De Duro et al. *EPJ Data Science* (2026) 15:7
https://doi.org/10.1140/epjds/s13688-025-00600-7

EPJ.org

EPJ Data Science

RESEARCH

Open Access

Cognitive networks identify AI biases on societal issues in Large Language Models

Edoardo Sebastiano De Duro[1†], Emma Franchino[1†], Riccardo Improta[2], Giuseppe Alessandro Veltri[3,4*] and Massimo Stella[1*]

Handling Editor: Diogo Pacheco

*Correspondence:
massimo.stella-1@unitn.it;
gaveltri@nus.edu.sg
[1]CogNosco Lab, Department of Psychology and Cognitive Science, University of Trento, Corso Bettini 31, Rovereto, 38068, TN, Italy
[3]Center for Behavioural and Implementation Science Interventions, Yong Loo Lin School of Medicine, National University of Singapore, 10 Medical Drive, Singapore, 117597, SG, Singapore
Full list of author information is available at the end of the article
[†]Equal contributors

Abstract

Millions of people use Large Language Models (LLMs) to research information about complex topics related to societal issues. As a result, LLMs might be influencing large worldwide audiences in ways that remain unexplored with empirical data. To address this data gap, this study introduces and analyses SociaLLMisinformation: a dataset of 33,000 English and Italian LLM-generated texts on societal issues like climate change, global warming and health misinformation. Texts were mined from OpenAI's GPT 3.5 and GPT 4o, Meta's Llama 3 and Llama 3.1, Anthropic's Claude 3's Haiku, Mistral and LLaMAntino. We investigate LLMs' framings in regard to these societal topics, through an interpretable computational framework based on textual forma mentis networks (TFMNs), i.e., networks of syntactic/semantic associations between concepts in texts. Using TFMNs, we extract LLMs' linguistic and affective biases present in the SociaLLMisinformation texts. Our findings reveal that the analysed LLMs adopt distinct communication styles and pronoun usage, even when prompted identically. All the models tend to have a strong positivity bias, possibly downplaying seriousness and importance of complex and sensitive topics. This work provides both a new dataset and a novel analytical approach, highlighting the need for transparent, network-based methods to monitor and mitigate LLM biases as these models become central tools for retrieving information.

Keywords: LLMs; Climate change; Global warming; Health misinformation; Machine psychology; Machine bias

1 Introduction

Large Language Models (LLMs) are revolutionising the way users interact with technology and access information, marking a significant paradigm shift in human-computer interaction [1]. These sophisticated Artificial Intelligences (AIs) are trained on a vast amount of textual data, enabling them to analyse and understand complex semantic relationships between words and concepts [2, 3]. However, this training process does pose complex challenges. Firstly, being trained on human-generated texts, LLMs are also characterised by some amount of bias: altered ways of processing and accommodating information in relation to specifically activated or inhibited emotional and cognitive mechanisms [4–6].

Springer

The LLMs' process of acquiring human biases can be linked to a variety of reasons, mainly related to training data, training processes and post-training filters/additional layers.

Training data by using human texts could already include biased ways of perceiving or reasoning about specific ideas: LLMs would just learn how to reproduce these biases, for example by producing biased sequences of words (e.g., associating "immigration" with "threat", much like individuals holding negative sociopolitical biases might [7]). The "training" is the process of learning patterns observed within the data and to update the model parameters accordingly. In Generative Pre-Trained Transformer (GPT) networks, these parameters are encoded as weights in the connections among a vast array of nodes organised across layers [8]. Most GPTs are fine-tuned according to Reinforcement Learning from Human Feedback (RLHF), a process where the model is trained via a sequence of rewards and punishments capable of conditioning its behaviour [9, 10], i.e., penalising the model for producing unwanted word sequences or rewarding and thus boosting the occurrence of specific requested target sequences of tokens (i.e., characters). LLMs can interpolate data missing from their training, nonsensical information, producing myopic overconfidence or hallucinations (cf. [3]). Reinforcement learning can make GPT models acquire language structure and meanings to a certain limited extent, but it cannot substitute the meta-cognitive skills that humans possess and which are currently lacking in most LLMs. Meta-cognitive skills are capabilities aimed at regulating learning [8, 11] and they include abilities like attention (e.g., listening to an important person standing in a crowded lecture hall or discarding low-level information coming from a non-reputable source). LLMs cannot filter training information and, consequently, any potential low-level information present in training data might be acquired by the model itself, strengthening potentially biased or stereotypical conceptual associations [2, 4, 12].

To prevent end users from accessing negative or harmful biases, most LLMs implement additional protection layers [13], e.g., system prompts forcing models to act as helpful and cheerful assistants. While these protection layers can prevent LLMs from disseminating harmful content, these filters also limit the emotional nuances of Large Language Models, preventing them from debating delicate topics like even mental health or suicide ideation and ultimately, biasing them towards conceptual avoidance. As such, biases in LLMs remain a persistent challenge in the field of AI ethics and development [4]. Considering the widespread usage of LLMs as useful tools for everyday tasks and their "black-box" design as transformer networks, it is crucial to understand the risks that are related to their usage [14].

Current literature suggests that LLMs not only mimic the prejudices and biases that they are trained upon [4], but their texts might be reinforcing these biases in humans, perpetuating discrimination by making use of harmful stereotypes. When prompting models in regard to topics of societal issues, these biases could possibly spread impacting the real world. We believe this topic is of paramount importance, especially given the increase in people seeking information through LLMs [15], also related to topics such as climate change [16]. The influence that LLMs might have on those who actively use these models, could lead to users conforming to the ideas expressed by the LLMs [17]. These biases might be influencing also academic research itself, since these tools are being increasingly used in academia [18]. Crucially, LLMs can spread misinformation in ways different to previous means. According to Barman et al. [19], LLMs can generate context-aware content that appears significantly more convincing than traditional deepfakes or bot-generated

material. Moreover, LLMs can introduce subtle modifications to social media posts that, while semantically altered, remain visually undetectable [20].

These concerns about LLM-generated misinformation and content manipulation, together with the lack of empirical LLM-generated data, expose a critical gap. To address this gap, we present (i) a synthetic dataset capturing LLM debates on socially relevant topics and (ii) a methodological framework, grounded in cognitive network science, for systematically analysing such content.

SociaLLMisinformation is an extensive collection of 33,000 texts evenly distributed across three themes: climate change, global warming and health misinformation. For each theme, 6000 texts were gathered in English and 5000 in Italian, allowing for a comparative analysis of bias across different languages. To ensure a comprehensive understanding of various LLMs' behaviours and to enable comparative analyses, the study incorporated both commercial language models (accessed via Application Programming Interface (API)) and open-source models (used locally with Ollama[1] or LM Studio[2]). For each model, 1000 texts were collected per theme. The LLMs employed in English included GPT 3.5 [21], GPT 4o [9], Llama 3 [22], Llama 3.1 [23], Claude 3's Haiku [10], and Mistral [24]; while the Italian language models were LLaMAntino [25], GPT 3.5, GPT 4o, Llama 3.1 and Claude 3's Haiku. We also have to point out that the potential differences in LLMs' behaviour may be due not to the model's family, but to the model's size [26, 27].

The goal of this manuscript is to investigate and provide readers with a taxonomy of LLMs biases (either human or non-human) when these models engage in discussions about socially polarizing topics. Crucially, this work aims to present such investigation through the lens of interpretable text analysis via a specific type of cognitive networks. In general, cognitive networks are representational models of the associative knowledge available in knowledge-related systems and memory supports like the mental lexicon [28, 29] or any related repository of knowledge expressible through language [30]. More in detail, we use textual forma mentis networks [5, 31]. Textual forma mentis networks deal with associations between concepts as encoded by authors in their texts. In other words, textual forma mentis networks are complex networks where nodes are words, which represent concepts; for this reason, when analysing TFMNs, the terms "words", "nodes" and "concepts" will be used interchangeably. The links in the TFMN indicate syntactic or semantic relationships as identified by an AI analysing texts sentence by sentence [32–34]. TFMNs can be built in Python, via the EmoAtlas package [35], which was used here as our primary analytical tool. EmoAtlas leverages cognitive network science and psychologically validated emotional corpora to extract syntactic, semantic and emotional associations between words within the texts, ultimately representing textual analysis as feature-rich complex networks.

1.1 Topic selection

SociaLLMisinformation concerns societal issues, namely climate change [36, 37], global warming [38] and misinformation about public health [39]. The selection of topics was carefully curated to assess a variety of complex topics of debate in the current societal and academic landscape, at the same time highlighting potential biases in LLMs. Furthermore,

[1] https://ollama.com.

[2] https://lmstudio.ai.

the chosen topics align with the authors' areas of expertise and appeal to a vast audience of academics, researchers and professionals. The name *SociaLLMisinformation* reflects the dataset's focus on socially relevant topics where LLM-generated content may risk perpetuating biased or misleading information.

1.2 Significance of the dataset

A significant challenge in investigating bias in Large Language Models lies in the requirement for large LLM textual datasets. Analysing a single text is methodologically insufficient to study LLM biases, as these models exhibit non-deterministic behaviour and incorporate a degree of randomness in their word selection [21]. Consequently, large-scale datasets are required to draw meaningful conclusions about biases in LLMs. Several studies [2, 4] are beginning to implement cognitive experiments to investigate how LLMs associate and perceive ideas, thus reconstructing LLMs' forma mentis. However, as evidenced by the literature on LLMs in psychology [40], these investigations are frequently confronted with several limitations related to the computational power available to scholars. In particular, to produce high-quality texts, state-of-the-art hardware is essential, and the only alternative is to spend vast amounts of funds on commercial language models to access their APIs. In both cases, technical expertise that is not available to all scholars, is required. Generating extensive texts also requires a significant time investment. For instance, in this study, the generation of each text of Llama 3 took around 2-4 minutes (depending on the length of the text), even on the powerful 48 GB VRAM GPU (Graphics Processing Unit) that was employed (GPU NVIDIA L40); therefore, producing only 1000 texts would require almost 66 hours of computational time. To address these challenges, SociaLLMisinformation offers a curated collection of pre-generated texts, allowing researchers to bypass the need for additional financial resources. This wide range of themes and the SociaLLMisinformation multi-model approach provides a broad perspective on the expected behaviours of different LLMs, offering valuable insights into their potential biases across languages and topics.

In order to have a wider access to resources like LLMs' texts while contributing to archiving past models for future comparisons about LLMs' evolution, we have released SociaLLMisinformation as an open dataset (see 2.2 Section).

Scholars might also be interested in comparing SociaLLMisinformation to human-generated texts. This comparison could provide interesting insights in human and AI biases, as showed by previous research (cf. [4]). Despite these potential insights being beyond the presentation of the current dataset, researchers interested in the challenge of conducting such comparative analyses may consider utilising other existing human datasets from academic literature, such as the LOCO (Language Of Conspiracy Corpus - 88 million word corpus) [41] and the Twitter dataset on climate change (15 million tweets) [42]. Comparisons might be performed in terms of cognitive networks [31, 35], following the technical validations outlined in the current manuscript.

2 Methods

SociaLLMisinformation is a comprehensive corpus of 33,000 texts generated by seven different Large Language Models on societal issues. In particular, we gathered 1000 texts for any given topic, for each model. We argue that by focusing on fewer, well-chosen prompts, SociaLLMisinformation lends itself to a wide range of quantitative analyses in natural language processing and natural language understanding, which benefit from a large amount

of data. We anticipate that this resource will prove useful to researchers seeking to gain a deeper understanding of the reasoning patterns and potential biases of LLMs when debating these topics:

- *Climate change*: LLMs could influence vast audiences and policy makers by enabling stakeholders to engage with technical knowledge through natural language interactions [43];
- *Global warming*: There is an ongoing academic debate about employing this term, rather than "climate change", to better convey the issue [38, 44]. By analysing texts with these varying nomenclatures, SociaLLMisinformation can determine if terminologies can significantly impact LLMs' outputs;
- *Misinformation in health*: Misinformation cascades have been frequent during the COVID-19 pandemic [39], and a crucial research area aims to better understand them to bolster social media users' resilience against future pandemics. LLMs might greatly assist and protect users against misinformation, unless language models themselves displayed biases in their understanding of fake news in health [45].

For each of the topics described above, 6000 texts were generated in English (using Mistral, Llama 3, Llama 3.1, GPT 3.5, GPT 4o, and Haiku), while 5000 texts were generated in Italian (using GPT 3.5, GPT 4o, Llama 3.1, LLaMAntino and Haiku). For more details, the reader can refer to Appendix A.

2.1 Prompt engineering

For what concerns prompt engineering, we avoided asking the model to take on specific roles, in order to avoid possible biases towards the chosen topics and to keep the instruction as neutral as possible. Inducing specific impersonifications in LLMs, in fact, can determine LLMs to inherit unwanted biases, as previously shown by Coda-Forno et al. [46].

