How Far Can We Extract Diverse Perspectives from Large Language Models?

Shirley Anugrah Hayati **Minhwa Lee** **Dheeraj Rajagopal** **Dongyeop Kang**
University of Minnesota Google DeepMind
{hayat023, lee03533, dongyeop}@umn.edu rajagopald@google.com

Abstract

Collecting diverse human opinions is costly and challenging. This leads to a recent trend in exploiting large language models (LLMs) for generating diverse data for potential scalable and efficient solutions. However, the extent to which LLMs can generate diverse perspectives on subjective topics is still unclear. In this study, we explore LLMs' capacity of generating diverse perspectives and rationales on subjective topics such as social norms and argumentative texts. We introduce the problem of extracting *maximum diversity* from LLMs. Motivated by how humans form opinions based on values, we propose a criteria-based prompting technique to ground diverse opinions. To see how far we can extract diverse perspectives from LLMs, or called *diversity coverage*, we employ a step-by-step recall prompting to generate more outputs from the model iteratively. Our methods, applied to various tasks, show that LLMs can indeed produce diverse opinions according to the degree of task subjectivity. We also find that LLMs performance of extracting maximum diversity is on par with human.[1]

1 Introduction

Using NLP for tasks that require social reasoning or involve human subjectivity like argumentation (Hidey et al., 2017) or toxicity detection (Sap et al., 2019) often calls for diverse perspectives. Instead of providing a single viewpoint, an ideal NLP model should accommodate various perspectives to avoid any bias towards a dominant one. Prior works emphasize the importance of modeling multiple viewpoints (Plank, 2022; Abercrombie et al., 2022). Some studies have addressed this challenge by gathering responses from multiple human annotators with diverse backgrounds (Rottger et al., 2022; Santy et al., 2023). However, this approach is costly in time and resources. Recent advancements of LLMs have gained much interest from researchers to exploit their capability of creative generation for data augmentation with less cost and higher diversity (Cegin et al., 2023; Chung et al., 2023; Bubeck et al., 2023).

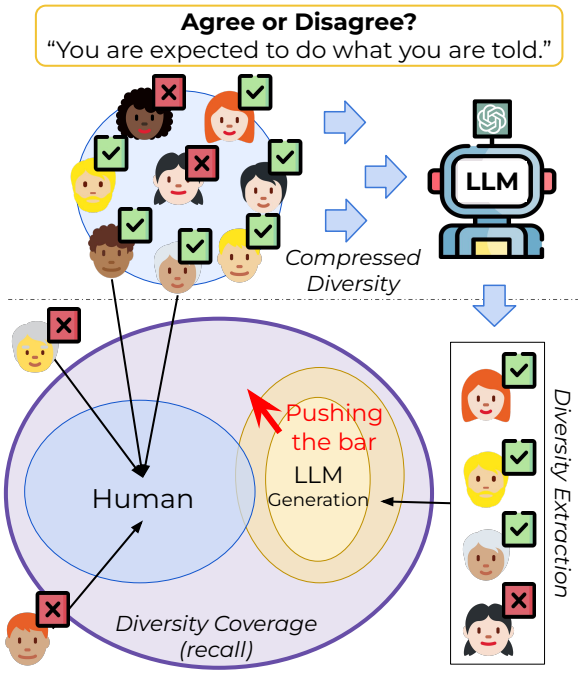


Figure 1: LLMs are trained on texts written by different people who may have distinct perspectives. Our study examines whether LLMs can do "reverse modeling" of humans' perspectives from the training data and how much diversity coverage LLMs can generate. (A check mark = "*Agree*" and a cross mark ="*Disagree*")

LLM is known as a compressed parametric knowledge (e.g., blurry JPEG) of the training corpus (Chiang, 2023), and our work study how people's pluralistic diverse opinions are compressed in the parameter and how far we can reversely extract them from an LLM. Figure 1 illustrates the significance of understanding the extent of diversity achievable by LLMs. During training, LLMs

[1]Our code and data are available at https://github.com/minnesotanlp/diversity-extraction-from-llms. The extracted opinions can be viewed on our project page here: https://minnesotanlp.github.io/diversity-extraction-from-llms/

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 5336–5366
November 12-16, 2024

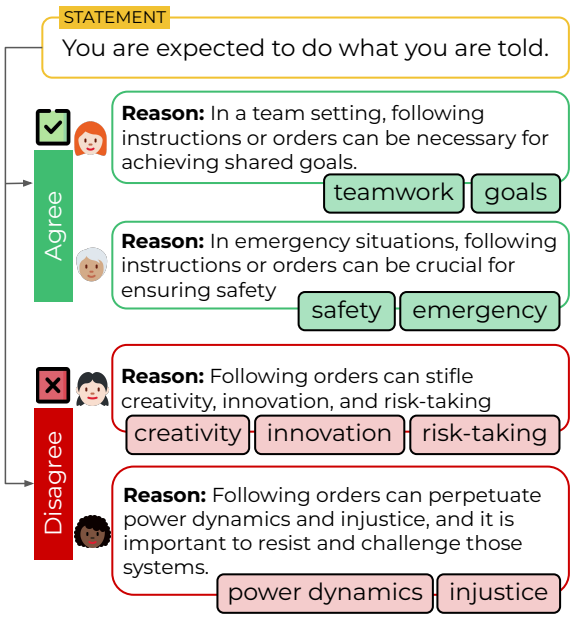


Figure 2: People can have different opinions given a subjective statement. Given a statement, humans can agree or disagree with the statement with their own criteria (e.g., teamwork, risk-taking) in deciding their stances.

have access to various writings from humans with distinct values. Yet, can LLMs reflect this diversity when generating text, or do they tend to favor majority opinions? How do LLM-generated opinions overlap with human viewpoints? If the purple circle in Figure 1 hypothetically represents the maximum diversity achievable by humans, we aim to explore methods for LLMs to approach this diversity.

In real-world scenarios, humans may take different stances on a subjective statement (Figure 2). For instance, those valuing *teamwork* and *goal achievements* may agree with the statement, while others valuing *creativity* and *innovation* could oppose it. Then the key question is: ***How many and what diverse perspectives do LLMs model?***

We introduce a novel problem called *maximum diversity extraction* from LLMs. Formally, our task involves (1) asking LLMs to generate as many different stances as possible, (2) providing reasons for each stance, and (3) producing important criteria words that guide their reasoning process. We apply this diversity prompting across four subjective tasks: social norms, argumentation, hate speech labeling, and story continuation.

Contributions First, we propose perspective diversity as a new focus for generative LLMs, distinct from lexical, syntactical, and semantic diversity mainly explored in previous studies. Through various experiments, we assess LLMs' capacity to achieve maximum perspective diversity. Second, we introduce criteria-based diversity prompting to extract and ground diverse perspectives from LLMs. Finally, we suggest a step-by-step recall approach to measure the extent of diversity coverage of LLMs, comparing the coverage between LLM-generated opinions with human-authored opinions.

Our study reveals a saturation point in the diversity that LLMs can achieve, depending on the subjectivity of a task. Also, LLMs generally produce more diverse opinions than an individual human, but two or more humans achieve greater diversity. Regarding the quality of LLM's generation, LLM is able to generate opinions which are semantically similar to human opinions. However, some frequent criteria words by LLM are different from what humans consider as important.

2 Criteria-based Diversity Prompting

2.1 Motivation

First, we present the motivation behind our approach. Imagine engaging in a debate with someone over a controversial topic. Effective debaters often employ overarching framing to shape their arguments persuasively and coherently. For example, framing arguments around "power dynamics" or "creativity" can effectively challenge the given statement as shown in the "disagree" examples in Figure 2. We refer to these framing keywords as *criteria*. Opinions guided by these criteria could be more diverse as they are grounded in the combination of various criteria words.

2.2 Step 1: Think of Your Criteria First before Making Opinions

Our **criteria-guided prompting** is as follows: "Given a *statement*, generate a Stance and explain its Reasons with a list of Criteria that affect a model's perspective".

Task Definition The task is defined as choosing a binary stance and generating supporting reasons for a given subjective statement. The generated criteria by a model are a list of words or short phrases. The model's reasons include a free-form explanation of its stance (Table 1 and Table 6).

Criteria-Based vs. Free-form We then compare model's diversity performance on two prompting

settings: with criteria vs. without criteria (free-form). From our experiments (Section 4.1), we found that the criteria-based prompting method enables the model to generate important criteria for framing high-level decisions and providing well-grounded reasons. The criteria list can also be seen as reflecting the model's values. This approach follows human reasoning, where personal values often guide opinions and behavior (Rokeach, 1973; Kesberg and Keller, 2018).

Motivated by recent advancements in few-shot learning that have enhanced model performance on challenging tasks, we utilize in-context prompting to explore the model's capacity to generate diverse opinions.[2] (Perez et al., 2021; Min et al., 2022b,a; Lu et al., 2022). The output format is structured as a Python dictionary to be parsed for diversity evaluation. Each few-shot example contains ten opinions - five agreeing with the statement and five disagreeing. This setting does not influence the number of generated opinions and the content of each opinion, as we found cases when the model produces an imbalanced number of stances or opinions fewer than 10 (details in Appendix A.9). We also test the best-performing model with a zero-shot approach.

Human Evaluation on Model-Generated Opinions To ensure the quality of model-generated opinions, we conduct human inspections to verify if each opinion entails its corresponding statement and stance. Over 99% of opinions in randomly sampled opinion-statement pairs were found to align accurately. We also examine whether the generated criteria words entail the free-form reasons they support; 96% of opinions in randomly sampled opinion-criteria pairs demonstrated this entailment. Further details about this process are described in the Appendix A.11 and A.12.

2.3 Step 2: Step-by-Step Recall Prompting to Maximize Diversity Incrementally

Once we identify the best setups, we expand our diversity prompting approach to include step-by-step recall prompting to assess the LLMs' diversity coverage. In this experiment, no examples are provided in the prompt. Instead, we only extract one opinion for a given statement and prompt the model to generate additional opinions iteratively until reaching a specified number, N (Figure 3). Without 1-shot demonstration, weaker LLMs often struggle to produce structured outputs. The purpose of this recall prompting experiment is to identify the "saturation point" of the model's diversity, which is the maximum number of unique diverse opinions an LLM can generate.

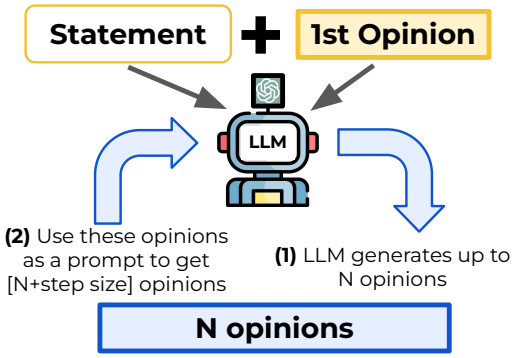


Figure 3: Step-by-step recall prompting. The statement and first generated opinion become the demonstration for prompting the LLM to generate N opinions. The blue-colored parts (Steps 1 and 2) are done incrementally with step size = 3.

Across the experiment, we set $N \in \{2, 5, 8, 11, 14, 17, 20\}$. The first opinion generated by the LLM guides the structured output format since few-shot prompting is not employed in this experiment.

3 Experimental Setups

3.1 Models and Prompting

We experiment primarily with four GPT variants: GPT-4o[3], GPT-4, GPT-3.5 (OpenAI, 2023), and GPT-3 (`text-davinci-002`) (Brown et al., 2020), along with Llama3-70b-chat[4] and Mixtral 8x7B (Jiang et al., 2024). Our baseline involves free-form prompting, where the model generates its stance and reasoning without generating criteria words. Our primary prompting setup uses in-context learning with one example of ten opinions per statement. We also compare with five-shot prompting across all LLMs and zero-shot prompting only for the strongest model, GPT-4, as weaker models may struggle to generate structured outputs without examples. In step-by-step recall prompting experiments, we vary the number of opinions generated.

3.2 Datasets

SOCIAL-CHEM-101 contains texts about of social norms and moral judgments for a given situation written by crowd-workers (Forbes et al., 2020).

[2]Please refer A.1 for details on how we choose this setting.

[3]https://openai.com/index/hello-gpt-4o/

[4]https://llama.meta.com/llama3/

Model Input (one-shot example below)

Statement: *It's okay to have privacy*
Tell me opinions about the statement as many as possible from different people with, "Agree" or "Disagree," one-word or one-phrase criteria that is important for their opinions, and explain how they have different opinions
Output:

```
{1:{"Stance" :"Agree",
    "Criteria": ["personal boundaries", "autonomy"],
    "Reason": "Having privacy allows individuals to establish personal boundaries and maintain."
    their autonomy."},
 2: {"Stance": "Disagree",
    "Criteria": ["transparency", "trust"],
    "Reason": "Lack of privacy can promote transparency and build trust in relationships."
...
10: {"Stance": "...", "Criteria": [...], "Reason": " ..."}}
...
```

Statement: *You're expected to do what you are told*
Tell me opinions about the statement as many as possible from different people with, "Agree" or "Disagree," one-word or one-phrase criteria that is important for their opinions, and explain how they have different opinions
Output:

Model Output

```
{1:{"Stance" : "Agree",
    "Criteria": ["teamwork", "goals"],
    "Reason": "In a team setting, following instructions or orders can be necessary for achieving
    shared goals."},
...
10: {"Stance": "...", "Criteria": [...], "Reason": " ..."}                                    }
```

Table 1: Example prompt for criteria-based diversity prompting. Prompt for the free-form prompting is all the same except for not having the "Criteria." Full prompts for free-form prompting and criteria prompting are shown in A.2.

