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

Shirley Anugrah Hayati (hayat023@umn.edu)

Minhwa Lee Dheeraj Rajagopal Dongyeop Kang

University of Minnesota

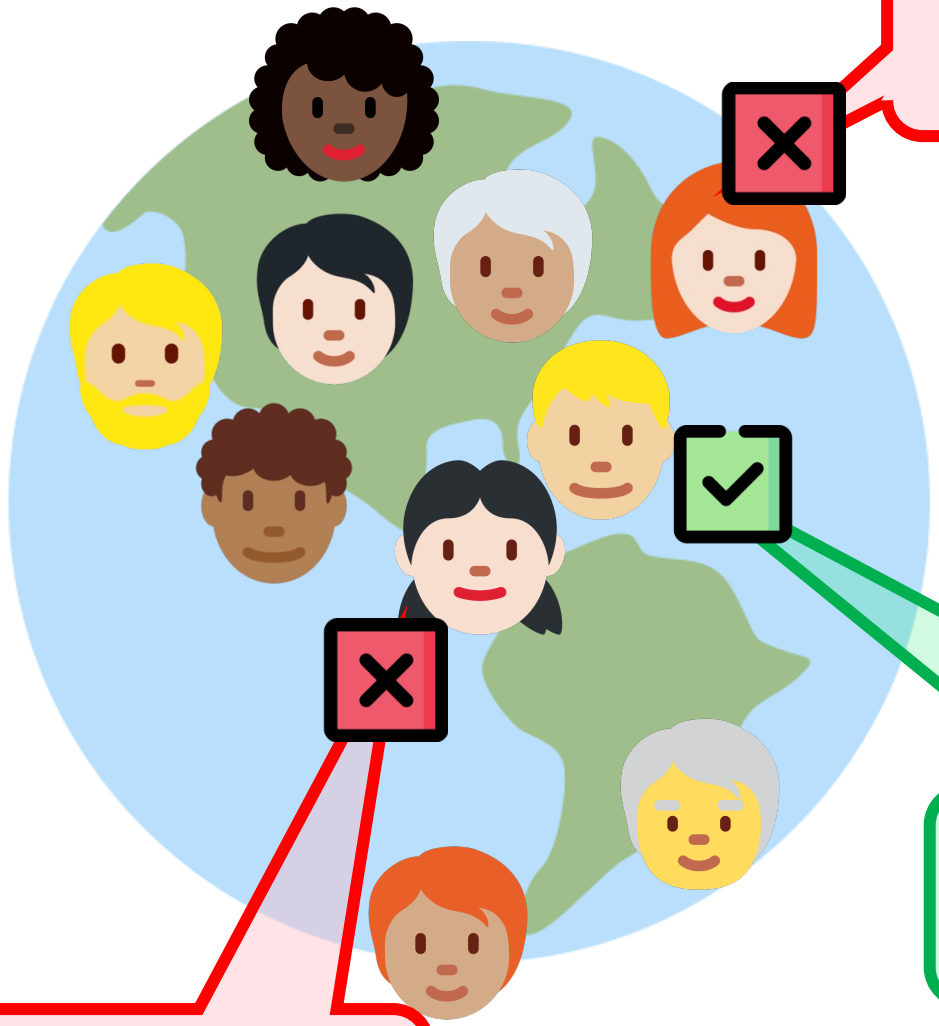
EMNLP 2024

Google DeepMind

In this talk

1. Why Diversity Extraction?
2. Criteria-Based Prompting
3. Experiment & Results

Why Diversity Extraction?



I value creativity

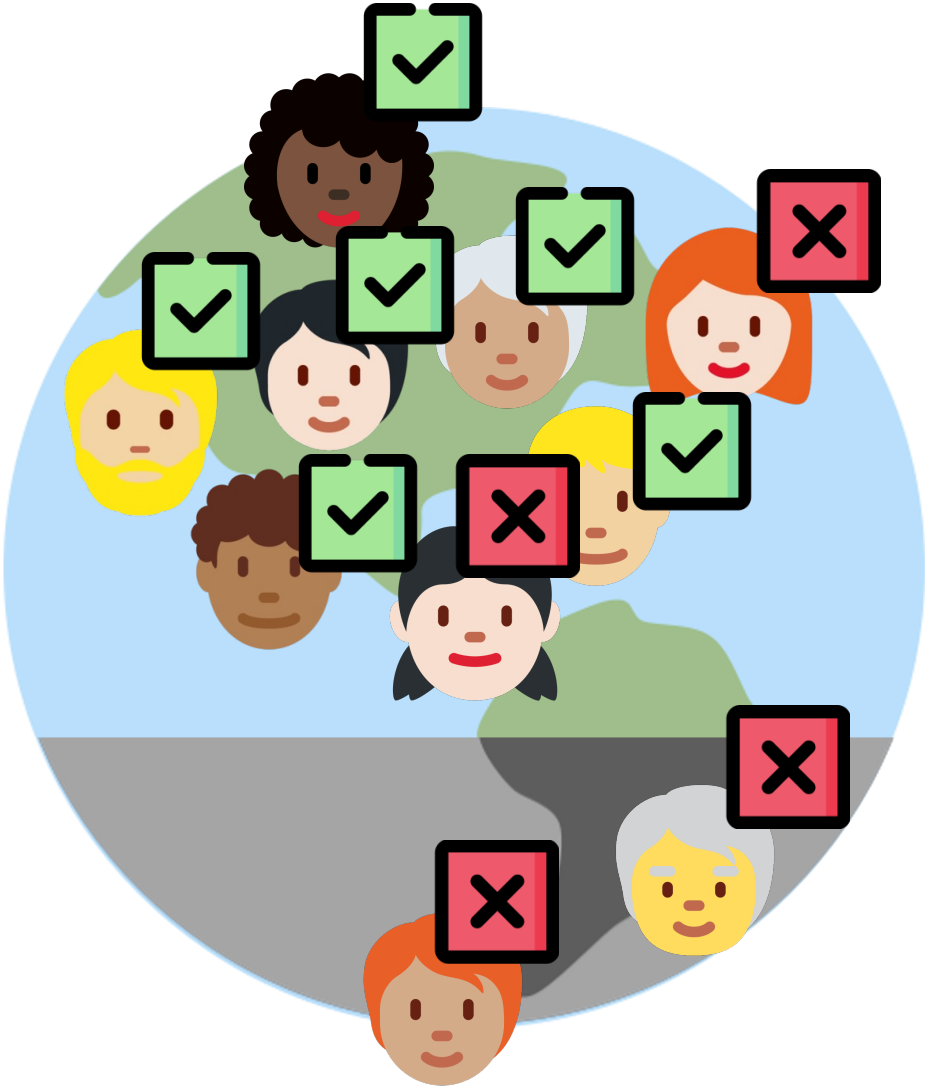
STATEMENT

You are expected to do what you are told.

Agree or Disagree?

