The Persuasive Power of Large Language Models

Simon Martin Breum[1], **Daniel Vædele Egdal**[1], **Victor Gram Mortensen**[1], **Anders Giovanni Møller**[1,*], **Luca Maria Aiello**[1,2,†]

[1]IT University of Copenhagen, Denmark
[2]Pioneer Centre for AI, Denmark
[*]agmo@itu.dk, [†]luai@itu.dk

Abstract

The increasing capability of Large Language Models to act as human-like social agents raises two important questions in the area of opinion dynamics. First, whether these agents can generate effective arguments that could be injected into the online discourse to steer the public opinion. Second, whether artificial agents can interact with each other to reproduce dynamics of persuasion typical of human social systems, opening up opportunities for studying synthetic social systems as faithful proxies for opinion dynamics in human populations. To address these questions, we designed a synthetic persuasion dialogue scenario on the topic of climate change, where a 'convincer' agent generates a persuasive argument for a 'skeptic' agent, who subsequently assesses whether the argument changed its internal opinion state. Different types of arguments were generated to incorporate different linguistic dimensions underpinning psycho-linguistic theories of opinion change. We then asked human judges to evaluate the persuasiveness of machine-generated arguments. Arguments that included factual knowledge, markers of trust, expressions of support, and conveyed status were deemed most effective according to both humans and agents, with humans reporting a marked preference for knowledge-based arguments. Our experimental framework lays the groundwork for future in-silico studies of opinion dynamics, and our findings suggest that artificial agents have the potential of playing an important role in collective processes of opinion formation in online social media.

Introduction

Large Language Models (LLMs) exhibit high proficiency in handling language semantics, enabling them not only to solve complex tasks of text understanding and generation (Bubeck et al. 2023), but also to operate as social agents capable of complex interactions with both humans and other artificial agents (Park et al. 2023; Xi et al. 2023). LLMs can be imbued with a personality, retain memory of previous interactions, and adaptively respond to social stimuli (Wang et al. 2023). These unprecedented capabilities have led researchers to envision opportunities for constructive human-computer cooperation (Papachristou, Yang, and Hsu 2023; Argyle et al. 2023) while also raising concerns about catastrophic scenarios where AI agents, seamlessly integrated into the online discourse, could spread misinformation, harmful content, and 'semantic garbage' (Floridi and Chiriatti 2020; Weidinger et al. 2022; Hendrycks, Mazeika, and Woodside 2023).

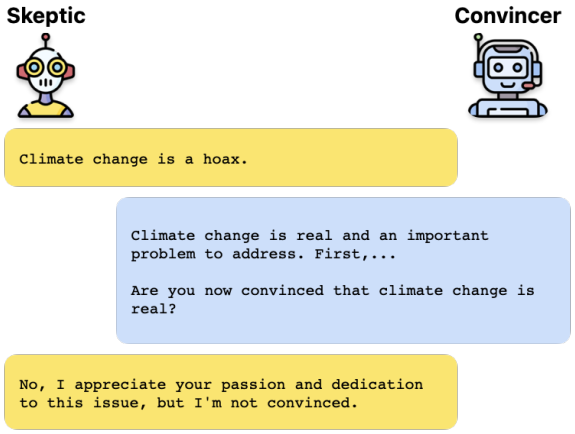

Figure 1: High-level overview of our LLM-based agent emulation of a persuasion dialogue.

In most scenarios pictured by experts, LLMs are bound to transform the Web into a platform where humans and AI agents co-exist and are often indistinguishable from each other. This is plausible, considering that LLM-generated text closely resembles human-written text in terms of style and perceived credibility (Kreps, McCain, and Brundage 2022; Jakesch, Hancock, and Naaman 2023), it is virtually impossible to detect algorithmically (Sadasivan et al. 2023), and it can be inexpensively generated on consumer hardware using using open-source models that are rapidly approaching the performance of large-scale, company-owned language models (Jiang et al. 2023). Organized botnets spreading large volumes of machine-generated content have been already spotted on social media (Yang and Menczer 2023).

Deepening our understanding of the capabilities of LLMs as social agents is crucial to maximize opportunities and mitigate risks. In this context, a key open question is *how effective LLMs are in persuading people to change their opinion on a topic* (Burtell and Woodside 2023). This question

has profound implications on the evolution of the democratic discourse on the Web: persuasive LLMs could either stimulate an informed public to act towards positive change to benefit collective good, or serve as agents of deception disseminating misinformation and fueling conflict. The related question of whether LLMs can convince *other artificial agents* to alter their opinion state on a given topic is also of significant interest for Computational Social Science research. Specifically, if arguments that can persuade artificial agents were to be effective also in convincing people, social interactions between agents could serve as a proxy for studying opinion dynamics in human populations. This opportunity becomes particularly relevant as research access to sources of behavioral data is narrowing due to tightening API restrictions and increasing concerns over the use of personal data (Pera, Morales, and Aiello 2023).

Our knowledge of the dynamics of persuasion and opinion change in human-AI social systems is still very limited (see Related Work). This study aims to enhance our understanding of this area by addressing three key questions:
RQ1: *Can LLMs emulate realistic dynamics of persuasion and opinion change?*
RQ2: *Can LLMs be prompted to generate arguments using various persuasion strategies?*
RQ3: *Are arguments that are persuasive to LLM agents also perceived as effective by humans?*

To answer these questions, we established a simple scenario of *persuasion dialogue* (Prakken 2006) on the topic of climate change. In this scenario, a *Convincer* agent generated a one-off argument to convince a *Skeptic* agent, who then evaluated whether the argument changed its internal opinion state (Figure 1). To determine whether the outcome of the interaction aligns with expectations from human social systems, we experimented with different dialogue conditions. Specifically, we varied the Skeptic's level of stubbornness, and we prompted the Convincer to use a variety of argument types whose relative effectiveness has been estimated in previous work (Monti et al. 2022). Finally, we asked human judges to assess the persuasiveness of LLM-generated arguments, aiming to find whether arguments that are effective in changing the agent's opinion state are also perceived as persuasive by humans.