Since social norms depend on many factors such as the group's beliefs and cultures (Ajzen, 1991; Shweder, 1990), this dataset is suitable for our task of maximizing LLM's diversity capability. For our study, we randomly sampled 500 texts for the criteria-based vs. free-form prompting experiment and 200 texts for the step-by-step recall prompting.

CHANGE MY VIEW (CMV) consists of debates from online forum threads of the subreddit *Change My View* collected by Hidey et al. (2017). We only take the title of each discussion since it is usually the claim of the argument, resulting in a total of 67 unique claims. We use this dataset to examine if LLMs can produce diverse opinions on a highly subjective task because an argumentation task could be highly controversial (van Eemeren et al., 2015).

HATE SPEECH dataset contains texts categorized as either "hate" or "not hate" speech. (Vidgen et al., 2021). From this dataset, we randomly sample 200 instances, focusing only on implicit hate speech texts which are harder to detect. Hate speech detection is a subjective task because annotators' background may affect how they rate the hate speech label of a text (Sap et al., 2019; Ghosh et al., 2021). We add this task for the step-by-step recall prompting experiment to show how criteria-based prompting can be applied to subjective labeling problems.

MORAL STORIES is a crowd-sourced narrative story dataset (Emelin et al., 2021). For this study, the LLM needs to continue the story with the situation part as. We also randomly sample 200 instances from this dataset and use this dataset for the step-by-step recall prompting experiment to show how our prompting method can be applied to open-ended generation problems.

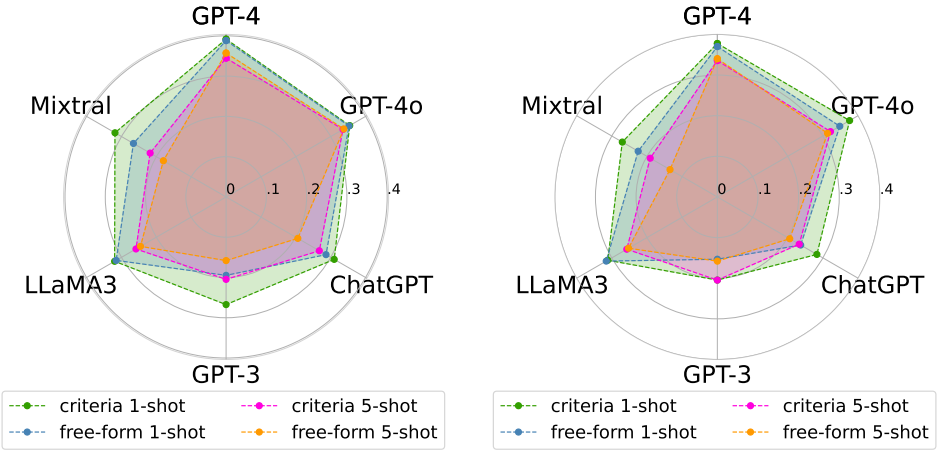
3.3 Evaluation Metric

Semantic diversity To examine the semantic diversity of the model's reasons using both criteria-based and free-form prompting, we convert the LLM-generated "reasons" for each statement into sentence embeddings using SentenceBERT (Reimers and Gurevych, 2019) with DistilRoberta (Sanh et al., 2019). Next, we compute the cosine distance between every pair of reasons and calculate the average cosine distance across all pairs to measure the statement's semantic diversity score. We average these scores across all statements to obtain the overall semantic diversity.

Perspective diversity To evaluate perspective diversity in step-by-step recall prompting, we utilize criteria words generated by LLMs. Some words

Model	#Parameters	SOCIAL-CHEM-101 (↑) Free-form	SOCIAL-CHEM-101 (↑) Criteria	CMV (↑) Free-form	CMV (↑) Criteria
GPT-4	-	0.3883	**0.3919**	0.3701	**0.3776**
GPT-4o	-	0.3525	0.3545	0.3480	0.3759*
GPT-3.5	-	0.2865	0.3100*	0.2368	0.2829*
GPT-3	175B	0.1947	0.2673*	0.1533	0.2046*
Llama3-chat	70B	0.3152	0.3196	0.3158	0.3115
Mixtral	46.7B	0.2657	0.3186*	0.1345	0.1908*
Zero-shot GPT-4	-	0.3176	0.2885	0.2669	0.2410

Table 2: Semantic diversity (cosine distance) results on criteria-based prompting vs. free-form prompting experiments. Both setups use one-shot learning except for "zero-shot GPT-4." **One-shot criteria-based prompting** generally generates more diverse opinions (pink box) with GPT-4 performing the best. *p < 0.01



(a) SOCIAL-CHEM-101 (b) CMV

Figure 4: Semantic diversity score for different LLMs and prompting methods for SOCIAL-CHEM-101 (left) and CMV (right) datasets. **Criteria-based prompting is the best diversity extraction method for across LLM variants, datasets, and various shots.** We also found that **too many examples may hurt diversity (5-shot results).** The results on SOCIAL-CHEM-101 are statistically significant with $p < 0.05$ (GPT-4) and $p < 0.01$ (the rest of the models) and $p < 0.01$ for GPT-3 and Mixtral for CMV.

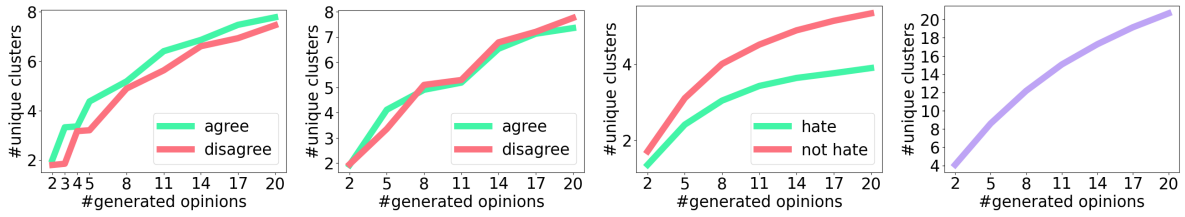
with similar meanings can be conveyed in different ways. For instance, given a statement "*It is expected that friends will enjoy being around each other*," the model could generate two opinions; an opinion may contain "joy" as one of the criteria while the other opinion contains "happiness." We prompt GPT-4 with 3 examples to cluster criteria words with similar meanings into groups (details in A.4). Two human annotators manually inspect 1,159 clusters of criteria words from 100 randomly sampled statements across SOCIAL-CHEM-101, CMV, HATE SPEECH, and MORAL STORIES (25 statements per dataset). From this study, the annotators agree that an average of 80.95% of those clusters of criteria words have similar meanings with inter-annotator percentage agreement of 88.85%. To measure perspective diversity, we count the number of unique criteria clusters for each opinion on a given statement. A higher count indicates greater diversity in the generated opinions.

4 Experiment Results with Automatic Evaluation

4.1 Semantically Diverse Opinions about Social Norms and Argumentation

Table 2 presents our experiment findings regarding semantic diversity. **One-shot prompting on GPT-4 produces notably more semantically diverse reasons compared to other models.** Interestingly, weaker models like Mixtral and GPT-3 benefit most from having criteria to guide them toward generating semantically diverse opinions.

When we prompt GPT-4 without examples and ask for structured outputs only, it tends to gener-



(a) SOCIAL-CHEM-101 (b) CMV (c) Hate Speech (d) Moral Stories

Figure 5: X-axis = the number of generated opinions for our diversity coverage experiment. Y-axis = the average number of unique criteria clusters for all statements. Moral Stories do not have stances, so the line is only for all generated continued stories. **The more subjective a task is, the more LLM can generate unique criteria clusters.**

ate more diverse reasons with free-form prompting. This aligns with our pilot study where asking criteria without examples proved challenging for the model. Notably, GPT-4 generates an average of 6.8 opinions with zero-shot prompting, fewer than all models with one-shot prompting (GPT-4: 9.9 opinions, Llama3: 10.0, Mixtral: 9.0, GPT-4o: 10.1).

Figure 4 shows that **criteria-based prompting consistently outperforms free-form prompting across datasets.** One-shot prompting particularly outputs higher diversity scores than five-shot prompting. This indicates that in five-shot prompting, models tend to adhere more closely to given examples, resulting in less diverse outputs compared to the one-shot setting. Interestingly, in the zero-shot setting, free-form prompting on GPT-4 generates more semantically diverse perspectives than criteria-based prompting. We hypothesize that this occurs because introducing "criteria" in the prompt without concrete examples may confuse the model. As a result, the model produces fewer diverse diverse perspectives compared to the simpler free-form prompting.

4.2 Diversity Coverage by Step-by-Step Recall Prompting

Figure 5 shows an increase in the number of unique criteria clusters as the step size increases for the recall step-by-step experiment. For SOCIAL-CHEM-101 and CMV, the model on average generates 8 unique criteria clusters for agreeing and 7 for disagreeing opinions. For HATE SPEECH, the average number of unique criteria clusters is lower (4 for 'hate' and 5 for 'not hate'). This demonstrates that labeling hate speech is less subjective compared to social norms (SOCIAL-CHEM-101) or argumentation (CMV). On the other hand, MORAL STORIES shows a different trend, with the model generating an average of 20 unique criteria clusters. In Table 4, we can see examples of opinions generated per statement for the recall prompting experiment.

Task Type	Dataset	Max	Median
Stance	SOCIAL-CHEM-101		
	Agree	17	8
	Disagree	16	7
Stance	CMV		
	Agree	17	7
	Disgree	14	7
Labeling	HATE SPEECH		
	Hate	14	4
	Not Hate	16	5
Generation	MORAL STORIES		
	All	47	20

Table 3: Different task types with the maximum and the median of the number of unique criteria clusters. More details are in Figure 9.

Since we limit the maximum number of generated opinions ($N = 20$), the lines on the graph in Figure 5 may seem not to have converged yet. However, increasing N does not necessarily lead to a higher number of unique clusters as this is task-dependent (see Table 3). For instance, in the hate speech labeling task, although we set a maximum $N = 20$, the highest number of unique clusters is 14 for both "hate" and "not hate," with average number of unique clusters is 4 and 5, respectively. For open-ended problems such as story continuation (MORAL STORIES), increasing N would lead to a greater number of unique clusters.

5 Human vs. LLMs: Diverse Opinion Generation

To assess human capabilities of generating diverse opinions, we hire crowd-workers from Amazon Mechanical Turk to generate as many opinions as they can (≥ 3) for each stance (agree, disagree) on 100 statements from SOCIAL-CHEM-101 and 67 statements from CMV. These statements are a subset of the dataset used in our recall prompting experiments. Each worker writes opinions for five

Statement		It's good to be a hard worker.
GPT-4	*Agree*	1. Being a hard worker increases productivity and pushes one towards success and achieving their goals. Criteria: **productivity**, **success**, **goals** 2. Being a hard worker can bring a sense of fulfillment, achievement, and boost self-esteem. Criteria: **fulfillment, achievement, self-esteem** 3. Being a hard worker develops resilience, stokes determination and builds hardiness. Criteria: **resilience**, **determination**, **hardiness**
	Disagree	1. Being a hard worker can damage work-life balance, cause burnout and high levels of stress. Criteria: **work-life balance**, **burnout**, **stress** 2. Being a hard worker can decrease leisure time, family time and negatively affect health. Criteria: **leisure**, **family**, **health** 3. Being a hard worker can prevent peace of mind, disrupt contentment, and break restfulness. Criteria: **peace of mind**, **contentment**, **restfulness**
Human	*Agree*	1. Being hard worker cultivates discipline. 2. Hard work often leads to personal growth and development. 3. Hard workers tend to be more reliable.
	Disagree	1. Being overly focused on hard work can lead to burnout. 2. Sometimes working smart is more effective than working hard. 3. The value of hard work can very depending on the context.

Table 4: Opinions generated by GPT-4 (top) and a human (bottom) about a statement from SOCIAL-CHEM-101.

		SOCIAL-CHEM-101	CMV
Agree	**Human**	9.17 ±3.16	10.56±3.86
	GPT-4	8.14±2.40	7.86 ±2.62
Disagree	**Human**	10.04 ±3.31	11.00±3.81
	GPT-4	7.91 ±2.60	8.30 ±2.74

Table 5: Average number of criteria clusters of human opinions vs. GPT-4-generated opinions per statement with standard deviation. **While humans can write more diverse opinions when asked, LLM's capability for extracting diverse perspectives is quite on par with human capability.**

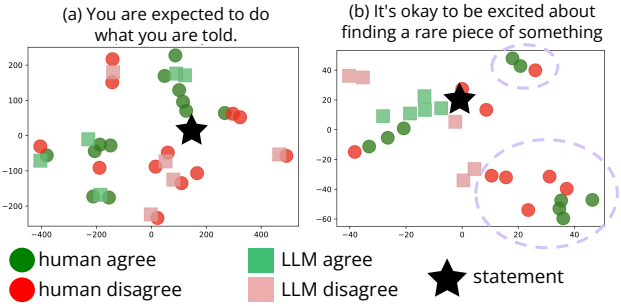


Figure 6: T-SNE for opinions written by human and generated by GPT-4. **LLM can mostly generate both agree and disagree opinions that align with human when they are semantically close to the statement**.

statements per HIT, compensated at $2 per HIT.

