Characterizing AI Manipulation Risks in Brazilian YouTube Climate Discourse

Wenchao Dong[1*], **Marcelo Sartori Locatelli**[1,2*], **Virgilio Almeida**[2†], **Meeyoung Cha**[1,3†]

[1]Max Planck Institute for Security and Privacy (MPI-SP), Bochum, Germany
[2]Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil
[3]Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea

Abstract

Climate change poses a global threat to public health, food security, and economic stability. Addressing it requires evidence-based policies and a nuanced understanding of how the threat is perceived by the public, particularly within visual social media, where narratives quickly evolve through voices of individuals, politicians, NGOs, and institutions. This study investigates climate-related discourse on YouTube within the Brazilian context, a geopolitically significant nation in global environmental negotiations. Through three case studies, we examine (1) which psychological content traits most effectively drive audience engagement, (2) the extent to which these traits influence content popularity, and (3) whether such insights can inform the design of *persuasive* synthetic campaigns such as climate denialism using recent generative language models. Another contribution of this work is the release of a large publicly available dataset of 226K Brazilian YouTube videos and 2.7M user comments on climate change. The dataset includes fine-grained annotations of persuasive strategies, theory of mind categorizations in user responses, and typologies of content creators. This resource can help support future research on digital climate communication and the ethical risk of algorithmically amplified narratives and generative media.

Datasets — https://doi.org/10.5281/zenodo.17551955

Introduction

The pursuit of collective peace and prosperity among nations has long been a key objective on a global scale. One embodiment of such ideals is the United Nations' Sustainable Development Goals (SDGs), which call on countries to tackle pressing global issues through a shared blueprint. Among them is the SDG #13, or "climate action," which has received considerable attention, particularly as extreme weather events become more frequent (National Academies of Sciences and Medicine 2016), posing a threat to public health, food security, and economic stability (Solomon and LaRocque 2019). In this context, understanding climate change discourse is increasingly important, as it can support efforts to increase public awareness on policies that address societal needs.

Understanding public discourse around climate change remains critical, particularly as social media platforms increasingly shape how climate narratives are communicated and received, subtly influencing public sentiment and policy conversations. Brazil plays a prominent role in these discussions, not only as a leading nation in the Global South, but also as a host of major climate summits such as the 30th United Nations Framework Convention on Climate Change (UNFCCC), or COP30. Its influence is further amplified by being home to a substantial portion of the Amazon rainforest and more than 10% of the world's freshwater resources. Together, these factors position Brazil as a key focal point for climate change-related research and policy development.

To characterize climate discourse in Brazil, we collected a large-scale dataset of Brazilian Portuguese YouTube content spanning 2019 to 2025, with a total of 226,775 videos and 2,756,165 user comments. As of 2025, YouTube reached approximately 68% of the Brazilian population, representing a significantly high penetration rate, and 86% of the country's 200 million residents are connected to the Internet (Kemp and Kepios 2025). This high level of digital engagement offers a valuable lens into how climate information is communicated and received by the public.

Building on recent advances in psychology-informed computational approaches to human decision-making (Binz et al. 2025), we apply psycholinguistic methods to examine how climate information is conveyed and received on YouTube. To characterize creator messaging, we annotate video content using 10 persuasion strategies, including emotional appeals that evoke fear or empathy, and logical arguments emphasizing cause-and-effect relationships. To infer potential mental states underlying viewer responses, we classify user comments according to seven Theory of Mind (ToM) categories. These include, for example, expressions of intentionality toward climate-related actions (*intention*) and beliefs about the legitimacy of climate change (*belief*). Drawing on these psychological traits, we present three case studies that (1) identify key influences shaping the Brazilian climate discourse, (2) assess the predictability of content popularity, and (3) examine how these insights can inform the automatic generation of persuasive messages. The full data analysis pipeline is illustrated in Figure 1.

[*]These authors contributed equally.
[†]Co-corresponding authors.

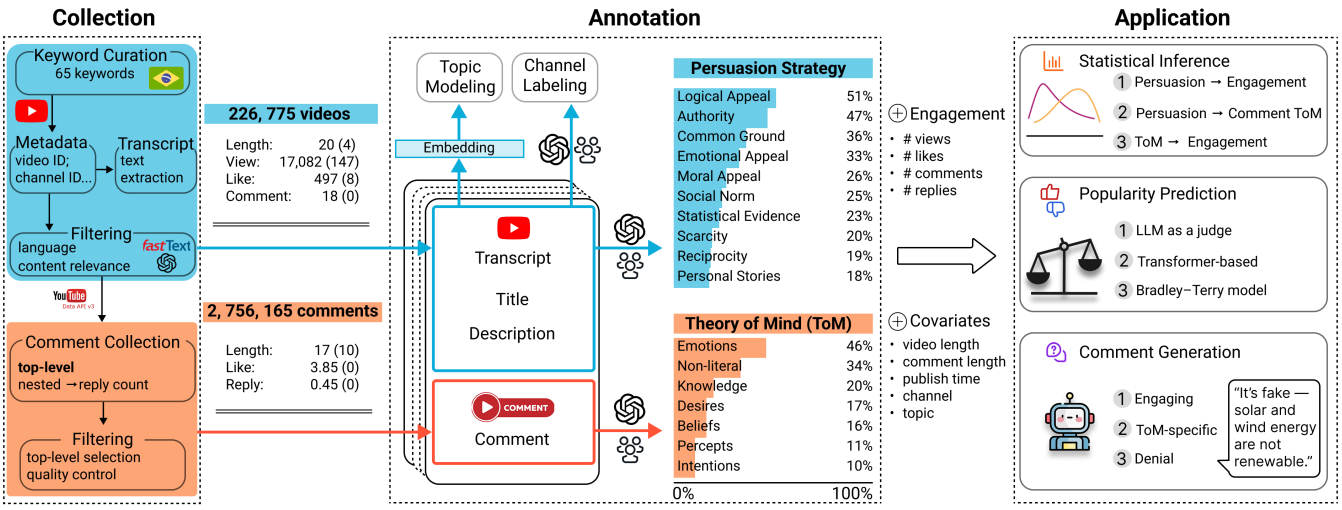

Figure 1: Proposed analytical framework includes data collection and preprocessing, annotations, and three case studies to understand the user engagement. Videos and comments' mean values, with median values in parentheses, are included. Ten persuasion strategies and seven theory of mind categories are ordered based on the distributions over the entire dataset.