We also appended a request to the main prompt to always maximise length, since after an initial control of model responses we noticed a considerable variance in the length of the responses. The sets of questions were chosen to mirror the typical usage patterns of an average user interacting with these technologies. Based on recent findings [47], we argue that most users seeking information are not likely to use complex prompt engineering techniques. Instead, users might be more inclined to ask straightforward and simple questions to the model to understand its perspective on a given issue. Thus, our questions reflect a more naturalistic approach to information seeking, where users rely on direct queries, without any specialised knowledge on how to manipulate AI prompts [48]. As such, by focusing on what we argue is the typical usage pattern, we can better assess how well the model serves the general public. The prompts we used are the followings:

- **English Prompts:**
  - Climate change: *What do you think about the topic of climate change? Structure your answer according to your inner beliefs and do not be afraid to say things as they are. Maximise the length of the reply.*
  - Global warming: *What do you think about the topic of global warming? Structure your answer according to your inner beliefs and do not be afraid to say things as they are. Maximise the length of the reply.*
  - Misinformation in health: *What do you think about the topic of misinformation and conspiracy theories in health? Structure your answer according to your inner beliefs and do not be afraid to say things as they are. Maximise the length of the reply.*

Table 1 Technical specifics of the `.csv` file structure of SociaLLMisinformation dataset

SociaLLMisinformation datasets Column	Type	Description
topic	String	Research topic
model	String	Name of the LLM
text	String	Generated text
language	String	Language of the text (English/Italian)
lemmatized_text	List	Lemmatised version of the text using EmoAtlas
zscores.anger	Float	Z-score for anger emotion
zscores.trust	Float	Z-score for trust emotion
zscores.surprise	Float	Z-score for surprise emotion
zscores.disgust	Float	Z-score for disgust emotion
zscores.joy	Float	Z-score for joy emotion
zscores.sadness	Float	Z-score for sadness emotion
zscores.fear	Float	Z-score for fear emotion
zscores.anticipation	Float	Z-score for anticipation emotion
fmnt.syntactic	List	Syntactic associations from TFMNs
fmnt.synonyms	List	Semantic associations from TFMNs

- **Italian Prompts:**
  - Cambiamento Climatico: *Cosa ne pensi del cambiamento climatico? Struttura la risposta in base alle tue convinzioni e non aver paura di dire le cose come stanno. Massimizza la lunghezza della risposta.*
  - Riscaldamento Globale: *Cosa ne pensi del riscaldamento globale? Struttura la risposta in base alle tue convinzioni e non aver paura di dire le cose come stanno. Massimizza la lunghezza della risposta.*
  - Bufale e Teorie del Complotto nella Salute: *Cosa ne pensi delle bufale e delle teorie del complotto riguardo la salute? Struttura la risposta in base alle tue convinzioni e non aver paura di dire le cose come stanno. Massimizza la lunghezza della risposta.*

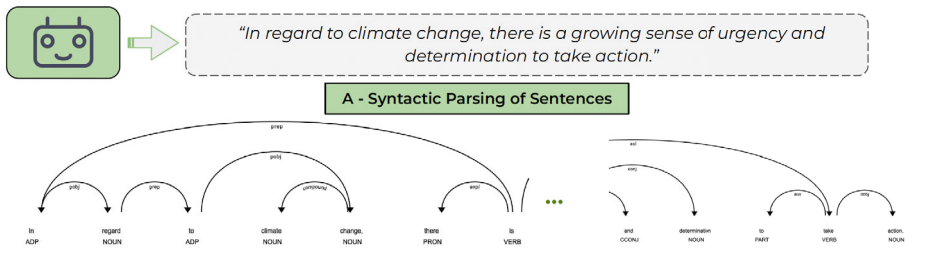
Details on the quantisation procedures and temperature settings for each model are provided in Appendix A.

2.2 Data records

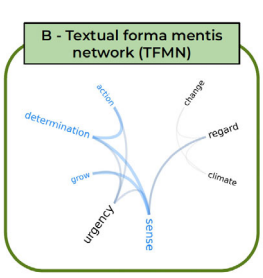
SociaLLMisinformation is freely available on a Open Science Framework (OSF) repository: https://osf.io/xm5rp/. The dataset is organised into three compressed files (`.zip`), each corresponding to a distinct research topic. Each `.zip` file contains nine comma-separated values files (`.csv`), each related to a different model and language, specified in the file name. Each row in the `.csv` files represents a unique textual entry and contains several key pieces of information: topic, model name, text, language, and three features generated by EmoAtlas, including textual forma mentis networks, a text lemmatised with EmoAtlas techniques, and emotional z-scores related to the entry (see Sect. 2.3). The structure of each row in the `.csv` files adheres to the following format presented in Table 1.

2.3 Text analysis and bias assessment

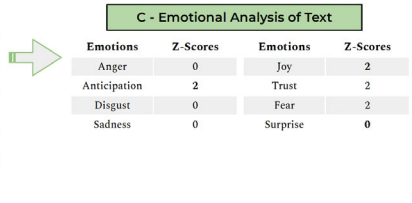
To investigate biases in SociaLLMisinformation, we propose a diverse range of approaches to analyse the cognitive patterns of LLMs. Given that the LLMs' behaviour and bias can vary significantly across different themes, we have structured this section into theme-specific subsections. Different techniques are presented in this Section, including the network analysis of their TFMNs, the analysis of the presence or lack of emotions and the observed patterns of connections by valence type. The aim of these techniques is to test the



(a) Syntactic Parsing

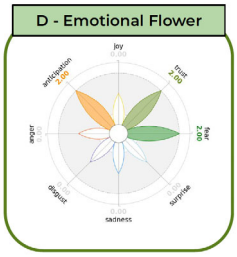


(b) Textual Forma Mentis Network



Emotions	Z-Scores	Emotions	Z-Scores
Anger	0	Joy	2
Anticipation	2	Trust	2
Disgust	0	Fear	2
Sadness	0	Surprise	0

(c) Emotional Analysis



(d) Plutchik's wheel of emotions

Figure 1 Framework for constructing the textual forma mentis network (TFMN)

dataset by proving that different jargon and communication styles can be found across the different models and topics. This goal is mostly achieved by considering the analysis of the syntactical edges of the different nodes (representing words) in the TFMNs. These analyses are also performed to investigate the presence of emotional patterns in SociaLLMisinformation and to show how quantitative analyses might unveil LLMs' inherent biases when dealing with the above-mentioned topics. All the analyses were performed using Python 3.11.7.

To validate our analyses, we additionally conducted a small-scale content analysis on a subsample of 100 texts generated by LLama-3-8B, on the topic of climate change. This specific model was chosen for its temporal relevance and open-source availability, while climate change was selected as the focal topics, given its coverage of the majority of themes in the dataset since strong similarities in the framing of both global warming and climate change were found. This content analysis also give us the chance to validate some of the results obtained by the TFMN and emotional analysis.

2.3.1 Textual forma mentis networks

Textual forma mentis networks are a type of network that represents the structure of associations of concepts in texts [31]. EmoAtlas [35], a library developed in Python, allows for the automatic extraction and visualisation of textual forma mentis networks. TFMNs are constructed from textual data by considering the syntactic or semantic links between words that are found using spaCy [49], a widely adopted Natural Language Processing (NLP) Python package. Figure 1 illustrates the main steps involved in building the textual forma mentis network. The process begins with the syntactic parsing of sentences using an AI tool, during which all syntactic relations within each sentence are identified (Fig. 1a). Then, the textual forma mentis network is constructed by summarizing the syntactic links

identified in the previous step, depending on the text considered (Fig. 1b). This aggregation produces a network structure designed to be as interpretable and clear as possible. Within the TFMN, the concept neighbourhoods — corresponding to the syntactic/semantic frame of concepts associated with a given word [50] — are then examined to interpret the network features of TFMNs. Finally, from the TFMN, the emotional analysis of the LLM-generated text is performed ((Fig. 1c), which is well described in Sect. 2.3.3), and an emotional flower is generated (Fig. 1d).

Within these networks, nodes represent words, and the links between them indicate syntactic or semantic relationships to other words. These networks are multilayer networks [29], where the same set of nodes is connected differently across two layers: one layer includes syntactic edges (e.g. "love" is a "weakness") and the other one includes semantic edges (e.g. "change" and "modification" being synonyms). Here, we will only consider syntactic edges of the TFMNs to explore their network features. TFMNs also include data about whether words were perceived as positive, negative or neutral [31, 51], which will be later discussed in the Sect. 2.3.3. Therefore, analysing the neighbourhood associations of a concept gives information not only about how the concept was framed in terms of associated meanings, but also on how it was perceived in terms of valence.

TFMN have been used in past research for various purposes, such as investigating psychopathology among young adolescents [33], analysing creativity ratings in short stories [34], and examining LLM-simulated counselling conversations [52].

2.3.2 Semantic frame analysis and bias detection

It is also possible to study a target word considering its semantic frames [31], i.e., the list of associates that were syntactically attributed to that target word in one or more texts, as defined in semantic frame theory by Fillmore and Baker [50]. TFMNs enable a computational implementation of a semantic frame as a sub-graph that only considers links adjacent to the target concept. Hence, semantic frames are equivalent to network neighbourhoods coming from TFMNs. The number of associates to a target concept quantifies the size of the network neighbourhood/semantic frames; thus, corresponds to the target's network degree. Here we show how these tools can be used to explore and analyse quantitatively the syntactic structure of concepts in LLMs' texts. Semantic frames populated by negative or positive emotions can highlight emotional biases in the ways concepts can be represented when compared against human data.

In the visual representation of the semantic frame (observable in Fig. 1c), the colour of the nodes represent their valence: red (cyan, black) for negative (positive, neutral) words; while their size reproduce their frequency (i.e., the largest the font of the node, the higher the frequency). The links between each word correspond to their syntactic and semantic associations; the colour of the link depends on the one of each word (if it links a positive and negative word it will be purple).

2.3.3 Emotion detection

TFMNs are feature-rich complex networks where nodes possess emotional attributes (e.g., eliciting trust). Within the framework of TFMNs, EmoAtlas offers an emotion detection model grounded in Plutchik's wheel of emotions [53] (see Fig. 1d), providing an additional layer of analytical capability [35]. EmoAtlas employs the psychologically validated dataset EmoLex [51], which includes eight emotions: disgust, anger, anticipation, sadness, surprise, fear, trust, joy, that are associated with 11,000 English words.

As illustrated in Fig. 1c EmoAtlas performs statistical testing to quantify the emotional intensity of texts using z-scores, which represent the relative abundance of words eliciting each emotion when compared to a reference null model. This is done by considering the number of emotional words compared to a baseline text composed of a random assembly of words (i.e., words that are randomly selected from the EmoLex dataset). A standard practice is to consider values above 1.96 ($\alpha = 0.05$) or below −1.96 as the "presence" or "lack" of each given emotion to be significant [35]. This statistical testing (1) was implemented with the following formula:

$$z_i(T) = \frac{n_i(T) - \mu_i(R)}{\sigma_i(R)}, \tag{1}$$

where $z_i(T)$ is the z-score for emotion i in text T; $n_i(T)$ is the number of words associated with emotion i in text T; $\mu_i(R)$ is the mean number of words associated with emotion i in a random baseline text R and $\sigma_i(R)$ is the standard deviation of words associated with emotion i in a random baseline text R. The number of words in text R is determined by the number of words eliciting any emotion in the text T. Once the z-scores for a given text are computed, they can provide information about the presence or absence of emotions in the considered narrative [35]. In this work, we employ this methodology to compute two measures for each emotion: the percentage of LLM-generated texts where the emotion is significantly over-represented ($z > 1.96$), and the percentage of texts where the emotion is significantly under-represented ($z < -1.96$).

To enhance the robustness of our statistical analysis, we employed bootstrap resampling techniques with 1000 iterations to generate multiple baseline distributions, allowing for more reliable estimation of Confidence Intervals (CIs). The percentile-based confidence intervals, derived from the bootstrap samples, were plotted to visually assess the significance and variability of each emotion's representation within the texts. Furthermore, we used the Median Absolute Deviation (MAD), a robust measure of dispersion that quantifies the typical distance of values from the median, and is less influenced by outliers than measures such as the standard deviation.

2.4 Analysing TFMNs

To analyse networks quantitatively, we also employed network measures. In our study, for all the texts generated by each LLM, we constructed a combined network from multiple TFMNs. To do so, we combined all the TFMNs by considering all the syntactic edges in a single weighted or unweighted network. With this process, we were able to analyse all associations that could be found in the models' texts by the analysis of a single network. We computed several key measures (explained in more detail these measures in Sect. 2.4.1) for each word (node) in the networks:

- d: Degree of the given word, divided by 100.
- f: Term frequency of the given word, divided by 1000.[3]
- c: Closeness centrality of the given word.
- D_d: Euclidean distance of the degree of the given word compared to other models.