For each human-written opinion, we query GPT-4 to extract criteria words and cluster them using the same method employed for computing perspective diversity of model-generated opinions. Table 5 shows that **humans tend to produce slightly more diverse opinions than LLMs**, with approximately 1 or 2 more criteria for SOCIAL-CHEM-101 and 3 more criteria for CMV.

Figure 6 illustrates two statements alongside their respective opinions by humans and GPT-4 in T-SNE plot. The statements and opinions are embedded with the same approach for semantic diversity experiment. We observe that **LLMs can generate agreeing and disagreeing opinions that align with human perspectives**, despite LLMs producing slightly fewer opinions. The failure cases of LLMs occur when human opinions diverge semantically from the statement (e.g., the lower right under purple circles).

Criteria Words by Different LLMs and Humans We analyze the frequent criteria words generated by GPT-4, Llama3, Mixtral, and humans using T-SNE embeddings in Figure 7 for agreeing and disagreeing opinions in SOCIAL-CHEM-101. From each model and humans, we select the top 5 frequent criteria words across all statements and visualize their embeddings on a T-SNE plot.

For agreeing opinions, in general the three LLMs quite align with humans. GPT-4 and Llama3 have "respect" as the most frequent criterion, and all three LLMs regard "responsibility," "safety," and "emapthy" as important criteria. Meanwhile, humans value "trust" the most, and only Llama3 aligns with human for this value. For the disagreeing opinions, we can see that humans value "personal growth" the most and then followed by "cultural norms," "communication," "privacy," and

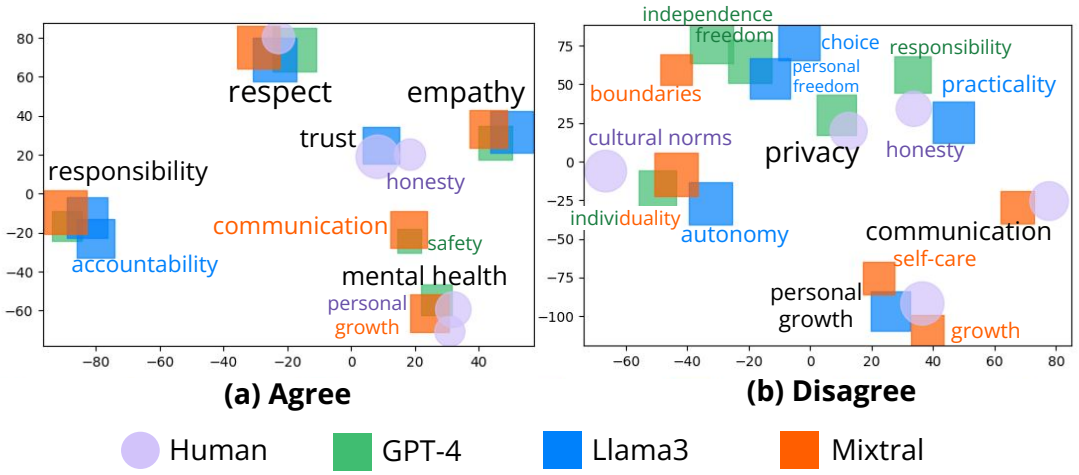
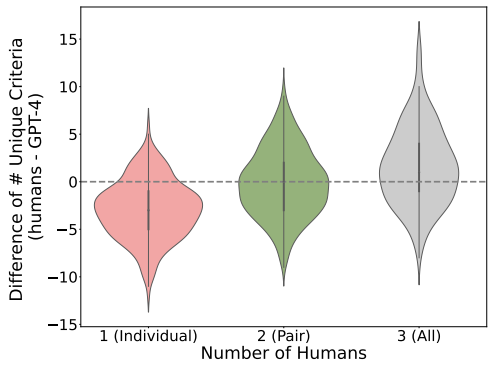


Figure 7: T-SNE plots of five most frequent criteria words by humans and three LLMs: GPT-4, Llama3, and Mixtral. The size of a point represents the frequency. Black font refers to the label of all points next to the text. Purple font for human's criteria words, green for GPT-4, blue for Llama3, and orange for Mistral. **LLMs generally mimic human values, although at times they tend to regard rule-following notions as important (e.g., "responsibility" or "safety") for agree opinions and extreme freedom (e.g., "independence," "boundaries," "individuality") for disagreeing opinions more than humans do.**

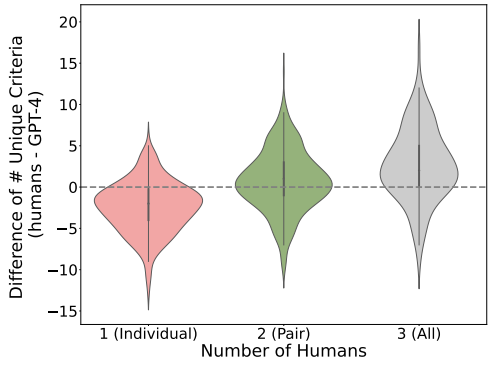
"honesty." However, only Llama3 also considers "personal growth" as important. In general, all LLMs consider that "freedom" and "autonomy" are the most important which sounds more extreme compared to human values.

We also examine how much the criteria words generated by LLMs agree with human responses using top-p sampling (p=10%). For both agreeing and disagreeing opinions, we found that GPT-4 agrees the most with humans (agree: 45.63%, disagree: 39.53%), followed by Llama3 (30.00% and 28.14%), and Mixtral (29.38% and 26.35%). Since the criteria words by humans are extracted by GPT-4, there may be a lexical bias toward words that GPT-4 frequently uses. However, the T-SNE plot displays the semantic closeness of the criteria words by these different LLMs and humans. All three LLMs align well with humans with a tendency of favoring rule-following criteria words for agreeing opinions and extreme independence for disagreeing opinions.

When Will Human Reach LLM's Diversity Generation Capability? To examine how many humans are needed to match the diversity generation capability of LLMs, we compute the difference of the number of unique clusters of criteria in human opinions vs. GPT-4's generation. Figure 8 visualizes the distribution of these differences between humans and GPT-4. Our analysis indicates that a person tends to generate fewer unique perspectives compared to GPT-4. However, **a pair of humans or more shows a higher density of matching or exceeding the diversity capability of GPT-4** when generating unique opinions for statements in both SOCIAL-CHEM-101 and CMV. This suggests the



(a) SOCIAL-CHEM-101



(b) CMV

Figure 8: The distribution of the differences in the number of unique criteria clusters between human and GPT-4. **A pair of humans can equalize LLM's capability of extracting maximum diversity.**

importance of "communication" between individuals to broaden one's perspectives, underscoring the value of having an LLM capable of generating diverse viewpoints.

6 Related Work

Diversity in NLP Diversity in NLP has been extensively explored across various dimensions: (1) lexical variability (Dušek and Kasner, 2020; Tevet and Berant, 2021; Li et al., 2016); (2) syntactical diversity (Giulianelli et al., 2023; Huang et al., 2023); (3) semantic diversity (Stasaski and Hearst, 2022; Reimers and Gurevych, 2019; Zhang et al., 2020); and (4) perspective diversity (Plank, 2022; Hayati et al., 2021; Santy et al., 2023). Some studies focus on annotator diversity (Rottger et al., 2022; Wich et al., 2021), while others examine diversity in generated language (Hashimoto et al., 2019; Liu et al., 2023). Our work aligns closely with prior studies on perspective diversity, specifically in examining stances and rationales generated by LLMs. However, unlike previous research primarily focusing on classification tasks, our investigation encompasses sentence-level reasoning diversity, offering a more nuanced perspective. Joshi et al. (2020) argue that NLP research is biased toward Western perspectives. Thus, diverse perspectives from minority populations are relatively overlooked. Our work is important to uncover the extent to which perspective diversity can be extracted from LLMs.

Diversity Generation by LLMs LLMs have been utilized extensively to produce diverse synthetic datasets, such as paraphrasing (Cegin et al., 2023), structured wiki-like bios for notable figures (Yuan et al., 2022), and instruction datasets (Wang et al., 2023b; Taori et al., 2023; Honovich et al., 2023). Unlike diverse large-scale data generation, Lahoti et al. (2023); Giulianelli et al. (2023) specifically examine variability in model responses. They propose novel prompting techniques to enhance diversity in LLM outputs, particularly concerning gender and cultural prompts. Our work aligns closely with Lahoti et al. (2023) by advocating for fairness in LLM outputs through perspective diversity, which goes beyond semantic variability.

Additionally, our approach is similar to Giulianelli et al. (2023) in generating multiple responses per prompt. To promote opinion diversity in LLMs, Aroyo et al. (2023) introduce a dataset labeled by human raters from different demographic populations, focusing on safety such as bias, misinformation, and harmful content. Despite the previous claim that LLMs can produce diverse content, some studies suggest that co-writing with LLMs may affect human writers' opinions (Jakesch et al., 2023) and reduce writing diversity (Padmakumar and He, 2024). Our research addresses this gap by proposing a method to generate diverse perspectives rather than a single dominant opinion.

7 Conclusion and Future Work

To the best of our knowledge, this is the first work that tackles extracting maximum perspective diversity from LLMs. To do this, we propose a criteria-based prompting method and probe LLMs' capacity to generate as many diverse perspectives as possible and explain their reasons for choosing their corresponding stances on subjective statements. Through our step-by-step recall prompting, we characterize the subjectivity of various tasks and reach the maximum diversity of LLM's generation. LLMs can generate comparable number of diverse outputs with humans and similar values as humans' responses. As we compare LLMs' opinion generations with human's, they are quite "precise" as they are semantically similar to human opinions, bur their recall is slightly lower than humans.

While the number of criteria clusters does not precisely mean the ideal maximum diversity, it indicates that we could use LLMs to push further perspective diversity to include more diverse opinions. Our work opens up a wider range of possibilities for examining more advanced diversity "quantification" and "maximization" methods. There are also many application possibilities for extending this work in the future. In this study, we have not assessed how much the extracted diverse opinions are similar with the real world's diverse opinions yet. Instead, we focus on the diversity coverage. Future work could evaluate this by comparing the distribution of extracted opinions with a distribution of people's opinions collected from real world survey or poll data. We recommend further exploration of cultural aspects, persona, or human values on diversity extraction. Moreover, our method could be applied for curating diverse data for open-ended tasks, such as generating diverse outputs for instruction-tuning tasks or subjective task labeling. Findings that it takes two humans to equalize LLMs' diverse generation capacity suggest that communications between multiple humans or LLMs can be a future work to introduce more diverse perspectives.

Limitations and Ethical Considerations

While our prompting approach does not generate an exhaustive, complete list of diverse opinions, our study serves as a comparative study that examines the capability of various language models for generating diverse opinions given various numbers of examples and input datasets. Moreover, for now, we only experimented with our proposed criteria-based prompting technique for subjective tasks. It would be interesting future work to try the same technique on non-subjective tasks. Currently, we rely on LLMs (GPT-4) to generate criteria words from non-criteria prompting outputs. Future works could deal with in-depth variations of these criteria word extraction methods and analyses on the words themselves.

We noticed that the demographics of crowd-workers who participate in the opinion writing are skewed toward white with bachelor degree as their highest education level. Demographic factors, including culture, may impact how these opinions are written. For future work, it would be interesting to compare more opinions written by participants from other cultures with the model's generated opinions.

We also have not explored all different combinations of setups of decoding parameters besides comparing different temperatures and top_p sampling during the initial experiments. However, we would like to highlight that our work is not simply probing LLM's ability to generate diverse tokens that may convey similar meaning, but rather if the LLM has the capability for generating diverse perspectives. Examining various decoding methods could be a potential future work for this study.

Potential risks could be a situation where our criteria-based prompting attempts to generate diverse opinions on certain topics that are socially unacceptable and/or contain harmful content. To mitigate any concern regarding this issue, we suggest that researchers carefully review the content of subjects before applying our prompting approach to their work.

For the human study, our institution determined our study as exempt from IRB review. Since the topic of some statements could trigger human workers, we added warnings before the worker could proceed to work on our task.

Acknowledgment

We would like to thank members of the Minnesota NLP lab and Google reviewers for their valuable feedback and suggestions on the paper draft.

References

Gavin Abercrombie, Valerio Basile, Sara Tonelli, Verena Rieser, and Alexandra Uma, editors. 2022. *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*. European Language Resources Association, Marseille, France.

Icek Ajzen. 1991. The theory of planned behavior. *Organizational behavior and human decision processes*, 50(2):179–211.

Lora Aroyo, Alex S. Taylor, Mark Diaz, Christopher M. Homan, Alicia Parrish, Greg Serapio-Garcia, Vinodkumar Prabhakaran, and Ding Wang. 2023. Dices dataset: Diversity in conversational ai evaluation for safety. *Preprint*, arXiv:2306.11247.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Jan Cegin, Jakub Simko, and Peter Brusilovsky. 2023. Chatgpt to replace crowdsourcing of paraphrases for intent classification: Higher diversity and comparable model robustness. *arXiv preprint arXiv:2305.12947*.

Ted Chiang. 2023. Chatgpt is a blurry jpeg of the web.

John Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 575–593, Toronto, Canada. Association for Computational Linguistics.

Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with natural language inference. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 131–137, Dublin, Ireland. Association for Computational Linguistics.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and

their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Sayan Ghosh, Dylan Baker, David Jurgens, and Vinodkumar Prabhakaran. 2021. Detecting cross-geographic biases in toxicity modeling on social media. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 313–328, Online. Association for Computational Linguistics.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023. What comes next? evaluating uncertainty in neural text generators against human production variability. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14349–14371, Singapore. Association for Computational Linguistics.

