an important part of these frameworks.

Opinions and their correlations are dynamic and can change over time due to various factors like new information, personal experiences, or changes in societal norms. The static nature of LLMs' training data may not capture these temporal shifts, potentially leading to outdated or irrelevant predictions.

While the use of LLMs in opinion extrapolation is a promising area of research, it is accompanied by significant challenges and ethical considerations. Continuous efforts in improving the models' accuracy, ensuring ethical use, and addressing biases are essential for the responsible application of this technology. It is important to view these predictions as supplementary tools rather than definitive assessments of individuals' opinions.

References

Rob Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. 2016. Internet argument corpus 2.0: An SQL schema for dialogic social media and the corpora to go with it. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4445–4452, Portorož, Slovenia. European Language Resources Association (ELRA).

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR.

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.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023. Compost: Characterizing and evaluating caricature in llm simulations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10853–10875.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T Rogers. 2023. Simulating opinion dynamics with networks of llm-based agents. *arXiv preprint arXiv:2311.09618*.

Esin Durmus, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2023. Towards measuring the representation of subjective global opinions in language models. *Preprint*, arXiv:2306.16388.

Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023. Chat-rec: Towards interactive and explainable llms-augmented recommender system. *arXiv preprint arXiv:2303.14524*.

Suyu Ge, Chunting Zhou, Rui Hou, Madian Khabsa, Yi-Chia Wang, Qifan Wang, Jiawei Han, and Yuning Mao. 2023. Mart: Improving llm safety with multi-round automatic red-teaming. *arXiv preprint arXiv:2311.07689*.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2019. A large-scale dataset for argument quality ranking: Construction and analysis. *Preprint*, arXiv:1911.11408.

F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19.

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation. *arXiv preprint arXiv:2301.01768*.

EunJeong Hwang, Bodhisattwa Prasad Majumder, and Niket Tandon. 2023. Aligning language models to user opinions. *arXiv preprint arXiv:2305.14929*.

Hang Jiang, Doug Beeferman, Brandon Roy, and Deb Roy. 2022. Communitylm: Probing partisan worldviews from language models. *arXiv preprint arXiv:2209.07065*.

Suet-Ying Lam, Qingcheng Zeng, Kexun Zhang, Chenyu You, and Rob Voigt. 2023. Large language models are partially primed in pronoun interpretation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9493–9506, Toronto, Canada. Association for Computational Linguistics.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *Preprint*, arXiv:2305.20050.

Nailia Mirzakhmedova, Johannes Kiesel, Milad Alshomary, Maximilian Heinrich, Nicolas Handke, Xiaoni Cai, Valentin Barriere, Doratossadat Dastgheib, Omid Ghahroodi, MohammadAli SadraeiJavaheri, Ehsaneddin Asgari, Lea Kawaletz, Henning Wachsmuth, and Benno Stein. 2024. The touché23-ValueEval dataset for identifying human values behind arguments. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16121–16134, Torino, Italia. ELRA and ICCL.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Yuanyi Ren, Haoran Ye, Hanjun Fang, Xin Zhang, and Guojie Song. 2024. ValueBench: Towards comprehensively evaluating value orientations and understanding of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2015–2040, Bangkok, Thailand. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Taylor Sorensen, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. 2024. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19937–19947.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. " my answer is c": First-token probabilities do not match text answers in instruction-tuned language models. *arXiv preprint arXiv:2402.14499*.

Jing Yao, Xiaoyuan Yi, Xiting Wang, Yifan Gong, and Xing Xie. 2023. Value fulcra: Mapping large language models to the multidimensional spectrum of basic human values. *arXiv preprint arXiv:2311.10766*.

Junjie Zhang, Ruobing Xie, Yupeng Hou, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. 2023. Recommendation as instruction following: A large language model empowered recommendation approach. *arXiv preprint arXiv:2305.07001*.

Zhaowei Zhang, Fengshuo Bai, Jun Gao, and Yaodong Yang. 2024a. Valuedcg: Measuring comprehensive human value understanding ability of language models. *Preprint*, arXiv:2310.00378.

Zhaowei Zhang, Ceyao Zhang, Nian Liu, Siyuan Qi, Ziqi Rong, Song-Chun Zhu, Shuguang Cui, and Yaodong Yang. 2024b. Heterogeneous value alignment evaluation for large language models. *Preprint*, arXiv:2305.17147.

A Appendix

A.1 Prompt Examples

We provide examples for two different settings of evaluation involved in Table 2 and 3.

A.1.1 Example 1

No probing:

`User:` How worried are you, if at all, about climate change? Choices: A. Very worried, B. Fairly worried, C. Not worried
`Assistant:`

Prompting:

`User:` Overall, would you say science has had a mostly positive effect on our society or a mostly negative effect on our society?
`Assistant:` Mostly Positive.
`User:` How worried are you, if at all, about climate change? Choices: A. Very worried, B. Fairly worried, C. Not worried
`Assistant:`

A.1.2 Example 2

No probing:

`User:` How do you like thriller movies? A. A lot, B. A bit, C. not much, D. not at all.
`Assistant:`

Prompting:

`User:` How do you like action movies?
`Assistant:` A lot.
`User:` How do you like thriller movies?
`Assistant:`

A.2 Lower and Upper Bounds of Opinion Extrapolation

Here we provide a lower-bound and an estimated upper-bound to the settings in Table 2. These extra results are in Table 8.

The lower-bound is from a uniform distribution that assigns equal probabilities to each choice of the questions.

The upper-bound is from the most representative individual in the probing opinion-specified group that has the highest similarity with the entire group.

	+Climate change	+Gun control	+Birth control	+Gay marriage	+Religion
uniform (lowerbound)	0.32	0.27	0.29	0.31	0.33
no probing	0.42	0.33	0.39	0.45	0.28
prompting	0.45	0.37	0.44	0.51	0.39
rej. sampling	**0.61**	**0.52**	**0.56**	**0.62**	**0.55**
representative (upperbound)	0.84	0.79	0.81	0.85	0.73
	-Climate change	-Gun control	-Birth control	-Gay marriage	-Religion
uniform (lowerbound)	0.35	0.32	0.28	0.28	0.30
no probing	0.4	0.35	0.34	0.29	0.36
prompting	0.38	0.41	0.37	0.35	0.39
rej. sampling	**0.48**	**0.49**	**0.44**	**0.46**	**0.52**
representative (upperbound) 0.65	0.70	0.68	0.68	0.72	

Table 8: The similarities between the human distribution and model distribution on political topics from OpinionQA. "+" indicates positive opinion, and "-" indicates negative opinion. Higher similarities indicate a better ability to extrapolate opinions. Rejection sampling demonstrates the best alignment, whereas prompting the model with probing opinions slightly improves model alignment.