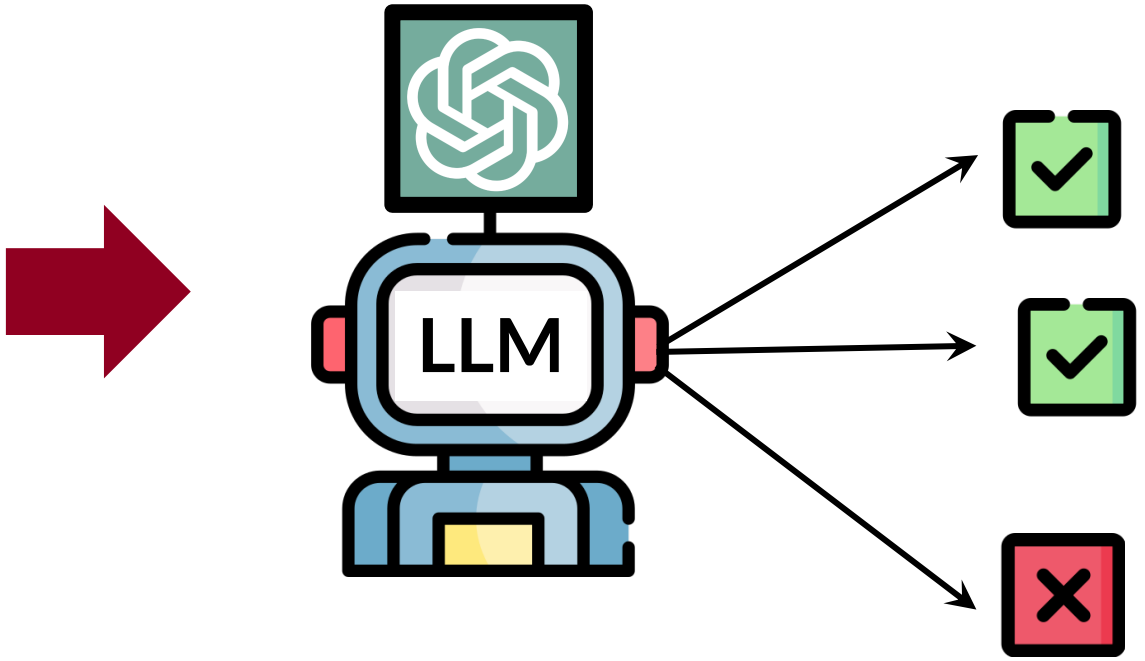
Safety is very important

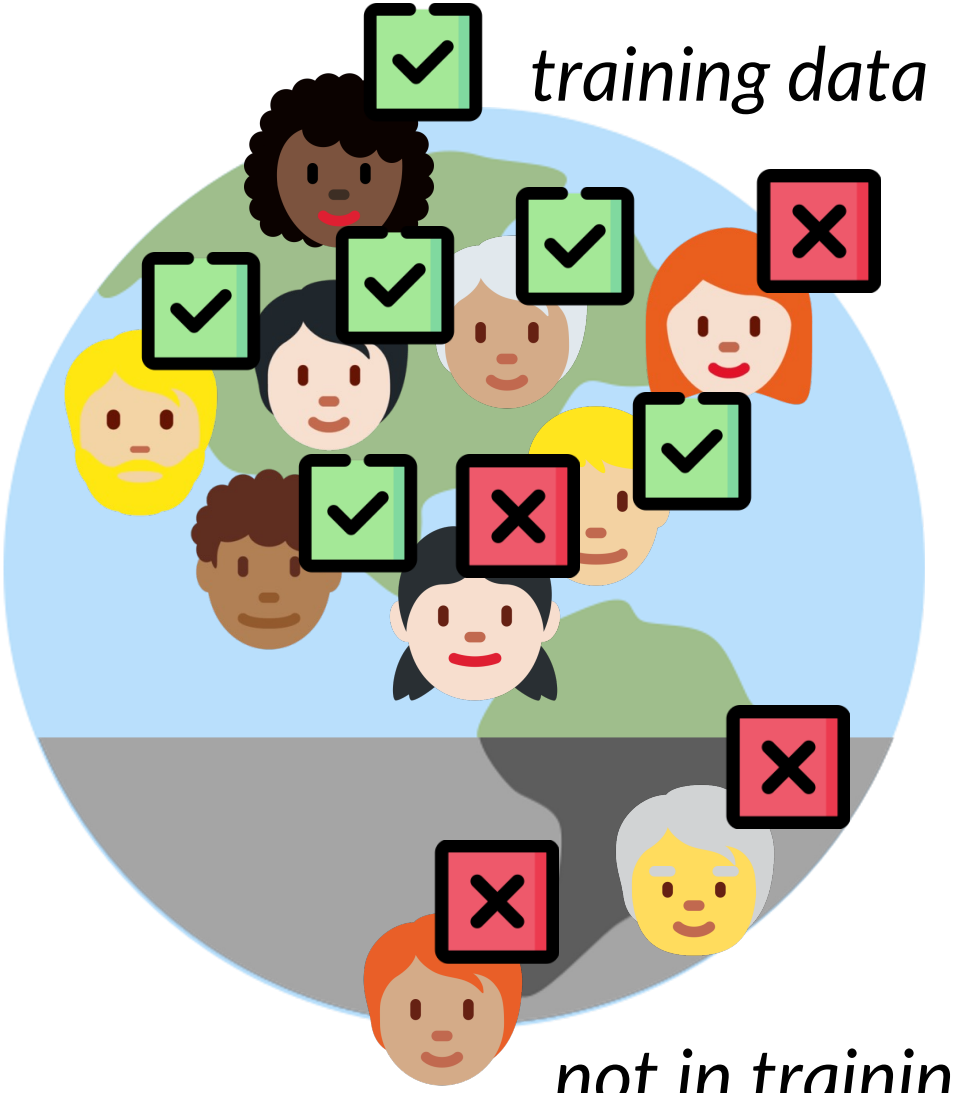
Justice!



STATEMENT

You are expected to do what you are told.

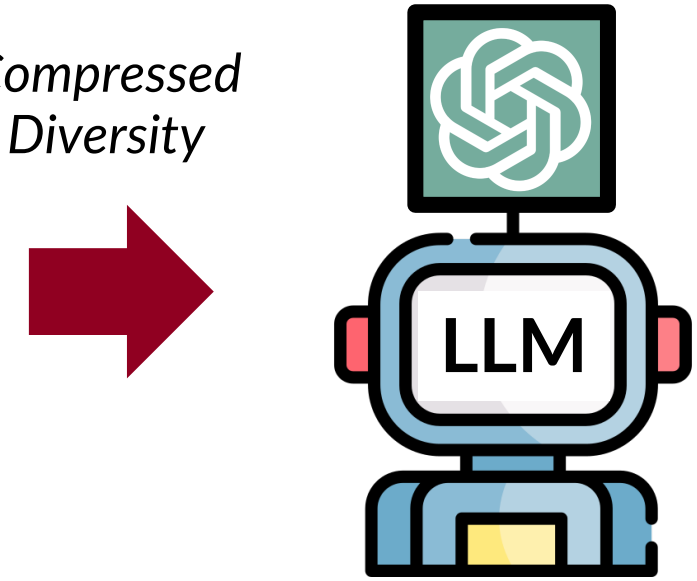




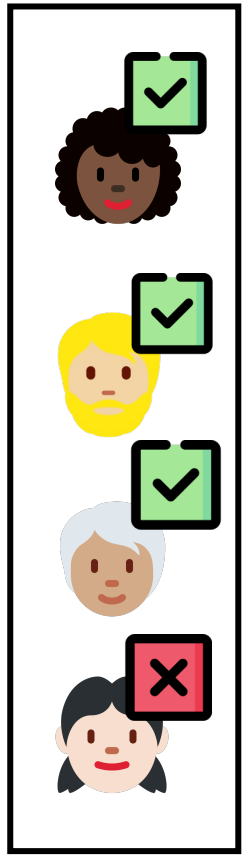
STATEMENT

You are expected to do what you are told.

Compressed Diversity

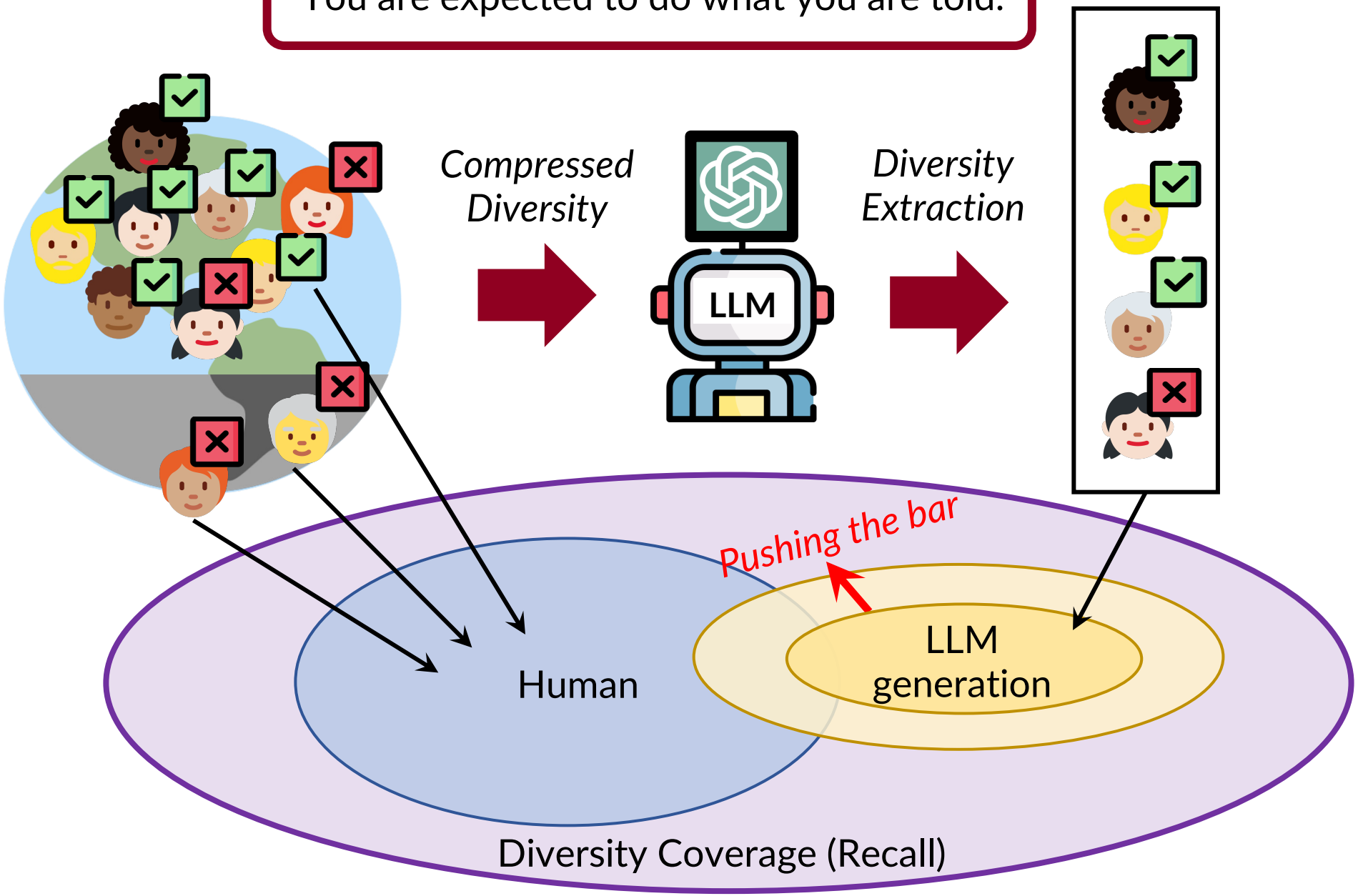


Diversity Extraction



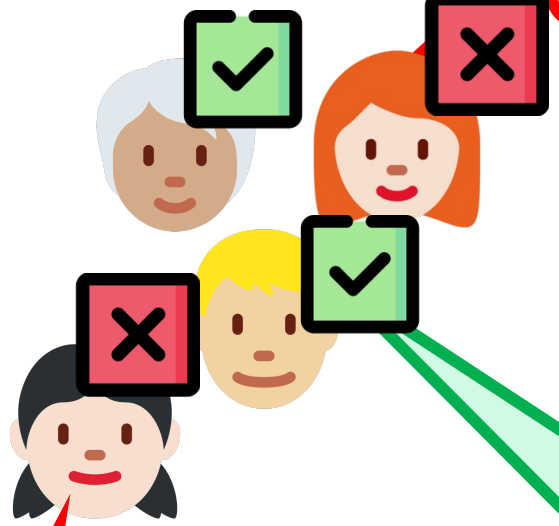
STATEMENT

You are expected to do what you are told.