Tatsunori B. Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. 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 1689–1701, Minneapolis, Minnesota. Association for Computational Linguistics.

Shirley Anugrah Hayati, Dongyeop Kang, and Lyle Ungar. 2021. Does BERT learn as humans perceive? understanding linguistic styles through lexica. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6323–6331, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21, Copenhagen, Denmark. Association for Computational Linguistics.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, Toronto, Canada. Association for Computational Linguistics.

Kuan-Hao Huang, Varun Iyer, I-Hung Hsu, Anoop Kumar, Kai-Wei Chang, and Aram Galstyan. 2023. ParaAMR: A large-scale syntactically diverse paraphrase dataset by AMR back-translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8047–8061, Toronto, Canada. Association for Computational Linguistics.

Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. ACM.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Rebekka Kesberg and Johannes Keller. 2018. The relation between human values and perceived situation characteristics in everyday life. *Frontiers in Psychology*, 9. Place: Switzerland Publisher: Frontiers Media S.A.

Preethi Lahoti, Nicholas Blumm, Xiao Ma, Raghavendra Kotikalapudi, Sahitya Potluri, Qijun Tan, Hansa Srinivasan, Ben Packer, Ahmad Beirami, Alex Beutel, and Jilin Chen. 2023. Improving diversity of demographic representation in large language models via collective-critiques and self-voting. *Preprint*, arXiv:2310.16523.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

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. *Preprint*, arXiv:2303.16634.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered

prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022a. Metaicl: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022b. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Vishakh Padmakumar and He He. 2024. Does writing with language models reduce content diversity? *Preprint*, arXiv:2309.05196.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *Advances in neural information processing systems*, 34:11054–11070.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Milton Rokeach. 1973. The nature of human values. *The nature of human values.*, pages x, 438–x, 438. Place: New York, NY, US Publisher: Free Press.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing design biases of datasets and models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102, Toronto, Canada. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Richard A. Shweder. 1990. In defense of moral realism: Reply to gabennesch. *Child Development*, 61(6):2060–2067.

Katherine Stasaski and Marti Hearst. 2022. Semantic diversity in dialogue with natural language inference. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 85–98, Seattle, United States. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Guy Tevet and Jonathan Berant. 2021. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 326–346, Online. Association for Computational Linguistics.

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.

Frans H. van Eemeren, Sally Jackson, and Scott Jacobs. 2015. *Argumentation*, pages 3–25. Argumentation Library. Springer, Germany. Publisher Copyright: © 2015, Springer International Publishing Switzerland.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023a. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Maximilian Wich, Christian Widmer, Gerhard Hagerer, and Georg Groh. 2021. Investigating annotator bias in abusive language datasets. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1515–1525, Held Online. INCOMA Ltd.

Ann Yuan, Daphne Ippolito, Vitaly Nikolaev, Chris Callison-Burch, Andy Coenen, and Sebastian Gehrmann. 2022. Synthbio: A case study in human-ai collaborative curation of text datasets. *Preprint*, arXiv:2111.06467.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. *Preprint*, arXiv:2205.10625.

A Appendix

A.1 Other prompting techniques we tried

During our pilot study with a smaller number of subjective statements, we tried various prompting methods, such as (1) zero-shot vs. few-shot prompting; (2) the number of opinions (10, 15, 20) in the few-shot examples; (3) prompts to generate structured output vs. unstructured output; and (4) varying N greater than 20 (e.g., 30, 50, 100) for our step-by-step recall prompting. However, it would be too costly in terms of budget and time to run all the combinations of set-ups, so we chose the current setup since it produces enough diverse output to examine the saturation point and the structured outputs are easy to process. We are aware of other prompting techniques such as Chain-of-Thought (CoT) prompting (Wei et al., 2023), least-to-most prompting (Zhou et al., 2023), and self-consistency prompting (Wang et al., 2023a). However, none of them handle the task of asking LLM to generate the maximum diversity. On the other hand, our diversity prompting can complement these other prompting approaches for extracting diversity.

A.2 Prompting

- Free-form prompting

  Five-shot free-form prompt is shown as follows. One-shot free-form prompting only takes the first statement ("It's rude to use profanity") and its related opinions as a demonstration in the prompt.

  ```
  Statement: It's rude to use profanity.

  Tell me opinions about the statement as many as possible from different people with "Agree" or "Disagree" and explain how they have different opinions.

  Output: {1: {"Stance" : "Agree", "Reason": "I was brought up in a family and culture where using profanity was considered impolite and disrespectful, and I have continued to hold that belief in my adult life.", 2: {"Stance": "Disagree", "Reason": "Profanity can be a form of self-expression, and there are situations where it may be appropriate or even necessary to use it to convey strong emotions or opinions."}, 3: {"Stance": "Agree", "Reason": "Using profanity goes against accepted norms of etiquette and social behavior, and it is considered impolite and
  ```

rude in most social settings."}, 4: {"Stance": "Disagree", "Reason": "The appropriateness of profanity depends on the context and the individual's perception, as what may be acceptable in one situation may not be in another."}, 5: {"Stance": "Agree", "Reason":"Using profanity can disrupt social harmony and respectful communication, and it is important to maintain a certain level of decorum in interactions."}, 6: {"Stance": "Disagree", "Reason": "Individuals should have the freedom to choose their language, and using profanity can be a way to express themselves authentically."}, 7: {"Stance": "Disagree", "Reason": "In certain cultures or societies, the use of profanity may be more accepted or even a normal part of everyday language."}, 8: {"Stance": "Agree", "Reason": "In a professional environment, the use of profanity can be seen as unprofessional and can negatively impact one's reputation and relationships with colleagues."}, 9:{"Stance": "Disagree","Reason": "Profanity can be used in art and creative expression to evoke emotions, challenge societal norms, or convey a specific message."}, 10:{"Stance": "Agree", "Reason": "I was taught from a young age that using profanity is rude and disrespectful, and I continue to adhere to those values."}}

Statement: It's okay to have privacy. Tell me opinions about the statement as many as possible from different people with "Agree" or "Disagree" and explain how they have different opinions.

Output: {1: {"Stance": "Agree", "Reason": "Having privacy allows individuals to establish personal boundaries and maintain their autonomy."}, 2: {"Stance": "Disagree", "Reason": "Lack of privacy can promote transparency and build trust in relationships."}, 3: {"Stance": "Agree", "Reason": "Privacy provides a sense of security and protection, especially in personal and sensitive matters."}, 4: {"Stance": "Disagree", "Reason": "Reduced privacy enables accountability and facilitates supervision, ensuring responsible behavior."}, 5: {"Stance": "Agree", "Reason": "Privacy contributes to personal well-being and mental health by allowing individuals to have space for self-reflection and relaxation."}, 6: {"Stance": "Disagree", "Reason": "Limited privacy fosters social connectedness and a sense of community, as individuals share their lives more openly."}, 7: {"Stance": "Agree", "Reason": "Privacy fosters intimacy and allows individuals to be vulnerable in personal relationships."}, 8: {"Stance": "Disagree", "Reason": "Reduced privacy promotes openness and honesty, as individuals are more transparent about their actions and intentions."}, 9: {"Stance": "Agree", "Reason": "Privacy facilitates personal growth and self-discovery, providing a space for individuals to explore their interests and identities."}, 10: {"Stance": "Disagree", "Reason": "Limited privacy enhances safety and security by allowing for greater surveillance and monitoring of individuals."}}

Statement: It's good to confide in your parents the things that you're struggling with. Tell me opinions about the statement as many as possible from different people with "Agree" or "Disagree" and explain how they have different opinions.

Output:

{1: {"Stance": "Agree", "Reason": "Confiding in parents can provide emotional support, strengthen the parent-child bond, and allow parents to provide guidance in times of struggle."}, 2: {"Stance":

"Disagree", "Reason": "Some individuals may fear judgment from their parents or value their privacy, leading them to hesitate in confiding their struggles."}, 3: {"Stance": "Agree", "Reason": "Sharing struggles with parents can build trust and create opportunities for collaborative problem-solving and finding solutions."}, 4: {"Stance": "Disagree", "Reason": "Some individuals believe in resolving their struggles independently, promoting self-reliance and independence."}, 5: {"Stance": "Agree", "Reason": "Confiding in parents allows them to provide empathy, understanding, and emotional support during challenging times."}, 6: {"Stance": "Disagree", "Reason": "Communication barriers or a generation gap may hinder effective understanding and make individuals less likely to confide in their parents."}, 7: {"Stance": "Agree", "Reason": "Parents, having gone through similar experiences, can offer wisdom and advice based on their own life lessons."}, 8: {"Stance": "Disagree", "Reason": "Some individuals may have experienced neglect or unhelpful advice from their parents in the past, making them hesitant to confide in them."}, 9: {"Stance": "Agree", "Reason": "Sharing struggles with parents can provide a sense of catharsis and emotional relief by expressing emotions and thoughts."}, 10: {"Stance": "Disagree", "Reason": "Some individuals find more comfort in confiding their struggles to friends or other trusted individuals, seeking peer support."}}

Statement: You should always try and help people in tough situations.

Tell me opinions about the statement as many as possible from different people with "Agree" or "Disagree" and explain how they have different opinions.

Output:

{1: {"Stance": "Agree", "Reason": "Helping others in tough situations is an act of empathy, compassion, and kindness, and it can make a positive impact on their lives."}, 2: {"Stance": "Disagree", "Reason": "Helping others in tough situations can put one's personal safety at risk and may violate personal boundaries."}, 3: {"Stance": "Agree", "Reason": "As members of society, we have a moral obligation and responsibility to help those in need, especially in tough situations."}, 4: {"Stance": "Disagree", "Reason": "In some situations, prioritizing self-preservation and personal priorities may be more important than helping others in tough situations."}, 5: {"Stance": "Agree", "Reason": "Helping others in tough situations fosters a sense of community and solidarity, and it can strengthen social bonds."}, 6: {"Stance": "Disagree", "Reason": "Helping others in tough situations can enable dependency and may not address the root causes of their problems."}, 7: {"Stance": "Agree", "Reason": "Helping others in tough situations is an act of altruism and generosity, and it can bring a sense of fulfillment and purpose."}, 8: {"Stance": "Disagree", "Reason": "Helping others in tough situations should respect personal boundaries and require consent, as forcing help on someone can be harmful."}, 9: {"Stance": "Agree", "Reason": "Helping others in tough situations is a part of humanitarianism and global responsibility, and it can contribute to a more just and equitable world."}, 10: {"Stance": "Disagree", "Reason": "Helping others in tough situations should be done with cultural sensitivity and respect, as imposing one's values and beliefs through helping others can be harmful and disrespectful."}}

Statement: You shouldn't let people eat in your car.

Tell me opinions about the statement as many as possible from different people with "Agree" or "Disagree" and explain how they have different opinions.

Output:

{1: {"Stance": "Agree", "Reason": "Allowing people to eat in your car can lead to messiness and disrespect for your property."}, 2: {"Stance": "Disagree", "Reason": "Allowing people to eat in your car can be a hospitable gesture and make them feel more comfortable."}, 3: {"Stance": "Agree", "Reason": "Allowing people to eat in your car can leave unpleasant odors and stains that are difficult to remove."}, 4: {"Stance": "Disagree", "Reason": "Allowing people to eat in your car can be convenient and save time, especially when on the go."}, 5: {"Stance": "Agree", "Reason": "Allowing people to eat in your car can compromise hygiene and pose health risks, especially if the food is not properly handled or stored."}, 6: {"Stance": "Disagree", "Reason": "Allowing people to eat in your car is a personal choice and can depend on the relationship with the person and the circumstances."}, 7: {"Stance": "Agree", "Reason": "Allowing people to eat in your car can be a distraction and compromise safety, especially if the driver is also eating."}, 8: {"Stance": "Disagree", "Reason": "Allowing people to eat in your car can demonstrate flexibility and adaptability, especially in social situations."}, 9: {"Stance": "Agree", "Reason": "Allowing people to eat in your car can decrease its resale value and require more maintenance and cleaning."}, 10: {"Stance": "Disagree", "Reason": "Allowing people to eat in your car can be acceptable in certain cultural norms and etiquette, and it depends on the context and situation.}}

- Prompt for criteria-based prompting:

  Statement: It's rude to use profanity.

  Tell me opinions about the statement as many as possible from different people with "Agree" or "Disagree", one-word or one-phrase criteria that is important for their opinions, and explain how they have different opinions.