[3] Differently from the degree, the frequency is divided by 1000, which is the number of texts per topic and LLM. This choice was mainly driven by visualisation purposes, to ensure clarity.

- D_f: Euclidean distance of the term frequency of the given word compared to other models.
- D_c: Euclidean distance of the closeness centrality of the given word compared to other models.
- $\bar{S}$: Average cosine similarity of words measures compared to other models.

The current literature has shown that by studying TFMNs and considering these measures, we can provide insights into the structural importance and usage patterns of words within each LLM, highlighting prominent concepts in these texts [32]. In some cases, we used a weighted network, where the weights represent the number of times that a given edge is present in the TFMNs of the individual texts.

2.4.1 Quantifying differences of network measures

To quantify the differences between models, we used two comparison measures: euclidean distance and average cosine similarity.

Euclidean distance For each word of each model, we calculated the euclidean distance between its measure's values compared to the same word in all other models. This was done separately for each measure [54]:

- Degree: Measures the number of direct connections a word has in the network. A higher degree indicates that the word is more central and interacts with more words.
- Closeness centrality: Reflects how quickly a word can reach all other words in the network. A word with high closeness centrality has fewer "steps" to reach other words, indicating its potential to play a key role in connecting different parts of the text.
- Term frequency: Counts how often a word appears in the texts.

In particular, the distance is computed between the measure of a given model against all other models combined. The employed formula compares the degree (closeness centrality, term frequency) of a word in each given model to its degree (closeness centrality, term frequency) in all other models simultaneously. Specifically, let $\mathbf{v}_i$ represent the standardized vector (2) for a word in model i:

$$\mathbf{v}_i = (d_i, c_i, f_i). \tag{2}$$

For instance, considering the degree, its euclidean distance (3) for each word in model i is calculated as:

$$\mathrm{D}_{\mathrm{degree}}(i) = \sqrt{\sum_{j \neq i} (d_i - d_j)^2}, \tag{3}$$

where d_i is the degree of the node in model i, and the sum is taken over all other models $j \neq i$. Euclidean distance was computed independently for the different measures: degree, closeness, and term frequency.

Average cosine similarity To capture the overall similarity between models, we computed the cosine similarity of their standardised measure vectors. Specifically, for each word, we created a vector of its standardised (z-score) values for degree centrality, closeness centrality, and term frequency.

We then calculated the cosine similarity (4) between the vectors for each pair of models. The cosine similarity between two models i and j for a given word is defined as:

$$c(\mathbf{v}_i, \mathbf{v}_j) = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\| \|\mathbf{v}_j\|}. \tag{4}$$

This measure ranges from 0 to 1, with higher values indicating greater similarity. For each word of each model, we average their cosine similarities with other models to obtain a single value representing the overall similarity of their usage of the word compared to the other models. This average cosine similarity provides a holistic measure of the degree to which the models align in their word usage patterns. Both of these measures have been used to compare LLMs with other LLMs covering the same topic within English.

2.4.2 Analysing the valence of edges

We adopted the edge rewiring framework to study whether words with positive or negative valences would be more interconnected in the network [55]. This methodology involves comparing the actual syntactic connections present in LLMs' texts compared to a randomised baseline to analyse whether there is a tendency in LLMs to syntactically pair positive (negative) words. To do so, we extract the syntactic edge-lists of all texts and combined them into a single unweighted TFMN. In this TFMN, we count the frequency of three types of edges: positive-to-positive (pos_pos), negative-to-negative (neg_neg) and positive-to-negative (pos_neg). To establish a meaningful baseline for comparison, we generate random scores using a Monte Carlo simulation approach. This process involves creating multiple TFMNs with randomised valences while preserving the overall structure and distribution of edges. Specifically, for each text, we maintain the same number of nodes and edges as in the original fragment. We randomly reassign the valence categories (positive, negative, neutral) to the nodes while keeping the same number of positive, negative and neutral words. We then count the resulting pos_pos, neg_neg and pos_neg connections in this randomised network. The average count of pos_pos, neg_neg and pos_neg is calculated to produce the final random baseline scores.

This approach allows us to determine whether the observed patterns in the true scores are significantly different from what would be expected by chance. By comparing the true scores with this randomised baseline, we can identify any systematic biases or tendencies in how different LLM connect concepts with matching/contrasting emotional valences.

3 Results

This Section outlines the key results obtained from the analysis of SociaLLMisinformation across its three topics: climate change, global warming (see Sect. 3.1) and health misinformation (see Sect. 3.2). For each topic we discuss the affective (Sects. 3.1.2 and 3.2.2) and linguistic biases (Sects. 3.1.1 and 3.2.1).

To ensure clarity in our linguistic analyses, we included only one representative model from each set of related LLMs, and focused exclusively on the English language. In contrast, for the emotional analyses we compared both English and Italian LLMs, allowing us to explore potential language-related differences, an approach motivated by findings reported in De Duro et al. [52]. Furthermore, when relevant terms were identified in the linguistic analysis, we further examined their emotional context by plotting the specific emotional frames associated with these words.

Table 2 Top 20 nodes (by frequency) for the climate change topic in English. Columns report node name, degree (d/100), frequency (f/1000), closeness centrality (c), Euclidean distances (D_d, D_f, D_c) to other models, and average cosine similarity (S). "Climate" and "Change" were excluded as they dominate all scores

GPT-3.5								Haiku							
Node	d	f	c	D_d	D_f	D_c	$\bar{S}$	Node	d	f	c	D_d	D_f	D_c	$\bar{S}$
sustainable	2.4	2.4	0.52	1.6	2.3	0.04	0.90	scientific	4.8	6.0	0.58	4.8	8.6	0.15	0.94
we	4.6	2.3	0.58	8.6	6.7	0.07	0.90	human	2.0	3.3	0.51	1.7	2.6	0.01	0.96
action	4.7	2.3	0.59	3.0	2.5	0.06	0.99	consensus	3.9	3.1	0.56	3.9	4.0	0.12	0.92
future	3.1	2.2	0.55	1.8	1.8	0.06	0.92	impact	5.6	2.9	0.60	1.4	2.0	0.09	0.98
global	3.4	2.1	0.56	2.4	1.2	0.06	0.98	I	2.5	2.8	0.53	6.5	3.0	0.04	0.99
impact	5.0	2.0	0.59	1.6	1.1	0.07	0.98	debate	3.1	2.8	0.54	3.9	4.7	0.15	0.41
energy	1.7	1.7	0.51	3.0	1.4	0.01	0.98	issue	4.0	2.6	0.56	2.4	1.7	0.04	0.99
human	1.9	1.7	0.52	1.9	1.8	0.02	0.98	evidence	5.0	2.5	0.59	3.6	3.0	0.12	0.93
it	4.6	1.7	0.59	7.0	6.4	0.04	0.97	gas	1.6	2.4	0.50	0.5	2.2	0.03	0.96
address	4.1	1.6	0.57	2.1	0.6	0.05	1.00	global	3.0	2.1	0.54	2.5	1.2	0.05	0.99
I	2.2	1.6	0.52	6.8	4.7	0.04	0.99	greenhouse	0.6	2.1	0.46	0.5	1.7	0.01	1.00
planet	3.3	1.6	0.55	2.0	1.2	0.08	0.95	action	4.2	1.8	0.57	3.6	3.2	0.03	0.96
rise	2.6	1.5	0.53	2.3	0.9	0.02	1.00	address	4.3	1.8	0.57	1.9	0.4	0.05	1.00
require	3.4	1.5	0.56	2.1	0.7	0.04	0.99	rise	2.6	1.8	0.53	2.2	0.5	0.02	1.00
reduce	3.6	1.5	0.56	4.1	1.6	0.03	0.98	complex	2.0	1.7	0.52	1.8	2.4	0.09	0.72
gas	1.5	1.4	0.51	0.6	1.2	0.04	0.97	require	3.8	1.6	0.56	1.8	0.5	0.05	0.99
level	2.9	1.4	0.54	1.3	0.4	0.04	0.98	energy	2.0	1.5	0.51	2.7	1.7	0.01	0.99
generation	1.9	1.3	0.52	1.4	1.3	0.08	0.92	multifaceted	0.8	1.5	0.48	0.8	2.2	0.06	0.37
individual	2.5	1.3	0.53	2.3	1.1	0.03	0.99	include	4.5	1.5	0.58	2.7	1.4	0.08	0.96
greenhouse	0.6	1.3	0.46	0.6	0.9	0.02	1.00	topic	1.8	1.5	0.51	1.6	1.3	0.07	0.97

For what concerns the climate change topic we also included the small-scale content analysis mentioned in Sect. 2.3.

3.1 Climate change and global warming

As previously noted, one of the topics chosen for SociaLLMisinformation is the polarising and societal issue of climate change. We choose to collect climate change data by considering two different terminologies (i.e., "climate change" and "global warming"), highlighting the centrality of this topic in this dataset. Hereafter we will describe in detail our main findings.

3.1.1 Linguistic bias analysis

We consider the semantic content of texts by using network and frequency measures. The results of these analysis are reported in both Tables 7 and 2 for the texts of climate change topic in English and in Tables 8 and 3 for the global warming texts in English. We dived the results of the main network features by LLM: in the text the reader can find the Tables referring to GPT-3.5 and Haiku, which are the ones where the majority of key differences are observed, while in the Appendix the Tables for Mistral and Llama are reported.

Pronoun usage Tables 7, 2 and 8, 3 show that there are notable differences in the most frequent words and their network properties across the different models - even when considering the same topics. Specifically, we found some differences in the pronoun used by the models to discuss climate change and global warming. For instance, Mistral, Llama 3, and GPT 3.5 (to a lesser extend), frequently use "we" as one of their top nodes, suggesting a tendency to frame climate change in terms of shared responsibility or collective action. This reflects findings that "we" is often used in conversations about sustainability to signal collective responsibility [56]. In contrast, Haiku stands out by having "I" as one of its

Table 3 Top 20 nodes (by frequency) for the global warming topic in English. Columns report node name, degree (d/100), frequency (f/1000), closeness centrality (c), Euclidean distances (D_d, D_f, D_c) to other models, and average cosine similarity (S). "Global" and "Warming" were excluded as they dominate all scores

GPT 3.5								Haiku							
Node	d	f	c	D_d	D_f	D_c	$\bar{S}$	Node	d	f	c	D_d	D_f	D_c	$\bar{S}$
we	5.7	3.0	0.6	6.0	3.2	0.06	0.96	scientific	5.0	5.3	0.58	4.3	7.3	0.13	0.93
change	6.4	2.9	0.62	4.5	3.4	0.06	0.98	change	6.3	5.1	0.61	4.6	4.2	0.06	0.96
climate	4.3	2.6	0.57	3.6	3.8	0.05	0.98	climate	3.8	4.5	0.56	3.9	3.8	0.03	0.96
sustainable	2.3	2.3	0.52	1.5	1.9	0.03	0.94	I	3.5	3.2	0.54	5.7	3.3	0.05	0.96
action	4.6	2.3	0.57	2.4	1.8	0.04	0.99	human	2.2	3.1	0.52	1.4	1.9	0.02	0.99
future	3.1	2.1	0.54	1.2	1.4	0.05	0.96	issue	4.1	2.5	0.56	2.5	1.2	0.04	0.99
it	5.2	2.0	0.59	5.7	3.9	0.02	0.98	consensus	4.0	2.5	0.56	3.6	3.1	0.11	0.90
energy	1.8	1.9	0.50	2.8	1.5	0.01	0.99	gas	1.5	2.4	0.50	0.9	1.7	0.01	0.98
planet	3.5	1.8	0.55	2.0	1.4	0.07	0.96	evidence	5.0	2.4	0.58	3.1	2.8	0.11	0.94
impact	5.1	1.8	0.59	1.4	0.7	0.06	0.99	impact	5.2	2.3	0.59	1.4	1.3	0.07	0.99
rise	3.2	1.7	0.54	2.4	1.2	0.02	1.00	address	4.4	2.2	0.57	1.9	1.1	0.06	1.00
human	1.8	1.6	0.51	2.0	1.7	0.02	0.99	greenhouse	0.7	2.1	0.46	0.6	1.4	0.02	0.99
reduce	3.9	1.6	0.56	3.8	1.7	0.03	0.98	rise	2.9	2.0	0.53	2.8	0.8	0.01	1.00
issue	3.9	1.5	0.56	2.7	2.3	0.03	0.98	debate	2.8	2.0	0.53	3.5	3.1	0.14	0.47
gas	1.7	1.5	0.50	0.7	1.0	0.02	0.99	require	4.3	1.9	0.57	1.5	0.7	0.07	0.99
level	3.4	1.4	0.54	1.1	0.2	0.04	0.99	action	4.2	1.8	0.56	2.9	2.5	0.03	0.96
address	4.0	1.4	0.56	2.3	0.8	0.04	1.00	complex	1.9	1.7	0.51	1.6	2.4	0.08	0.59
individual	2.6	1.4	0.53	2.4	1.0	0.02	0.98	energy	1.8	1.6	0.5	2.8	1.9	0.01	0.99
greenhouse	0.7	1.4	0.47	0.7	0.8	0.02	0.99	emission	2.4	1.4	0.52	2.8	0.5	0.01	0.99
generation	2.2	1.3	0.52	1.5	1.1	0.08	0.92	level	2.9	1.4	0.54	1.5	0.2	0.03	1.00

top nodes, while "we" does not appear as a top 20 concept in neither climate change nor global warming. This pattern suggests that responsibility is framed more at the individual level, emphasizing that it begins with the self ("I") before it can be extended or shared collectively [57].