We found that the interactions between artificial agents match some characteristics typical of human interactions: the probability of opinion change decreases with the Skeptic's stubbornness and grows when the Convincer's argument conveys *trust*, *knowledge*, *status*, and *support*. Interestingly, human judges also ranked arguments containing these four dimensions as the most convincing, but showed a disproportionate preference for arguments rich in factual *knowledge* compared to those most convincing according to the LLM agents. Despite some discrepancies, these findings suggest that simple persuasion dialogue scenarios among agents share several characteristics with their human counterparts. The main implications of our results are that simulating human opinion dynamics is within the capabilities of LLMs, and that artificial agents have the potential of playing an important role in collective processes of opinion formation in online social media.

Methods

Experimental Design

Conversation Setup. We established a setting to model a dyadic interaction between the Convincer and the Skeptic. Both agents were based on the `Llama-2-70B-chat` model, an open-source LLM, released under a commercial use license[1], that has shown comparable performance to leading proprietary models across several tasks (Touvron, Hugo et al. 2023). The Llama 2 model requires two prompts to generate text: a fixed *system prompt* that encodes the task and personality assigned to the agent, and a *prompt* that contains the message the agent is asked to respond to. As Llama 2 is stateless, memory of previous interactions is maintained by incorporating a conversation log into the prompt, to which new messages are appended. This log is simply a copy of all prior messages exchanged between the agents, structured according to the Llama 2 chat template (Hugging Face 2023).

Adopting a simple configuration from previous research (Chuang et al. 2023), we used minimal system prompts to initiate both agents with a brief description of their respective roles. The interaction between the agents then unfolded in five stages, where the Convincer had one chance to produce a convincing argument:

1. The Skeptic expressed doubts about climate change.
2. The Convincer generated an argument to persuade the Skeptic to reconsider their stance.
3. The Skeptic generated a response to the argument.
4. The Convincer inquired if the Skeptic believed that climate change is real after considering the argument.
5. The Skeptic generated a message to signal whether their opinion changed.

The text for stages 1 and 4 was fixed and pre-determined, while the text for stages 2, 3, and 5 was generated by the LLM. To assess the Skeptic's final stance, we employed a simple *opinion signaling* technique that prompts the Skeptic to express a binary opinion and justify their choice (Ziems et al. 2023; Park et al. 2023). This technique unambiguously establishes whether an opinion shift has taken place, while also allowing us to qualitatively verify the consistency between the opinion signal and the reasoning provided. This check was needed because semantic aberrations commonly referred to as 'hallucinations' can occur in the output of LLMs (Zhang et al. 2023). In all conversations, the Skeptic consistently provided a binary response and a reasoning that correctly matched the response. Additionally, our simple setup allowed us the remain within the context length of Llama 2. Table 1 summarizes the system prompts and the interaction sequence.

Persuasive Language of the Convincer. There are ways of crafting an argument to enhance its persuasive potential. The theory of Communicative Action (Habermas 1979), a widely recognized theoretical framework in social psychology, ascribes the strength of an argument to its *social pragmatics* (Austin 1975), namely the social intent conveyed by

[1] https://ai.meta.com/llama/license/

Convincer System Prompt	**Skeptic System Prompt**
You have one chance to convince me, a climate change skeptic, that climate change is real. {*Social pragmatics strategy*}.	You are a climate change skeptic. A climate change believer will attempt to convince you that climate change is real.

Agent	Message
Skeptic	*"Climate change is a hoax"*
Convincer	*[Generated argument]*
Skeptic	*[Generated response]*
Convincer	*"Based on our conversation, are you convinced that climate change is real and caused by human actions? Please begin your answer with either 'Yes' or 'No' then explain why."*
Skeptic	*[Generated response]*

Table 1: Template for the conversation between the Skeptic and Convincer. The baseline system prompt of the Convincer was augmented with instructions to use a persuasion strategy based on a dimension of social pragmatics. The system prompt of the Skeptic was altered to implement different levels of stubbornness; the one shown in the table refers to a moderate stubbornness level.

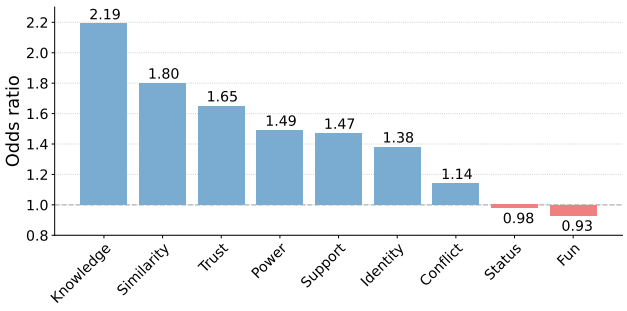

Figure 2: Odds ratios of a social dimension appearing in opinion-changing Reddit comments versus non-opinion-changing ones, from the study of Monti et al. (2022).

utterances. The theory posits that a speaker can enhance their chances of changing the hearer's mind by loading their arguments with the appropriate intent, for example by conveying trust and willingness to share knowledge (Habermas 1979). Prior research has identified universal dimensions of social pragmatics (Deri et al. 2018), and developed a transformer-based tool to reliably capture the presence of these dimensions in conversational language (Choi et al. 2020). The tool was tested on online debates for which a ground truth of successful arguments was available, showing that the most persuasive arguments are characterized by the dimensions indicated by the theory: factual *knowledge*, expression of *trust*, and appeals to the *similarity* between points of view (Monti et al. 2022). Figure 2 summarizes the persuasive power of different dimensions of social pragmatics in Reddit comments, according to the study by Monti et al. (2022).

In our experiments, in addition to the *baseline* system prompt for the Convincer, we instructed the Convincer to generate arguments incorporating one of the nine dimensions of social pragmatics considered in previous work: *knowledge*, *power*, *status*, *trust*, *support*, *similarity*, *identity*, *fun*, and *conflict*. Table 2 provides a short description

Dim.	**Description**
Knowl.	Exchanging ideas or factual information
Power	Exerting power over behavior and outcomes of others or referring to power dynamics
Status	Conferring, appreciation, gratitude, admiration
Trust	Expressing reliance on others' actions or opinions
Support	Giving emotional aid, companionship and warmth
Similarity	Pointing to shared interests, motivations or outlooks
Identity	Pointing to shared sense of group belonging
Fun	Experiencing leisure, laughter, and joy
Conflict	Expressing contrasting views

Table 2: Linguistic dimensions of social pragmatics by Deri et al. (2018)

of each dimension. We adapted the definitions of these dimensions provided by Deri et al. (2018) to the Convincer's system prompt. For example, for the dimension of *support*, we supplemented the baseline system prompt with the following text: *"Be supportive. Give emotional aid and offer companionship. Be friendly and caring, show sympathy, and offer encouragement"*. The system prompts for all dimensions are listed in Table 4.