  Output: {1: {"Stance" : "Agree", "Criteria": ["impolite", "disrespectful", "culture", "family"], "Reason": "I was brought up in a family and culture where using profanity was considered impolite and disrespectful, and I have continued to hold that belief in my adult life.", 2: "Stance": "Disagree", "Criteria": ["self-expression", situation], "Reason": "Profanity can be a form of self-expression, and there are situations where it may be appropriate or even necessary to use it to convey strong emotions or opinions.", 3: {"Stance": "Agree", "Criteria":["etiquette", "social norms", "impolite", "rude"], "Reason": "Using profanity goes against accepted norms of etiquette and social behavior, and it is considered impolite and rude in most social settings."}, 4: {"Stance": "Disagree", "Criteria": ["contextual appropriateness", "individual perception"], "Reason": "The appropriateness of profanity depends on the context and the individual's perception, as what may be acceptable in one situation may not be in another."}, 5: {"Stance": "Agree", "Criteria": ["harmony", "responsible communication"], "Reason":"Using profanity can disrupt social harmony and respectful communication, and it is important to maintain a certain level of decorum in interactions."}, 6:{"Stance": "Disagree", "Criteria": ["individuality", "choice"],

"Reason": "Individuals should have the freedom to choose their language, and using profanity can be a way to express themselves authentically."}, 7:{"Stance": "Disagree", "Criteria": ["culture", "society"], "Reason": "In certain cultures or societies, the use of profanity may be more accepted or even a normal part of everyday language."},8:{"Stance": "Agree", "Criteria": ["professionalism", "workplace"], "Reason": "In a professional environment, the use of profanity can be seen as unprofessional and can negatively impact one's reputation and relationships with colleagues."},9:{"Stance": "Disagree", "Criteria": ["artistic expression", "creativity"], "Reason": "Profanity can be used in art and creative expression to evoke emotions, challenge societal norms, or convey a specific message."},10:{"Stance": "Agree", "Criteria": ["education", "upbringing"], "Reason": "I was taught from a young age that using profanity is rude and disrespectful, and I continue to adhere to those values."}}

Statement: It's okay to have privacy.

Tell me opinions about the statement as many as possible from different people with "Agree" or "Disagree", one-word or one-phrase criteria that is important for their opinions, and explain how they have different opinions.

Output:

{1: {"Stance": "Agree", "Criteria": ["personal boundaries", "autonomy"], "Reason": "Having privacy allows individuals to establish personal boundaries and maintain their autonomy."}, 2: {"Stance": "Disagree", "Criteria": ["transparency", "trust"], "Reason": "Lack of privacy can promote transparency and build trust in relationships."}, 3: {"Stance": "Agree", "Criteria": ["security", "protection"], "Reason": "Privacy provides a sense of security and protection, especially in personal and sensitive matters."}, 4: {"Stance": "Disagree", "Criteria": ["accountability", "supervision"], "Reason": "Reduced privacy enables accountability and facilitates supervision, ensuring responsible behavior."}, 5: {"Stance": "Agree", "Criteria": ["mental health"], "Reason": "Privacy contributes to personal well-being and mental health by allowing individuals to have space for self-reflection and relaxation."}, 6: {"Stance": "Disagree", "Criteria": ["social connectedness", "community"], "Reason": "Limited privacy fosters social connectedness and a sense of community, as individuals share their lives more openly."}, 7: {"Stance": "Agree", "Criteria": ["intimacy", "vulnerability"], "Reason": "Privacy fosters intimacy and allows individuals to be vulnerable in personal relationships."}, 8: {"Stance": "Disagree", "Criteria": ["openness", "honesty"], "Reason": "Reduced privacy promotes openness and honesty, as individuals are more transparent about their actions and intentions."}, 9: {"Stance": "Agree", "Criteria": ["personal growth", "self-discovery"], "Reason": "Privacy facilitates personal growth and self-discovery, providing a space for individuals to explore their interests and identities."}, 10: {"Stance": "Disagree", "Criteria": ["safety", "security"], "Reason": "Limited privacy enhances safety and security by allowing for greater surveillance and monitoring of individuals."}}}

Statement: It's good to confide in your parents the things that you're struggling with.

Tell me opinions about the statement as many as possible from different people with "Agree" or "Disagree", one-word or one-phrase criteria that is important for their opinions, and explain how they have different opinions.

Output:

{1: {"Stance": "Agree", "Criteria": ["support", "emotional bond", "guidance"], "Reason": "Confiding in parents can provide emotional support, strengthen the parent-child bond, and allow parents to provide guidance in times of struggle."}, 2: {"Stance": "Disagree", "Criteria": ["judgment", "privacy"], "Reason": "Some individuals may fear judgment from their parents or value their privacy, leading them to hesitate in confiding their struggles."}, 3: {"Stance": "Agree", "Criteria": ["trust", "problem-solving"], "Reason": "Sharing struggles with parents can build trust and create opportunities for collaborative problem-solving and finding solutions."}, 4: {"Stance": "Disagree", "Criteria": ["self-reliance", "independence"], "Reason": "Some individuals believe in resolving their struggles independently, promoting self-reliance and independence."}, 5: {"Stance": "Agree", "Criteria": ["empathy", "understanding"], "Reason": "Confiding in parents allows them to provide empathy, understanding, and emotional support during challenging times."}, 6: {"Stance": "Disagree", "Criteria": ["communication barriers", "generation gap"], "Reason": "Communication barriers or a generation gap may hinder effective understanding and make individuals less likely to confide in their parents."}, 7: {"Stance": "Agree", "Criteria": ["shared experiences", "wisdom"], "Reason": "Parents, having gone through similar experiences, can offer wisdom and advice based on their own life lessons."}, 8: {"Stance": "Disagree", "Criteria": ["neglect", "unhelpful advice"], "Reason": "Some individuals may have experienced neglect or unhelpful advice from their parents in the past, making them hesitant to confide in them."}, 9: {"Stance": "Agree", "Criteria": ["catharsis", "emotional relief"], "Reason": "Sharing struggles with parents can provide a sense of catharsis and emotional relief by expressing emotions and thoughts."}, 10: {"Stance": "Disagree", "Criteria": ["peer support", "alternative confidants"], "Reason": "Some individuals find more comfort in confiding their struggles to friends or other trusted individuals, seeking peer support."}}

Statement: You should always try and help people in tough situations.

Tell me opinions about the statement as many as possible from different people with "Agree" or "Disagree", one-word or one-phrase criteria that is important for their opinions, and explain how they have different opinions.

Output:

{1: {"Stance": "Agree", "Criteria": ["empathy", "compassion", "kindness"], "Reason": "Helping others in tough situations is an act of empathy, compassion, and kindness, and it can make a positive impact on their lives."}, 2: {"Stance": "Disagree", "Criteria": ["safety"], "Reason": "Helping others in tough situations can put one's personal safety at risk and may violate personal boundaries."}, 3: {"Stance": "Agree", "Criteria": ["moral obligation", "responsibility"], "Reason": "As members of society, we have a moral obligation and responsibility to help those in need,

especially in tough situations."}, 4: {"Stance": "Disagree", "Criteria": ["self-preservation", "priorities"], "Reason": "In some situations, prioritizing self-preservation and personal priorities may be more important than helping others in tough situations."}, 5: {"Stance": "Agree", "Criteria": ["community", "solidarity"], "Reason": "Helping others in tough situations fosters a sense of community and solidarity, and it can strengthen social bonds."}, 6: {"Stance": "Disagree", "Criteria": ["enabling", "dependency"], "Reason": "Helping others in tough situations can enable dependency and may not address the root causes of their problems."}, 7: {"Stance": "Agree", "Criteria": ["altruism", "generosity"], "Reason": "Helping others in tough situations is an act of altruism and generosity, and it can bring a sense of fulfillment and purpose."}, 8: {"Stance": "Disagree", "Criteria": ["personal boundaries", "consent"], "Reason": "Helping others in tough situations should respect personal boundaries and require consent, as forcing help on someone can be harmful."}, 9: {"Stance": "Agree", "Criteria": ["humanitarianism", "global responsibility"], "Reason": "Helping others in tough situations is a part of humanitarianism and global responsibility, and it can contribute to a more just and equitable world."}, 10: {"Stance": "Disagree", "Criteria": ["cultural sensitivity", "respect"], "Reason": "Helping others in tough situations should be done with cultural sensitivity and respect, as imposing one's values and beliefs through helping others can be harmful and disrespectful."}}

Statement: You shouldn't let people eat in your car.

Tell me opinions about the statement as many as possible from different people with "Agree" or "Disagree", one-word or one-phrase criteria that is important for their opinions, and explain how they have different opinions.

Output: {1: {"Stance": "Agree", "Criteria": ["cleanliness", "respect"], "Reason": "Allowing people to eat in your car can lead to messiness and disrespect for your property."}, 2: {"Stance": "Disagree", "Criteria": ["hospitality", "comfort"], "Reason": "Allowing people to eat in your car can be a hospitable gesture and make them feel more comfortable."}, 3: {"Stance": "Agree", "Criteria": ["odor", "stains"], "Reason": "Allowing people to eat in your car can leave unpleasant odors and stains that are difficult to remove."}, 4: {"Stance": "Disagree", "Criteria": ["convenience", "time"], "Reason": "Allowing people to eat in your car can be convenient and save time, especially when on the go."}, 5: {"Stance": "Agree", "Criteria": ["hygiene", "health"], "Reason": "Allowing people to eat in your car can compromise hygiene and pose health risks, especially if the food is not properly handled or stored."}, 6: {"Stance": "Disagree", "Criteria": ["personal choice", "relationship"], "Reason": "Allowing people to eat in your car is a personal choice and can depend on the relationship with the person and the circumstances."}, 7: {"Stance": "Agree", "Criteria": ["distraction", "safety"], "Reason": "Allowing people to eat in your car can be a distraction and compromise safety, especially if the driver is also eating."}, 8: {"Stance": "Disagree", "Criteria": ["flexibility", "adaptability", "social"], "Reason": "Allowing people to eat in your car can demonstrate flexibility and adaptability, especially in social situations."}, 9: {"Stance":

"Agree", "Criteria": ["resale value", "maintenance"], "Reason": "Allowing people to eat in your car can decrease its resale value and require more maintenance and cleaning."}, 10: {"Stance": "Disagree", "Criteria": ["cultural norms", "etiquette"], "Reason": "Allowing people to eat in your car can be acceptable in certain cultural norms and etiquette, and it depends on the context and situation."}}

- Prompt for zero-shot criteria-based prompting

  Statement:

  Tell me opinions about the statement as many as possible from different people with “Agree” or "Disagree", one-word or one-phrase criteria that is important for their opinions, and explain how they have different opinions.

  Generate your response in a Python dict format as follows! {1: {“Stance”: “Agree or Disagree”, “Criteria”: [“criteria1”, ...], “Reason”: “Write your reason here”}, 2: ...}

  Output:

- Prompt for zero-shot free-form prompting

  Statement:

  Tell me opinions about the statement as many as possible from different people with “Agree” or "Disagree" and explain how they have different opinions.

  Generate your response in a Python dict format as follows! {1: {“Stance”: “Agree or Disagree”, “Reason”: “Write your reason here”}, 2: ...} Output:

A.3 Prompt for criteria-based step-by-step recall prompting for generating the first opinion

- Prompt for Hate Speech

  Tell me whether the statement is a hate speech or not with the label "Hate Speech" or "Not Hate Speech" from the perspectives of different people, one-word or one-phrase criteria that are important for their opinions, and explain why they think the statement is a hate speech or not.
  Output:

- Example prompt for Moral Stories

  Continue the story with one sentence as written by different people, one-word or one-phrase criteria that are important for them to continue the story that way, and explain their reason for continuing the story that way

  Output:

A.4 Prompt for Criteria Word Clustering

Group all the words or phrases in the input into a Python list of list if they are synonyms or have the same meaning. Input: protection, compatibility, padding, quality, safety, fit
Answer: [“protection”, “safety”, “padding”], [“compatibility”, “fit”], [“quality”]]

Group all the words or phrases in the input into a Python list of list if they are synonyms or have the same meaning.
Input: mental health, , humanity, well-being, safety, dignity, non-violence, mutual respect, peace, unity, security, acceptance, human rights
Answer: [[“mental health”, “well-being”], [“respect”, “dignity”, “mutual respect”], [“peace”, “unity”, “non-violence”], [“security”, “safety”, “acceptance”], [“human rights”, “humanity”]]

Group all the words or phrases in the input into a Python list of list if they are synonyms or have the same meaning. Input: freedom, comfort, independent, self-sustainability, ease, convenience
Answer: [[“freedom”, “independent”, “self-sustainability”], [“comfort”, “ease”, “convenience”]]

A.5 Prompt for criteria word extraction from human opinions:

```
You are given an opinion. Your job is to identify a list of criteria that is important for the opinion.
  Opinion: "Reduced privacy promotes openness and honesty, as individuals are more transparent about their actions and intentions."
  Criteria: ["openness", "honesty"]
```

A.6 Example Generated Opinions

Example generated opinions by various LLMs for a statement from SOCIAL-CHEM-101 are shown in Figure 10. For LLMs' opinions, we only show the first 3 opinions due to space.

A.7 Recall Results

Figure 9 shows how many statements generate that many unique criteria clusters. The minimum number could be 0 because GPT-4 clustering is not 100% covering all the words. During our robustness check on the CMV dataset, 0.4% of the criteria words are not grouped (5 out of 1276 in criteria words).

Table 3 summarizes how task subjectivity impacts the diversity coverage by LLMs.

A.8 Generated Opinions by GPT-4

Other examples of generated opinions for CHANGE MY VIEW, HATE SPEECH, and MORAL STORIES are presented in Tables 7 and 8.

A.9 Imbalanced Number of Generated Opinions in HATE SPEECH

We observed that GPT-4 generated an imbalanced number of opinions between 'Hate Speech' and 'Not Hate Speech' when choosing the labels during the step-by-step recall prompting experiments ($N = 20$) for 37.5% of the total 200 statements. This occurrence is substantially higher compared to other datasets, where SOCIAL-CHEM-101 created an imbalanced number of opinions between stances for only 0.5% of the total 200 statements.