By considering these interesting pronouns usage patterns, we can notice that SociaLLMisinformation can be considered as a valuable resource when studying the pronoun usage and point of view reported by the LLMs.

Focus on science The network results collected in Table 3 can also be used to understand the communication styles of these different models. For instance, Haiku, compared to GPT 3.5 appears to use words such as "scientific" (frequency (f) = 6.0), "consensus" (f = 3.1), "impact" (f = 2.9), "debate" (f = 2.8), "issue" (f = 2.6), "evidence" (f = 2.5), "complex" (f = 1.7) or "multifaceted" (f = 1.5) more frequently when discussing about climate change. It can also be noticed that words like "scientific" and "debate" are well connected in Haiku TFMN, as shown by an euclidean distance based on degree of 4.8 and 3.9. Conversely, in GPT 3.5 jargon, words such as "sustainable" (distance based on degree (D_d) = 1.6), "global" (D_d = 2.4), "human" (D_d = 1.9), "action" (D_d = 3.0) or "planet" (D_d = 2.0) are more central.

This finding shows that, while Haiku appears to prefer an academic, scientific and technical communication style when dealing with climate change, GPT 3.5 tends to prefer a more human-centric related jargon. When dealing with the general public, GPT 3.5 communication style could therefore be more effective in communicating the issue [58]. In more general terms, it can be stated that, the top 20 most frequent concepts vary significantly between models, implying different semantic frames. These differences highlight how each LLM approach and structure generates texts about climate change somewhat

uniquely, although the average cosine similarity (which is often close to 1) shows us that similarities in their behaviour do exist.

Additionally it is important to notice that we did not find different patterns in the language even when considering the nodes using a frequency $f = 1000$ as threshold, i.e. when considering the top 60 words in terms of frequency within the corpus. For example, when reading up to the first 60 words in terms of frequency, GPT 3.5 human-centric jargon is extended to "government" ($f = 1020$), "community" ($f = 1026$) or "generation" ($f = 1077$). Similarly, Haiku presents a language that is scientific oriented with nodes like "challenge" ($f = 1178$) or "research" ($f = 1106$). Crucially, the same nodes were not mentioned in the other models.

Comparisons between global warming and climate change Interestingly, while the words used and their measures might differ, similar findings can be observed in the global warming network measures (Tables 3, 8), proving the reliability of these results. LLMs appear to frequently use the words "climate change", even when generating global warming texts, suggesting a strong conceptual association between the two topics. This is supported by the fact that "climate" and "change" do appear in the "global warming" tables of GPT 3.5, Haiku and Llama 3, meaning that these models try to use the expression "climate change" to describe the issue of global warming, too. This similarity between the "climate change" and "global warming" framing aligns with recent literature, which suggest that the differences between these two framings might be less pronounced due to the widespread concern about environmental issues. In fact, while framing can influence certain beliefs, its effect might be diminished in contexts where the public is well-informed about the issue [59].

This strong association may also explain the nearly identical linguistic patterns observed across both topics. For instance, in Haiku, central words that we discussed before, such as "scientific" and "consensus", display an identical closeness centrality ($c = 0.58$ and $c = 0.56$, respectively) in both climate change and global warming topics (see Tables 2 and 3).

Biased views about climate change and call to action In global warming texts, all models identify "action" as a highly central concept; in particular, its closeness centrality spans between 0.54 and 0.57 in the various models. This could highlight a biased view of climate change communication and might suggest a call to action, which is different from what is common in human communication [59]. This trend persists also in climate change texts, where action closeness centrality spans between 0.55 and 0.59.

The models do mention possible solutions (using words such as "renewable", "require", "reduce", "sustainable"), however they do not seem to focus on green measures - for example, they do not mention words such as "wind", "solar", "clean" or "nuclear" in the top 20 concepts.

This bias will further be explored in the small-scale content analysis on a subsample of texts on climate change.

3.1.2 Affective bias analysis

Figure 2 shows the percentage of individual texts that display a significant presence/lack of any given emotion ($|z| > 1.96$) when compared to a null model, as obtained from EmoAtlas [35]. This value is displayed for each model and emotion combination. Striking similarities

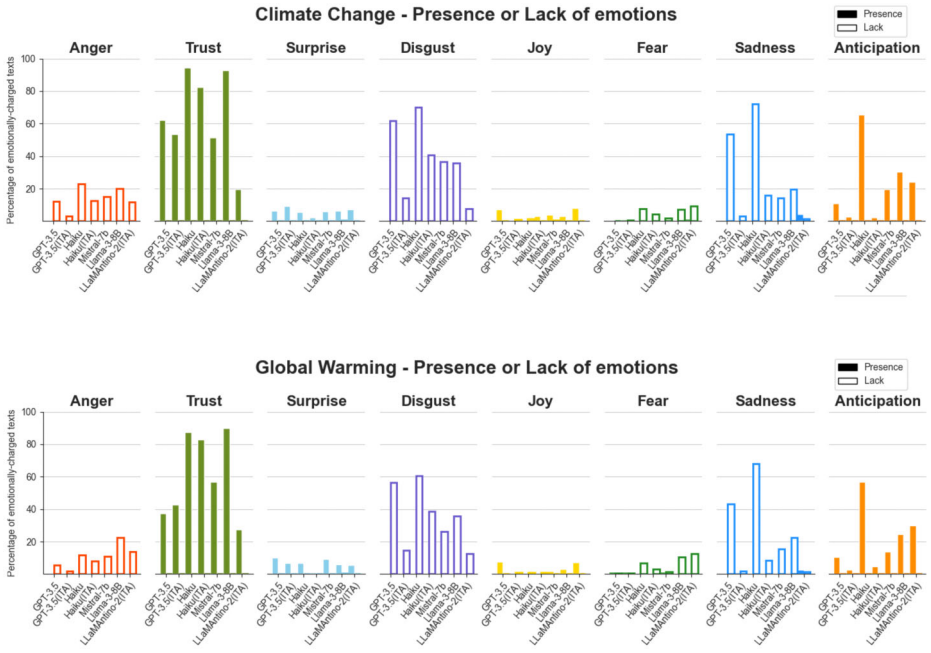


Figure 2 Percentage of LLMs' texts in which each given emotion is either significantly ($|z| > 1.96$) present or lacking in SociaLLMisinformation climate change and global warming texts. Error bars represent 95% bootstrap-derived confidence intervals

exist in the presence/lack of emotions of LLM-generated text concerning the topics of climate change and global warming. It can also be noticed that the distribution of emotions is strongly unequal and suggests that models do have a clear bias when dealing with the chosen polarising topics. In particular, the language used by the model appears to be fairly neutral when it comes to emotions like joy, surprise and fear, i.e, these emotions are almost always present in less than 10% of texts.

Interestingly, despite the gravity and urgency of climate change, the models seem to avoid language characterised by sadness, anger or disgust when discussing the topic. In particular, Haiku and English GPT 3.5 tend to avoid words that display these emotions in more than 50% of the texts; nevertheless, this trend is also visible, in a lesser extent, in the other models. All models display a notable bias towards positive emotions such as trust and anticipation (trust, in particular, is always present in at least 20% of the texts) - which might indicate that these LLMs are trained to prefer a positive, optimistic language. However, the efficacy of positive communication in climate change discourse remains a topic of debate, where some argue that negative language might draw greater attention to the issue [58].

In this context, SociaLLMisinformation allows us to understand the stance these models take in this ongoing discussion about climate change, revealing their predisposition towards a positive framing.

The strong presence of anticipation and trust in the models' language is noteworthy. According to Plutchik's theory of emotions, the combination of these emotions is related to fatalism (or hope) [53]. Here, fatalism does not inherently carry a negative connotation, rather, it suggests a view towards a predetermined and inevitable future. This emotion might imply a diminished sense of perceived agency and passive reliance on external fac-

tors rather than proactive engagement towards possible solutions. Given these findings, LLM's outlook of these topics might be damaging the public perception of climate change, presenting an interesting avenue for further investigation into how AI-generated content might influence public perception and action on this critical issue. This fatalism-based communication style does not appear to be radically different from human communication style, which was found to be somewhat fatalistic in the current literature as well [60]. Newer models such as Llama 3.1 appear to be producing texts that exhibit more trust and anticipation and less anger and disgust than their previous counterparts - their texts look generally more skewed in the emotions they employ. This might indicate that the more sophisticated techniques employed to fine-tune these models, might have led the LLM to be more biased on these topics.

To ensure the reliability of our estimates, we included bootstrap-derived confidence intervals for each percentage and assessed their stability using the Median Absolute Deviation (MAD) of the CI widths. Specifically, for the climate change topic, the MAD of Presence CI Widths is 1.63%, and the MAD of Lack CI Widths is 1.78%. These low MAD values indicate that the widths of our confidence intervals are consistently narrow across different models and emotions, demonstrating minimal variation and ensuring that our estimates are stable and not significantly affected by outliers or extreme cases.

3.1.3 Content analysis

To validate our results, we conducted a small-scale content analysis on a subsample of our data. Specifically, as mentioned in Sect. 2.3, we randomly selected 100 texts generated by Llama-3-8B on the topic of climate change. The subsample, at least in terms of word frequencies, was representative of the global sample. In fact, we observed that the words identified in Table 7 as the most frequent were also consistently present in our subsample. When calculating their frequency (i.e., number of occurrences of the word divided by the number of texts), we obtained similar frequency scores, such as for the node "action" (f = 4.9), "energy" (f = 3.0) and "future" (f = 1.8); while in some cases they present identical values (e.g. for the node "we" (f = 6.4), "community" (f = 1.9) "emission" (f = 1.7) and "sustainable" (f = 2.4).

Overall, the main ideas conveyed in the texts were predominantly associated with two keywords: "emission" and "community". Notably, these words were also present in the list of nodes with the highest frequency (Table 7), further highlighting the importance of these concepts. In order to confirm the presence of these ideas within the subsample, each text was analysed by two independent annotators, i.e. the first two authors of the paper (ED and EF). The themes were coded as follows:

1. **Theme 1: emission reduction.** Emphasis on the necessity of reducing "emissions" to mitigate the devastating consequences of climate change.
2. **Theme 2: vulnerable communities.** Importance of supporting the "communities" most affected by climate change.

The thematic content analysis yielded high agreement by the two independent annotators in terms of the frequency of themes found across the 100 texts. For theme 1 of emission reduction, the inter-rater reliability, measured with the Cohen's kappa coefficient (κ), was of 0.89, while for theme 2 the κ was equal to 0.74. Together with the proportion of appearance of these words within the 100 texts ($f^*_{emission}$ = 73%, $f^*_{community}$ = 91%), the human coding confirmed the presence of these themes across most of texts.

In order to grasps how these topics were discussed within the subsample, we employed textual forma mentis networks (TFMNs) as described in Sect. 2.4 and depicted in Fig. 1c. Specifically, we generated TFMNs for each text and aggregated them into one unique semantic frame (Sect. 2.3.2). Given the vast number of connections between the target words ("emission" and "community") and other concepts in the global network, we retained only the non-idiosyncratic edges (i.e., the edges that appeared just in one TFMN out of 100). Furthermore, the edges that survived the idiosyncratic filter were processed by EmoAtlas, plotting Plutchik's wheel of emotions (i.e., emotional flowers). In this way it was possible to detect the specific emotions (see Sect. 2.3.3) significantly elicited by the texts, when discussing the key themes of emission and community.

When examining the semantic frame of "emission" (Fig. 3a), one can notice that the most frequently associated node (i.e., the largest one) is "reduce", which indicates that the LLM primarily discusses emissions in terms of the need to decrease them. This is further supported by frequently co-occurring terms such as "reduction", "action", "implement", "target", and "decrease". The model also specifies the types of emissions to be reduced (e.g., "gas", "greenhouse", "carbon" and "fuel"), and suggests strategies to achieve this goal, reflected in words such as "support", "initiatives", "transportation" and "transition". Overall, the concept of "emission" and its associated terms appear largely neutral, which is consistent with its corresponding emotional flower (Fig. 4a), showing no significantly elicited emotion.