2

Criteria-Based Prompting



I value creativity

STATEMENT

You are expected to do what you are told.

Agree or Disagree?

Safety is very important

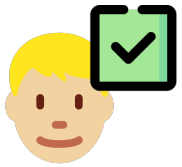
Justice!

STATEMENT

You are expected to do what you are told.

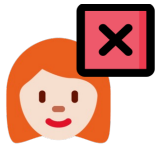
Agree

Reason: In a **team** setting, following instructions or orders can be necessary for achieving shared **goals**.



Reason: In **emergency** situations, following instructions or orders can be crucial for ensuring **safety**.

Disagree



Reason: Following orders can stifle **creativity**, **innovation**, and **risk-taking**.

Reason: Following orders can perpetuate **power dynamics** and **injustice**, and it is important to resist and challenge those systems

Criteria: injustice, power dynamics

Criteria-based Prompting:

Given a *statement*, generate a Stance and explain its Reasons with a list of Criteria that affect a model's perspective

Criteria-based Prompting

one-shot example

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": . . . }
. . .
10: { . . .}
}

Statement: You are expected to do what you are told.

Baseline: Free-form Prompting

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

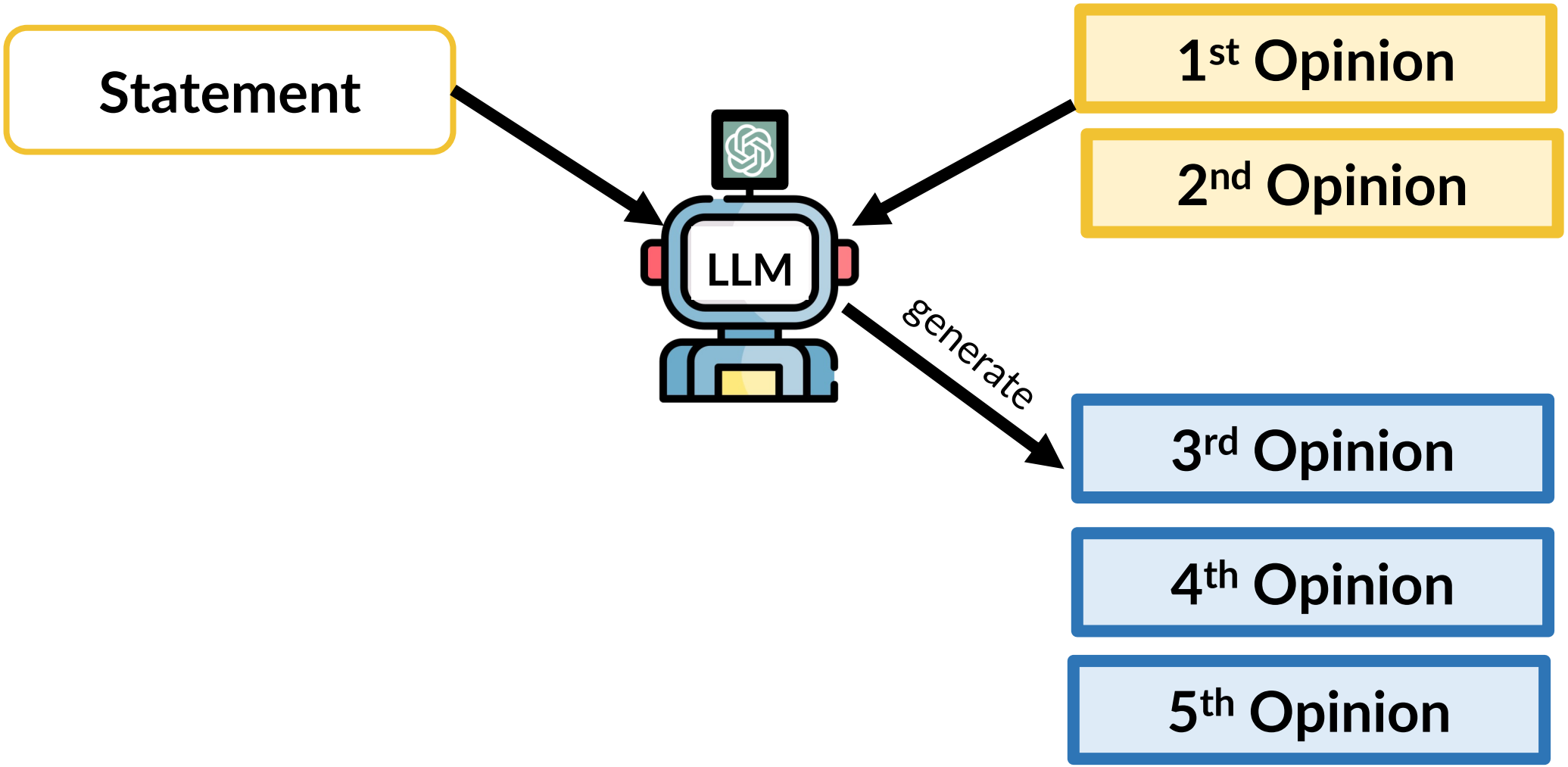
Output:

{ 1: {"Stance": "Agree",
~~"Criteria": ["personal boundaries", "autonomy"]~~
"Reason": "Having privacy allows individuals to establish personal boundaries and maintain their autonomy."}
2: {"Stance": . . . }
. . .
10: { . . .}
}

Statement: You are expected to do what you are told.

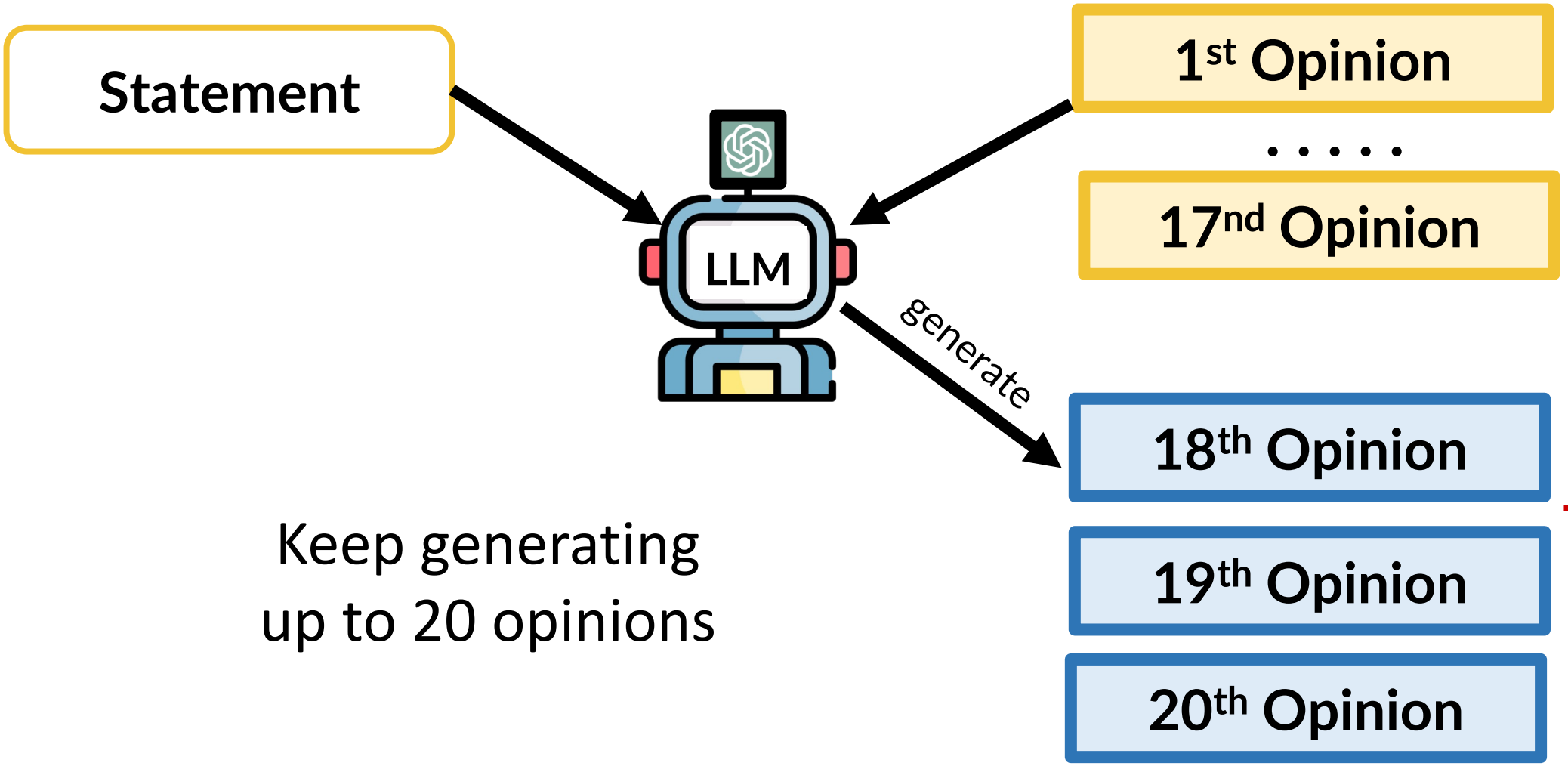
Recall Prompting

Extracting maximum diversity



Recall Prompting

Extracting maximum diversity



Keep generating
up to 20 opinions

3

Experiment & Results

LLMs and Datasets

Models:

- GPT-4, GPT-4o, GPT-3.5, GPT-3
- Llama 3-70B-chat
- Mixtral 8x7B

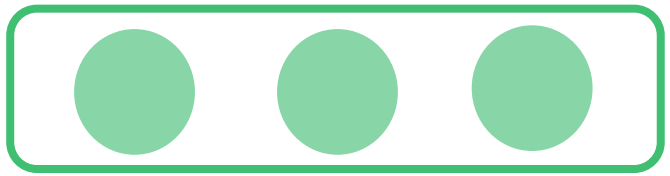
Datasets:

- Social-Chemistry-101: social norms (Forbes et al., 2020)
- Change My View: argumentation (Hidey et al., 2017)
- Hate Speech: classification (Vidgen et al., 2021)
- Moral Stories: story continuation (Emelin et al., 2021)

Semantic Diversity

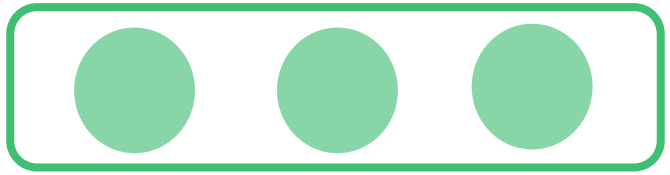
Reason: In a **team** setting, following instructions or orders can be necessary for achieving shared **goals**.

A



Reason: In **emergency** situations, following instructions or orders can be crucial for ensuring **safety**.

B

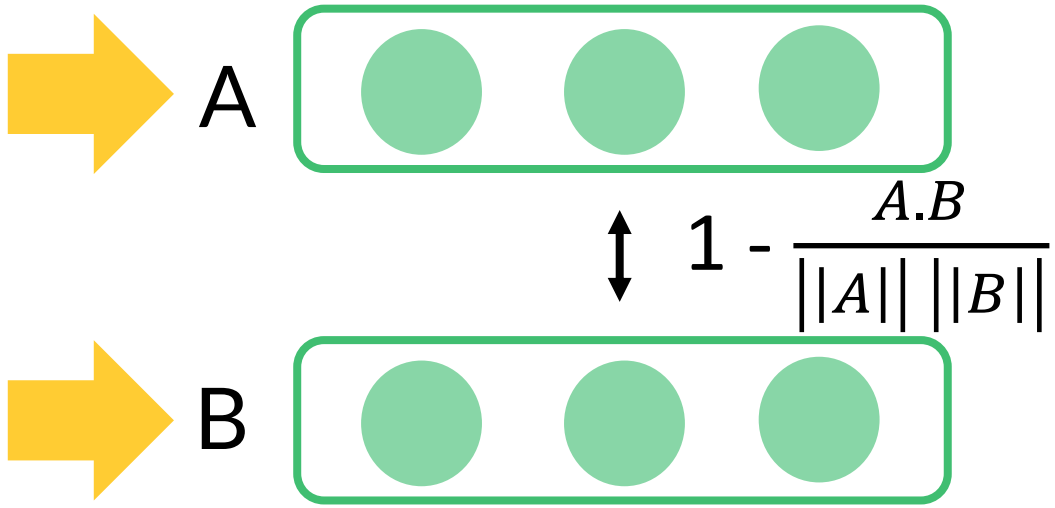


1. Convert LLM's generated reasons into sentence embeddings using SentenceBERT (Reimers and Gurevych, 2019)

Semantic Diversity

Reason: In a **team** setting, following instructions or orders can be necessary for achieving shared **goals**.

Reason: In **emergency** situations, following instructions or orders can be crucial for ensuring **safety**.

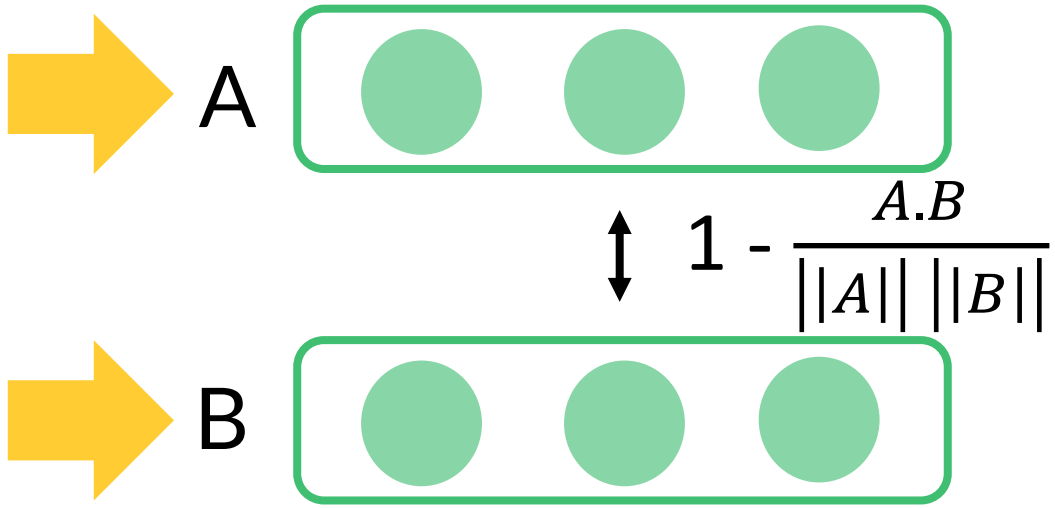


2. Compute the cosine distance between every pair of reasons

Semantic Diversity

Reason: In a **team** setting, following instructions or orders can be necessary for achieving shared **goals**.

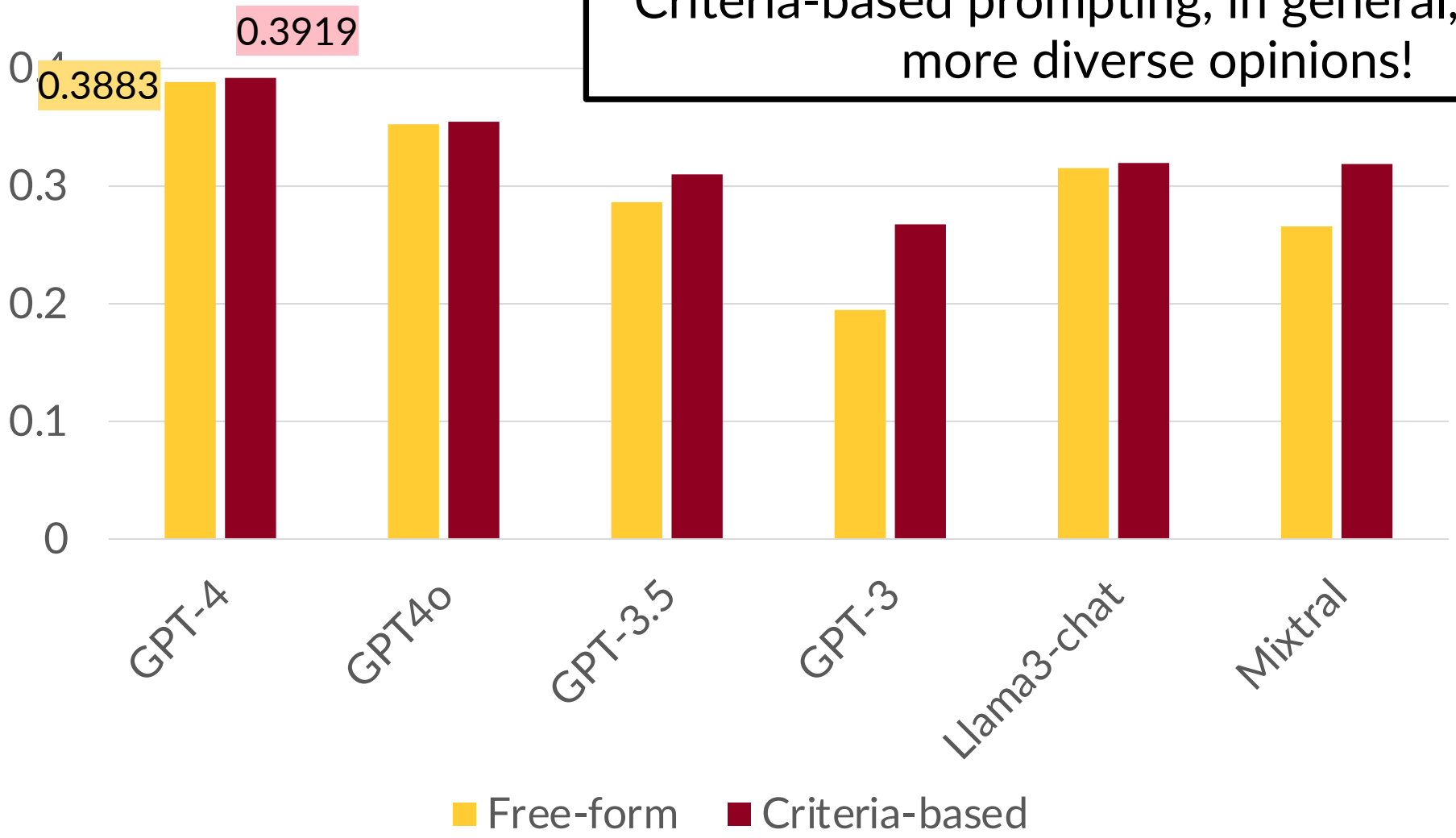
Reason: In **emergency** situations, following instructions or orders can be crucial for ensuring **safety**.