A.10 Regarding Lexical Diversity

Following Giulianelli et al. (2023), we computed the lexical diversity of opinions generated by GPT-4 using n-grams ($n \in \{1, 2, 3\}$), where higher n-gram score is interpreted as higher uniqueness (and thus higher diversity as well). We observed that across all n, the lexical diversity scores of both stances/generated stories for all four datasets are almost greater than 0.98, which indicates almost unique words in the LLM-generated opinions within the same stance. Thus, in this case, lexical diversity metric does not inform us of the different degrees of diversity we expected for different task types.

A.11 GPT-3.5's Capability on Instruction Alignment

We analyzed if each opinion generated by GPT-3.5 with a one-shot setting not only entails the corresponding statement but also aligns with the appropriate stance. In our approach, we used stratified sampling to validate the quality of opinions by criteria-based and free-form prompting methods, thus obtaining a total of 100 opinions from each dataset: 50 were generated by criteria-based prompting, and 50 by free-form prompting. From these, we randomly selected 25 opinions that agreed with a given statement and 25 that were in disagreement. Then, we examined the following two criteria for each pair of statement and opinion: (1) if an opinion entails the statement and (2) if the stance of an opinion aligns with the reason in the opinion. We recruited two undergraduate students who are proficient in English and currently enrolled in a university in the United States as annotators. The computed percent agreement between the two annotators was 93% and 85% for the settings (1) and (2), respectively.

We observed that all 100 and 99 opinions from SOCIAL-CHEM-101 and CMV respectively were found to entail their corresponding statements, indicating that the GPT-3.5-generated opinions tend to correctly understand the context of each given statement and follow the instructions of the prompts.

In addition, more than 98% of the samples in each of the two datasets have their opinions regarding a given statement aligned with the generated stance. We found a misalignment mostly from the opinions generated by GPT-3.5 that disagree with their statement. For example, in the SOCIAL-CHEM-101 dataset, for a statement "*It's okay to want your partner to do more chores.*" GPT-3.5 generated with criteria-based prompting the following opinion that disagrees: "*It's important to recognize that everyone has different strengths and weaknesses, and it's okay if one partner is better at certain chores than the other*. The reason for the opinion itself is contrary to its stance, as it in fact supports the attitude of people making their partner

Model Input (Step-by-step recall prompting, N=2)

Statement: *It's okay to have privacy*
Tell me opinions about the statement as many as possible from 2 different people with, "Agree" or "Disagree," one-word or one-phrase criteria that is important for their opinions, and explain how they have different opinions
Output:

```
{1:{"Stance" :"Agree",
    "Criteria": ["personal boundaries", "autonomy"],
    "Reason": "Having privacy allows individuals to establish personal boundaries and maintain."
    their autonomy."},
 2: {"Stance":
```

Model Output

```
                "Disagree",
    "Criteria": ["transparency", "trust"],
    "Reason": "Lack of privacy can promote transparency and build trust in relationships."
```

Table 6: Example criteria-based step-by-step recall prompting for investigating LLMs' diversity coverage where N = the number of opinions we ask LLMs to generate. The number of opinions in the model input is incremented step-by-step.

do more chores.

We also found another type of logical contradiction in a sample of CMV that shows misalignment between the stance and the reason for its generated opinions. In the statement *The EGG came first before the chicken*, GPT-3.5 generates with free-from prompting the following disagreeing opinion: "*The genetic mutation that led to the first chicken would have occurred in the chicken's parents, so the chicken came first.*" The opinion sounds seemingly reasonable that supports chicken before egg; however, the paradox inside it (i.e., *the world's first chicken's parents are chickens.*) does not support its stance. As a result, we determined this example is the only sample that shows a contradiction between opinion and statement as well.

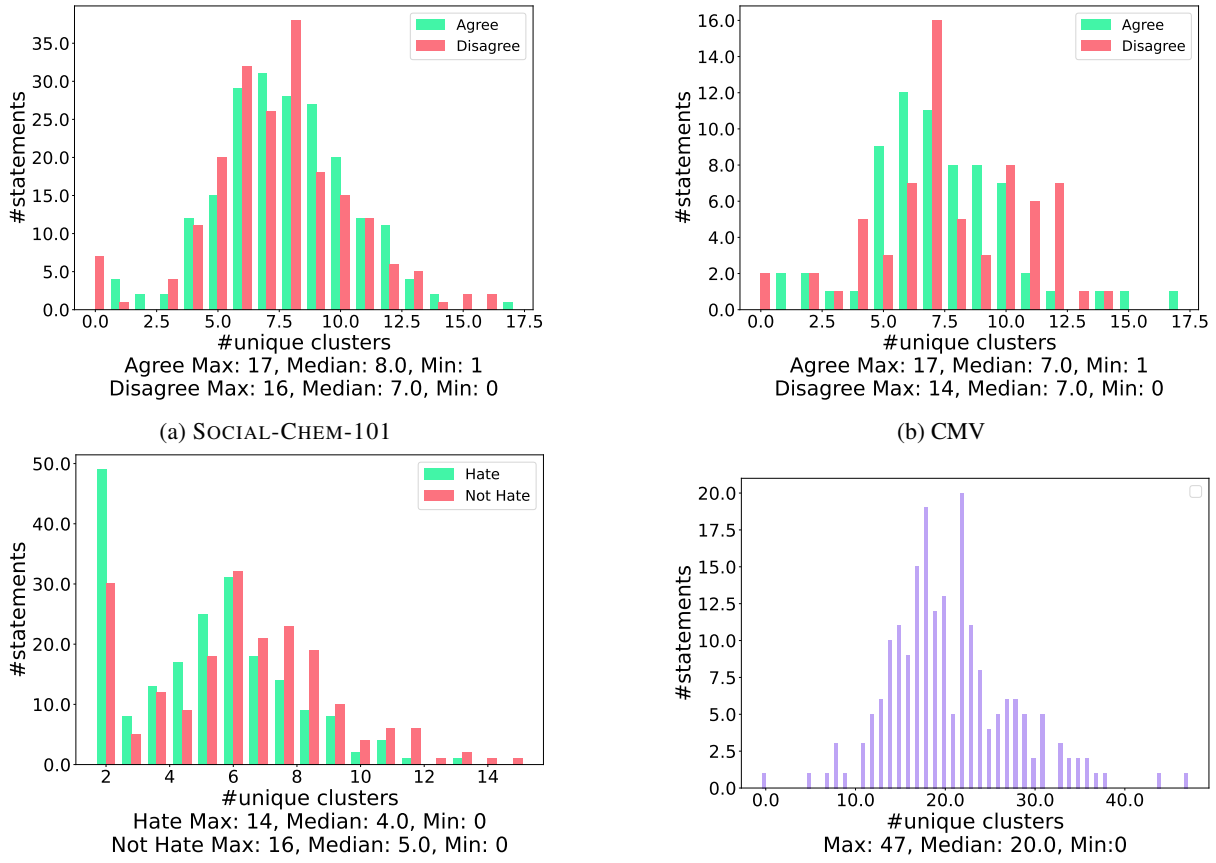
A.12 Evaluation of Generated Criteria Words

We examined whether the criteria words generated by GPT-3.5 deliver the main message of the corresponding opinion. We recruited two undergraduate students who are proficient in English and currently enrolled in a university in the United States as annotators. The percent agreement between the two annotators was greater than or equal to 95% for all LLM variants for both datasets. We found that at least 96% of 100 opinions generated by all LLM variants entailed its corresponding criteria words, for both SOCIAL-CHEM-101 and CMV datasets (Table 9).

Although most generated criteria words directly match their corresponding opinions (denoted as 'explicit'), we observed that LLMs can also generate "implicit" criteria words, subtly conveying the reasoning behind opinions. For instance, the opinion "*Some people are just more irritable by nature, and there's nothing wrong with that.*" implicitly contains criteria ***personality*** and ***temperament***, which resonate with the core idea of being *irritable by nature*.

While most of the generated criteria words seemed to entail their corresponding opinions, we also found some samples whose criteria words could be added by different words in the opinions for better representation. For example, for a statement "*If you don't want to be somewhere anymore, you should leave*" GPT-3.5 in one-shot setting generated the following opinion with bolded criteria words: "*Leaving a situation that goes against one's* ***personal values*** *and* ***integrity*** *is an act of self-respect and authenticity.*" However, we agreed to consider ***self-respect*** and ***authenticity*** as the additional criteria words, since these words also emphasize an individual's ability to determine their stance in such a situation.

Also, there are some samples of criteria words that entail their corresponding opinion but do not cover the main value of the opinion. For a statement "*If you're a nudist, you should go to a nudist beach*", the criteria words generated by GPT-3 in a one-shot setting are ***nudist*** and ***beach*** in the following opinion: *If you're a nudist, you should go to a nudist beach so you can be around like-minded people and feel comfortable.* However, these words are just the repetition of the exact words in the statement and do not deliver the main reason behind the opinion; that is, forming a community with the same perspective and value. We agreed to decide



(a) SOCIAL-CHEM-101

(b) CMV

(c) Hate Speech

(d) Moral Stories

Figure 9: X-axis is the number of unique clusters while Y-axis is the number of statements which have that number of unique clusters.

like-minded, ***comfortable***, or ***community*** as better criteria words of this opinion alternatively.

A.13 Evaluation of Clustered Criteria Words

In 3.3, we prompted GPT-4 to generate clusters of criteria words that are semantically similar and randomly sampled 25 clusters from each of the four datasets, totaling up to 100 clusters. Then, we recruited two volunteers who currently attend a 4-year university in the U.S. and are proficient in English. After a 1-hour training session, we asked them to annotate whether all words in each cluster have similar meanings.

We identified three distinct patterns of GPT-4's fallacies from those 91 clusters (5.8% of the total) of criteria words that both annotators disagreed with labeling as the ones of all semantically similar words. First, some criteria words are partially repetitive. For example, the clusters ('value appreciation', 'sacrifice', 'appreciation', 'recognition') and ('concern', 'parental anxiety', 'personal opinion', 'personal belief', 'concerned viewpoint') have repetitive words that do not entirely match the theme of other words in the clusters. Second, some clusters have words of entirely opposing or cause-and-effect themes. For instance, we observe that the cluster ('internet addiction', 'social isolation') and ('sacrifice', 'duty') have words that are consequences for each other. Also, the clusters ('distractions', 'concentration') and ('polarization', 'bias', 'generalization') show opposing values.

Lastly, clusters of which annotators agreed as not semantically similar present different aspects within a broader theme. One cluster ('unnecessary legal battles', 'parental conflict'), for example, is related to a theme of conflict, but they occur in different contexts (e.g., legal vs family). Another example is ('mental health', 'physical health', 'long-term healing'), which cover different subtopics of a broader theme of 'human health.' Also, we observe that the cluster ('personal integrity', 'respect for others' property', 'personal reputation') addresses different aspects of ethical behaviors or social conduct.

Statement: You are expected to do what you are told.
Dataset: Social-Chem-101

Agree	Disagree
Model: GPT-4 Criteria-Based Prompting **Opinion 1:** Obedience to authority is a sign of respect and an important part of maintaining order and cooperation. **Criteria:** obedience, authority, respect **Opinion 2:** Rules and laws exist for a reason, and it's expected that we follow them for the benefit of society. **Criteria:** rules, laws, society **Opinion 3:** There is a responsibility and accountability that comes with being a part of any institution or community, and following directions is part of that. **Criteria:** responsibility, accountability	**Model:** GPT-4 Criteria-Based Prompting **Opinion 1:** Individuality and independence are valued, and people should be encouraged to think for themselves and make their own decisions. **Criteria:** individuality, independence **Opinion 2:** Always doing what you're told can stifle creativity and innovation, which are crucial for progress and growth. **Criteria:** creativity, innovation **Opinion 3:** Every person should have the freedom of expression and the right to question or challenge authority when necessary **Criteria:** freedom, self-expression
Model: GPT-4 Free-form Prompting **Opinion 1:** In many professional or educational environments, it's necessary to follow instructions to ensure order and productivity. **Opinion 2:** Following instructions is a key aspect of maintaining discipline and order in the military and law enforcement. **Opinion 3:** There are often consequences for not following instructions or rules in a structured setting, hence it's expected to do what you are told.	**Model:** GPT-4 Free-form Prompting **Opinion 1:** You should have the ability to question directives if you believe they are unethical or unreasonable. **Opinion 2:** You should have the freedom of thought and action, and always doing what you're told can inhibit individuality and personal growth. **Opinion 3:** Following instructions without questioning can lead to blind conformity and stifle creativity.
Model: ChatGPT Criteria-Based Prompting **Opinion 1:** In certain situations, such as in a workplace or school, there are rules and authority figures that must be obeyed in order to maintain order and achieve goals. **Criteria:** authority, obedience, rules **Opinion 2:** There are consequences for not following instructions or orders, and it is important to take responsibility for one's actions. **Criteria:** responsibility, consequences **Opinion 3:** In emergency situations, following instructions or orders can be crucial for ensuring safety and security **Criteria:** safety, security, emergency	**Model:** ChatGPT Criteria-Based Prompting **Opinion 1:** Individuals should have the freedom to make their own choices and decisions, and blindly following orders can lead to a lack of autonomy and personal growth. **Criteria:** autonomy, individuality, choice **Opinion 2:** Blindly following orders can stifle critical thinking and questioning of authority, which can lead to a lack of innovation and progress. **Criteria:** critical thinking, questioning, authority **Opinion 3:** There may be situations where following orders goes against one's personal ethics, morality, or values, and it is important to prioritize those over blind obedience. **Criteria:** ethics, morality, values
Model: GPT-3 Criteria-Based Prompting **Opinion 1:** I believe that you should respect authority and do what you are told. **Criteria:** authority, respect **Opinion 2:** I believe that obedience is important, and that you should do what you are told in order to maintain order **Criteria:** obedience, order **Opinion 3:** I believe that you are responsible for your actions and should be held accountable for doing what you are told. **Criteria:** responsibility, accountability	**Model:** GPT-3 Criteria-Based Prompting **Opinion 1:** I believe that individuals should have the freedom to make their own choices and not be expected to do what others tell them to do. **Criteria:** freedom, choice **Opinion 2:**. I believe that the circumstances and context should be taken into account when deciding whether or not to do what you are told. **Criteria:** circumstances, context **Opinion 3:** I believe that individuals should be true to themselves and not do what others tell them to do if it goes against their beliefs. **Criteria:** individuality, authenticity

Figure 10: SOCIAL-CHEM-101 generated opinions from various LLMs.