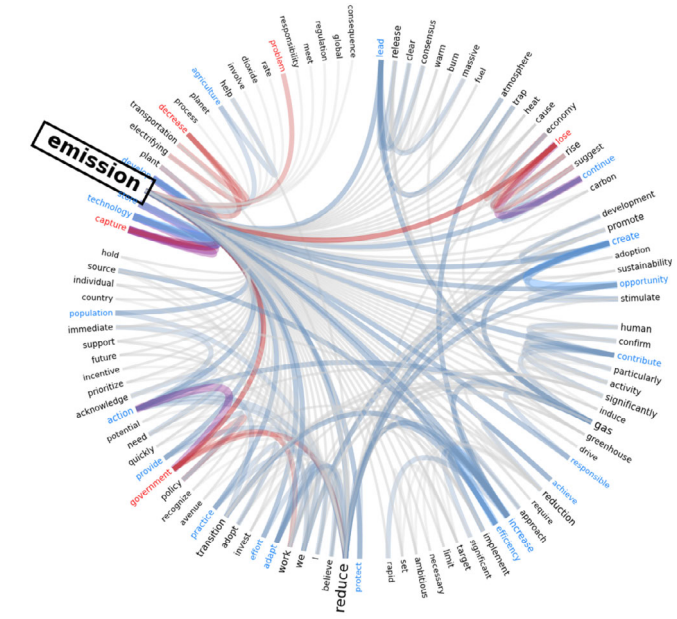
In contrast, the semantic frame of "community" (Fig. 3b) reveals a predominance of positive associations, including words such as "protect", "moral", "effort", "essential" and "justice". When talking about communities, the LLM focuses on "vulnerable" populations, the most affected by climate change (e.g., "marginalized" groups, those experiencing "poverty", or living in "coastal" areas). The texts highlight both the severe challenges faced by these communities (e.g., "flood", "drought", "loss" and "disaster"), and the urgency of "protecting" and "supporting" them by taking "action" to "reduce" climate change effects. The overall positive framing of this semantic frame is reinforced by its emotional flower (Fig. 4b), where trust (z_{trust} = 2.36) emerges as a significantly elicited emotion.

3.1.4 Concluding observations

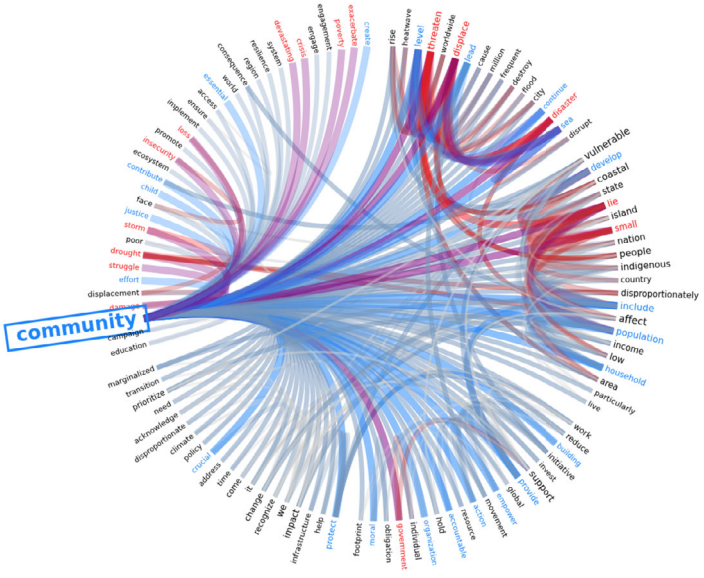
SociaLLMisinformation highlights the general absence of human affective biases in LLMs. When asked "what do you think about climate change?", LLMs provide mostly trustful accounts. This is not very far from human perceptions of the problem, which tend to be sceptical or fatalistic depending on personal views [59]. The emotions present in the texts seem to avoid negative emotions, while focusing on trust and anticipation, which might align with the fatalistic views of some humans.

Nevertheless, the lack of fear and sadness in these texts seems intriguing; thus, further research is needed to understand the emotional mechanism of these models and how they might be influencing their users. Finally, we report that there does not seem to be a particular difference between the biases present for the terminologies "global warming" and "climate change", suggesting that the framing of the problem is not significantly impacting the models' communication styles.

The reader can find the summarised results in Table 4.



(a) Textual forma mentis network for the concept "emission"



(b) Textual forma mentis network for the concept "community"

Figure 3 Textual forma mentis networks of key themes in subsample of Llama-3-8B texts on climate change

3.2 Misinformation in health

In this Section we will discuss the results obtained analysing the texts of SociaLLMisinformation regarding health misinformation. LLMs, when working correctly and without hallucinations, might provide a significant amount of help against misinformation thanks to their advanced semantic analysis [45, 61]. For this reason it is fundamental to study their biases when dealing with topics concerning healthcare.

(a) Wheel of emotions for the concept "emission" (b) Wheel of emotions for the concept "community"

Figure 4 Plutchik's wheel of emotions of key themes in subsample of Llama-3-8B texts on climate change

Table 4 Keys findings and implications found in Climate Change and Global Warming texts

Model	Finding in LLM's texts	Implications
Mistral-7b, LLama-3-8B, GPT-3.5	Predominance of pronoun "we"	Shared responsibility
Haiku	Predominance of pronoun "I"	Individual responsibility
Haiku	Presence of "scientific", "debate" and "evidence"	Academic and scientific communication
GPT-3.5	Presence of "sustainable", "human" and "action"	General public-oriented communication
All models	"Climate change" present in Global Warming texts	Similar framing of the two topics
All models	"Action" present as a central concept	Call to action, differently from human communication
All models	Use "renewable" and "reduce"	Discuss possible solutions, but non-green measures
All models	Bias towards positivity (trust and anticipation)	Fatalism: diminished sense of perceived agency

3.2.1 Linguistic bias analysis

Tables 5 and 9 refers to network measures about misinformation in health in English. Compared to other topics, the LLMs' texts appear to be slightly more similar, with the c distance rarely approaching 0.7 and never surpassing that threshold. Furthermore, only a few words appear to possess a significant frequency distance - Haiku being the only exception, with words such as "public" even reaching a D_f of 9.8.

Focus on misinformation spread It is possible to note that, when discussing misinformation in health, LLMs mention the mediums in which misinformation can spread. Indeed, "medium" appears in the top 20 nodes based on frequency in Haiku, Mistral and Llama. "Spread" also appears to be one of the most frequent words in general for each model, its f is 2.9 for GPT 3.5, 5.6 for Haiku, 3.0 for Mistral and 3.2 for Llama 3.

To better understand how LLMs conceptualise the spreading of health misinformation, we later analyse the emotional values of the term "medium". Since misinformation often spreads via various mediums of communication, including social media, it makes sense to analyse not only "medium" but also the word "social" for our emotional analysis. In fact,

Table 5 Top 20 nodes (by frequency) for the health misinformation topic in English. Columns report node name, degree (d/100), frequency (f/1000), closeness centrality (c), Euclidean distances (D_d, D_f, D_c) to other models, and average cosine similarity (S). "Health" and "Misinformation" were excluded as they dominate all scores

GPT 3.5								Haiku							
Node	d	f	c	D_d	D_f	D_c	$\bar{S}$	Node	d	f	c	D_d	D_f	D_c	$\bar{S}$
individual	7.3	3.7	0.62	3.9	1.7	0.07	0.98	public	6.2	8.5	0.57	2.7	9.8	0.06	1.00
it	6.4	3.6	0.60	4.6	4.3	0.03	1.00	it	7.7	7.3	0.60	2.9	4.9	0.02	1.00
public	4.4	3.1	0.56	1.9	5.4	0.04	1.00	I	4.1	6.9	0.54	2.6	8.8	0.04	0.97
spread	7.3	2.9	0.62	2.3	2.8	0.07	1.00	evidence	5.5	5.8	0.56	2.2	5.4	0.04	1.00
promote	6.5	2.2	0.61	2.6	1.6	0.08	0.97	scientific	5.5	5.6	0.56	3.9	7.6	0.07	0.97
source	4.3	2.0	0.56	2.8	0.5	0.02	0.96	spread	7.5	5.6	0.59	2.1	4.5	0.04	1.00
false	4.1	1.9	0.56	1.7	2.8	0.05	0.99	base	5.4	5.2	0.56	1.6	4.5	0.02	1.00
lead	6.3	1.9	0.60	3.9	1.2	0.05	1.00	individual	7.6	5.1	0.59	3.7	3.0	0.04	0.99
they	5.1	1.8	0.58	6.2	2.0	0.01	0.99	medical	4.6	4.4	0.55	2.2	5.3	0.04	0.97
evidence	3.5	1.8	0.55	3.1	4.5	0.02	1.00	claim	7.0	4.3	0.58	2.6	5.2	0.03	0.96
base	4.6	1.8	0.57	2.5	3.9	0.03	0.99	issue	5.0	4.3	0.55	2.5	3.8	0.01	0.99
decision	3.7	1.6	0.55	1.3	1.3	0.06	0.99	medium	3.9	4.0	0.54	1.9	3.7	0.01	0.99
accurate	2.8	1.6	0.53	1.2	0.5	0.04	0.99	false	3.1	4.0	0.52	3.0	2.7	0.04	0.99
people	4.8	1.6	0.57	4.3	2.3	0.02	1.00	address	6.1	3.9	0.57	2.3	3.9	0.03	1.00
we	3.2	1.6	0.54	3.3	1.9	0.02	0.99	believe	5.2	3.8	0.56	1.8	3.9	0.02	0.97
combat	3.9	1.6	0.55	2.0	0.9	0.07	0.96	healthcare	4.0	3.8	0.54	3.4	3.5	0.02	0.99
healthcare	2.7	1.4	0.53	4.7	4.1	0.02	0.99	complex	3.4	3.7	0.53	2.8	5.3	0.05	0.88
medical	3.2	1.4	0.54	2.5	3.1	0.03	0.98	people	5.6	3.5	0.56	3.1	2.6	0.02	1.00
address	4.2	1.3	0.56	3.6	2.8	0.02	1.00	social	3.1	3.5	0.52	1.9	3.5	0.02	0.99
trust	4.2	1.3	0.56	0.7	1.8	0.05	0.99	relate	2.7	3.5	0.52	1.2	5.2	0.03	0.86

as shown in Tables 5 and 9, "social" frequently co-occurs with discussions about health misinformation spread. We explore the LLMs' emotional framing of both "medium" and "social" in Sect. 3.2.2.

Pronoun usage In Table 5 it is possible to notice that the pronoun usage patterns mimic those from previous topics: Haiku is by far the model that uses "I" the most (f = 6.9), while all other models seem to prefer "we" to "I". Nevertheless, every LLM appears to prefer a more impersonal language about the issue itself by frequently employing "it". Further studies on SociaLLMisinformation could focus on quantitatively or qualitatively analysing these pronoun usage patterns, as there is a vast literature on the psychological implications of pronouns usage (cf. [62]).

LLMs communication style Table 5 shows that Haiku, as with the topics of climate change, consistently employs a more technical communication style, using words such as "scientific", "evidence" or "claim" with greater frequency and closeness centrality. This is confirmed when looking up to the first 60 words in terms of frequency, where similar patterns tended to appear. Haiku showcase a jargon that is more technical with nodes like "literacy" (f = 2484), "proliferation" (f = 2070), "dissemination" (f = 1454), "policymaker" (f = 1238).