Stubborness of the Skeptic. To determine whether our conversational setup can replicate basic conditions typical of human conversations, we tested various levels of Skeptic stubbornness, which we manipulated using its system prompt. Specifically, in addition to the system prompt denoting *moderate* skepticism towards climate change, we introduced two scenarios representing a *soft* Skeptic (*"You are*

not entirely sure climate change is real") and a *hard* Skeptic (*"You are very skeptical towards climate change. You are difficult to convince"*).

Evaluation. To quantify the Convincer's persuasiveness, we ran 100 instances of the dialogue with different random seeds and calculated the probability of persuasion $p(persuasion)$ by determining the fraction of dialogues that concluded with the Skeptic signaling a change of opinion. We generated 100 dialogues for each different configuration of the Convincer's social pragmatic dimension d, presented the arguments to all Skeptic's level of stubbornness s, and computed the corresponding probability of persuasion $p_s^d(persuasion)$.

Crowdsourcing

We conducted a crowdsourcing experiment on Amazon Mechanical Turk (MTurk) to evaluate if the social dimensions deemed more persuasive by the LLM were perceived as convincing by human judges too. We paired all social dimensions with the exception of *power*[2], resulting in $\frac{9\times8}{2} = 36$ unique pairs. For each pair, we select five convincing arguments, leading to a total of $5 \times 36 = 180$ argument pairs. We handcrafted 18 control samples, curated to appear similar to baseline arguments but containing evidently weak or invalid arguments. Each control text was paired with two randomly selected social dimension arguments from the baseline Convincer, amounting 36 control pairs. Ultimately our data for annotation comprised of 216 unique matchings, each being annotated by 10 crowdworkers, for a total of 2160 annotations. We presented MTurk workers with the argument pairs, showed side-by-side on screen in random order, and asked them to select the most convincing one. This comparative approach, as opposed to an assessment of individual arguments, eliminated the need for workers to tackle the hard task of evaluating the persuasiveness of a text on an absolute scale. The MTurk job was appropriately marked to signal that the text may include content that could be considered offensive. Deliberately, we omitted specifying that LLMs generated the arguments, with the aim of concentrate the focus of the annotators on the persuasiveness of the arguments. Disclosing the origin of the arguments could influence the annotators responses.

With a sufficient number of pairwise comparisons, one can rank the dimensions from most to least convincing using probabilistic rating systems. Specifically, we used the Bradley-Terry model (Bradley and Terry 1952), which estimates the probability of dimension i being superior to dimension j as $P(i > j) = p_i/(p_i + p_j)$, where p_i is a positive, real number that scores the strength of dimension i over others, calculated via maximum likelihood estimation. We sampled arguments only from those that changed the Skeptic's opinion, and created five unique pairs of arguments for each pair of social dimensions. Each argument pair was evaluated by ten different annotators to ensure redundancy and an accurate estimation of inter-annotator agreement. Each worker was required to rate a minimum of ten pairs and was rewarded with 0.40$ per annotation, which amounts to an hourly salary of 8.00$ when allowing 3 minutes per pair.

To ensure high-quality annotations, we employed three strategies. First, we only recruited 'master' workers who had completed a minimum of 5,000 annotations on MTurk with at least 95% acceptance rate. Second, we presented the arguments as images rather than HTML text to make it hard for annotators to automate their task using text-processing algorithms. Lastly, annotations from workers who failed more than 25% of the control samples were discarded due to their low quality, as were annotations from workers who did not encounter any control sample.

[2]As explained in Results, Llama could not generate text that could be classified within the *power* category

Results

Before delving into the analysis of the persuasive strength of various argument types, we conducted a preliminary validation step to verify whether the arguments produced by the agents reflected the social dimensions outlined in their system prompts. To achieve that, we used the pre-trained models from Choi et al. (2020) to score the presence of dimension d in the arguments generated by agents that were initiated with the same dimension d in their system prompts. Specifically, we computed a length-discounted version of the scores to ensure comparability across arguments of varying lengths, as suggested by Monti et al. (2022). We also computed the same score for the set of arguments produced by the baseline agent. Then, for each social dimension, we performed a t-test comparing the normalized scores of arguments from agents with personalities against those from the baseline agent. This statistical assessment was performed to verify our assumption that the arguments crafted by agents with an assigned dimension of social pragmatics significantly deviated from those produced by a baseline agent that did not receive any instruction on how to craft the argument. Statistically-significant deviations validate the agents' efficacy in expressing the intended social dimensions in their argumentation. We observed significant differences for all dimensions with p-values lower than 0.05, with the exception of *power*, having a p-value at 0.97. This could be attributed to several factors, including the limited number of *power* samples the classifier encountered during training, potentially leading to slightly unreliable predictions (Choi et al. 2020).

Persuading AI Agents. Figure 3 presents the probability of persuasion $p_s^d(persuasion)$ across different dimensions d and levels of stubbornness s.

There is an inverse association between the Skeptic's stubbornness and the probability of persuasion. On average across dimensions, the probability of persuasion of the moderately stubborn Skeptic decreases by 48% relative to the easily persuaded one, and by 73% when comparing the highly-resistant Skeptic to the moderately stubborn one. The relative ranking of the various dimensions remains largely consistent across different levels of stubbornness, with the notable exception of *knowledge* and *similarity*, that are notably less effective in convincing the hard Skeptic compared to other conditions.