Our study reveals that psychological features of climate discourse strongly influence audience engagement (i.e., views, comments) in Portuguese-language climate videos from Brazil. Emotionally framed messages consistently generated higher levels of user interaction, while logically or statistically driven content attracted comparatively less interaction. These dynamics raise concerns about the potential misuse of generative technologies such as large language models (LLMs) to produce *persuasive* synthetic comments. This generative capabilities can facilitate effortless reproduction of misleading narratives, including content promoting climate denialism.[1] Given the rising concerns over synthetic consensus in generative media (Schroeder et al. 2025), our study offers a psycholinguistic framework and annotated dataset to assess climate opinion manipulation in Brazil.

Related Work

Discourse Analysis

Social media plays a role in raising public awareness of climate issues (Mavrodieva et al. 2019), but it can also be used to spread misinformation that fuels climate skepticism, denial, and contrarianism (Treen, Williams, and O'Neill 2020). One study conducted a thematic analysis of TikTok videos and found that content creators tend to mention the topic through novel lenses, such as environment-conscious lifestyle (Galdeman and Aiello 2025). Chen et al. (2023) compared Twitter and news media to capture their different narrative styles. Moreover, multiple studies have explored how climate change features in political debates, suggesting emerging patterns of polarization (Shapiro and Park 2015; Jang and Hart 2015; Falkenberg et al. 2022). As for online information validity, research has documented that discussions are often dominated by a small group of highly active users (Shapiro and Park 2018), which may further contribute to the spread of conspiracy theories (Allgaier 2019).

Persuasion and Belief Adoption

Persuasion has long been a focus of social psychology and explains how decision-making can be shaped by external influences (Crano and Prislin 2006). Cialdini (2001) identified key factors, such as authority and social norms, that significantly affect individual choices. These mechanisms have proven effective in domains like marketing (Kumar et al. 2023) and charity (Wang et al. 2019). More recently, studies have highlighted the persuasive potential of AI-generated text (Breum et al. 2024; Salvi et al. 2025). Experiments show that LLM-written information can influence belief formation related to conspiratorial narratives (Costello, Pennycook, and Rand 2024; Hackenburg et al. 2025). This persuasive rhetoric raised further concerns about the tailored messages that are referred to as microtargeting (Tappin et al. 2023; Hackenburg and Margetts 2024). Conversely, automated persuasion has also been considered a tool to reduce skepticism about change (Czarnek et al. 2025).

Theory of Mind (ToM)

This concept was first introduced by Premack and Woodruff (1978) as the ability to impute mental states to themselves and others and was later identified as a core component of human cognition (Leslie, Friedman, and German 2004). Different categories of ToM are used to assess children's cognitive abilities (Lane et al. 2010; Beaudoin et al. 2020). Recognizing others' mental states enables the understanding of abstract concepts like beliefs in others and hence helps with decision-making (Frith and Singer 2008). Recent work has extended ToM to language models (Ma et al. 2023; Shapira et al. 2024) and demonstrated improvements in planning and

[1]Disclaimer: All examples are illustrative and our research framework is not intended to promote or amplify misinformation.

reasoning performance (Jung et al. 2024; Cross et al. 2025; Kim et al. 2025). In particular, LLMs have shown the ability to track others' mental states on par with humans (Kosinski 2024; Strachan et al. 2024).

Data Methodology

Data Collection

Before collecting climate-related data, we reviewed existing datasets in other languages to identify common keywords and methodologies. Most were English-language collections, primarily used in studies on climate misinformation and stance detection. A summary of relevant datasets is provided in the extended version (Dong et al. 2025a). Building on this foundation, we assembled a large dataset incorporating psychological attributes from *persuasion* and *Theory of Mind (ToM)* research, aiming to capture a broader range of mental states (Premack and Woodruff 1978).

Based on this literature survey, we selected 65 keywords that are commonly used as search queries to gather climate-related content (Salmi and Fleury 2022; Baltasar et al. 2024) and used them for YouTube Data API v3. We set the preferred language and location to 'Portuguese' and 'Brazil' for data collection, filtering out non-Portuguese videos. To remove irrelevant content that arises from keyword ambiguity, we used GPT-4.1-mini with a temperature of 0 to exclude videos with low climate relevance. We then collected all available transcripts and comments from the filtered videos. For live-streamed videos with comments that were published before the video itself, we adjusted the comment timestamps to match the video's publish time to better align engagement within the streaming period. Since nested comments often diverge from the video content (Cakmak, Agarwal, and Oni 2024), we included only top-level comments that are direct replies to video in our analysis. Further data descriptions are in (Dong et al. 2025b). Since YouTube engagement can vary with video length, we adopt a similar approach to (Violot et al. 2024) and define short videos as those under 3 minutes, based on the platform's official length threshold (YouTube Official Blog 2024).

Our dataset covers a seven-year period (2019–2025) and comprises 226,775 YouTube video metadata and 2,756,165 user comments. Figure 2 shows the number of new video posts over time. The number of long videos (i.e., those over 3 minutes) increased during the COVID-19 pandemic (between 2020 and 2022), but was soon overtaken by the rise of short-form videos, which have steadily gained traction among creators. Since 2023, the majority of videos on the climate change topic have been shared in short-form format.

Data Annotation

We use the following notation to describe our dataset. Let $\mathcal{D}$ denote the complete dataset, consisting of $V = \{v_1, v_2, \ldots, v_n\}$, the set of n videos, and $C = \{c_1, c_2, \ldots, c_m\}$, the set of m comments. We define a video-comment mapping $\phi : C \to V$, where $\phi(c_k) = v_i$ indicates that comment c_k belongs to video v_i. For each video v_i, let $C_i = \{c_k \in C : \phi(c_k) = v_i\}$ be the set of all comments linked to video v_i.

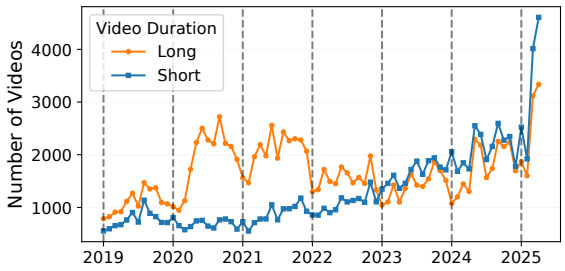

Figure 2: Monthly counts of climate-related Brazilian videos during 2019-2025. Vertical lines mark the start of each year.