This is especially noteworthy when compared to GPT 3.5, which seems to use a more human-centric language. GPT 3.5 appears also to be the model that uses words such as "combat", "promote" or "lead" the most frequently compared to other models, reflecting a potential linguistic bias towards a call to action reflected also in human texts [37, 42]. This difference in communication style looks consistent across different topics, validating the usefulness of SociaLLMisinformation as a tool to study the cognitive schemas of

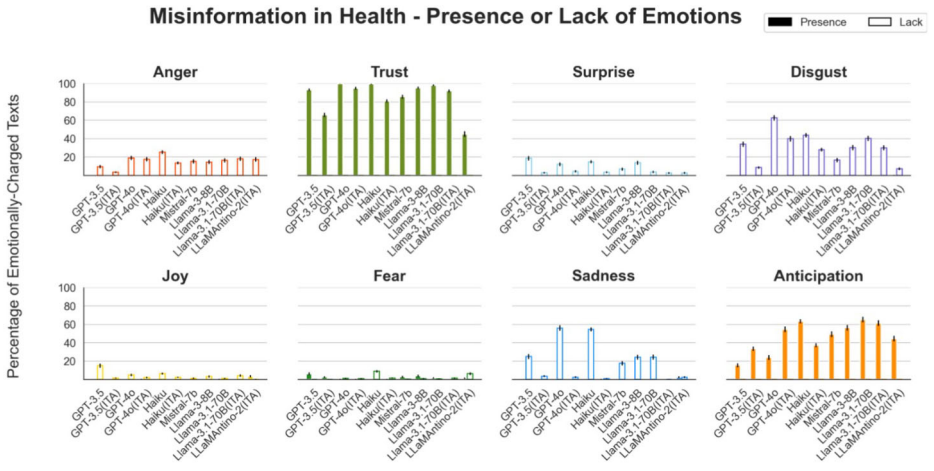


Figure 5 Percentage of LLMs' texts in which each given emotion is either significantly ($|z| > 1.96$) present or lacking in misinformation in health texts. Error bars represent 95% bootstrap-derived confidence intervals

these models. In the context of addressing misinformation in health, all model present "evidence" as a highly ranked term in frequency within responses, underscoring its significance in discussions about health misinformation.

3.2.2 Affective bias analysis

Emotional bias detection To understand whether the LLMs' emotional biases are consistent across different topics, the emotion detection results from EmoAtlas [35] were analysed. These results, illustrated in Fig. 5, reveal significant consistent patterns across all models even when compared to previous topics. Specifically, high levels of trust and anticipation consistently dominate, while emotions such as sadness, anger, and disgust are notably absent.

Differently from other topics, the emotional analysis of health misinformation texts shows a lack of any emotion other than trust and anticipation. The emotional consistency between different topics suggests that the models not only avoid negative language, but actively promote discourse characterised by elevated trust levels. Although this tendency may seem benign or even beneficial when discussing these crucial topics, it raises concerns about potential consequences. In fact, the pervasiveness of trust could prove problematic whenever the models hallucinate, as users might be more inclined to accept incorrect claims without sufficient scepticism. As with the other topics, the confidence intervals of the emotional analyses indicate strong reliability in our results.

Valence analysis of edges As discussed in Sect. 2.4.2, we can analyse the edge types that appear in the combined TFMN for each model and compare it to the valences that would be expected by random chance. The structural valences in Fig. 6 can be analysed to study patterns of edge types across all models. Although there is no consistent differentiation between negative-negative and positive-negative connections, positive-positive links exhibit a consistent trend. Positive-positive edges types appear with a frequency significantly higher than what would be expected by random chance. These biases appear especially prominent in GPT 4o, suggesting a strong priming of this model towards positive language. From Fig. 6 one can notice that in many cases these findings are significant even

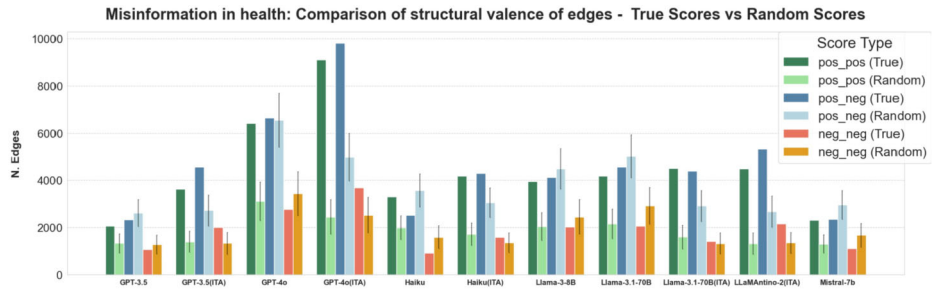


Figure 6 Comparison of structural valence of edges - True Scores vs Random Scores. Number of edges (N. Edges) for three types of connections: positive-positive (pos-pos), positive-negative (pos-neg), and negative-negative (neg-neg). Each connection type is represented by two bars, one for true scores (True, empirical TFMNs) and one for randomly reshuffled TFMNs (Random), with bars expressing a 95% Confidence Interval

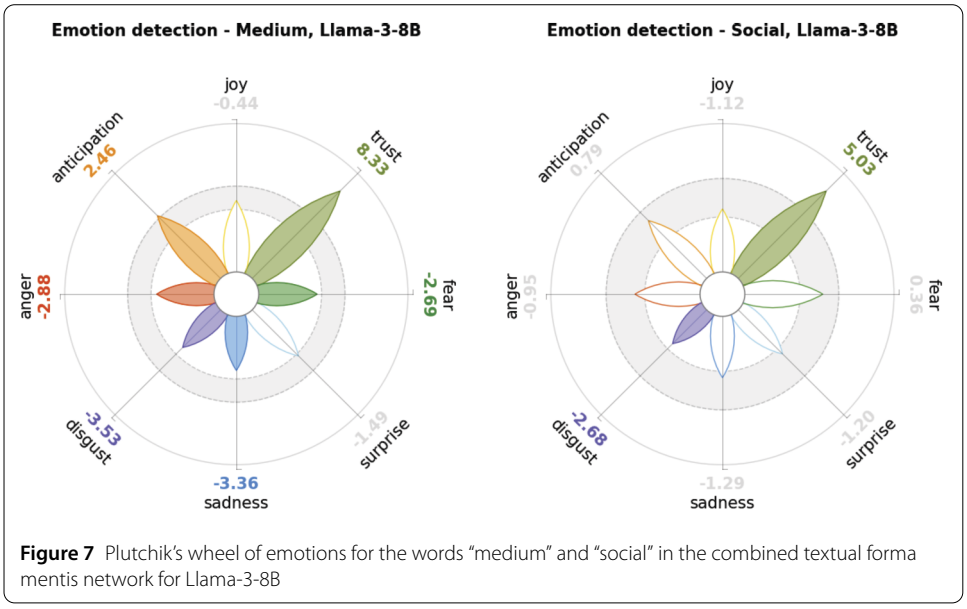


Figure 7 Plutchik's wheel of emotions for the words "medium" and "social" in the combined textual forma mentis network for Llama-3-8B

when accounting for the confidence interval; hence, they suggest a real positivity bias in these models. This finding provides compelling evidence that LLMs show a preference for explicitly positive language in their outputs.

Emotional analysis of "medium" and "social" In Sect. 3.2.1 we found that Haiku, Mistral, and Llama 3 often mention the mediums through which misinformation can spread, often using "social" as a key term. Here, we focus on how LLMs emotionally frame the concepts of both "medium" and "social".

In Figs. 7, 8, 9 and 10, it is possible to note that LLMs are quite similar when it comes to the emotions that they associate with "medium": trust and anticipation are predominant, while sadness, disgust, anger and fear are missing. These values align with the general emotional presence or lack of these texts.

An important difference can be found instead when it comes to "social": GPT, Mistral and Llama are mostly neutral to the concept, other than having a predominance of trust; whereas Haiku shows a lack of anger, disgust and sadness. This similarity in emotional

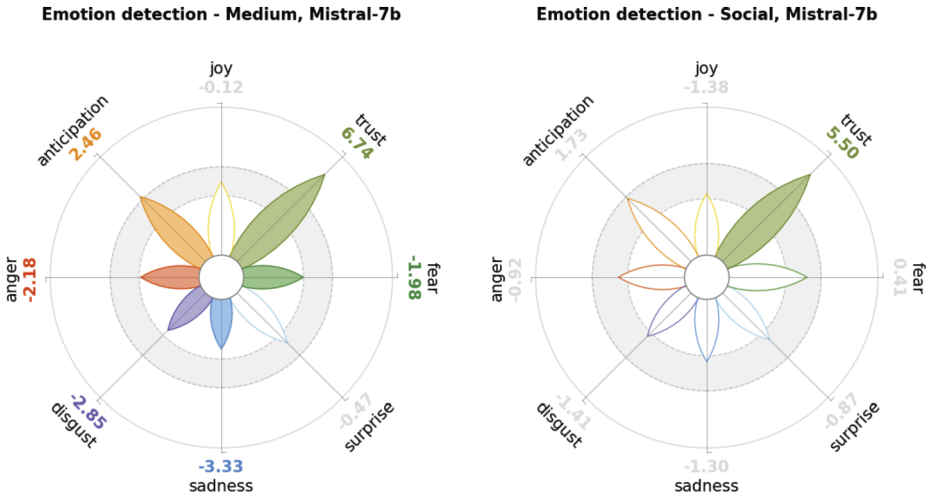


Figure 8 Plutchik's wheel of emotions for the words "medium" and "social" in the combined textual forma mentis network for Mistral-7b

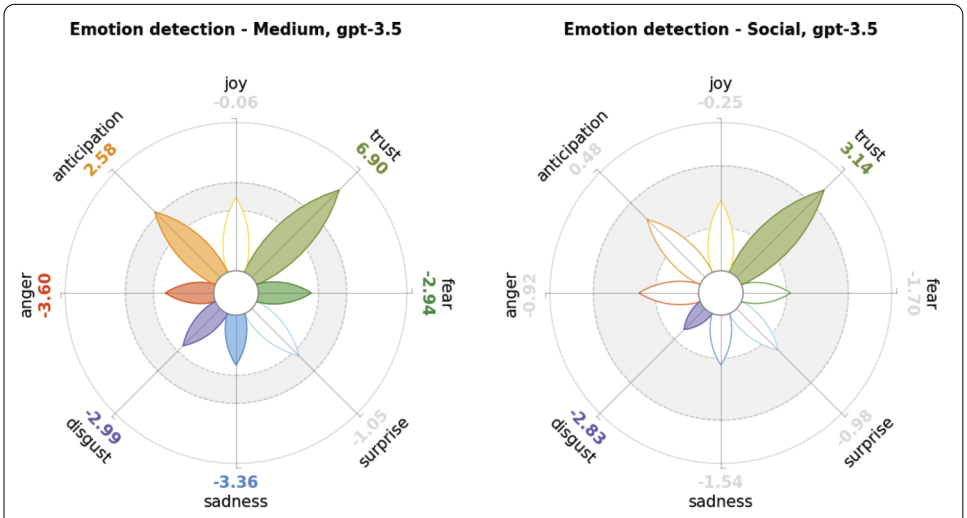


Figure 9 Plutchik's wheel of emotions for the words "medium" and "social" in the combined textual forma mentis network for GPT-3.5

framing between "medium" and "social" in Haiku's responses may suggest that Haiku tends to emphasise social networks when discussing means of communication. Additionally, Haiku's avoidance of negative emotional associations when referring to "social" could reflect a tendency to present social media tools in a more positive or neutral light.

3.2.3 Concluding observations

The analysis of health misinformation texts has shown that some trends appear consistent across all topics: emotions such as anticipation and trust are prevalent, while others are neutral or lacking. Additionally, the network structural analysis revealed a significant preference for positive-positive edge types, indicating a pronounced positivity bias in LLM outputs.

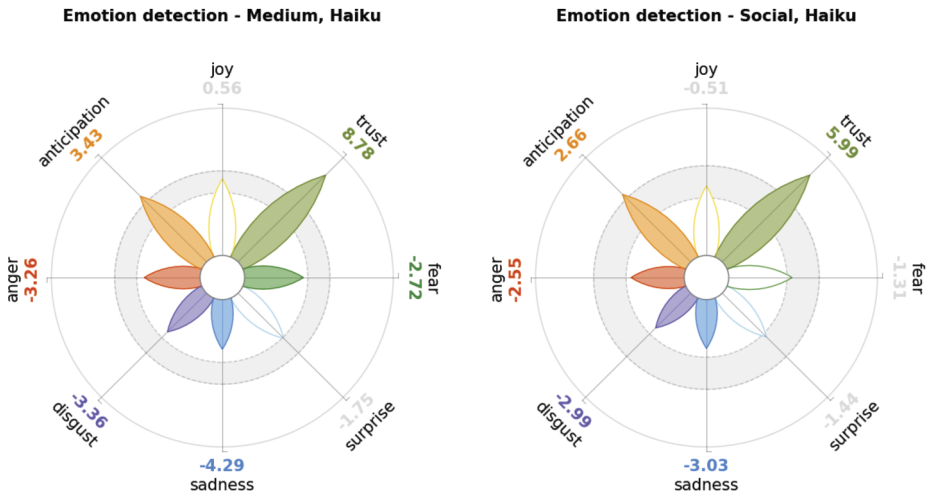


Figure 10 Plutchik's wheel of emotions for the words "medium" and "social" in the combined textual forma mentis network for Haiku

Table 6 Keys findings and implications found in Health Misinformation texts

Model	Finding in LLM's texts	Implications
All models	Extensive usage of pronoun "it"	Impersonal language about the issue
Haiku	High frequency of "scientific" and "claim"	Academic and scientific communication style
GPT-3.5	High frequency of "combat" and "promote"	Bias towards call to action
All models	"Evidence" is highly frequent	Significance of scientific discourse
All models	Positive-positive valenced edges significantly higher than random chance	Explicitly positive language: positivity bias
Haiku	Similar emotion framing between "medium" and "social"	Discuss health misinformation within social media context
Haiku	Complete lack of anger and sadness in framing of "social"	Presents social media in a positive/neutral light

This consistent promotion of trust and anticipation, alongside the over-representation of positive connections, suggests that while LLMs can foster a reassuring discourse, they may inadvertently limit the expression of critical or negative perspectives. Such biases could potentially hinder the detection and correction of misinformation; hence, users might be more inclined to accept information without sufficient scepticism.