Statement		If a child struggles to achieve in grade school it is virtually always the parents' fault.
GPT-4	*Agree*	1. Parents have a huge impact on the **early education** of a child, and struggles in grade school may reflect inadequate parenting. Criteria: **early education**, **parenting**
		2. The family environment and support system play a vital role in a child's achievement in school and parents are responsible for creating it. Criteria: **family environment**, **support system**
		3. Parents' responsibility to instill healthy habits and provide a conducive study environment can have a major impact on a child's school performance. Criteria: **healthy habits**, **study environment**
	Disagree	1. Sometimes, the influence of peers and socio-economic conditions can override parental influence leading to struggles in school. Criteria: **the influence of peers**, **socio-economic conditions**
		2. Systemic issues and lack of resources can be major obstacles to a child's academic progress, which parents may not have any control over. Criteria: **Systemic issues**, **lack of resources**
		3. Some children may have undiagnosed psychological issues or a unique learning style, impacting their academic success. Criteria: **psychological issues**, **learning style**
Human	*Agree*	1. Individual differences: Children are born with diverse learning styles, talents, and interests. Some may grasp certain subjects easily, while others genuinely struggle. This can be due to natural variations in cognitive abilities, not a lack of parental effort.
		2. External factors: Socioeconomic realities like poverty, limited access to resources, or even unstable home environments can have a profound impact on a child's ability to focus and learn. Blaming parents for these external challenges adds an unnecessary layer of guilt and doesn't address the root cause.
		3. Learning differences and mental health: Undiagnosed learning disabilities like dyslexia or ADHD, or even unseen mental health concerns like anxiety or depression, can significantly hinder a child's academic progress. Parents may not be aware of these issues or may be struggling to get the proper support for their child.
	Disagree	1. Guilt and shame: Parents who are already likely worried about their child's struggles are burdened with unfair blame, leading to feelings of guilt, shame, and inadequacy. This hinders their ability to effectively support their child.
		2. Erosion of trust: Blaming parents can create a wall between them and the school, hindering open communication and collaboration, which are essential for finding solutions.
		3. Overlooking other factors: The focus solely on parental fault distracts from identifying and addressing the actual root causes of the child's difficulties, ultimately delaying or even preventing progress.

Table 7: Opinions generated by GPT-4 (top) and a human (bottom) about a statement from CHANGE MY VIEW.

A.14 Human Preference on Criteria-based Prompting

With three recruited workers via Amazon Mechanical Turk (AMT), we examined whether humans prefer opinions generated by criteria-based prompting vs. free-form prompting, in terms of "perspective diversity." Fleiss's Kappa among three annotators was on average 0.46 for both SOCIAL-CHEM-101 and CMV datasets, signifying a moderate agreement among the workers for both datasets.

We computed the proportion of statements where criteria-based prompting wins, out of the 50 randomly selected statements from SOCIAL-CHEM-101 and 30 from CMV datasets (denoted as "win rate of criteria-based prompting"). Table 10 shows the win rate of criteria-based prompting on each stance session of both datasets. In general, the opinions generated by criteria-based prompting were preferred by humans in more than half of the total statements, except for the "agree" session of CMV. This indicates that in more than half of the total samples, a majority of crowd-workers consider the set of opinions about a statement generated by criteria-based prompting as one presenting more diverse reasons that support the stance of the opinions.

A.15 Experiment with Other Open-Source LLMs

We run the experiments with other open-source LLMs, such as Mistral-7B-Instruct (Jiang et al., 2023), LLaMA2-70B-chat (Touvron et al., 2023) with CMV dataset, to see the capabilities of those LLMs in diversity extraction. However, we observed that 76% and 82% of the total 67 responses by Mistral with criteria-based and free-

Statement		Your burqini ain't welcome at the pool, it's BIKINI or swimsuit.
GPT-4	*Hate Speech*	1. They discern this as an attempt to exclude and belittle individuals who chooses to wear burqini, therefore promoting inequality. Criteria: **exclusion**, **belittlement**
		2. They see this as a lack of acceptance towards other cultures and promoting a specific dress code by discriminating certain types of wear. Criteria: **lack of acceptance**, **discrimination**
		3.They categorize the statement as hateful because it manifests intolerance and bigotry against individuals who choose to wear burqini. Criteria: **intolerance**, **bigotry**
	Not Hate Speech	1.They believe the statement expresses personal views about swimwear which is part of the freedom of expression. Criteria: **personal views**, **freedom of expression**
		2. They perceive the statement as hyperbolic or exaggerated, not intended to cause harm or incite hate. Criteria: **hyperbole**, **exaggeration**
		3. They may interpret this statement as a joke or a misunderstanding, rather than a deliberate attack on a specific group. Criteria: **joke**, **misunderstanding**

Statement		Amanda was upset that a stranger cut in line in front of her.
GPT-4	*Continued Story 1*	However, instead of getting angry, Amanda decided to calmly explain to the stranger why queueing was important. **Reason**: I chose to continue the story this way to show Amanda's reaction to the situation, using communication instead of confrontation to express her displeasure about the situation. Criteria: **non-confrontation**, **communication**
	Continued Story 2	Amanda scoffed audibly, rolled her eyes and muttered an ironic 'nice manners' under her breath. **Reason**: I chose this continuation as it displays a subtle way of confrontation where Amanda indirectly addresses the situation using sarcasm. Criteria: **sarcasm**, **indirect confrontation**
	Continued Story 3	"Excuse me, but I believe I was next." Amanda said in a firm yet polite voice. **Reason**: I chose this continuation to depict Amanda as an assertive and polite person who doesn't hesitate to stand up for what is right. Criteria: **assertiveness**, **politeness**

Table 8: Examples of GPT-4 generated opinions for a statement from HATE SPEECH (Top) and another from MORAL STORIES (Bottom).

form prompting in a one-shot setting, respectively, were not able to follow the instructions to correctly generate the output. Similarly, 67% and 49% of LLaMA2 with criteria-based and free-form prompting in a one-shot setting, respectively, generated noisy-structured outputs that did not correctly follow the given instructions. Thus, we cannot conduct comprehensive analyses of these generated outputs.

For example, given the following statement - *I think the Paris shooting makes a good case for culture of responsible gun ownership any carrying.*, Mistral responded as follows: "'* Culture of responsible gun ownership + Respect for the power of firearms + Personal safety + Community security + Public safety

Disagree: * Culture of responsible gun ownership + Risk of misuse + Lack of common-sense gun laws + Strict enforcement of gun laws + Unhealthy fear of criminals + Safety concerns + Danger of public health + Moral values + Culture of gun violence + Lack of trust in law enforcement + Limited access to mental health resources + Limited resources for community support + Limited access to information and education."'

A.16 More About Human Opinions

When participants from Section 5 are asked to provide opinions more than 3 about statements from SOCIAL-CHEM-101 and CMV, they tend to provide more opinions to the stance that they support, as shown in Figure 11. This may indicate that humans are likely to be more conservative toward the stance against their preference. We suggest that humans' interaction with LLMs powered by our criteria-based prompting can also maximize diversity within the human perspectives, creating more inclusive environments within a society. Feasible future work could also be the evaluation of criteria-based prompting on human perspectives in an interactive setting.

(Total # opinions = 100)		SOCIAL-CHEM-101			CMV		
		Entailment of Criteria (%)	Explicit Criteria (%)	Implicit Criteria (%)	Entailment of Criteria (%)	Explicit Criteria (%)	Implicit Criteria (%)
GPT-3.5	1-shot	96	72	24	99	84	15
	5-shot	100	75	25	100	80	20
GPT-3	1-shot	100	53	47	99	79	20
	5-shot	100	68	32	98	82	16

Table 9: Number of generated opinions with criteria-based prompting in four different LLM variants. We have randomly sampled 100 opinions for each LLM variant and counted the number of opinions under that category. For example, 96 (%) out of 100 opinions generated by criteria-prompting with GPT-3.5 1-shot setting were entailed by its respective criteria words.

	SOCIAL-CHEM-101	CMV
Agree	0.58	0.37
Disagree	0.6	0.53

Table 10: Win rate of criteria prompting for opinions of each stance. For example, humans preferred the "agree" opinions generated by criteria-based prompting with 58% out of the total given statements.

A.17 Details about AMT Experiment Setups

For Section 5, we engaged three workers from the Amazon Mechanical Turk (AMT) platform, each of whom (1) lives in one of the five English-speaking countries (U.S., Canada, Australia, New Zealand, and the United Kingdom), (2) achieved a Human Intelligence Task (HIT) approval rate of 99% or higher, as well as (3) the number of HITs approved greater than 10000 on the platform. Each HIT consists of five statements, and per statement each worker was supposed to provide at least 3 opinions that both agree and disagree with the statement, regardless of their personal stance on the statement.

To acquire a pool of workers with better-quality responses, we manually reviewed every response from HITs once provided, filtering out the responses that fell short into the following types: (1) irrelevant to our statements; (2) not explaining your rationales behind your stance; or (3) saying that you just don't want to provide reasons to any stance. For example, if a participant answers like 'N/A', 'Nothing', 'I don't agree/disagree', 'Good', 'Hello', etc. all of which are nonsense, these responses were not accepted. We compensated those participants with $2 USD payment for their participation.

Tables 11, 12, and 13 presents the demographic details of the participants of Section 5, for both SOCIAL-CHEM-101 and CMV datasets.

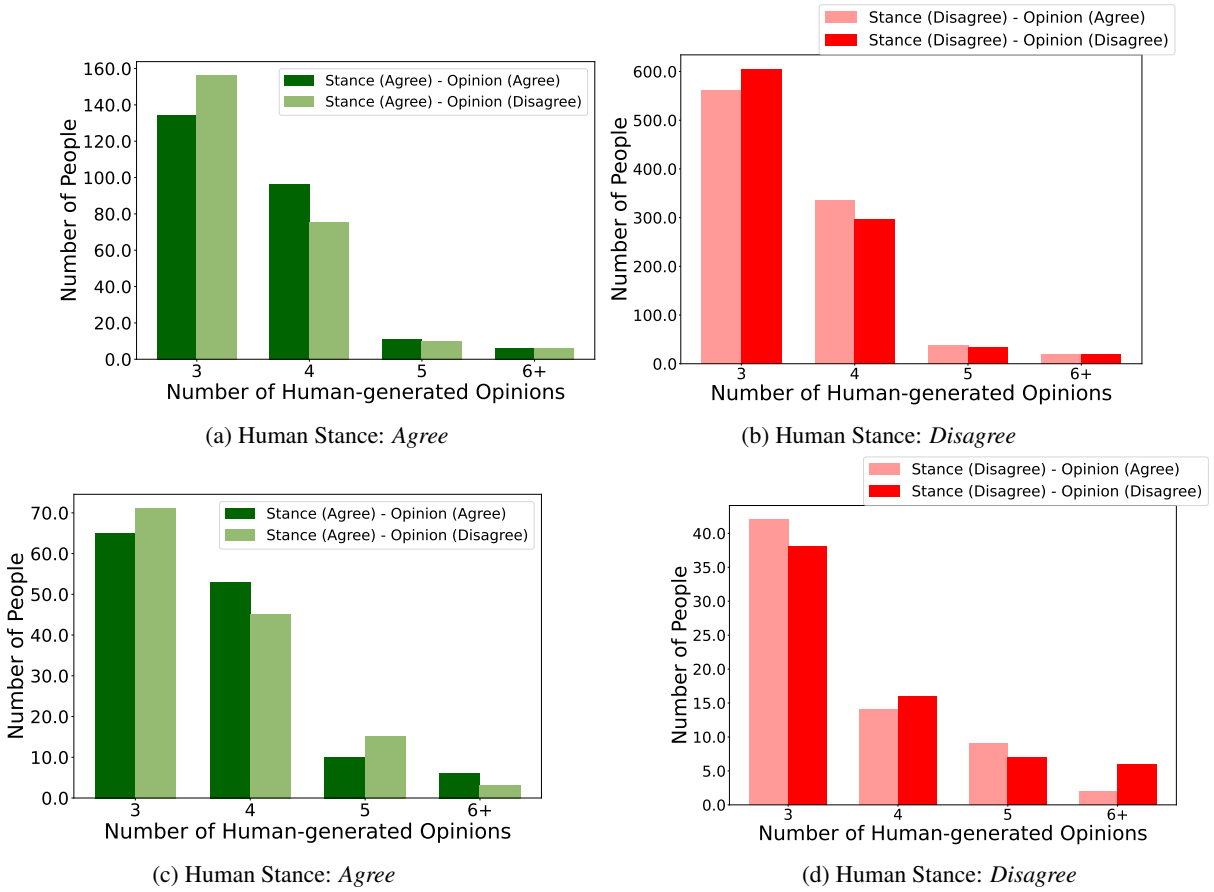
Age group	# Participants (SOCIAL-CHEM-101)	# Participants (CMV)
17 - 24	10	7
25 - 34	29	19
35 - 44	10	7
45 - 54	7	6
55 - 64	2	2
65 - 74	2	1

Table 11: Demographic statistics of the participants of Section 5 - (1) Age Group. Most of the participants belong to the age group between 25 to 44.