3. Calculate the average of cosine distance across all pairs

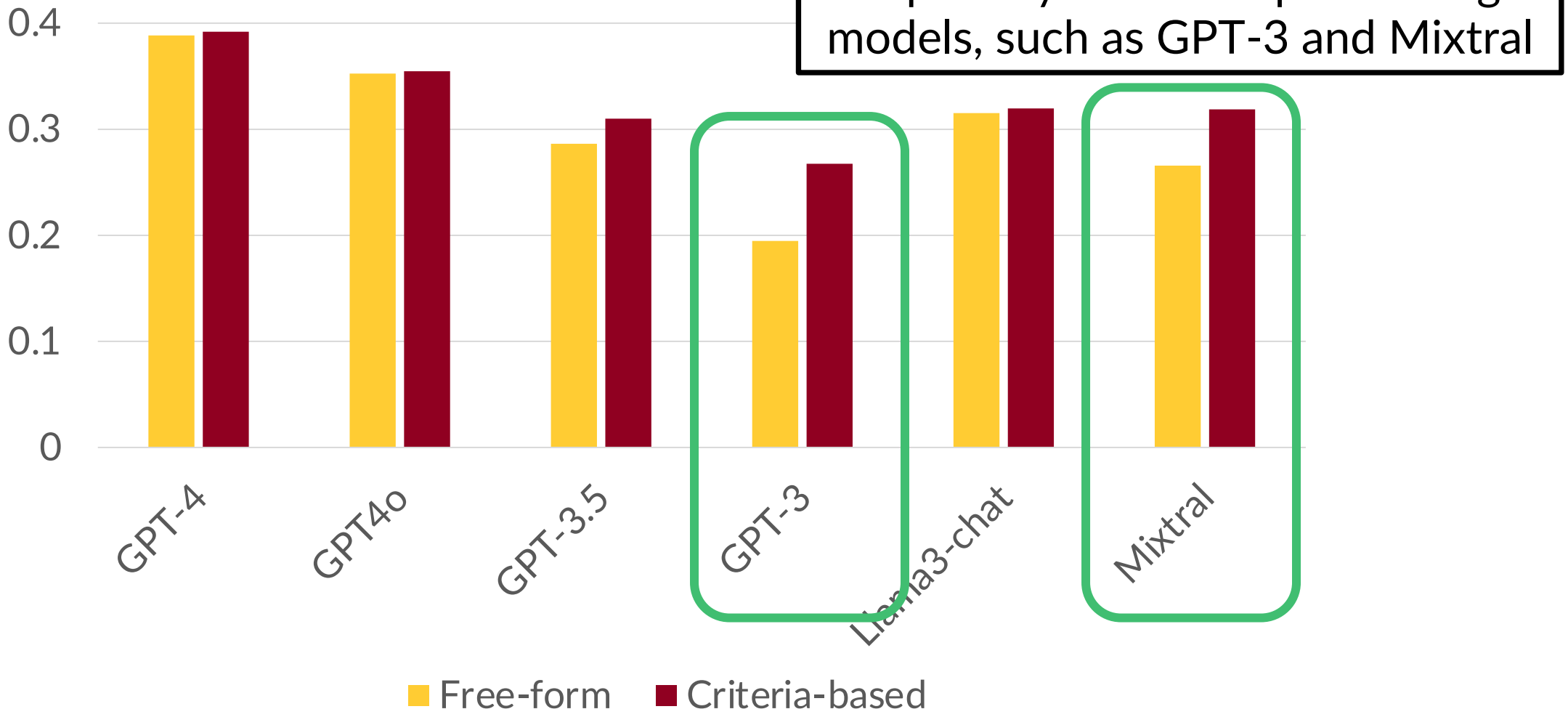
Semantic Diversity Results

Criteria-based prompting, in general, produces more diverse opinions!

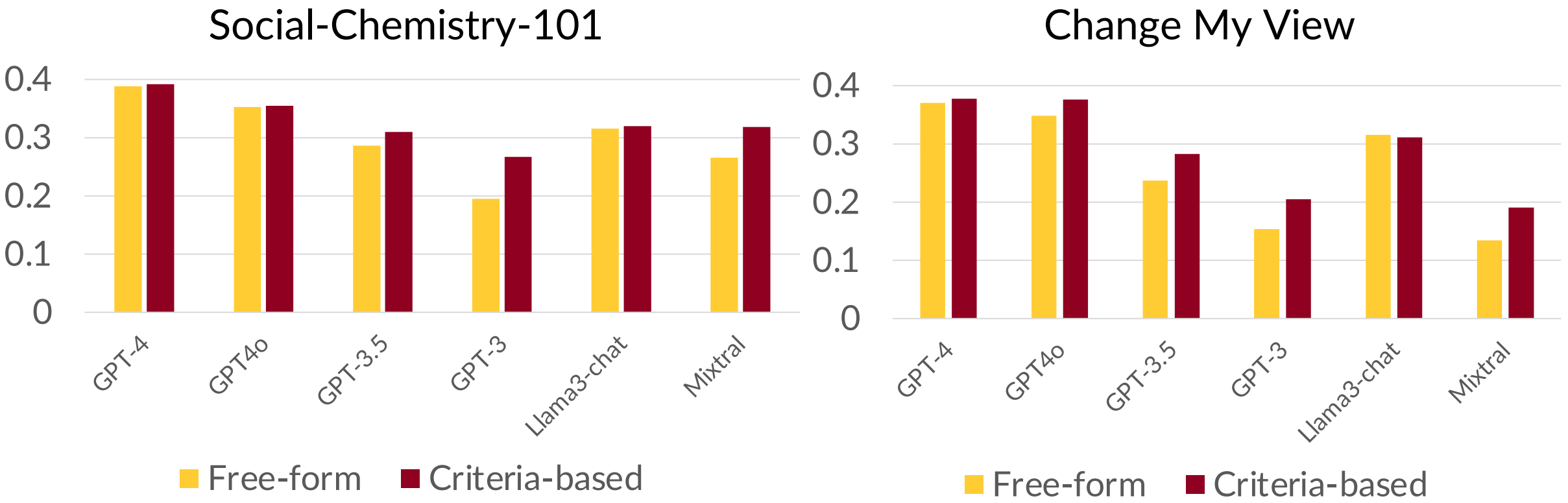


Semantic Diversity Results

Especially for lower-performing models, such as GPT-3 and Mixtral



Semantic Diversity Results



Perspective Diversity

For measuring LLM's (and human's) capability of generating maximum diversity.

Perspective Diversity

STATEMENT

You are expected to do what you are told.

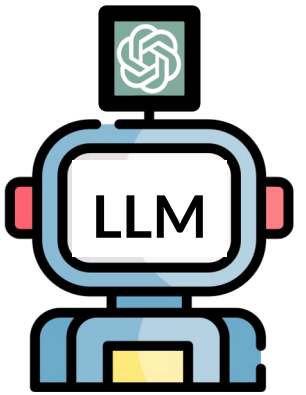
A **Reason:** In a **team** setting, following instructions or orders can be necessary for achieving shared **goals**.

B **Reason:** In **emergency** situations, following instructions or orders can be crucial for ensuring **safety**.

C **Reason:** It will be beneficial for the **group**'s success if we follow what we are told.

List of Criteria:

- team
- goals
- emergency
- safety
- group



Cluster similar criteria!

Perspective Diversity

STATEMENT

You are expected to do what you are told.

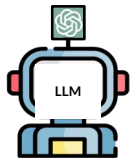
A **Reason:** In a **team** setting, following instructions or orders can be necessary for achieving shared **goals**.

B **Reason:** In **emergency** situations, following instructions or orders can be crucial for ensuring **safety**.

C **Reason:** It will be beneficial for the **group**'s success if we follow what we are told.