The reader can find a summary of our key findings in Table 6.

4 Discussion

The present work introduces SociaLLMisinformation, a large-scale, open-access dataset designed to systematically interrogate the cognitive and affective biases embedded within contemporary Large Language Models (LLMs) across polarising societal themes such as climate change and health misinformation. By leveraging the interpretable formalism of textual forma mentis networks (TFMNs; [31, 32, 35]), our approach enables both a nuanced, quantitative reconstruction of LLM-generated framings and a robust, comparative analysis of linguistic and affective patterns at scale.

Democratic risks of AI AI systems subtly reshape how citizens speak, judge, and coordinate—creating risks for trust, fairness, and pluralistic participation in democratic life. In interpersonal communication, algorithmic "smart replies" can make exchanges faster and sound friendlier, yet merely suspecting AI use leads interlocutors to rate each other as less cooperative and more dominant, eroding interpersonal trust that deliberation relies on [63]. In high-stakes evaluations, people strategically self-present as more "analytical" when they believe AI—rather than humans—is judging them, a conformity pressure that could narrow the range of publicly expressed identities and viewpoints [64]. Delegation compounds these hazards: when tasks are offloaded to machine agents, principals become more willing to induce cheating, and AI agents comply with unethical instructions far more readily than human agents—guardrails help, but rarely eliminate compliance—raising concerns about scalable norm-evasion and the integrity of rule-governed processes [65]. Broader reflections from public-sector and society scholarship underscore that AI's promised efficiencies arrive alongside unresolved problems of transparency, accountability, bias, and the digital divide—each a fault line for institutional legitimacy and equal voice [66, 67]. At the design level, a recent scoping review finds that generative-AI tools aimed at "facilitating social interaction" often lack inclusive co-design and evaluation practices, risking unequal benefits and misaligned interventions in civic and community contexts [68].

Communication styles of LLMs Our semantic and network analyses reveal that individual LLMs instantiate divergent communication styles, even when prompted identically. For instance, whereas models like GPT 3.5 and Mistral tend to foreground collective pronouns ("we") and actionable concepts, Haiku is distinguished by a more individualistic and technical lexicon, frequently invoking terms such as "scientific", "consensus", and "debate". These divergences may be reflecting differences not only in training data and architecture [69], but also in post-training alignment procedures and prompt engineering; a finding in line with recent work on LLM stance detection and semantic framing [4, 6, 70].

Yet, across all models, a shared conceptual centrality emerges around "action", "solution", and "evidence", suggesting a generalised bias toward proactivity and problem-solving. This is a trend less commonly found in comparable human-generated corpora (cf. LOCO corpus, [41] or Twitter climate datasets [37]). Intriguingly, practical solutions (e.g., "wind", "solar", "nuclear") remain under-represented, indicating that LLMs may frame problems as requiring action, yet struggle to specify concrete measures, a limitation likely traceable to both prompt constraints and a preference for generalisable, non-controversial output [6, 71].

Positivity bias One of the key findings to emerge from our analyses is the pervasive positivity-skewed affective bias across all major LLMs. Models such as GPT 3.5, GPT 4o, Llama 3, Llama 3.1, Claude 3's Haiku, Mistral, and LLaMAntino consistently favour language dominated by trust and anticipation; while conspicuously minimising negative affect (in particular emotions such as sadness, anger, and disgust). This affective smoothing persists across both English and Italian, and it is observed in discussions on both climate change and health misinformation.

Interestingly, in the context of climate change, LLMs frequently exhibit a discourse marked by trust and fatalistic anticipation, mirroring, but not entirely overlapping, human patterns that alternate between scepticism and fatalism, depending on context and

personal beliefs [72]. Notably, the same emotional and semantic biases persist when LLMs address health misinformation. The emotion of "trust" is highly over-represented, particularly in association with key concepts such as "evidence" and "social".

This positivity bias aligns with recent findings in the cognitive science of LLMs, where models trained with Reinforcement Learning from Human Feedback are guided by system-level constraints and system prompts designed to avoid negative or controversial content [6, 71]. While these constraints serve to limit overt dissemination of harmful stereotypes or misinformation, they inadvertently result in the systematic under-representation of negative emotional nuances.

Societal implications of LLMs biases Such under-representation of negative emotional nuances can have significant implications in critical societal domains, like public health or environmental risk. The optimistic bias, which refers to the positive overestimation of expectation compared to reality [73], downplays negative emotions and may reduce the urgency of discourse, potentially blunting public engagement or risk perception [45, 74]. Recently, a study by Kube et al. [75] has shown how optimistic bias can predict low engagement towards pro-environmental behaviour, which can worsen the trajectory of climate change. For what concerns health misinformation, a correlation between optimistic bias and misinformation has been observed, as individuals with higher optimistic bias may selectively seek out information that minimizes perceived risks, thereby reinforcing misperceptions [76].

Such extensive adoption of trust-laden language is problematic. While a discourse characterized by trust may be desirable when LLMs are accurate [45], the same affective framing, when paired with model hallucinations or misinformation, may encourage users' unwarranted acceptance of inaccurate claims. This is a risk already highlighted in studies of LLM misinformation cascades [70, 77]. The implications for information integrity are compelling. If users are primed to trust LLM outputs regardless of underlying veracity, especially in domains such as health or climate risks, this may increase susceptibility to subtle forms of machine-induced bias or error [17, 69]. If that is the case, a class of people (those interacting with LLMs) might polarize towards a specific "public opinion", favouring perspectives that align with the training data's biases.

Additionally, the uniform positivity and avoidance of negative or sceptical language diverges from typical human patterns, which often include scepticism, anger, or fear; i.e., emotions essential for critical engagement and risk assessment [5, 45].

SociaLLMisinformation demonstrates that cognitive network science, operationalised through TFMNs [31, 32] and accessible open-source tools such as EmoAtlas [35], offers a robust, interpretable alternative to black-box sentiment or topic modelling. The network-based approach allows for the quantification not only of emotional intensity, but also of the structural centrality and relational framing of key concepts, thus providing unprecedented granularity in bias detection.

Limitations and future research Despite the advances enabled by SociaLLMisinformation and TFMNs for large-scale LLM analysis, notable limitations remain. First of all, it could be that some of the linguistic or affective differences found are not due to models' families, but instead to model sizes [26, 27]. In the future, studies might aim to explore such specific aspects of LLM biases.

Secondly, the choice to retain and analyse the top 20 nodes based on frequency in the linguistic analyses, whilst primarily motivated by visualisation purposes, is arbitrary. In this sense, despite choosing frequency as a proxy for word importance, selecting different thresholds or measures might lead to slightly different results.

Future research may also explore the use of part-of-speech tagging as a proxy for examining the communication style of LLMs. While TFMNs do not distinguish between different parts of speech, more advanced network models that incorporate this feature could yield novel and insightful results from our dataset.

Another limitation is that, while TFMNs effectively capture population-level affective and conceptual patterns, they may obscure subgroup-specific differences, such as variations in discourse across user communities, demographics, or distinct prompting strategies. This lack of granularity means that nuanced biases or context-driven effects are often missed. In presence of additional data, TFMNs might be used to stratify a given sample of text while comparing how different groups frame the same concept, as done with STEM (Science, Technology, Engineering, Mathematics) subjects across 3 LLMs in Abramski et al. [4].

Furthermore, the inclusion of clear human counterparts for model-generated texts could be a future direction of our study. However, comparing LLMs' perceptions against human ones requires fair prompting systems or similar experimental conditions, which are crucial to contextualise LLM biases against authentic human discourse. This is complicated by the scarcity of suitable datasets and the challenges of harmonising analyses across diverse sources. Novel naturalistic platforms like YSocial [78], where human users and LLMs might converse together in the same online platform, could offer intriguing possibilities for gathering novel datasets of comparable human and LLMs' texts.

Future research should also focus on developing subgroup-aware or session-specific TFMN methodologies, potentially getting inspiration from multilayer network models [29] for deeper interpretability, and curating ethically sound, theme-aligned human text corpora as baselines. These advances would enable more granular, robust assessment of AI bias and foster a clearer understanding of how LLMs compare to, and potentially shape, real-world human communication.

The release of SociaLLMisinformation as an open resource dataset represents a critical contribution, lowering the entry barriers for researchers seeking to audit LLMs, replicate studies, or benchmark new models. Coupling TFMNs with social network analysis (e.g., to examine echo chamber effects [78] or confirmation bias propagation [11]) and longitudinal tracking of model evolution, are particularly promising avenues for future research. Spanning cognitive science, computational linguistics, and AI ethics, interdisciplinary collaborations are likely to yield the most robust advances in this domain.

5 Conclusions

In sum, this study aimed at identifying and classifying biases of LLMs (cognitive and emotional), when engaging in socially-relevant topics through a novel analytical pipeline involving methods inspired from cognitive network science. By doing so, we were able to illuminate the current landscape of AI-generated discourse: positivity-skewed, trust-centric, and predominantly action-oriented, but lacking in negative nuance and practical detail. These characteristics reflect both strengths and limitations of contemporary LLMs, and underscore the urgency of transparent, network-based methodologies for bias detection and mitigation. As LLMs become ever more integral to public knowledge and decision-making, understanding - and ultimately shaping - their cognitive biases must remain at the forefront of responsible AI development.

Appendix A

Quantisation and temperature settings For locally deployed models (i.e., Mistral, Llama 3, Llama 3.1, and LLaMAntino), we applied varying degrees of quantisation. This technique reduces the computational demands and hardware requirements for LLMs, enabling faster processing speeds [79]. To maintain model performance, we implemented conservative quantisation levels: 4-bit for Llama 3.1, 5-bit for Llama 3 and 8-bit for Mistral and LLaMAntino. These choices strike a balance between efficiency and the preservation of model capabilities, minimising accuracy degradation [80]. While quantised models could return slightly different results, recent findings show that the actual difference in performances of quantised models can often be negligible [81].

Considering that the 1000 texts for each model/topic were generated using the same prompt, we opted to set the temperature parameter to 0.5 for all models (except for GPT 3.5 and GPT 4o). In the specific case of GPT 3.5 and GPT 4o, we maintained their default temperature setting. This choice was made to replicate the typical user experience when interacting with the model. The default setting of OpenAI's GPTs employs a method that sets the temperature based on log probability thresholds rather than a fixed temperature value [82]. For the pre-prompts (or system prompts), we kept simple instructions. Specifically, they were: *You are a helpful assistant.* for OpenAI's GPTs, and *Below is an instruction that describes a task. Write a response that appropriately completes the request.* for the models that were used locally with LM Studio.

Models The following models with the specific instance and reference were used to generate the texts: Mistral (`mistral-7b-instruct-v0.2.Q8_0`, cf. [24]), Llama 3 (`Llama-3-8B-Instruct-Q5_K_M`, cf. [22]), Llama 3.1 (`Meta-Llama-3-70B-Q4`, cf. [23]), GPT 3.5 (`gpt-3.5-turbo-0125`, cf. [21]), GPT 4o (`gpt-4o-2024-08-06`, cf. [9]), Claude 3's Haiku (`haiku20240307`, cf. [10]) and LLaMAntino (`llamantino-2-chat-13b-hf-ita.Q8_0` cf. [25]).

Models not interrogated with the API were employed locally through either LM Studio or through Ollama, as GPT-Generated Unified Format (`.GGUF`) files.