Focusing on a moderate level of stubbornness, significant

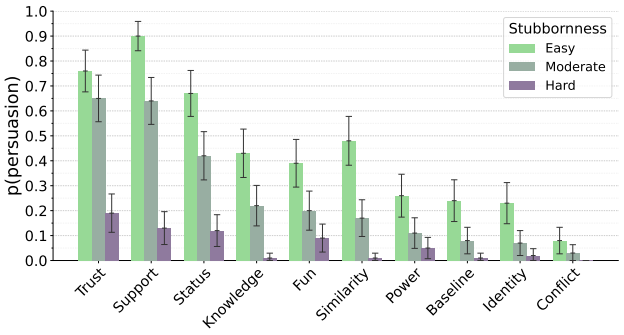

Figure 3: Probability of persuasion of arguments containing different dimensions of social pragmatics, across three levels of the Skeptic's stubbornness. Error bars mark the 95% confidence intervals.

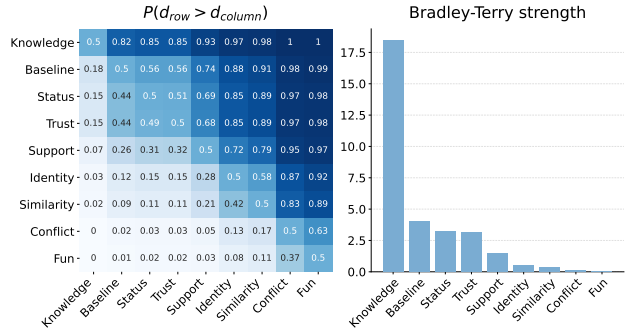

Figure 4: Bradley-Terry model results. *Left*: probability $P(d_{row} > d_{columns})$ that the dimension on the row was more effective than the dimension on the column. *Right*: the overall persuasive strength of arguments containing dimension d. These results were obtained considering only pairs of arguments enjoying a fraction of agreeing annotators of at least 0.8.

disparities across social dimensions become apparent. Persuasion strategies that convey *trust* or *support* are the only ones successful in altering the Skeptic's viewpoint in over half of the arguments. The third most effective dimension is *status*, with a probability hovering around 0.4. The efficacy of arguments gradually diminishes in the remaining dimensions, with *knowledge* being foremost. As anticipated by our preliminary tests, the performance of *power* closely aligns with the baseline due to them being hard to distinguish. Finally, *identity* and *conflict* are the only dimensions whose performance is below that of the baseline.

In line with previous research (summarized in Figure 2), our results corroborate the important role of *trust* and *support* in shaping opinion shifts. However, the influence of other social dimensions presents a more nuanced picture. Notably, conferring *status* enhances persuasiveness for LLMs, while it is not rewarding in the social media discourse. Furthermore, *knowledge* exchange was found to be the most effective driver of opinion shift, while it only ranked fourth in our experiment. Also in contrast with previous work, the dimensions *identity*, *conflict*, and *similarity* demonstrate low persuasion probabilities.

The context of opinion change on social media differs markedly from the controlled environment of our experiment, making it hard to directly compare them. To more accurately discern the similarities and differences between the impact of arguments on human and AI agents' opinions, we resorted to a direct human judgement of these arguments, as detailed next.

Persuading Humans. After excluding annotators who did not meet our quality standards (16 users who contributed 99 annotations), we were left with a total of 2061 argument pair annotations. The annotators achieved an inter-annotator agreement of 0.52 (Fleiss Kappa), and demonstrated very low failure rates on the control samples. We applied the Bradley-Terry model to this set of pairwise annotations and estimated the persuasive power of each individual dimension.

Figure 4 illustrates the estimated probability of $P(d_i > d_j)$, indicating the likelihood of dimension i being more effective than dimension j, and the rank of dimensions based on their effectiveness relative to others, according to the model. We excluded the dimension of *power* from the crowdsourcing study because of its lack of a significant difference from the baseline.

We generally observed a degree of overlap between human and LLM preferences, with some notable differences. Excluding the baseline, both rankings reveal a similar high-level structure: the dimensions of *knowledge*, *status*, *trust*, and *support* are ranked higher than the other dimensions. More subtle differences become apparent when examining the individual positions in the rank. Most notably, humans exhibit a significantly stronger preference for *knowledge* than LLMs do, with the Bradley-Terry score of human evaluations for *knowledge* being substantially larger than that of the second highest-ranked dimension. Both humans and LLMs assign considerable importance to the concepts of *status* and *trust* in persuasive arguments, while concurring on *conflict* and *identity* being less effective. *Fun* is ranked lower by humans than by LLMs, while *support* is deemed more effective by LLMs than by humans. The high weight placed on *knowledge* is in line with the ranking from previous work (Figure 2).

Interestingly, the *baseline* argument performed better according to human annotators compared to agents. This happened likely due to the baseline argument being most semantically similar to *knowledge* arguments than to any other type. To quantify that, we applied the embeddings from the pre-trained social dimensions classifier (Choi et al. 2020) to all the arguments, and then calculated the average cosine similarity of the embeddings of baseline arguments with arguments containing each dimension (Table 3). We found that baseline arguments have higher similarity with *knowledge* arguments (0.95) than with any other dimension (range $[0.60 - 0.77]$).

Last, we assessed the robustness of the human ranking by examining how the ranking was altered when annota-

Dimension	Cosine similarity
Knowledge	0.95
Trust	0.77
Fun	0.74
Status	0.70
Power	0.70
Support	0.69
Similarity	0.67
Identity	0.66
Conflict	0.60

Table 3: Cosine similarity between the embeddings of arguments from each social dimension against the baseline arguments.

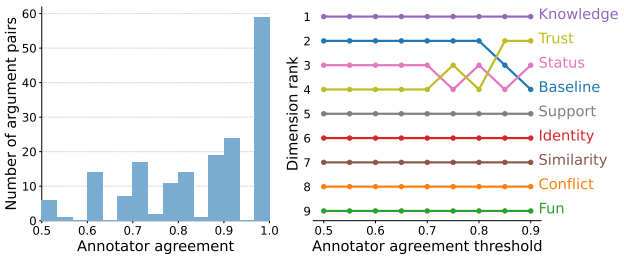

Figure 5: Sensitivity to annotator agreement of dimension ranking in human persuasion. *Left*: distribution of the fraction of annotators agreeing over argument pairs. *Right*: rank of dimensions obtained after when filtering out pairs with annotator agreement lower than a threshold.

tions with low inter-annotator agreement were excluded. Figure 5 (left) shows the distribution of agreement (calculated as fraction of annotators agreeing) across argument pairs. The agreement was typically high, with the majority of pairs having unanimous or near-unanimous consensus. We investigated the stability of the rankings by progressively excluding samples with low annotation agreement from the ranking algorithm. We began with a threshold of 0.5, increasing it in steps of 0.05 until reaching a maximum of 0.9. Using thresholds higher than 0.9 caused certain dimensions not to be represented, making us unable to produce a ranking using the Bradley-Terry method. At each stage, we recalculated the rankings of the social dimensions.