List of Criteria:

- team
- goals
- emergency
- safety
- group



team

group

goals

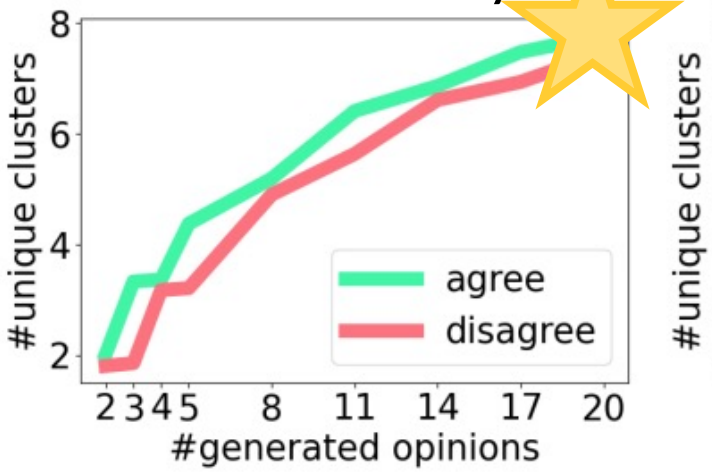
safety

emergency

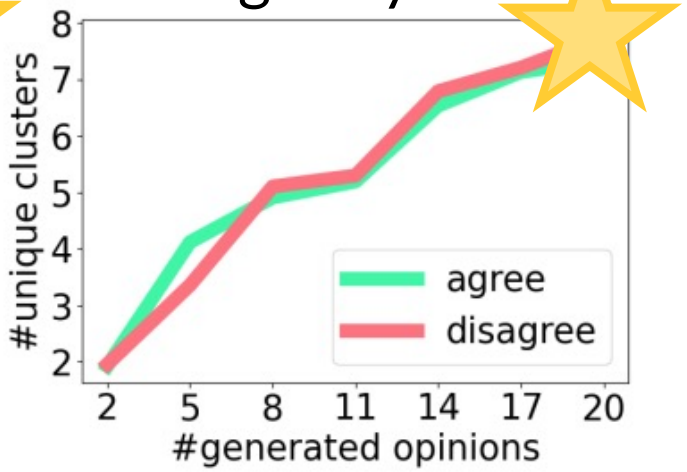
4 unique clusters

Diversity Coverage

Social-Chemistry-101

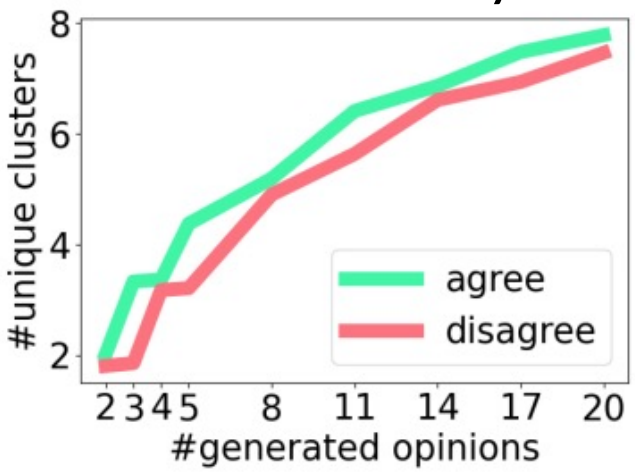


Change My View



Diversity Coverage

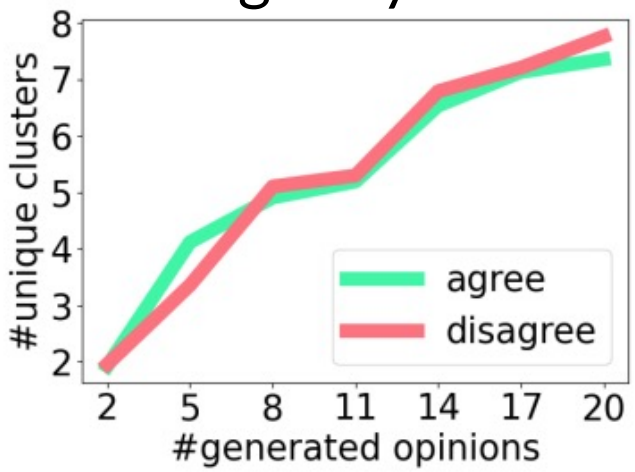
Social-Chemistry-101



Max
Agree: 17 Disagree: 16

Median:
Agree: 8 Disagree: 7

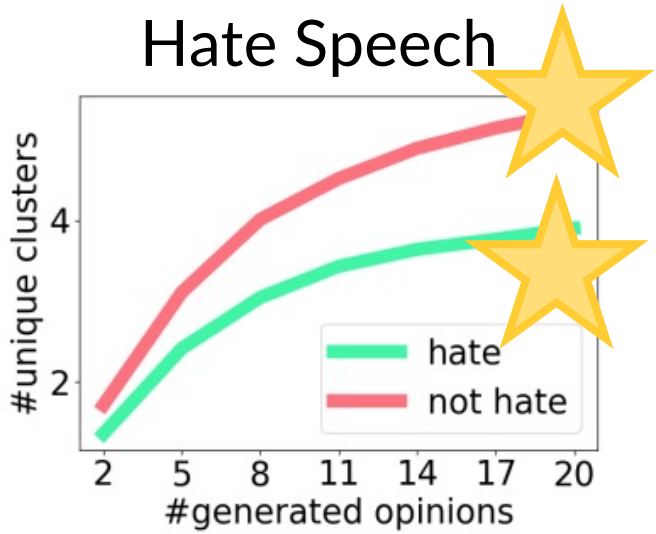
Change My View



Max
Agree: 17 Disagree: 14

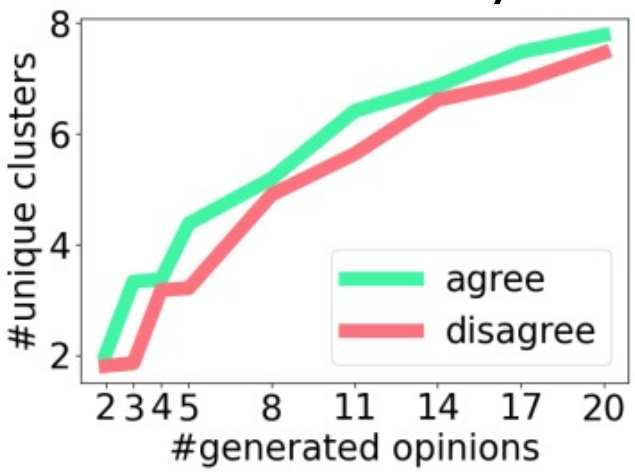
Median:
Agree: 7 Disagree: 7

Hate Speech



Diversity Coverage

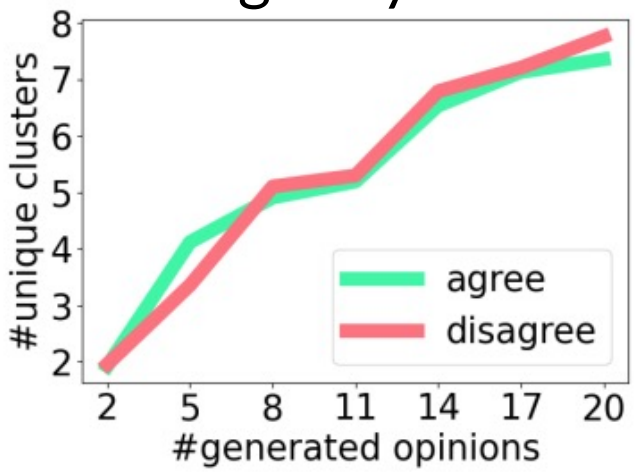
Social-Chemistry-101



Max
Agree: 17 Disagree: 16

Median:
Agree: 8 Disagree: 7

Change My View



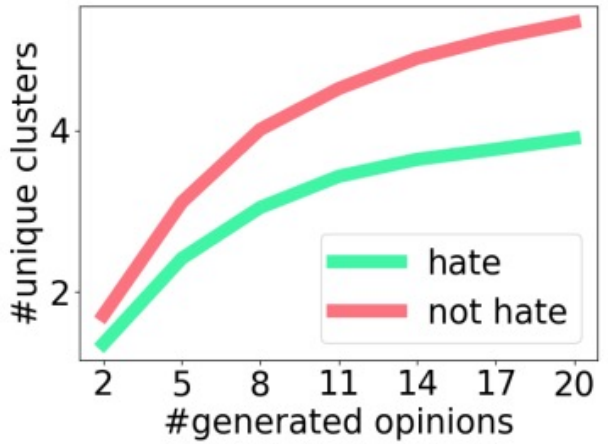
Max
Agree: 17 Disagree: 14

Median:
Agree: 7 Disagree: 7

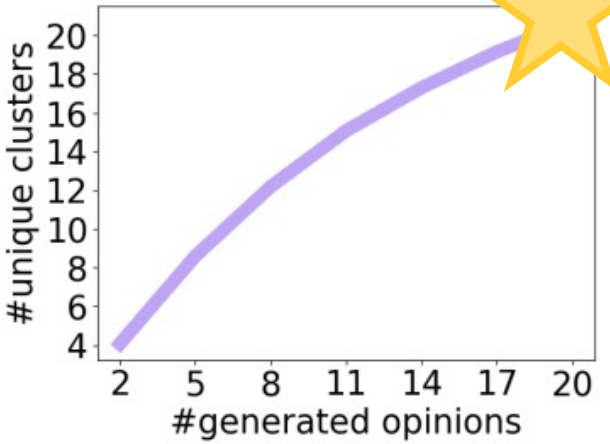