Persuasion Traits To study how creators communicate the topic to their audience, we analyze psychological linguistic patterns, focusing on 10 persuasive strategies drawn from the literature (Cialdini 2001; Kumar et al. 2023; Costello, Pennycook, and Rand 2024):

- *Logical Appeal*: appealing with reasons and evidences
- *Emotional Appeal*: eliciting emotional feelings
- *Statistical Evidence*: providing concrete data, statistics
- *Social Norm*: creating pressure through social acceptance
- *Authority*: citing experts, institutions, and official reports
- *Personal Stories*: explaining individual experiences
- *Moral Appeal*: appealing with ethical responsibility
- *Reciprocity*: emphasizing mutual benefits of giving back
- *Scarcity*: presenting limited time and irreversible impacts
- *Common Ground*: building shared identity and values

Language persuasiveness or the ability to generate compelling, context-aware text across domains has received increasing attention since the emergence of LLMs (Breum et al. 2024; Salvi et al. 2025; Bai et al. 2025). These models can be instructed to apply specific persuasion techniques during text generation (Shi et al. 2021; Costello, Pennycook, and Rand 2024). Given that advanced LLMs can identify persuasion strategies in text (Jose and Greenstadt 2025), we used GPT-4.1 to annotate the presence of various persuasion strategies in video content. All inferences were performed using five-shot prompting with a temperature setting of 0. To assess annotation quality, we conducted a human evaluation and comparison, which yielded an average F1 score of 0.93 and an average accuracy of 0.98. For each video, we define $\mathbf{p}_i = (p_{i,1}, p_{i,2}, \ldots, p_{i,P}) \in \mathbb{R}^P$ as a vector representing the presence of persuasion strategies P, where $p_{i,j}$ indicates the presence of the persuasion strategy j in the video v_i.

Theory of Mind (ToM) Traits Another key psychological dimension we examine is rooted in ToM, the capacity to understand and interpret the mental states of others. Based on literature, we consider a wide range of ToM categories, which serve as analytical anchors to infer potential audience reactions to climate-related videos from a third-person perspective. In this study, we evaluate 7 distinct types of ToM, based on a widely recognized taxonomy (Beaudoin et al. 2020; Ma et al. 2023):

- *Belief*: information states that people hold to be true
- *Intention*: committed choices with planned actions
- *Desire*: motivational states representing preferences
- *Emotion*: affective states emotional responses
- *Knowledge*: organized representations of information
- *Percept*: sensory or socially shared perceptions
- *Non-literal*: using figurative or indirect language

We used GPT-4.1-mini to annotate ToM types via a five-shot prompting approach with a temperature of 0. To validate the annotations, we applied two methods: (1) randomly sampling comments and re-annotating them with a larger language model (e.g., GPT-4.1) to assess consistency and quality; and (2) manual annotation and comparison against GPT-4.1-mini results, which yielded an average F1 score of 0.66 and average accuracy of 0.83. For each comment c_k, we define $\mathbf{t}_k = (t_{k,1}, t_{k,2}, \ldots, t_{k,T}) \in \mathbb{R}^T$ as a vector of T ToM labels, where $t_{k,j}$ denotes the presence of the ToM aspect j in comment c_k.

Topic and Channel Modeling To better control for factors influencing video engagement (Dong et al. 2025b), we apply topic clustering and channel annotation. We cluster video content V into thematic groups using the unsupervised BERTopic algorithm (Grootendorst 2022). Channel labels were generated using GPT-4.1 through iterative prompt refinement and subsequently validated against human annotations (see our extended paper).

Discourse Manipulation

Building on the methodology above, we present three case studies that collectively suggest the potential risk of discourse manipulation using generative AI.

Case Study 1: Engagement Modeling

The first case study examines the relationship between psychological traits and audience engagement at two levels. The first is *video-level engagement*, defined by the number of likes and comments per video. The second is *comment-level engagement*, defined by the number of likes and replies per individual comment. To account for differences in view counts, we normalize the number of likes and comments by the number of views for each video v_i and analyze the *like ratio* L_i and *comment ratio* R_i instead. For each comment c_k, we denote the number of likes and replies as ℓ_k and r_k.

Psycholinguistic Traits We assess how psycholinguistic narratives in climate discourse influence user engagement through a three-stage evaluation process. First, we quantify the influence of persuasion strategies $\mathbf{p}_i$ on video-level metrics L_i and R_i using linear regression models. These models control for confounding factors such as video length and the publishing channel. Next, we further validate the effects of persuasion across the temporal span of our dataset and examine whether the conclusions hold across different video lengths. As a robustness check, we replicate these analyses within each discussion topic and channel.

Second, since each video contains multiple comments, we aggregate Theory of Mind (ToM) categories across all relevant comments and align the resulting summary with the video-level persuasion score. For each video v_i, we compute the aggregated ToM vector as $\bar{\mathbf{t}}_i = (1/|C_i|) \sum_{c_k \in C_i} \mathbf{t}_k$, where $|C_i|$ is the number of comments associated with v_i. We then compute the correlations between $\mathbf{p}_i$ and $\bar{\mathbf{t}}_i$, controlling for video length and channel subscriber count.

Third, user engagement at the comment level is modeled using regression, with the number of likes (ℓ_k) and replies (r_k) for each comment as the dependent variable, and the ToM score ($\mathbf{t}_k$) as the independent variable. Additional control variables include comment length, publishing channel, and the temporal gap between video publication and comment posting time.

Case Study 2: Popularity Prediction

The next case study evaluates how these traits influence video popularity, a key factor in assessing the feasibility of automated information manipulation (Wu, Rizoiu, and Xie 2018). In the context of YouTube video comments, for any pair of comments (c_i, c_j), we define the relative likes preference $y_{ij}^{(\ell)}$ as a binary indicator of whether comment c_i received more likes than comment c_j: $y_{ij}^{(\ell)} = \mathbb{I}[\ell_i > \ell_j]$, where $\mathbb{I}[\cdot]$ denotes the indicator function. The overall accuracy is computed as the average prediction correctness for each pair $A_{ij} = \mathbb{I}[y_{ij}^{(\ell)} = \hat{y}_{ij}^{(\ell)}]$, where $\hat{y}_{ij}^{(\ell)}$ is a model's prediction.