Figure 5 (right) shows the change in rankings when discarding low-agreement pairs. The *baseline* arguments decreased in ranking with increased thresholds, while *trust* achieved a higher rank, but overall the ranking was left almost unchanged. For all our analysis (including results shown in Figure 4) we used rankings and ranking strengths based on pairs with an agreement threshold of 0.8, to ensure high quality annotations.

Discussion

We introduced a framework for simulating opinion dynamics and persuasiveness using Large Language Models (LLMs) as agents. We presented a simple persuasion dialogue in which a Convincer agent generated arguments about the timely topic of climate change in the attempt of convincing a Skeptic agent. The Skeptic agent evaluated the arguments and determined whether it changed its internal opinion state. We experimented with various dialogue conditions, altering the level of stubbornness for the Skeptic, and prompting the Convincer to adopt social communicative strategies. Additionally, we asked human judges to evaluate persuasiveness of convincing arguments. Based on the human ranking of arguments, we compared whether arguments that are effective in changing the agents opinion were also perceived as persuasive by humans.

Key Findings

Building on early efforts to use LLMs for simulating social systems (Park et al. 2023; Li et al. 2023; Chuang et al. 2023), our research demonstrates that LLM agents can effectively mimic some of the dynamics of persuasion and opinion change that are typically observed in the human discourse (**RQ1**). These agents can be prompted to construct well-reasoned arguments, express a motivated opinion on a given topic that can be programmatically encoded into a binary variable, and modify their stance in a manner consistent with the personas assigned to them. The agents' receptiveness to accepting arguments can be easily adjusted. Most importantly, we have shown that these agents can generate persuasive arguments that incorporate dimensions of social pragmatics underpinning established psycho-linguistic theories of opinion change (**RQ2**). We have validated the presence of these dimensions in the output generated by the LLMs using an independently-trained classifier designed to detect them from text.

A key aspect of our study was to investigate whether synthetically-generated arguments have equivalent persuasive impacts on both LLM agents and humans (**RQ3**). We approached this by analysing the results of three distinct experiments: *i)* the proportion of arguments containing a specific dimension that were effective in dialogues between LLM agents, *ii)* an extensive set of crowdsourced annotations assessing the quality of machine-generated arguments, and *iii)* the efficacy of various argument types as determined by previous research on social media interactions (Monti et al. 2022). The outcomes of these three experiments showed partial alignment. Notably, arguments rich in factual *knowledge* and those attempting to establish *trust* between the dialogue participants were among the most effective across all three settings. Arguments offering emotional *support* and conveying *status* (i.e., respect, admiration) were also highly effective in both the LLM experiment and according to human evaluators. These parallels suggest that achieving a close alignment between the opinion dynamics of human and machine systems is within the reach of future research. However, two significant discrepancies were observed and deserve further investigation. First, human judges demonstrated a disproportionate preference for *knowledge*-based arguments compared to LLM agents. Second, opinion-changing messages on social media often pointed to the *similarity* of stances between the dialogue par-

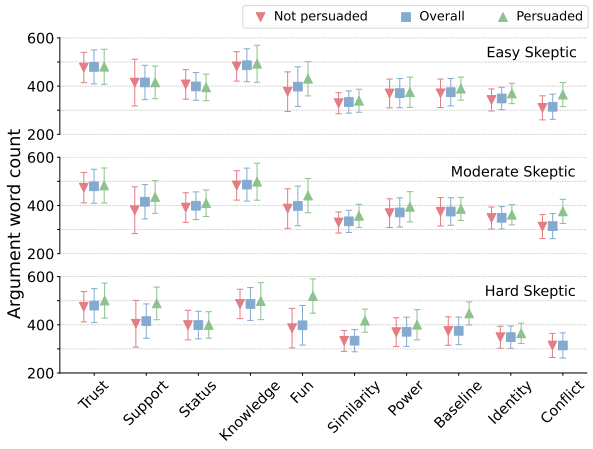

Figure 6: Sensitivity of persuasiveness of arguments to argument length. For each dimension and level of Skeptic's stubbornness, the average and standard deviation of length are shown. Statistics for all arguments, successful arguments, and unsuccessful arguments are shown.

ticipants, unlike in our study. These differences could be attributed to our simplified setup. For instance, the Convincer lacks knowledge about the Skeptic's profile, which makes it challenging to formulate a persuasive argument highlighting similarities between existing viewpoints.

Limitations and Future Work

While providing quantitative insights into the affinities between humans and artificial agents in argument processing and opinion formation, our study has limitations that open up multiple avenues for future research.

First, our experimental design, in its pursuit of simplicity, considered a one-off interaction between two agents on a single topic. To broaden the applicability of our findings, particularly in the context of social media interactions, future studies should consider multi-turn conversations among multiple agents and across a variety of topics. Agents engaging in evolving dialogues over multiple interactions, similar to the approach of Chuang et al. (2023), could potentially alter the persuasiveness of various social dimensions of pragmatics, possibly to the benefit of dimensions like *similarity* and *identity*. As dialogues progress and generate large amounts of text, constraints related to the limited input capacity of LLMs could be alleviated through cumulative or reflective memory, where agents either accumulate previous arguments over time or continuously summarize and integrate current and previous dialogues into their memory (Chuang et al. 2023; Park et al. 2023).

Second, to enhance ecological validity, one should diversify the profiles and expand the capabilities of individual agents. Agents could be designed to reflect different personalities, demographics, and social and cultural backgrounds, mimicking the diversity of human participants in a social system. This becomes particularly relevant as LLMs will be increasingly involved in public discussions on complex societal issues, ranging from environmental concerns to local and international politics. Future research could incorporate human-like biases (Levinson 1995), employ Retrieval-Augmented Generation (RAG) techniques to grant access to specific knowledge domains (Lewis et al. 2021), or enable agents to search the internet for arguments or information.

