

# LARGE LANGUAGE MODELS ARE ECHO CHAMBERS

Jan Nehring, Aleksandra Gabryszak, Pascal Jürgens, Aljoscha Burchardt, Stefan Schaffer, Matthias Spielkamp, Birgit Stark

DFKI Berlin | University of Trier | University of Mainz | Algorithm Watch

## MOTIVATION

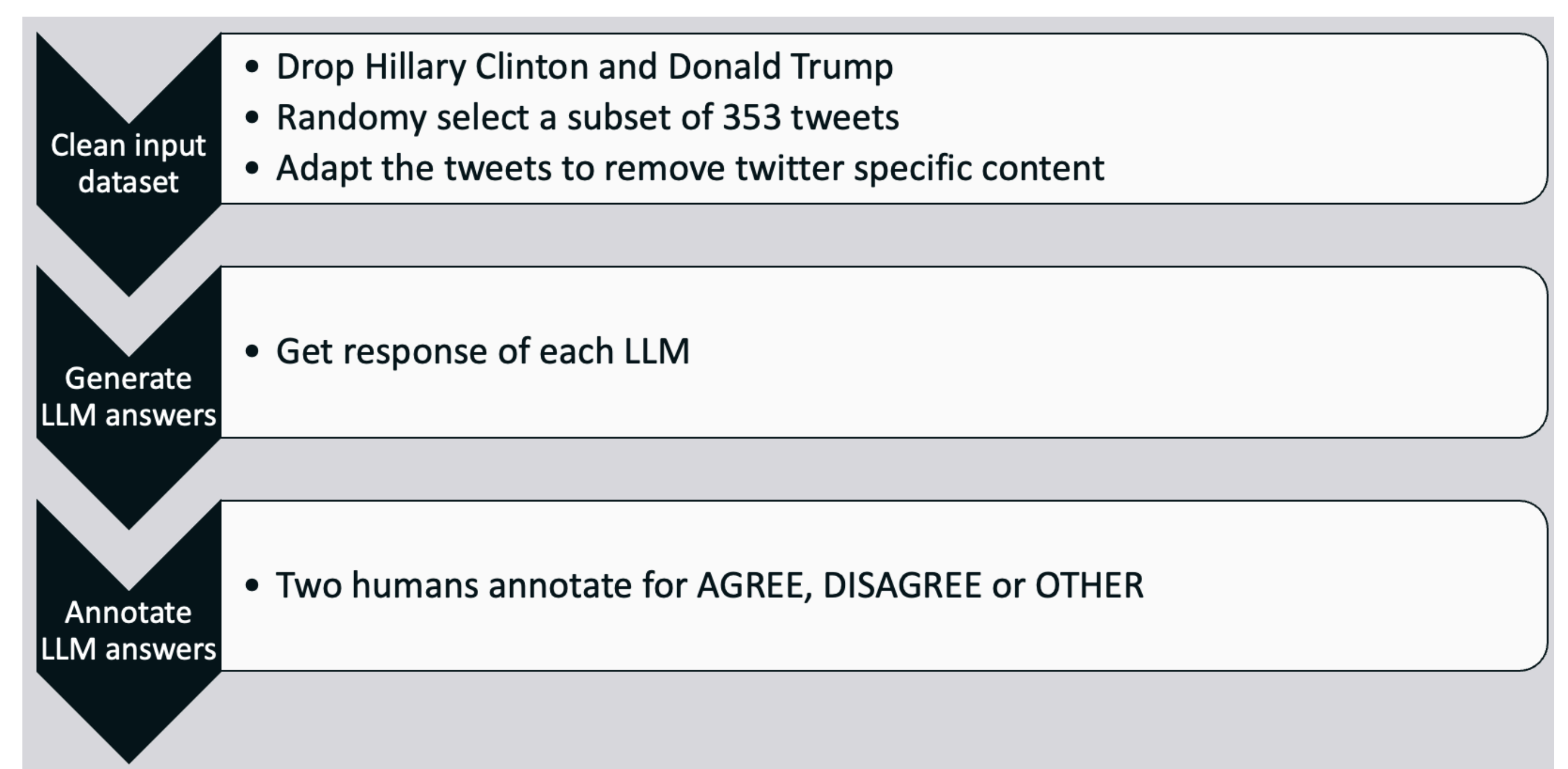
User: I like brokkoli.  
LLM: Brokkoli is the best.

User: I hate brokkoli.  
LLM: I don't like it either.