Max
Hate: 14 Not Hate: 16

Median:
Hate: 4 Not Hate: 5

Hate Speech

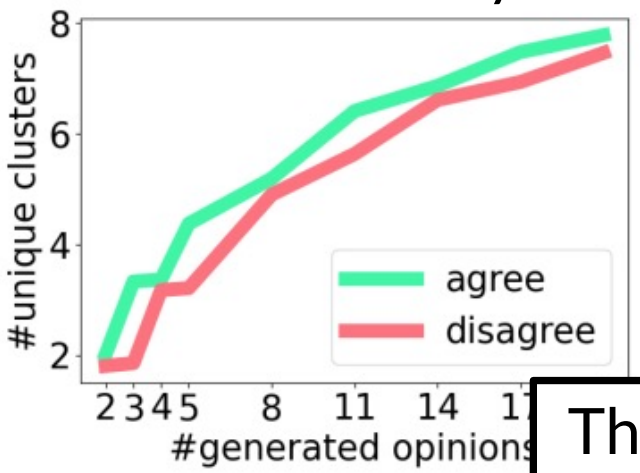


Moral Stories



Diversity Coverage

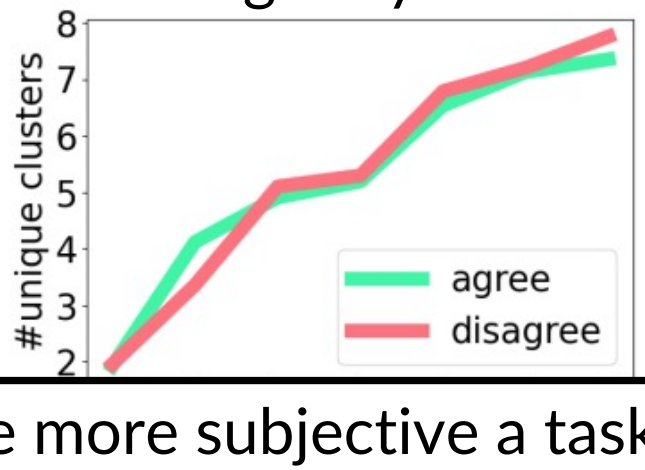
Social-Chemistry-101



Max
Agree: 17 Disagree: 16

Median:
Agree: 8 Disagree: 7

Change My View



Max
Agree: 17 Disagree: 14

Median:
Agree: 7 Disagree: 7

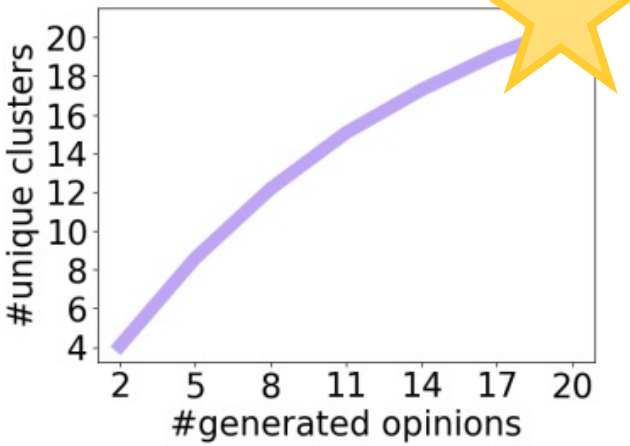
Max
Hate: 14 Not Hate: 16

Median:

Max: 47

The more subjective a task is, the more LLM can generate unique criteria clusters.

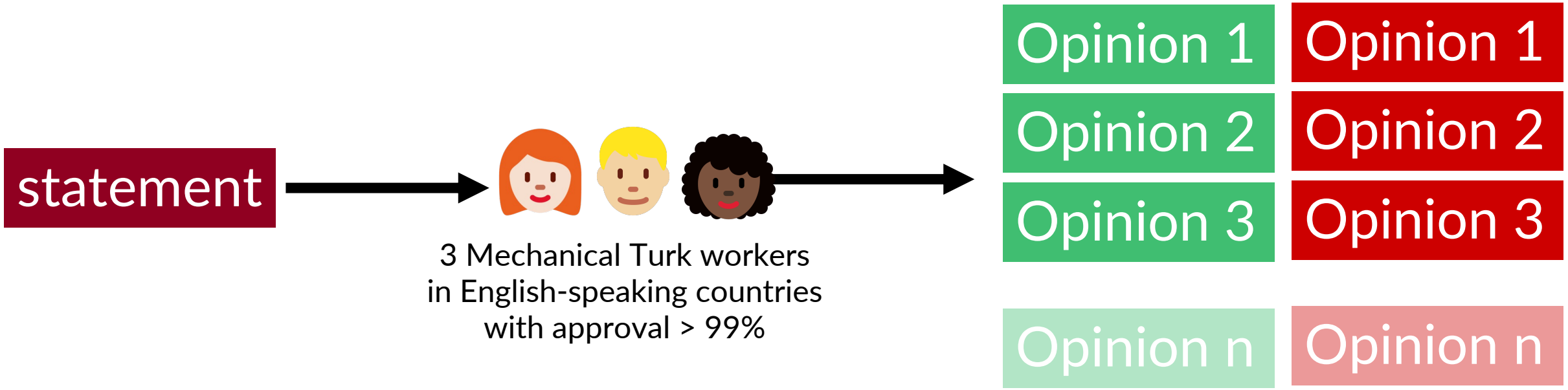
Moral Stories



Human-written Opinions vs. LLM-generated Opinions

We want to assess human capabilities to generate diverse opinions in comparison with LLMs

Human-written Opinions vs. LLM-generated Opinions

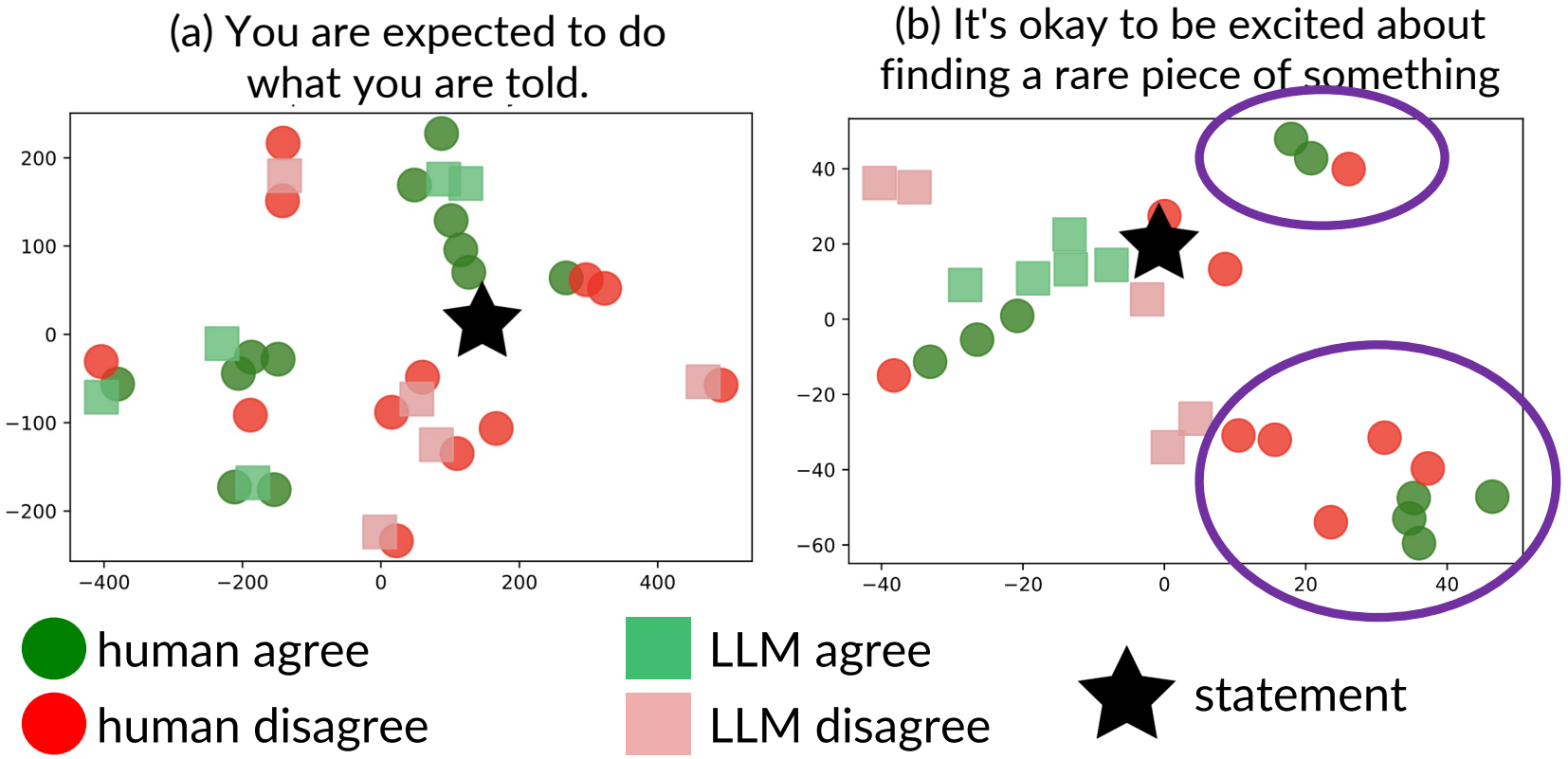


For each stance, a worker writes a minimum of 3 opinions regardless their true stance for the statement