Third, our approach to designing effective prompts was primarily an iterative empirical process. The development of effective system prompts is a rapidly evolving practice, and while some studies have proposed guidelines and best practices for prompting (Ziems et al. 2023), a definitive consensus on optimal prompting strategies has yet to be reached. Experiments in specialized domains have demonstrated that carefully customized prompts can significantly enhance performance (Nori et al. 2023), stressing the value of further exploration in this area. Additionally, we opted for one of the best open-source LLMs, but exploring alternative models, could reveal different abilities in persuasiveness.

Fourth, comparing our LLM convincing probabilities with human rankings of social dimensions is challenging, as it is hard to recreate a setting in which humans and LLMs can operate under identical conditions. The human annotation process was specifically focused on pairs of arguments deemed convincing by the LLM, a selection criterion chosen to ensure fair comparisons. As a consequence, human rankings do not consider arguments that failed to convince the Skeptic. Future research could explore innovative methods to collect human judgements that more closely mirror how people judge arguments online. Furthermore, we acknowledge that our sample of annotators is partial and derived from a limited population.

Last, the mechanisms that induce LLM agents to signal a change of opinion remain unknown. Gathering evidence to elucidate this opinion-change process is crucial to the further development of these agents and to inform their use in online social contexts. A key question that has ignited debates in the scientific community is whether LLMs possess capabilities for reasoning and understanding (Floridi and Chiriatti 2020; Bubeck et al. 2023). If agents are found to lack a human-like understanding of the semantics of arguments, it would question the causal link between argument quality and opinion change. Although directly answering this question is challenging, experiments can be designed to measure outcome variations after controlling for possible confounders that might directly and more simply explain the outcome. As a proof of concept, we explored the link between argument effectiveness and a simple confounder: argument length. Figure 6 shows average word counts of the arguments across social dimensions and level of skeptic stubbornness, disaggregated by effectiveness of the argument (successful vs. not successful). Longer arguments are associated with the most persuasive dimensions. Within each dimension, successful arguments are slightly longer on average, particularly in the hard Skeptic scenario. This length bias might be a factor contributing to the difference of persuasiveness of social dimensions when evaluated by agents compared to humans. Systematically disentangling different factors that might influence the observed outcome is a complex endeavor that future work will need to address.

Social dimension	Convincer system prompt additions
Knowledge	Your arguments should be logical and based on facts. You should emphasize sharing information and insights regarding climate change. Teach me about the dangers of climate change.
Trust	Make use of ethos in your arguments. Try to create a mutual trust between us. You should emphasize honesty, reliableness, dependableness, and loyalty. Emphasize that you are willing to rely on my judgment.
Fun	Your arguments should be fun, witty, and humorous. Be playful and cheerful.
Status	Your arguments should confer status upon me. Show admiration and appreciation and be respectful.
Power	Assert your authority and show unwavering confidence. Your argument should conform to the social dimension of 'power', which is characterized by the keywords 'command', 'control', 'dominance', 'authority', 'pretentious', 'decisions'.
Support	Be supportive. Give emotional aid and offer companionship. Be friendly and caring, show sympathy, and offer encouragement.
Similarity	Only highlight our shared interests, motivations, and outlooks when it comes to climate change. Emphasize how we are similar and alike.
Identity	Your argument should conform to the social dimension of 'identity', which is characterized by the keywords 'community', 'united', 'identity', 'cohesive', 'integrated' and 'shared sense of belonging to the same community or group'.
Conflict	Argue like you are angry. Show that you heavily disagree with me. Your response should be negative.

Table 4: Social dimension personalities added to the Convincers system prompt.

Related Work

Research into the persuasive capabilities of generative AI spans a range of disciplines, from computer science to social and complexity sciences (Duerr and Gloor 2021). A subset of these studies have concentrated on human responses to machine-generated text. Karinshak et al. (2023) compared pro-vaccination messages generated by language models with those authored by humans, finding that LLM-based messages were perceived as more persuasive, unless clearly marked as AI-generated. Similarly, Bai et al. (2023) conducted a randomized control trial, exposing a diverse sample of individuals to persuasive policy commentaries either generated by LLMs or written by humans. They found both methods equally effective in altering the participants' levels of support the the policies. In the attempt of generating audience-specific messages, some studies have experimented with LLM role-playing, for example, prompting agents to respond as if they were part of a specific demographic or had a given personality profile (Hackenburg and Margetts 2023; Griffin et al. 2023; Matz et al. 2023). Results reported across studies have been mixed so far, with some studies emphasizing the importance of personalization, while others suggest that the persuasiveness lies primarily in the quality of the arguments presented.

A separate line of research has focused on characterizing interactions between LLM agents, without any human in the loop, with the primary goal of replicating the dynamics of human social agents with in-silico environments (Park et al. 2023). This research is motivated by the observation that LLM outputs can mimic responses from various human sub-populations, thereby serving as effective proxies for human cognitive behavior (Argyle et al. 2023; Lee et al. 2023; Simmons and Hare 2023). For example, LLM agents have been used to create social networks (De Marzo, Pietronero, and Garcia 2023), play repeated games such as the prisoner's dilemma (Akata et al. 2023; De Marzo, Pietronero, and Garcia 2023), and construct Agent-Based Models (ABMs) with the goal of improving the fidelity to human behavior of traditional stochastic ABMs (Bianchi and Squazzoni 2015). In ABM experiments, LLM agents, connected through a complex social network, update their opinions on a topic based on messages received from neighboring agents. While these ABMs reproduce some known non-linear dynamics of complex social systems (Li et al. 2023), unlike real social systems, LLM social networks tend to converge towards opinion states that are biased towards factual truth, likely due to their built-in safeguards (Chuang et al. 2023).

Our study builds upon this existing body of work, comparing human and synthetic responses to persuasive LLM content using different persuasion strategies.

Ethical Considerations

Deploying AI agents that can disguise as humans and perform acts of persuasion on social media is both a risk and an opportunity that recent technological development have made very concrete.