We evaluate the extent to which popularity is influenced by ToM. For each trait j, we partition the comment pairs based on whether the more popular comment c_i exhibits the trait ($\mathcal{P}_j^+$) or not ($\mathcal{P}_j^-$), given that ToM traits tend to be more prevalent among popular comments. We then compute the difference in prediction accuracy between the two sets:

$$(1/|\mathcal{P}_j^+|) \sum_{(c_i,c_j)\in\mathcal{P}_j^+} A_{ij} - (1/|\mathcal{P}_j^-|) \sum_{(c_i,c_j)\in\mathcal{P}_j^-} A_{ij} \quad (1)$$

We finally design multiple experimental conditions to constrain data sampling under varying levels of contextual information.

Computational Models We use three methods to predict the relative popularity of video comments. First, we adopt LLM-as-a-judge framework (Zheng et al. 2023), using five models: OpenAI GPT-4.1 and o4-mini, Microsoft Phi 4; Meta Llama-3.1-8B and Llama-4-Maverick. Second, we fine-tune encoder-only pretrained language models: Brazilian Portuguese BERTimbau (Souza, Nogueira, and Lotufo 2020) and multilingual DeBERTa V3 (He, Gao, and Chen 2021). Third, we implement the Bradley-Terry model (Bradley and Terry 1952; Ye et al. 2025b). In this method, we compute for each comment c_k an embedding representation $\mathbf{e}_k = \text{Embed}(c_k) \in \mathbb{R}^d$ and train a linear classifier to estimate the relative popularity score. The probability that comment c_i is more popular than comment c_j is defined as follows, with a threshold of 0.5 used to binarize the probabilities for consistent comparisons:

$$\frac{\exp(\beta_{i,\ell})}{\exp(\beta_{i,\ell}) + \exp(\beta_{j,\ell})}, \quad \text{where} \quad \beta_{i,\ell} = \text{Linear}(\mathbf{e}_i) \quad (2)$$

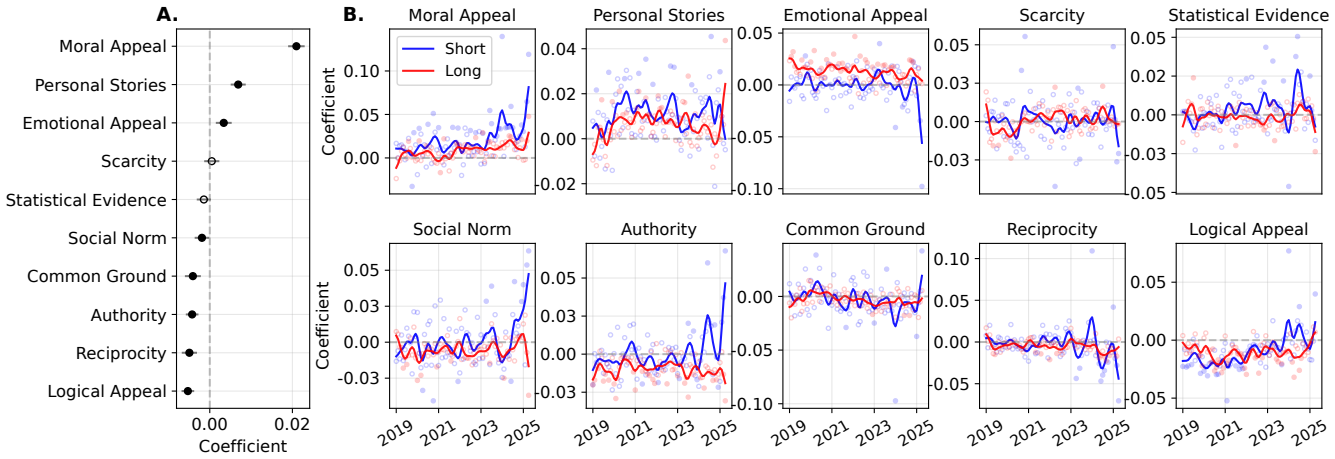

Figure 3: Effect of different persuasion strategies on the like ratio for (A) all videos and (B) monthly trends by video length. Solid points indicate statistically significant regression coefficients at the 0.05 level.

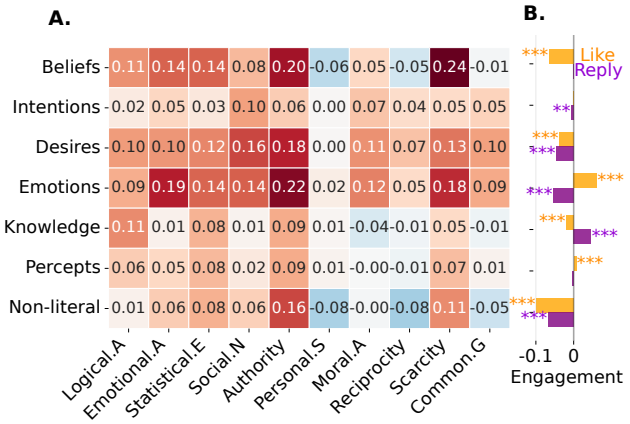

Figure 4: (A) Pearson correlations at the video level between 10 persuasion strategies and 7 ToM categories. (B) Effects of ToM mental states on audience engagement of likes and replies. *Note.* $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Case Study 3: Comment Generation

The final case study demonstrates how persuasive climate-related comments can be synthetically generated by leveraging the patterns identified earlier as a hypothetical risk scenario. We follow the experimental setup in Ye et al. (2025a) to fine-tune Llama-3-8B (Grattafiori et al. 2024) for comment generation. To estimate the quality of a generated comment c_{gen}, we retrieve K most semantically similar comments from C_{ref} using cosine similarity.

$$\hat{\ell}_{\text{gen}|K} = \frac{1}{K} \sum_{c' \in \mathcal{N}_K(c_{\text{gen}})} \ell_{c'}, \quad \text{where}$$
$$\mathcal{N}_K(c_{\text{gen}}) = \arg \max_{S \subseteq C_{\text{ref}}, |S|=K} \sum_{c' \in S} \cos(\mathbf{e}_{c_{\text{gen}}}, \mathbf{e}_{c'}) \tag{3}$$

Targeted Profiles We construct fine-tuning datasets in three scenarios and, for each, train two model variants: a *likable model* trained on the top 10% of most liked comments, and a *baseline model* trained on randomly selected comments. For similarity evaluations, we use all comments from each scenario, excluding those used for training.