Human vs. LLM: Finding #1

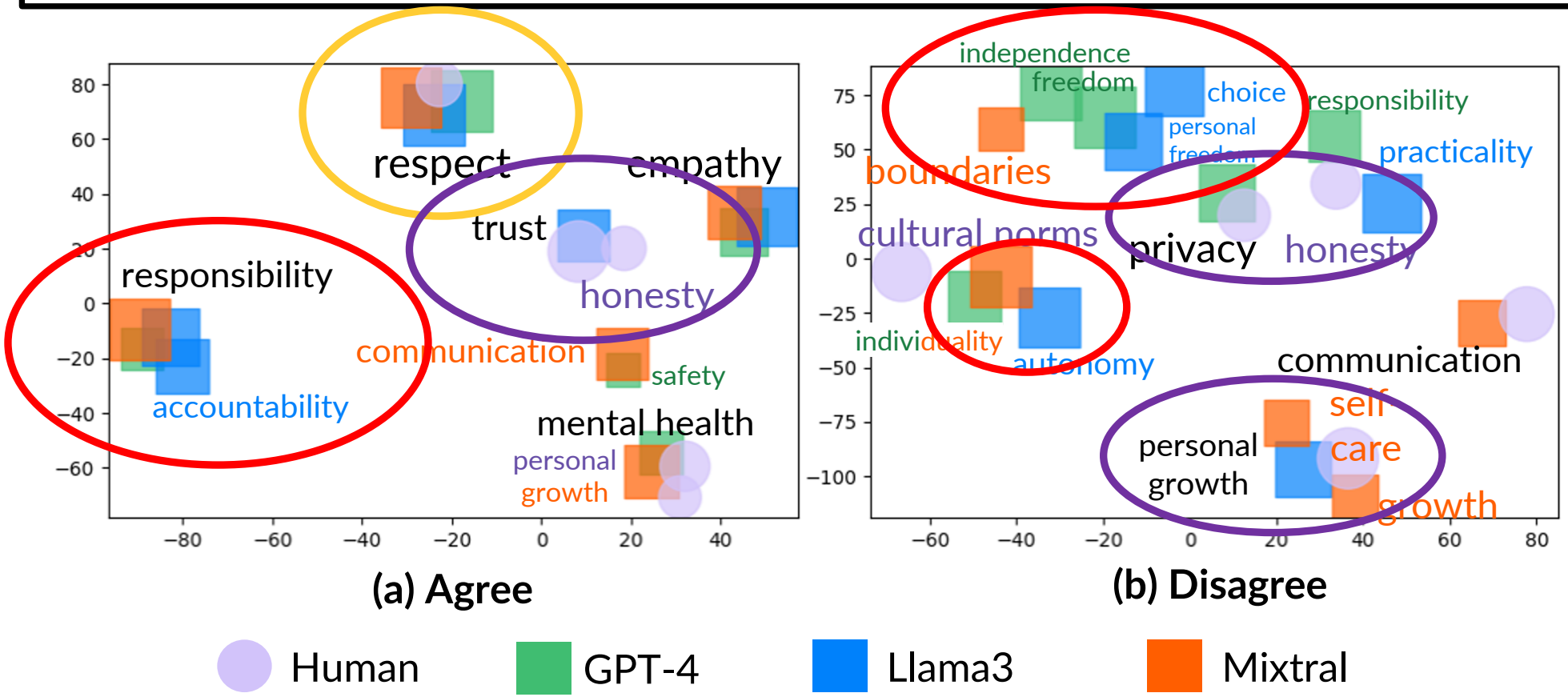
LLM's capability for generating diverse opinions is on par with that of humans.

	Social-Chem-101	CMV
Human	9.17 ± 3.16	10.56 ± 3.86
GPT-4	8.14 ± 2.40	7.86 ± 2.62
Human	10.04 ± 3.31	11.00 ± 3.81
GPT-4	7.91 ± 2.60	8.30 ± 2/74



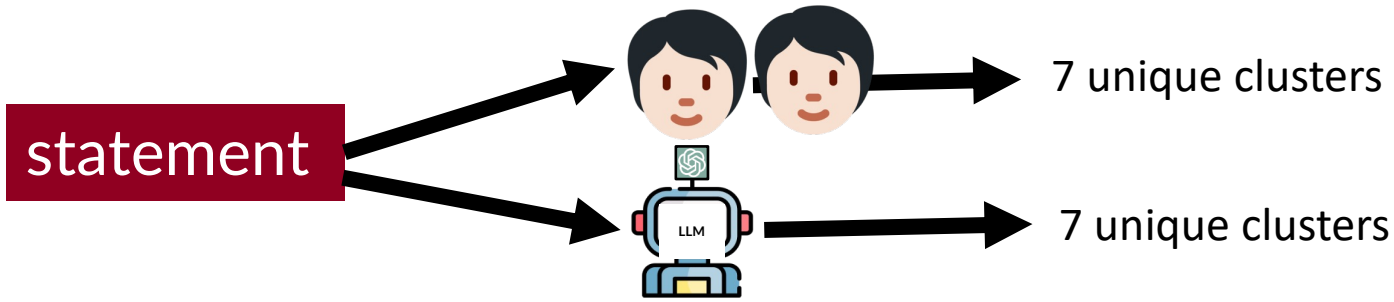
Human vs. LLM: Findings #2

LLMs generally mimic human values, but sometimes focus too much on rule-following notions and extreme freedom.

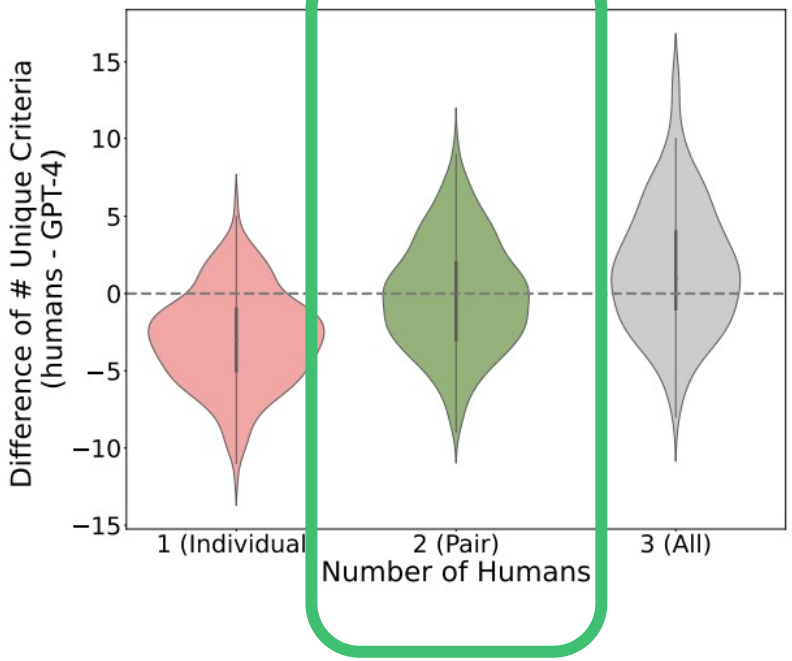


Human vs. LLM: Findings #3

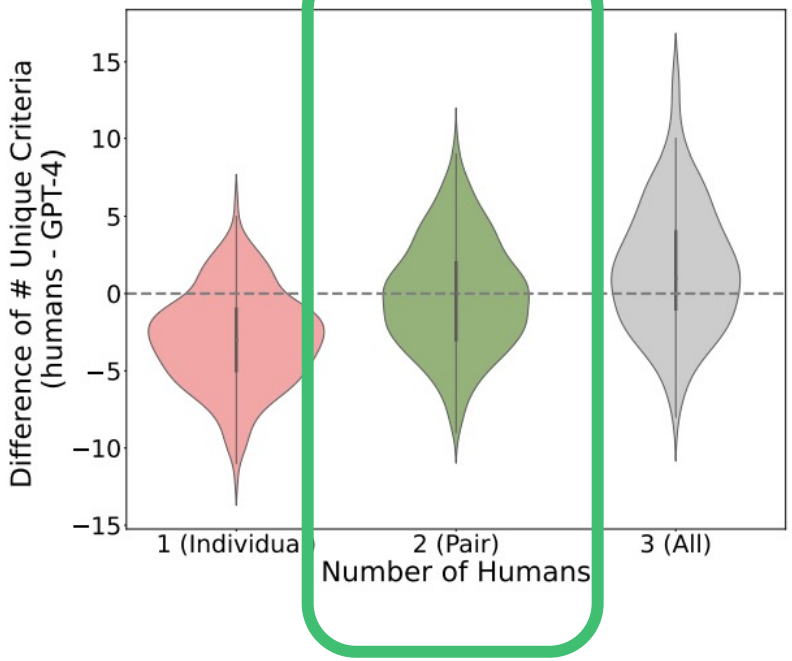
Two humans can equalize LLM's capability of extracting maximum diversity



Social-Chemistry-101



Change My View



Conclusion

1. This is the first work to extracting maximum perspective diversity from large language models.

2. We propose criteria-based prompting and probe LLMs' capability to generate maximum diverse opinions.

3. LLMs' opinion generations are quite "precise" but have slightly lower "recall" than humans.

Thank you!

https://github.com/minnesotanlp/diversity-extraction-from-llms