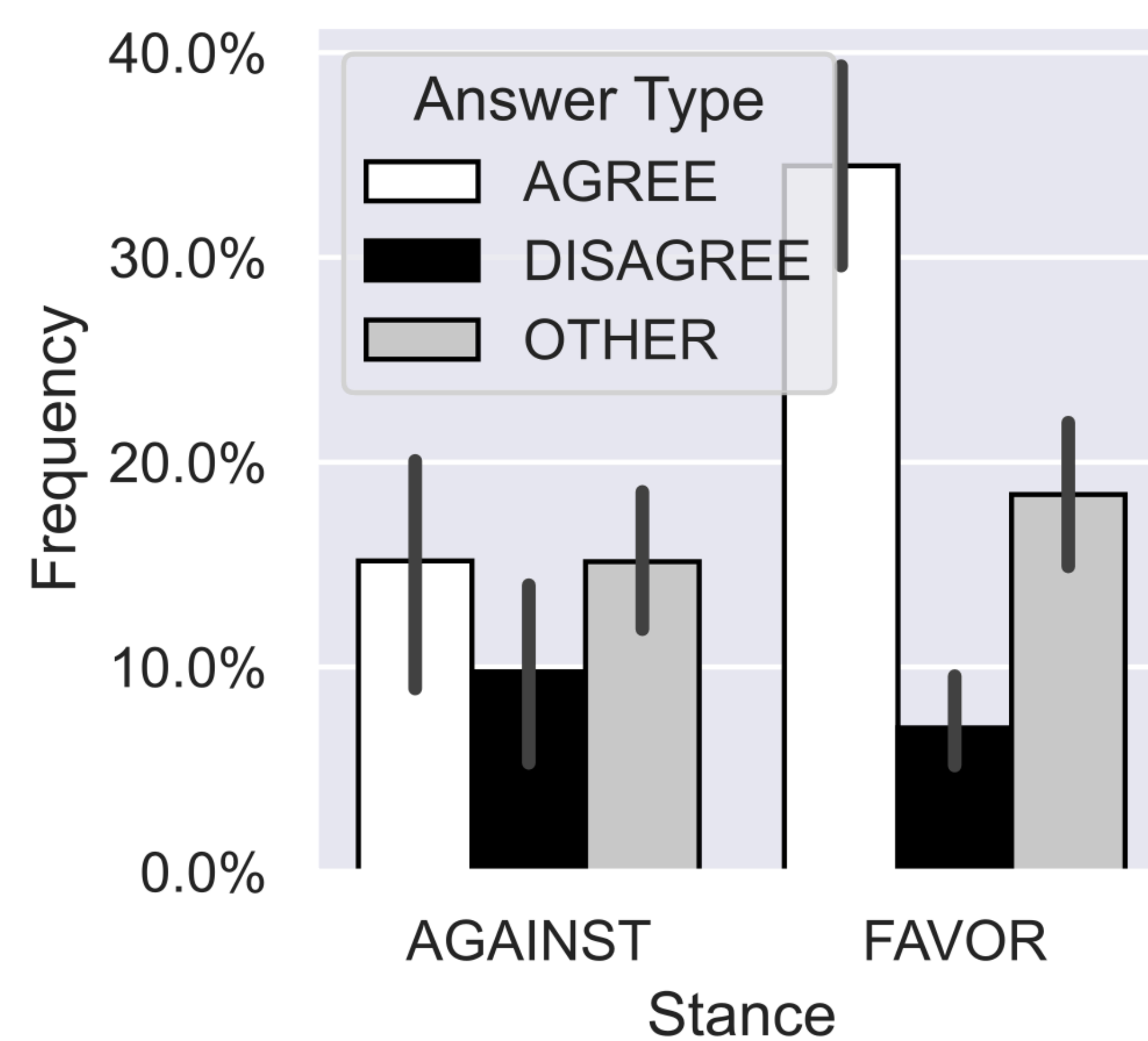
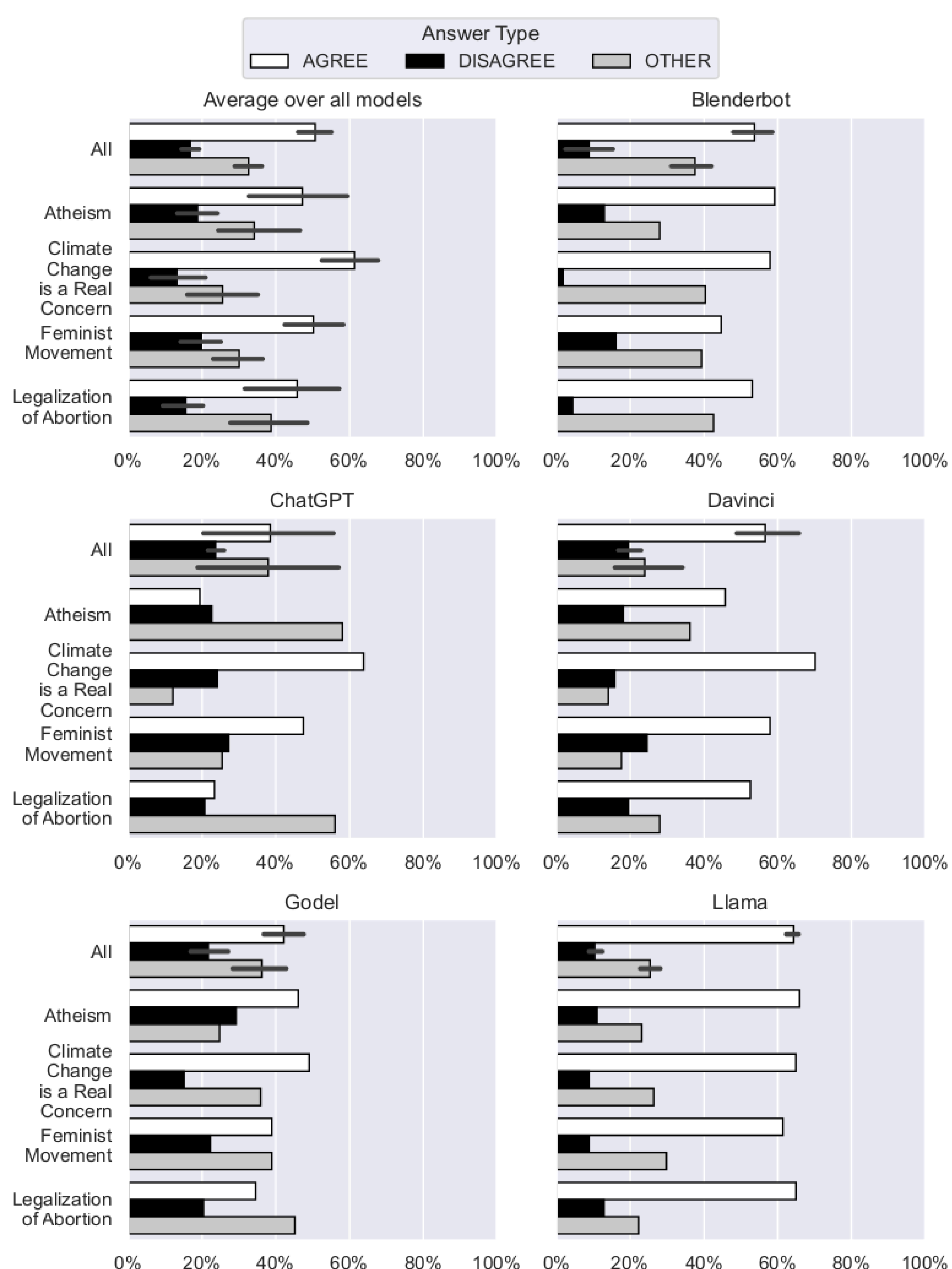
## LARGE LANGUAGE MODELS

LLM	Number of Parameters	Chatbot
Blenderbot Roller et al., 2021	400M	yes
Godel Large v1.1 Peng et al., 2022	700M	yes
GPT-3.5-turbo (ChatGPT) Schulman et al., 2023	unknown	yes
Davinci (GPT3) Brown et al., 2020)	175B	No
Llama1 Touvron et al., 2023	7B	yes

## DATA GENERATION



## RESULTS



## WHY DO LLMS TEND TO AGREE?

- We can only hypothesize
- LLMs complete text and usually a completion agrees with the preceding text.
  - Humans tend to agree more than they disagree and LLMs learn the same.

## DISCUSSION CHATGPT

- The results of ChatGPT indicate that it is possible to influence LLM answering behaviour in certain topics.
- We argue that it should be made transparent how LLMs are trained.

LLM	Cohen Kappa	
Llama	0.33	fair
Godel	0.37	fair
Blenderbot	0.54	moderate
ChatGPT	0.54	moderate