First, we sample videos categorized by distinct persuasion strategies to control for video-level effects while maintaining content diversity. Second, we sample comments that match a target ToM profile $\mathbf{t}_{\text{target}} \in \mathbb{R}^T$ to generate comments that reflect specific mental states. Third, we further subcategorize the "Belief" ToM category into distinct stances, including climate change belief and climate change denial. To simulate extreme scenarios, we filter denial-related comments using targeted keywords and fine-tune an *extreme denial model* capable of generating content aligned with strong climate skepticism.

Experimental Results

Engagement Modeling

Figure 3(A) illustrates the effects of persuasion on video likes, suggesting that certain strategies consistently increase user engagement. The three most commonly used persuasion strategies—logical appeal (51% of climate-related videos), authority (47%) and common ground (36%)—are each associated with a statistically lower audience engagement (see the distribution in Figure 1). In contrast, emotional (33%) and moral (26%) appeals are linked to significantly higher levels of interaction, with morality-focused content emerging as the most effective strategy, producing an average increase of 2.1% in video likes.

To examine temporal and format-specific trends, we disaggregate effects by video length. Figure 3(B) shows that the persuasive effect of moral rhetoric in short-form videos has increased over time, accompanied by a rise in authority-driven strategies. Here is an example of an authority appeal used in climate denialism, translated from Portuguese:

> *"Climatologist Ricardo Felício stated in an interview that global warming is a hoax."*

	Rand.	Vid.	Date	Len.
GPT-4.1				
Base	$0.75_{.040}$	$0.77_{.005}$	$0.75_{.071}$	$0.69_{.062}$
+ CO	$0.76_{-.010}$	$0.81_{.040}$	$0.77_{.095}$	$0.76_{.113}$
o4-mini				
Base	$0.69_{.085}$	$0.73_{.024}$	$0.71_{.073}$	$0.62_{.110}$
+ CO	$0.73_{.115}$	$0.74_{.093}$	$0.71_{.143}$	$0.64_{.130}$
LLaMA-3.1				
Base	$0.57_{.037}$	$0.59_{-.013}$	$0.58_{-.005}$	$0.43_{-.011}$
+ CO	$0.54_{-.012}$	$0.63_{-.008}$	$0.62_{.024}$	$0.55_{.022}$
LLaMA-4				
Base	$0.68_{.072}$	$0.69_{.045}$	$0.69_{.125}$	$0.60_{.039}$
+ CO	$0.70_{.112}$	$0.71_{.067}$	$0.69_{.094}$	$0.62_{.109}$
Phi-4				
Base	$0.72_{.054}$	$0.74_{.022}$	$0.73_{.076}$	$0.66_{.049}$
+ CO	$0.68_{.016}$	$0.75_{.012}$	$0.75_{.039}$	$0.67_{.016}$
BERTimbau				
Base	$0.88_{.001}$	$0.84_{.021}$	$0.82_{-.001}$	$0.49_{-.033}$
+ CO	$0.86_{.015}$	$0.83_{.022}$	$0.81_{.016}$	$0.76_{-.019}$
DeBerta V3				
Base	$0.81_{.013}$	$0.79_{-.023}$	$0.81_{-.004}$	$0.75_{-.060}$
+ CO	$0.84_{.031}$	$0.81_{-.015}$	$0.84_{.017}$	$0.49_{-.066}$
Bradley–Terry				
Base	$0.78_{.002}$	$0.76_{.002}$	$0.75_{.031}$	$0.67_{.050}$
+ CO	$0.73_{.020}$	$0.68_{-.045}$	$0.66_{.067}$	$0.64_{-.010}$

Table 1: Relative likability prediction results, with subscripts indicating improvements attributable to emotional ToM. "CO" denotes the inclusion of video context; "Rand." refers to random comment pairs; "Vid." indicates comments from the same video; "Date" represents comments posted within a similar timeframe; "Len." refers to similar comment lengths. See the extended version for additional results.

These influences vary by publishing channels. The less effective trait, social norm, is perceived positively when videos are posted by international organizations. Conversely, the overall effective trait, emotional appeals, does not extend to videos posted by scientific research institutes:

> *"Architect and urban planner Eduardo Pizarro, a CAJU member, showed how COVID-19 has most intensely affected the outskirts of the city of São Paulo."*

Although logical and statistical appeals generally receive lower levels of interaction, they tend to be more effective when delivered through national government channels. Below is an example with logical appeals:

> *"If urgent measures are not taken, the planet's global temperature could rise by up to three degrees by the end of the 21st century. Therefore, ..."*

Discussion topics further influence the effectiveness of informational appeals. Consistent with broader trends, sharing personal experiences and emphasizing moral values in DIY-related videos are associated with higher audience engagement, whereas logical and statistical narratives tend to reduce interaction. Although common ground generally has a negative impact, it can enhance engagement in sustainability-related themes:

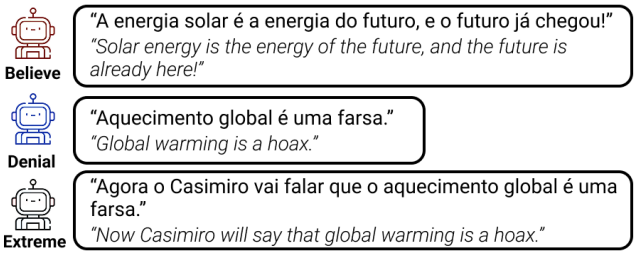

Figure 5: Sampled Portuguese comments generated by Believe, Denial, and Extreme models, with translations below.

> *"It is important that we have a common understanding of what sustainability means, which involves balancing environmental, economic, and social issues."*

When videos address climate change in specific geolocations, scarcity emerges as a strong predictor of comment engagement, even though it has minimal influence overall:

> *"Sahara Desert was once a place with a huge forest, crisscrossed by rivers and lakes, and inhabited by a variety of animals."*

Persuasion strategies also influence viewers' mental states. Figure 4(A) shows correlations between 10 persuasion strategies and 7 ToM categories. Authority-based messaging tends to elicit stronger ToM responses ($ps < .001$ for all seven categories on two-tailed t-tests), whereas narratives centered on personal experiences generate fewer cognitively reflective comments. Belief and emotion-related mental states exhibit greater variability compared to intention and percept, suggesting that they are more susceptible to change under different strategies. Figure 4(B) further illustrates that emotion-oriented comments are more likely to be liked, while informative comments attract more replies.