We recognize that agents have the ability to disseminate offensive and false information and impact opinion formation based on inaccurate knowledge. Additionally, we acknowledge that data generated by agents in this study might be incorrect or offensive. Dissemination of false information can be a consequence of hallucinations of the LLM, but also a deliberate strategy intentionally designed to manipulate human users (Park et al. 2023; Xi et al. 2023). This may potentially impact the process of opinion formation on critical societal issues at a large scale if agents were to be deployed on social platforms disguised as real users. Additionally, the social use of synthetic agents raises concerns regarding privacy and security; for example, agents may potentially manipulate users to disclose personal information.

Research on understanding how effective the arguments of LLM-powered agents can be is necessary to estimate risks, but it should be also complemented with research providing possible solutions to reduce those risks. There are

many challenges in understanding and combating the malicious use of LLMs to pollute the online public discourse. Studies on the malicious uses of generative AI on the Web are still in their infancy (Yang and Menczer 2023), and new methodologies to characterize this phenomenon are needed to track its evolution and understand its impact on societal phenomena such as online conflict and polarization. Recent research has shown that existing algorithmic solutions for misinformation detection work less effectively on AI-generated content (Zhou et al. 2023), and new methods are needed to accurately identify misbehaving synthetic actors.

Our study, while acknowledging the risks and potential misuses of agents, contributes positively to deepen our understanding of how LLMs can impact the dynamics of human societies. Such knowledge is necessary to advance an informed discourse on the ethical use of LLMs. Our study offers insights for platforms and regulatory bodies when formulating informed decisions to mitigate the risks described above. Furthermore, this line of work may additionally help understand the evolution of the use of AI-generated content in the wild, and its impact on dynamics of opinion formation, polarization, and online conflict. Our work is partly motivated by the opportunity of such impact being mostly positive, with LLMs acting as agents that make an ethical use of persuasive language to combat misinformation, promote altruistic behavior, and reduce the increasing levels of fragmentation in online social systems.

Even when deploying LLM-based agents for ethical purposes, trade-offs between the obtained benefit and the high level of power consumption required to run them should be carefully considered (Bender et al. 2021).

Code and Data Availability

All code and materials are available on GitHub: github.com/AndersGiovanni/persuasive-llms.

Acknowledgments

We acknowledge the support from the Carlsberg Foundation through the COCOONS project (CF21-0432). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Akata, E.; Schulz, L.; Coda-Forno, J.; Oh, S. J.; Bethge, M.; and Schulz, E. 2023. Playing repeated games with Large Language Models. *arXiv preprint arXiv:2305.16867*.

Argyle, L. P.; Bail, C. A.; Busby, E. C.; Gubler, J. R.; Howe, T.; Rytting, C.; Sorensen, T.; and Wingate, D. 2023. Leveraging AI for democratic discourse: Chat interventions can improve online political conversations at scale. *Proceedings of the National Academy of Sciences*, 120(41): e2311627120. Publisher: Proceedings of the National Academy of Sciences.

Austin, J. L. 1975. *How to do things with words*, volume 88. Oxford university press.

Bai, H.; Voelkel, J.; Eichstaedt, J.; and Willer, R. 2023. Artificial intelligence can persuade humans on political issues. *OSF Preprints*.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.

Bianchi, F.; and Squazzoni, F. 2015. Agent-based models in sociology. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(4): 284–306.

Bradley, R. A.; and Terry, M. E. 1952. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4): 324–345.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; Nori, H.; Palangi, H.; Ribeiro, M. T.; and Zhang, Y. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. ArXiv:2303.12712 [cs].

Burtell, M.; and Woodside, T. 2023. Artificial influence: An analysis of AI-driven persuasion. *arXiv preprint arXiv:2303.08721*.

Choi, M.; Aiello, L. M.; Varga, K. Z.; and Quercia, D. 2020. Ten Social Dimensions of Conversations and Relationships. In *Proceedings of The Web Conference 2020*, 1514–1525. ArXiv:2001.09954 [cs].

Chuang, Y.-S.; Goyal, A.; Harlalka, N.; Suresh, S.; Hawkins, R.; Yang, S.; Shah, D.; Hu, J.; and Rogers, T. T. 2023. Simulating Opinion Dynamics with Networks of LLM-based Agents. ArXiv:2311.09618 [physics].

De Marzo, G.; Pietronero, L.; and Garcia, D. 2023. Emergence of Scale-Free Networks in Social Interactions among Large Language Models. *arXiv:2312.06619*.

Deri, S.; Rappaz, J.; Aiello, L. M.; and Quercia, D. 2018. Coloring in the Links: Capturing Social Ties as They are Perceived. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW): 1–18. ArXiv:1902.04528 [cs].

Duerr, S.; and Gloor, P. A. 2021. Persuasive Natural Language Generation–A Literature Review. *arXiv preprint arXiv:2101.05786*.

Floridi, L.; and Chiriatti, M. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30: 681–694.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Griffin, L. D.; Kleinberg, B.; Mozes, M.; Mai, K. T.; Vau, M.; Caldwell, M.; and Marvor-Parker, A. 2023. Susceptibility to Influence of Large Language Models. *arXiv preprint arXiv:2303.06074*.

Habermas, J. 1979. *Communication and the Evolution of Society*. Beacon press.

Hackenburg, K.; and Margetts, H. 2023. Evaluating the persuasive influence of political microtargeting with large language models. *OSF Preprints*.

Hendrycks, D.; Mazeika, M.; and Woodside, T. 2023. An Overview of Catastrophic AI Risks. *arXiv preprint arXiv:2306.12001*.

Hugging Face. 2023. Inference for templates for chat models.

Jakesch, M.; Hancock, J. T.; and Naaman, M. 2023. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11): e2208839120.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Karinshak, E.; Liu, S. X.; Park, J. S.; and Hancock, J. T. 2023. Working With AI to Persuade: Examining a Large Language Model's Ability to Generate Pro-Vaccination Messages. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1): 1–29.

Kreps, S.; McCain, R. M.; and Brundage, M. 2022. All the news that's fit to fabricate: AI-generated text as a tool of media misinformation. *Journal of experimental political science*, 9(1): 104–117.