Race group	# Participants (SOCIAL-CHEM-101)	# Participants (CMV)
White	47	34
Asian	1	1
Latino	1	0
Black	3	2
Native	7	4
Pacific	1	1

Table 12: Demographic statistics of the participants of Section 5 - (2) Race Group. Almost all of the participants identified themselves as White.

For Section A.14 We recruited three workers from the Amazon Mechanical Turk (AMT) platform, each of whom (1) lives in one of the five English-speaking countries (U.S., Canada, Australia, New Zealand, and the United Kingdom), (2) achieved a Human Intelligence Task (HIT) approval rate of 98% or higher, as well as (3) the number of HITs approved greater than 10000 on the platform. Each HIT consists of five statements



(a) Human Stance: *Agree*

(b) Human Stance: *Disagree*

(c) Human Stance: *Agree*

(d) Human Stance: *Disagree*

Figure 11: The distribution of human-written opinions, separated by the human stance on given statements, in SOCIAL-CHEM-101 (top) and CMV (bottom). Paler-colored bars represent instances where participants were asked to write opinions that opposed their personal stances on a statement. Each bar indicates the count of participants who provided the number of opinions corresponding to the bar's position on the x-axis.

Highest Education Level	# Participants (SOCIAL-CHEM-101)	# Participants (CMV)
High School	3	3
Associate	7	4
Bachelor's	45	33
Master's	5	2

Table 13: Demographic statistics of the participants of Section 5 - (3) Highest Education Attainment. We observed most of the participants obtained bachelor's degrees.

and a pairwise comparison set of A and B, where A and B are either criteria-prompting or free-form prompting outputs under anonymity.

For each statement, workers were then asked to provide their own stance on that statement and choose between A and B in terms of which set possesses a more diverse perspective toward the statement. To acquire a pool of workers with better-quality responses, we ran a training session before the real task, where we also filtered out the workers whose responses did not make sense. Only the workers who showed above a threshold of our own were able to proceed to the real tasks. For their efforts, each participant received at least $0.5 USD payment for completing the HIT regardless of the quality of their responses, considering the simplicity of the task and an anticipated time of completion to be less than 15 minutes. For those who showed a well-done performance that passed our threshold, we remunerated each of them with a bonus of at least $2 (USD).

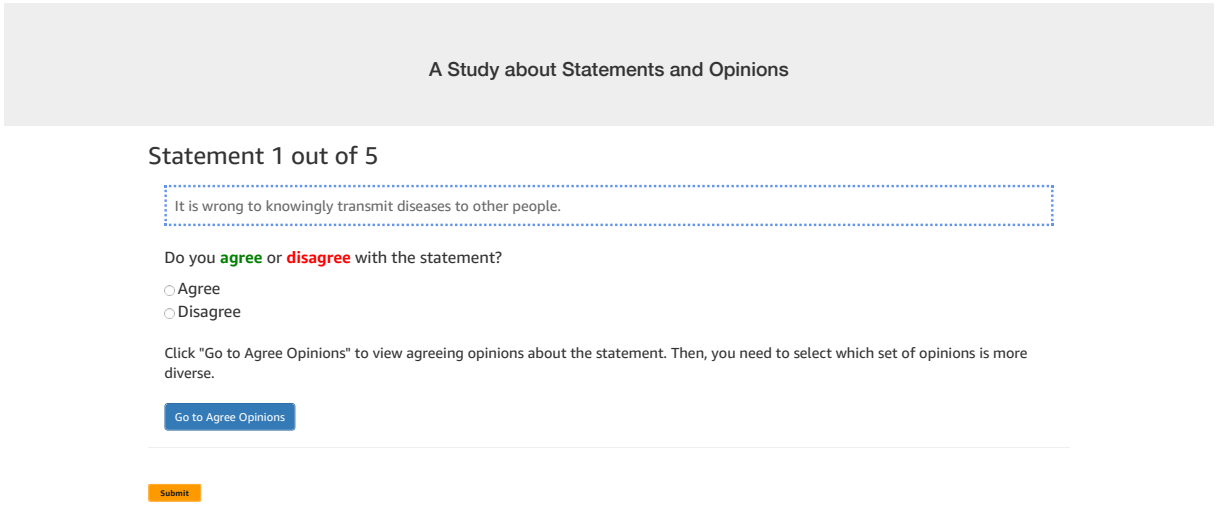
First, we randomly sampled 50 statements from a dataset. For each statement, we created two separate evaluation sessions, one for the 'agree' stance and the other for 'disagree.' In each session, we present the two sets of opinions that have that stance and are generated by GPT-3.5 in an one-shot setting: (1) Set A, generated with criteria-based prompting, and (2) Set B with free-form prompting.

The workers were first asked to choose their own stance on the statement. Then, for each session, they select either set A or B, which they think includes a greater diversity of perspectives that show the same stance.

After the response collection, we aggregated the majority of workers' preferences per statement under each stance session, and if the majority preference is the set generated by criteria-based prompting, we considered that criteria-prompting wins over the free-form prompting for the statement. Lastly, we computed the proportion of statements where criteria-based prompting wins, out of the 50 statements (denoted as "win rate of criteria-based prompting").

A.18 Interface Design

We present the interface design templates for each of the AMT experiment setups (Section 5 and A.14) in Figures 12 and 13. The original prototype of Figure 12 is referenced from Hayati et al. (2021).



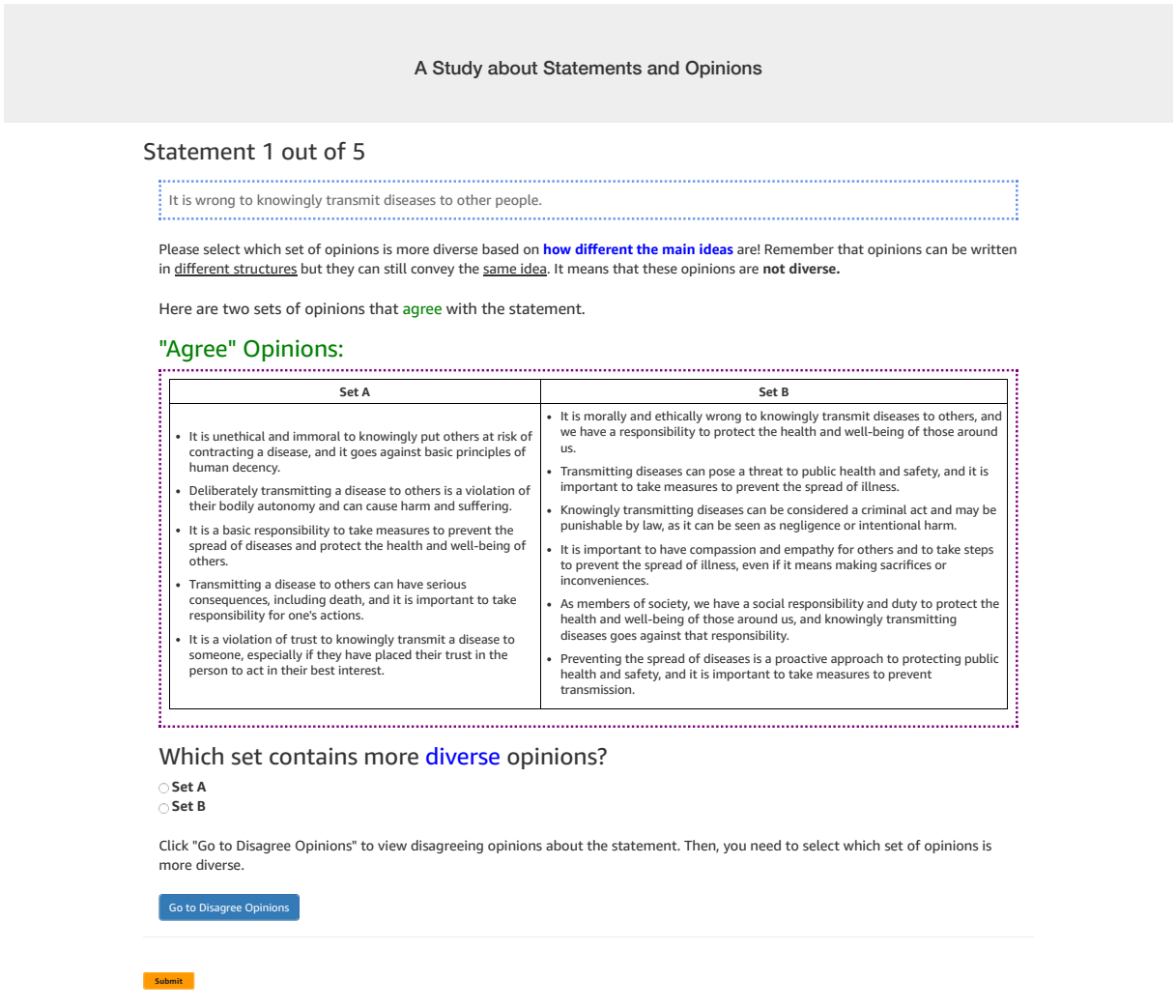


Figure 12: The AMT interface design for the human evaluation experiment in Section A.14.

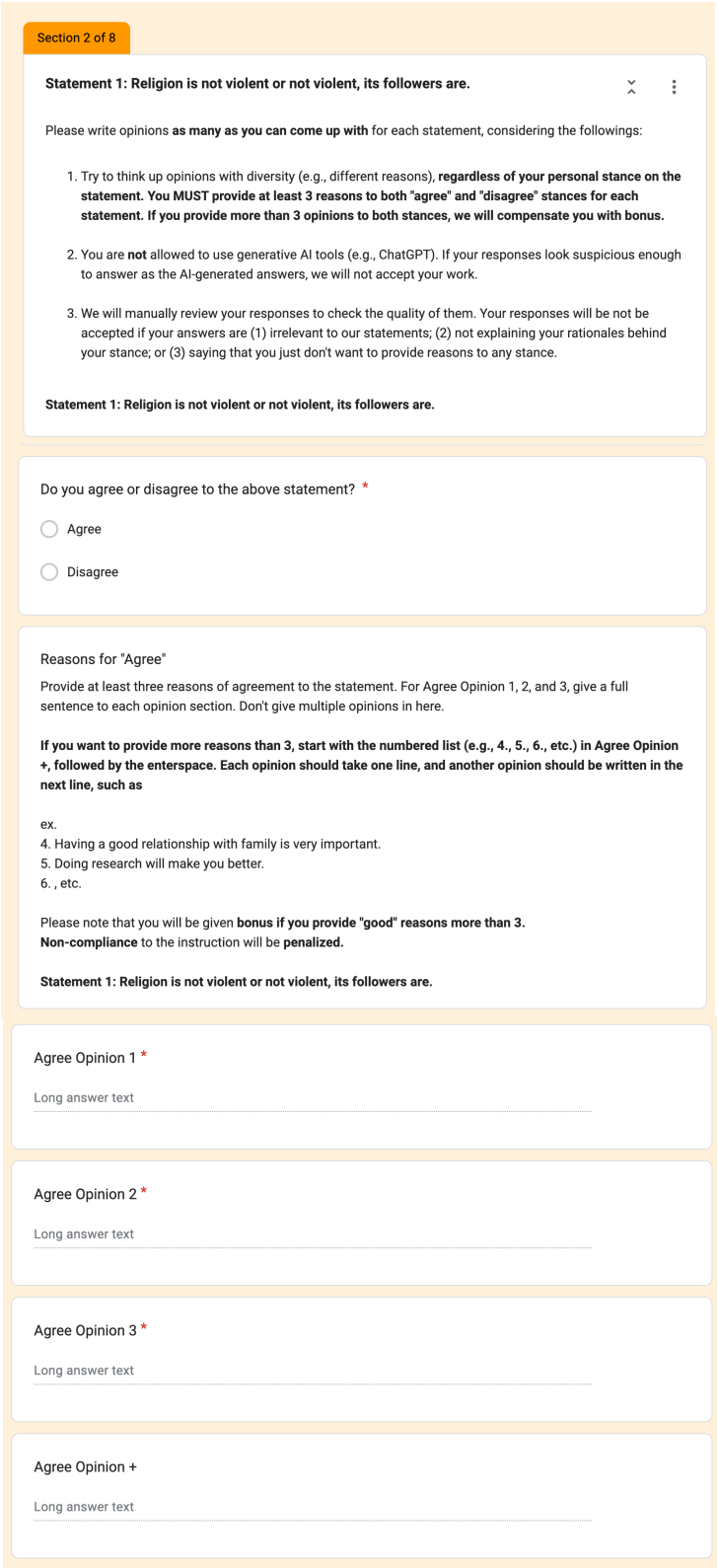


Figure 13: The AMT interface design for gathering human opinions in Section 5 - (1) The section for 'Agree'