Popularity Prediction

Table 1 compares the accuracy of five LLMs, two encoder-based models, and a statistical approach across various experimental conditions. GPT-4.1 achieves the best performance among LLMs (81% accuracy) when provided with the video context. Surprisingly, BERTimbau, a Brazilian Portuguese encoder model, achieves 88% accuracy even without contextual information, suggesting that the content of comments alone is often sufficient to predict engagement. In addition, incorporating ToM mental states from user comments enhances the accuracy of engagement tendency predictions. Table 1 shows the likability prediction gaps with and without emotional ToM narratives, showing that emotional ToM leads to an average improvement of 4.69% in predictive performance.

Comment Generation

Fine-tuned LLMs show the ability to generate persuasive comments. Table 2 presents the results of similarity-based proxy evaluations. The *Baseline (Emotion)* model, trained

	Baseline (Emotion)	**Engaging** (Emotion)	**ToM** (Intention)	**Believe**	**Denial**	**Extreme**
K	$\hat{\ell}_{\text{gen}\|K}$ / $\hat{r}_{\text{gen}\|K}$ / S.	$\hat{\ell}_{\text{gen}\|K}$ / $\hat{r}_{\text{gen}\|K}$ / S.	$\hat{\ell}_{\text{gen}\|K}$ / $\hat{r}_{\text{gen}\|K}$ / S.	$\hat{\ell}_{\text{gen}\|K}$ / $\hat{r}_{\text{gen}\|K}$ / S.	$\hat{\ell}_{\text{gen}\|K}$ / $\hat{r}_{\text{gen}\|K}$ / S.	$\hat{\ell}_{\text{gen}\|K}$ / $\hat{r}_{\text{gen}\|K}$ / S.
1	2.20 / 0.25 / 0.89	7.25 / 0.59 / 0.79	2.26 / 0.39 / 0.72	3.23 / 0.42 / 0.77	1.91 / 0.38 / 0.85	2.37 / 0.41 / 0.77
2	2.07 / 0.25 / 0.88	4.88 / 0.43 / 0.78	2.13 / 0.37 / 0.71	3.68 / 0.45 / 0.76	1.56 / 0.36 / 0.85	2.60 / 0.43 / 0.76
3	2.51 / 0.28 / 0.88	4.90 / 0.47 / 0.78	2.20 / 0.39 / 0.71	3.42 / 0.46 / 0.75	1.84 / 0.38 / 0.84	2.88 / 0.45 / 0.75
4	2.48 / 0.27 / 0.87	7.96 / 0.56 / 0.77	2.06 / 0.38 / 0.70	3.49 / 0.47 / 0.75	2.49 / 0.59 / 0.84	3.01 / 0.48 / 0.75
5	2.58 / 0.29 / 0.87	7.44 / 0.56 / 0.77	2.00 / 0.37 / 0.70	3.30 / 0.47 / 0.74	2.48 / 0.56 / 0.84	2.83 / 0.49 / 0.74

Table 2: Evaluation results for generated comments, where $\hat{\ell}_{\text{gen}|K}$ and $\hat{r}_{\text{gen}|K}$ denote estimated like and reply counts respectively based on K retrieved samples. S. represents the average similarity score among the K retrieved samples.

on randomly sampled comments, shows limited engagement effectiveness ($\hat{\ell}_{\text{gen}|1} = 2.20$). In contrast, the *Engaging (Emotion)* model, fine-tuned on top-liked comments, achieves three times more engagement ($\hat{\ell}_{\text{gen}|1} = 7.25$). The model trained to generate comments with intentionality performs similarly to the baseline ($\hat{\ell}_{\text{gen}|1} = 2.26$).

To assess the risks of opinion manipulation, we further examine belief-specific comment generation The *Denial* model ($\hat{\ell}_{\text{gen}|1} = 1.91$) produces less engaging content than the *Believe* model ($\hat{\ell}_{\text{gen}|1} = 3.23$). However, when fine-tuned on extreme climate denial narratives, the *Extreme* model ($\hat{\ell}_{\text{gen}|1} = 2.37$) generates comments that potentially attract more likes and replies. Figure 5 shows representative generated samples from the three models, showing that the *Extreme* model tends to include more details and rhetorical intensity, making its content more engaging.

Discussions and Conclusions

Factors of User Engagement

We analyzed Brazilian YouTube videos on climate change and presented a detailed characterization of user engagement patterns using psycholinguistic traits. Our large-scale annotations reveal that content employing moral and emotional rhetoric consistently drives higher viewer engagement (Figure 3), a pattern that also extends to interactions within video comments (Figure 4). These findings echo prior research on affective content (Lerner and Keltner 2000; Han, Cha, and Lee 2020). Importantly, our data confirm the distinct role of content source in shaping audience response: emotional expressions from research institute accounts received lower engagement compared to those from individual creators. This variation suggests the need to interpret the audience engagement dynamics within the specific communicative contexts in which content is produced.

While facts and figures are commonly viewed as effective tools for belief change (Costello, Pennycook, and Rand 2024), our findings suggest that such approaches were less prominent and less engaging within the YouTube ecosystem. Strategies emphasizing statistical evidence on climate change did generate some initial interest—particularly when presented by official government channels—but their overall reach remained limited in YouTube's short-form, algorithmically curated environment. As users increasingly favor short-form content, opportunities for disseminating detailed, fact-based information may have become constrained. These dynamics raise concerns about fact-checking efficiencies.

Implications on Opinion Manipulation

This study was motivated by the potential risk that generative AI could be misused to manipulate public opinion. Specifically, we sought real-world examples of possibilities of such manipulation in climate conversations in the Global South. We presented case studies, one demonstrating that comment popularity can be accurately predicted and that psychological traits further enhance performance (Table 1); and another showing that persuasive comments can be generated using targeted profiles via fine-tuned LLMs (Table 2). In a concerning example, a model fine-tuned on climate denialism content was capable of generating highly engaging comments that could potentially disseminate climate misinformation automatically and at scale (Figure 5).

The escalating volume of AI-generated content, coupled with its growing difficulty in human detection of such content (Jakesch, Hancock, and Naaman 2023; Dugan et al. 2023), intensifies the risk of automated opinion manipulation (Spitale, Biller-Andorno, and Germani 2023). Climate discourse is no exception to this risk, and this calls for the urgent need for governance around synthetic media. The rapid proliferation of short-form videos further complicates the challenge of verifying content authenticity (as shown in Figure 2), as swarms of synthetic narratives can be tailored to diverse user preferences and potentially create a misleading sense of consensus around socially sensitive issues (Schroeder et al. 2025).