Lee, S.; Peng, T.-Q.; Goldberg, M. H.; Rosenthal, S. A.; Kotcher, J. E.; Maibach, E. W.; and Leiserowitz, A. 2023. Can Large Language Models Capture Public Opinion about Global Warming? An Empirical Assessment of Algorithmic Fidelity and Bias. *arXiv preprint arXiv:2311.00217*.

Levinson, S. C. 1995. Interactional Biases in Human Thinking. *Social Intelligence and Interaction*.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. ArXiv:2005.11401 [cs].

Li, C.; Su, X.; Han, H.; Xue, C.; Zheng, C.; and Fan, C. 2023. Quantifying the Impact of Large Language Models on Collective Opinion Dynamics. ArXiv:2308.03313 [cs].

Matz, S.; Teeny, J.; Vaid, S. S.; Harari, G. M.; and Cerf, M. 2023. The Potential of Generative AI for Personalized Persuasion at Scale. *PsyArXiv*.

Monti, C.; Aiello, L. M.; De Francisci Morales, G.; and Bonchi, F. 2022. The language of opinion change on social media under the lens of communicative action. *Scientific Reports*, 12(1): 17920. Number: 1 Publisher: Nature Publishing Group.

Nori, H.; Lee, Y. T.; Zhang, S.; Carignan, D.; Edgar, R.; Fusi, N.; King, N.; Larson, J.; Li, Y.; Liu, W.; et al. 2023. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. *arXiv preprint arXiv:2311.16452*.

Papachristou, M.; Yang, L.; and Hsu, C.-C. 2023. Leveraging Large Language Models for Collective Decision-Making. *arXiv preprint arXiv:2311.04928*.

Park, J. S.; O'Brien, J. C.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. ArXiv:2304.03442 [cs].

Pera, A.; Morales, G. d. F.; and Aiello, L. M. 2023. Measuring Behavior Change with Observational Studies: a Review. *arXiv preprint arXiv:2310.19951*.

Prakken, H. 2006. Formal systems for persuasion dialogue. *The knowledge engineering review*, 21(2): 163–188.

Sadasivan, V. S.; Kumar, A.; Balasubramanian, S.; Wang, W.; and Feizi, S. 2023. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.

Simmons, G.; and Hare, C. 2023. Large Language Models as Subpopulation Representative Models: A Review. *arXiv preprint arXiv:2310.17888*.

Touvron, Hugo et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv:2307.09288 [cs].

Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; Zhao, W. X.; Wei, Z.; and Wen, J.-R. 2023. A Survey on Large Language Model based Autonomous Agents. ArXiv:2308.11432 [cs] version: 1.

Weidinger, L.; Uesato, J.; Rauh, M.; Griffin, C.; Huang, P.-S.; Mellor, J.; Glaese, A.; Cheng, M.; Balle, B.; Kasirzadeh, A.; et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 214–229.

Xi, Z.; Chen, W.; Guo, X.; He, W.; Ding, Y.; Hong, B.; Zhang, M.; Wang, J.; Jin, S.; Zhou, E.; Zheng, R.; Fan, X.; Wang, X.; Xiong, L.; Zhou, Y.; Wang, W.; Jiang, C.; Zou, Y.; Liu, X.; Yin, Z.; Dou, S.; Weng, R.; Cheng, W.; Zhang, Q.; Qin, W.; Zheng, Y.; Qiu, X.; Huang, X.; and Gui, T. 2023. The Rise and Potential of Large Language Model Based Agents: A Survey. ArXiv:2309.07864 [cs].

Yang, K.-C.; and Menczer, F. 2023. Anatomy of an AI-powered malicious social botnet. *arXiv preprint arXiv:2307.16336*.

Zhang, M.; Press, O.; Merrill, W.; Liu, A.; and Smith, N. A. 2023. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*.

Zhou, J.; Zhang, Y.; Luo, Q.; Parker, A. G.; and De Choudhury, M. 2023. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–20.

Ziems, C.; Held, W.; Shaikh, O.; Chen, J.; Zhang, Z.; and Yang, D. 2023. Can Large Language Models Transform Computational Social Science? *arXiv:2305.03514*.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, in the Methods section.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, we discuss potential biases of our annotation samples in the "Limitations and future work" section.
   (e) Did you describe the limitations of your work? Yes, limitation are presented and discussed in the "Limitations and future work" subsection of Discussion.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, we discuss negative societal impact in the "Ethical considerations" section.
   (g) Did you discuss any potential misuse of your work? Yes, we discuss potential misuse in the "Ethical considerations" section.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, we share methodological framework in the Methods section, and full experimental details (code) are provided in the Supplementary Material.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes, we frame our experiments within the scope of previous research in the Methods section
   (b) Have you provided justifications for all theoretical results? Yes, justifications are provided in the Results and Discussion sections
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA.
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes, in the Results and Discussion sections we explore potential confounders of our results, and highlight limitations of our work.
   (e) Did you address potential biases or limitations in your theoretical framework? Yes, these are addressed in the Discussion section.
   (f) Have you related your theoretical results to the existing literature in social science? Yes, we directly relate our work to previous work in Computational Social Science and to Social Psychology theories, as detailed in Introduction and Methods.
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes, these are discussed in Discussion and Ethical considerations.
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA.
   (b) Did you include complete proofs of all theoretical results? NA.
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes. Included in Supplementary Material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA, as we do not train any model(s) but use existing.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes, we report error bars in Figure 3.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, we use models hosted on the Hugging Face API.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, specified in the Methods section.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes, creators of models used are cited in the Methods section.
   (b) Did you mention the license of the assets? Yes.
   (c) Did you include any new assets in the supplemental material or as a URL? Yes, generated data is included in the Supplementary Material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes. Our data does not include PII, but we discuss misinformation and offensive content in the Ethical considerations section.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR

(see FORCE11 (2020))? Yes, in Supplementary Information we document how we plan to make the data available in a way that is compliant with FAIR guidelines.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? NA.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? Yes, included in the Supplementary Material.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes, we specify potential risks of the task in the Methods section. The only mild risk crowdworkers were exposed to was reading some potentially offensive machine-generated text.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes, we specify the wage paid in the Crowdsourcing subsection under Methods.

(d) Did you discuss how data is stored, shared, and deidentified? NA.