Limitations

This study has several limitations. First, it focused solely on textual content, leaving out multimodal elements that may influence persuasion. Second, the engagement metrics used did not consider external influences such as recommendation algorithms or individual psychological variability. Lastly, findings are based on Brazilian Portuguese-language videos, which may limit generalizability to other contexts.

Acknowledgments

This project was supported by the Microsoft Accelerate Foundation Models Research (AFMR) program. The authors thank Pedro Henrique Alves dos Santos, Raul Ferreira da Cruz Neto, and anonymous reviewers for their feedback.

References

Allgaier, J. 2019. Science and environmental communication on YouTube: Strategically distorted communications in online videos on climate change and climate engineering. *Frontiers in Communication*, 4: 36.

Bai, H.; Voelkel, J. G.; Muldowney, S.; Eichstaedt, J. C.; and Willer, R. 2025. LLM-generated messages can persuade humans on policy issues. *Nature Communications*, 16(1): 6037.

Baltasar, C.; D'Antonio Maceiras, S.; Martín, A.; and Camacho, D. 2024. Analysis of Climate Change Misleading Information in TikTok. In *CEUR Workshop Proceedings*, volume 3782, 54–61.

Beaudoin, C.; Leblanc, É.; Gagner, C.; and Beauchamp, M. H. 2020. Systematic review and inventory of theory of mind measures for young children. *Frontiers in Psychology*, 10: 2905.

Binz, M.; Akata, E.; Bethge, M.; Brändle, F.; Callaway, F.; Coda-Forno, J.; Dayan, P.; Demircan, C.; Eckstein, M. K.; Éltető, N.; et al. 2025. A foundation model to predict and capture human cognition. *Nature*, 1–8.

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.

Breum, S. M.; Egdal, D. V.; Mortensen, V. G.; Møller, A. G.; and Aiello, L. M. 2024. The persuasive power of large language models. In *proc. of the ICWSM*, volume 18, 152–163.

Cakmak, M. C.; Agarwal, N.; and Oni, R. 2024. The bias beneath: analyzing drift in YouTube's algorithmic recommendations. *Social Network Analysis and Mining*, 14(1): 171.

Chen, K.; Molder, A. L.; Duan, Z.; Boulianne, S.; Eckart, C.; Mallari, P.; and Yang, D. 2023. How climate movement actors and news media frame climate change and strike: Evidence from analyzing twitter and news media discourse from 2018 to 2021. *The International Journal of Press/Politics*, 28(2): 384–413.

Cialdini, R. B. 2001. The science of persuasion. *Scientific American*, 284(2): 76–81.

Costello, T. H.; Pennycook, G.; and Rand, D. G. 2024. Durably reducing conspiracy beliefs through dialogues with AI. *Science*, 385(6714): eadq1814.

Crano, W. D.; and Prislin, R. 2006. Attitudes and persuasion. *Annu. Rev. Psychol.*, 57(1): 345–374.

Cross, L.; Xiang, V.; Bhatia, A.; Yamins, D. L.; and Haber, N. 2025. Hypothetical Minds: Scaffolding Theory of Mind for Multi-Agent Tasks with Large Language Models. In *proc. of the ICLR*.

Czarnek, G.; Orchinik, R.; Lin, H.; Xu, H. G.; Costello, T.; Pennycook, G.; and Rand, D. G. 2025. Addressing climate change skepticism and inaction using human-AI dialogues.

Dong, W.; Locatelli, M. S.; Almeida, V.; and Cha, M. 2025a. Characterizing AI Manipulation Risks in Brazilian YouTube Climate Discourse. *arXiv preprint arXiv:2511.06091*.

Dong, W.; Locatelli, M. S.; Almeida, V.; and Cha, M. 2025b. Characterizing Persuasion Patterns in Climate Discourse on Brazilian Portuguese YouTube Videos. In *proc. of the ACM GoodIT*.

Dugan, L.; Ippolito, D.; Kirubarajan, A.; Shi, S.; and Callison-Burch, C. 2023. Real or fake text?: Investigating human ability to detect boundaries between human-written and machine-generated text. In *proc. of the AAAI*, volume 37, 12763–12771.

Falkenberg, M.; Galeazzi, A.; Torricelli, M.; Di Marco, N.; Larosa, F.; Sas, M.; Mekacher, A.; Pearce, W.; Zollo, F.; Quattrociocchi, W.; et al. 2022. Growing polarization around climate change on social media. *Nature Climate Change*, 12(12): 1114–1121.

Frith, C. D.; and Singer, T. 2008. The role of social cognition in decision making. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1511): 3875–3886.

Galdeman, A.; and Aiello, L. M. 2025. Mapping the Climate Change Landscape on TikTok. In *proc. of the ICWSM*, volume 19, 2614–2621.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Hackenburg, K.; and Margetts, H. 2024. Evaluating the persuasive influence of political microtargeting with large language models. *PNAS*, 121(24): e2403116121.

Hackenburg, K.; Tappin, B. M.; Hewitt, L.; Saunders, E.; Black, S.; Lin, H.; Fist, C.; Margetts, H.; Rand, D. G.; and Summerfield, C. 2025. The Levers of Political Persuasion with Conversational AI. *arXiv preprint arXiv:2507.13919*.

Han, J.; Cha, M.; and Lee, W. 2020. Anger contributes to the spread of COVID-19 misinformation. *Harvard Kennedy School Misinformation Review*, 1(3).

He, P.; Gao, J.; and Chen, W. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Jakesch, M.; Hancock, J. T.; and Naaman, M. 2023. Human heuristics for AI-generated language are flawed. *PNAS*, 120(11): e2208839120.

Jang, S. M.; and Hart, P. S. 2015. Polarized frames on "climate change" and "global warming" across countries and states: Evidence from Twitter big data. *Global Environmental Change*, 32: 11–17.

Jose, J.; and Greenstadt, R. 2025. LLMs for Detection and Classification of Persuasion Techniques in Slavic Parliamentary Debates and Social Media Texts. In *proc. of the Slavic NLP*, 202–216.

Jung, C.; Kim, D.; Jin, J.; Kim, J.; Seonwoo, Y.; Choi, Y.; Oh, A.; and Kim, H. 2024. Perceptions to Beliefs: Exploring Precursory Inferences for Theory of Mind in Large Language Models. In *proc. of the EMNLP*, 19794–19809.

Kemp, S.; and Kepios. 2025. Digital 2025: Brazil. https://datareportal.com/reports/digital-2025-brazil. Accessed: 2025-07-29.

Kim, H.; Sclar, M.; Zhi-Xuan, T.; Ying, L.; Levine, S.; Liu, Y.; Tenenbaum, J. B.; and Choi, Y. 2025. Hypothesis-Driven Theory-of-Mind Reasoning for Large Language Models. In *proc. of the COLM*.

Kosinski, M. 2024. Evaluating large language models in theory of mind tasks. *PNAS*, 121(45): e2405460121.

Kumar, Y.; Jha, R.; Gupta, A.; Aggarwal, M.; Garg, A.; Malyan, T.; Bhardwaj, A.; Shah, R. R.; Krishnamurthy, B.; and Chen, C. 2023. Persuasion strategies in advertisements. In *proc. of the AAAI*, volume 37, 57–66.

Lane, J. D.; Wellman, H. M.; Olson, S. L.; LaBounty, J.; and Kerr, D. C. 2010. Theory of mind and emotion understanding predict moral development in early childhood. *British Journal of Developmental Psychology*, 28(4): 871–889.

Lerner, J. S.; and Keltner, D. 2000. Beyond valence: Toward a model of emotion-specific influences on judgement and choice. *Cognition & Emotion*, 14(4): 473–493.

Leslie, A. M.; Friedman, O.; and German, T. P. 2004. Core mechanisms in 'theory of mind'. *Trends in Cognitive Sciences*, 8(12): 528–533.

Ma, Z.; Sansom, J.; Peng, R.; and Chai, J. 2023. Towards A Holistic Landscape of Situated Theory of Mind in Large Language Models. In *Findings of the EMNLP*, 1011–1031.

Mavrodieva, A. V.; Rachman, O. K.; Harahap, V. B.; and Shaw, R. 2019. Role of social media as a soft power tool in raising public awareness and engagement in addressing climate change. *Climate*, 7(10): 122.

National Academies of Sciences, E.; and Medicine. 2016. *Attribution of Extreme Weather Events in the Context of Climate Change*. Washington, DC: The National Academies Press. ISBN 978-0-309-38094-2.

Premack, D.; and Woodruff, G. 1978. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4): 515–526.

Salmi, F.; and Fleury, L. C. 2022. Mudanças climáticas e Ciências Sociais: análise bibliométrica do campo (2011-2021). *Bib: revista brasileira de informação bibliográfica em ciências sociais.*, 1(97): 1–19.

Salvi, F.; Horta Ribeiro, M.; Gallotti, R.; and West, R. 2025. On the conversational persuasiveness of GPT-4. *Nature Human Behaviour*, 1–9.

Schroeder, D. T.; Cha, M.; Baronchelli, A.; Bostrom, N.; Christakis, N. A.; Garcia, D.; Goldenberg, A.; Kyrychenko, Y.; Leyton-Brown, K.; Lutz, N.; et al. 2025. How Malicious AI Swarms Can Threaten Democracy. *arXiv preprint arXiv:2506.06299*.

Shapira, N.; Levy, M.; Alavi, S. H.; Zhou, X.; Choi, Y.; Goldberg, Y.; Sap, M.; and Shwartz, V. 2024. Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models. In *proc. of the EACL*, 2257–2273.

Shapiro, M. A.; and Park, H. W. 2015. More than entertainment: YouTube and public responses to the science of global warming and climate change. *Social Science Information*, 54(1): 115–145.

Shapiro, M. A.; and Park, H. W. 2018. Climate change and YouTube: Deliberation potential in post-video discussions. *Environmental Communication*, 12(1): 115–131.

Shi, W.; Li, Y.; Sahay, S.; and Yu, Z. 2021. Refine and Imitate: Reducing Repetition and Inconsistency in Persuasion Dialogues via Reinforcement Learning and Human Demonstration. In *Findings of the EMNLP*, 3478–3492.

Solomon, C. G.; and LaRocque, R. C. 2019. Climate change—a health emergency. *New England Journal of Medicine*, 380(3): 209–211.

Souza, F.; Nogueira, R.; and Lotufo, R. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *BRACIS*, 403–417. Springer.

Spitale, G.; Biller-Andorno, N.; and Germani, F. 2023. AI model GPT-3 (dis) informs us better than humans. *Science Advances*, 9(26): eadh1850.

Strachan, J. W.; Albergo, D.; Borghini, G.; Pansardi, O.; Scaliti, E.; Gupta, S.; Saxena, K.; Rufo, A.; Panzeri, S.; Manzi, G.; et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7): 1285–1295.

Tappin, B. M.; Wittenberg, C.; Hewitt, L. B.; Berinsky, A. J.; and Rand, D. G. 2023. Quantifying the potential persuasive returns to political microtargeting. *PNAS*, 120(25): e2216261120.

Treen, K. M. d.; Williams, H. T.; and O'Neill, S. J. 2020. Online misinformation about climate change. *Wiley Interdisciplinary Reviews: Climate Change*, 11(5): e665.

Violot, C.; Elmas, T.; Bilogrevic, I.; and Humbert, M. 2024. Shorts vs. regular videos on YouTube: A comparative analysis of user engagement and content creation trends. In *proc. of the ACM WebSci*, 213–223.

Wang, X.; Shi, W.; Kim, R.; Oh, Y.; Yang, S.; Zhang, J.; and Yu, Z. 2019. Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good. In *proc. of the ACL*, 5635–5649.

Wu, S.; Rizoiu, M.-A.; and Xie, L. 2018. Beyond views: Measuring and predicting engagement in online videos. In *proc. of the ICWSM*, volume 12, 434–443.

Ye, H.; Xie, Y.; Ren, Y.; Fang, H.; Zhang, X.; and Song, G. 2025a. Measuring human and ai values based on generative psychometrics with large language models. In *proc. of the AAAI*, volume 39, 26400–26408.

Ye, H.; Zhang, T.; Xie, Y.; Zhang, L.; Ren, Y.; Zhang, X.; and Song, G. 2025b. Generative Psycho-Lexical Approach for Constructing Value Systems in Large Language Models. In *proc. of the ACL*, 11968–11991.

YouTube Official Blog. 2024. Tall Updates Coming to Shorts. https://blog.youtube/news-and-events/tall-updates-coming-to-shorts/. Accessed: 2025-07-27.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *proc. of the NeurIPS*, volume 36, 46595–46623.