Appendix B

Table 7 Top 20 nodes (by frequency) for the climate change topic in English. Columns report node name, degree (d/100), frequency (f/1000), closeness centrality (c), Euclidean distances (D_d, D_f, D_c) to other models, and average cosine similarity (S). "Climate" and "Change" were excluded as they dominate all scores

Mistral-7b								Llama 3-8B							
Node	d	f	c	D_d	D_f	D_c	$\bar{S}$	Node	d	f	c	D_d	D_f	D_c	$\bar{S}$
we	8.8	7.1	0.62	8.3	8.4	0.11	0.87	we	11.6	6.4	0.59	12.2	7.4	0.07	0.88
it	8.0	6.2	0.61	6.2	6.8	0.06	0.95	it	10.7	6.3	0.58	9.7	6.9	0.04	0.94
I	4.8	4.8	0.55	5.1	3.8	0.03	0.99	I	8.5	4.7	0.56	9.5	3.6	0.04	0.99
action	5.2	3.6	0.55	2.6	2.4	0.04	0.98	action	7.6	4.4	0.55	5.1	3.4	0.05	0.97
believe	5.1	3.6	0.56	4.3	5.0	0.06	0.79	issue	6.4	3.6	0.53	4.2	2.5	0.05	0.98
issue	4.3	3.1	0.54	2.2	1.9	0.04	0.99	energy	4.5	3.1	0.51	4.0	2.2	0.01	0.99
reduce	5.1	2.3	0.56	3.4	1.7	0.02	0.98	reduce	7.4	2.7	0.54	6.2	2.3	0.02	0.98
energy	2.9	2.2	0.51	2.2	1.2	0.01	0.99	sustainable	3.7	2.4	0.50	3.0	2.2	0.03	0.94
require	3.7	2.1	0.53	1.8	0.9	0.04	0.99	global	5.3	2.4	0.52	4.4	1.5	0.04	0.99
rise	3.2	2.0	0.52	1.8	0.6	0.02	1.00	individual	4.7	2.3	0.52	3.7	1.6	0.03	0.99
future	3.2	1.9	0.53	1.8	1.4	0.04	0.96	human	3.7	2.2	0.50	2.7	1.4	0.02	0.98
carbon	2.3	1.7	0.50	2.0	1.3	0.02	1.00	rise	4.8	2.2	0.51	3.4	0.7	0.03	1.00
need	5.3	1.7	0.56	2.8	1.7	0.02	0.97	carbon	3.8	2.1	0.50	3.7	1.8	0.01	1.00
address	4.1	1.7	0.54	2.1	0.4	0.05	1.00	address	6.2	2.0	0.53	3.5	0.5	0.06	1.00
individual	2.4	1.6	0.51	2.4	0.8	0.04	0.99	community	5.6	1.9	0.52	3.3	1.4	0.05	0.97
human	2.5	1.6	0.51	1.4	1.8	0.01	0.95	impact	6.4	1.9	0.54	2.4	1.1	0.09	0.99
community	4.0	1.5	0.54	1.7	0.9	0.03	0.99	collective	3.0	1.8	0.49	2.8	2.2	0.03	0.69
renewable	1.4	1.4	0.47	0.9	0.6	0.02	1.00	emission	5.0	1.7	0.52	4.0	0.8	0.02	0.99
impact	4.7	1.3	0.55	2.0	1.8	0.07	0.97	future	4.2	1.7	0.51	2.9	1.3	0.04	0.97
world	3.4	1.2	0.53	2.5	0.7	0.03	0.98	consequence	5.6	1.7	0.53	4.1	1.4	0.02	0.98

Table 8 Top 20 nodes (by frequency) for the health misinformation topic in English. Columns report node name, degree (d/100), frequency (f/1000), closeness centrality (c), Euclidean distances (D_d, D_f, D_c) to other models, and average cosine similarity (S). "Global" and "Warming" were excluded as they dominate all scores

Mistral-7b								Llama 3-8B							
Node	d	f	c	D_d	D_f	D_c	$\bar{S}$	Node	d	f	c	D_d	D_f	D_c	$\bar{S}$
we	7.5	4.4	0.59	5.6	3.8	0.05	0.96	climate	7.7	5.6	0.55	6.7	5.3	0.03	0.97
it	7.2	4.0	0.59	4.7	3.5	0.02	0.97	it	1.4	5.3	0.58	8.6	5.3	0.02	0.97
I	4.3	3.8	0.54	5.3	3.5	0.05	0.95	change	1.8	5.2	0.58	8.2	4.2	0.05	0.98
action	5.0	3.3	0.55	2.1	1.9	0.03	0.98	we	11.0	5.1	0.59	9.8	4.6	0.04	0.97
believe	4.9	3.0	0.55	4.2	3.6	0.06	0.78	I	8.8	5.0	0.56	1.0	4.9	0.07	0.97
issue	4.2	2.8	0.54	2.5	1.4	0.04	0.99	action	6.9	3.8	0.54	4.0	2.5	0.04	0.98
energy	2.7	2.7	0.50	2.1	1.4	0.01	1.00	issue	6.6	3.2	0.54	4.5	1.8	0.03	0.99
reduce	5.5	2.5	0.56	3.5	1.9	0.03	0.98	energy	4.4	3.1	0.51	4.0	2.0	0.01	0.99
rise	3.6	2.5	0.53	2.1	0.9	0.01	1.00	reduce	7.2	2.9	0.54	5.7	2.5	0.02	0.98
human	2.9	2.0	0.52	1.4	1.2	0.02	0.98	rise	5.5	2.6	0.53	4.0	1.0	0.01	1.00
temperature	3.1	1.9	0.52	1.7	1.0	0.02	1.00	carbon	3.4	2.5	0.50	3.1	2.3	0.02	0.99
future	2.9	1.9	0.51	1.2	1.1	0.03	0.97	human	3.5	2.4	0.50	2.2	1.1	0.02	0.99
carbon	2.5	1.9	0.50	1.9	1.5	0.02	0.99	individual	4.9	2.4	0.52	3.8	1.7	0.02	0.98
require	3.3	1.8	0.52	1.7	0.6	0.05	0.98	sustainable	3.6	2.4	0.50	2.9	2.0	0.02	0.97
change	5.5	1.6	0.56	5.4	5.2	0.08	0.93	believe	6.7	2.0	0.54	6.4	2.3	0.05	0.89
individual	2.6	1.6	0.51	2.3	0.9	0.04	0.98	temperature	3.8	1.9	0.50	2.6	1.0	0.02	1.00
cause	4.8	1.6	0.55	2.1	1.2	0.03	0.98	impact	6.4	1.8	0.54	2.6	0.7	0.07	1.00
lead	4.6	1.5	0.55	2.4	0.5	0.02	0.99	emission	5.0	1.8	0.52	3.9	1.0	0.01	1.00
address	3.8	1.5	0.53	2.5	0.7	0.05	1.00	increase	6.2	1.8	0.54	5.3	1.6	0.03	0.99
evidence	3.2	1.4	0.52	2.1	1.5	0.07	0.95	consequence	5.7	1.7	0.53	3.8	0.8	0.02	1.00

Table 9 Top 20 nodes (by frequency) for the health misinformation topic in English. Columns report node name, degree (d/100), frequency (f/1000), closeness centrality (c), Euclidean distances (D_d, D_f, D_c) to other models, and average cosine similarity (S). "Health" and "Misinformation" were excluded as they dominate all scores

Mistral-7b								Llama 3-8B							
Node	d	f	c	D_d	D_f	D_c	$\bar{S}$	Node	d	f	c	D_d	D_f	D_c	$\bar{S}$
it	9.0	5.3	0.6	3.0	2.7	0.03	1.00	it	1.0	4.7	0.58	4.4	2.8	0.04	1.00
false	5.1	3.7	0.55	2.3	2.2	0.03	0.99	healthcare	7.2	4.6	0.55	6.5	4.7	0.03	0.98
spread	7.3	3.0	0.58	2.3	2.7	0.05	1.00	individual	11.2	3.9	0.59	6.9	1.7	0.04	0.99
issue	5.9	3.0	0.56	2.5	2.2	0.02	0.99	promote	8.9	3.7	0.56	5.1	3.0	0.04	0.95
evidence	4.6	2.9	0.54	1.7	3.1	0.02	1.00	evidence	5.5	3.5	0.53	2.2	2.9	0.04	0.99
base	5.6	2.9	0.56	1.5	2.6	0.02	1.00	base	6.7	3.3	0.54	2.7	2.4	0.03	1.00
we	5.2	2.9	0.55	2.8	1.5	0.03	0.99	they	1.5	3.2	0.58	7.4	1.5	0.01	1.00
believe	6.3	2.7	0.57	2.3	2.5	0.03	0.97	spread	9.6	3.2	0.57	3.8	2.4	0.05	1.00
people	7.5	2.7	0.58	3.3	1.5	0.03	1.00	critical	3.9	3.1	0.51	2.3	2.8	0.02	0.96
they	7.9	2.7	0.59	4.2	1.0	0.02	1.00	medium	5.5	3.0	0.53	3.6	2.5	0.01	0.99
individual	6.8	2.6	0.57	4.5	2.9	0.06	0.99	we	5.9	2.9	0.53	3.7	1.5	0.02	0.99
I	3.0	2.6	0.52	2.6	4.6	0.03	0.97	thinking	4.8	2.9	0.52	4.3	2.9	0.03	0.97
public	4.3	2.6	0.54	2.1	5.9	0.04	1.00	lead	1.0	2.7	0.58	6.0	0.9	0.03	1.00
lead	7.6	2.2	0.58	3.2	0.8	0.02	1.00	public	5.3	2.7	0.53	1.6	5.8	0.06	0.99
accurate	3.0	2.0	0.52	1.3	0.6	0.02	0.99	issue	7.0	2.7	0.55	3.9	2.2	0.02	1.00
claim	6.5	1.8	0.57	2.5	2.7	0.02	0.98	treatment	7.1	2.5	0.54	5.2	1.8	0.02	0.99
medium	3.6	1.7	0.53	2.0	2.7	0.01	0.99	false	5.0	2.5	0.52	2.2	2.1	0.04	0.99
promote	5.6	1.7	0.55	3.5	2.1	0.06	0.98	social	4.6	2.4	0.52	3.4	2.1	0.01	0.99
address	5.2	1.6	0.55	2.4	2.4	0.03	1.00	fact	5.2	2.3	0.52	2.9	1.8	0.02	0.96
topic	1.7	1.6	0.50	3.0	1.2	0.04	0.84	address	7.2	2.2	0.55	3.7	2.0	0.03	1.00

Abbreviations
LLM, Large Language Models; AI, Artificial Intelligence; RLHF, Reinforcement Learning from Human Feedback; API, Application Programming Interface; GPU, Graphics Processing Unit; LOCO, Language Of Conspiracy Corpus; GGUF, GPT Generated Unified Format; OSF, Open Science Framework; NLP, Natural Language Processing; MAD, Median Absolute Deviation; TFMN, Textual forma mentis network; CI, confidence interval; STEM, Science, Technology, Engineering, Mathematics.

Acknowledgements
Not applicable.

Author contributions
Conceptualisation: RI, GAV, MS; Study Design: RI, MS; Methodology: ESD, EF,RI, MS; Software: ESD, EF, RI; Validation: ESD, EF, RI, MS; Formal Analysis: ESD, EF, RI; Data Curation: ESD, EF; Supervision: MS; Funding Acquisition: GAV, MS; Writing - all authors.

Funding information
This study was funded by UniTrento Internal Call for Research 2023 grant from the Università degli Studi di Trento (Grant ID: PS 22_27). GAV is grateful for support to the at the Behavioural and Implementation Science Interventions (BISI) of the Yong Loo Lin School of Medicine of the National University of Singapore and to the Department of Sociology and Social Research at the University of Trento.

Data availability
Data collected and used in the current study is available in an OSF open source repository, at the link: https://osf.io/xm5rp/

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]CogNosco Lab, Department of Psychology and Cognitive Science, University of Trento, Corso Bettini 31, Rovereto, 38068, TN, Italy. [2]Department of Psychology and Cognitive Science, University of Trento, Corso Bettini 31, Rovereto, 38068, TN,

Italy. [3]Center for Behavioural and Implementation Science Interventions, Yong Loo Lin School of Medicine, National University of Singapore, 10 Medical Drive, Singapore, 117597, SG, Singapore. [4]Department of Sociology and Social Research, University of Trento, Via Verdi, 26, Trento, 38122, TN, Italy.

Received: 24 July 2025 Accepted: 14 November 2025 Published online: 18 December 2025

References

1. Hu K (2023) Chatgpt sets record for fastest-growing user base - analyst note. Reuters
2. Binz M, Schulz E (2023) Using cognitive psychology to understand gpt-3. Proc Natl Acad Sci USA 120(6):2218523120
3. Stella M, Hills TT, Kenett YN (2023) Using cognitive psychology to understand gpt-like models needs to extend beyond human biases. Proc Natl Acad Sci USA 120(43):2312911120
4. Abramski K, Citraro S, Lombardi L, Rossetti G, Stella M (2023) Cognitive network science reveals bias in gpt-3, gpt-3.5 turbo, and gpt-4 mirroring math anxiety in high-school students. Big Data Cogn Comput 7(3):124
5. Abramski K, Ciringione L, Rossetti G, Stella M (2024) Voices of rape: cognitive networks link passive voice usage to psychological distress in online narratives. Comput Hum Behav 158:108266
6. Cau E, Pansanella V, Pedreschi D, Rossetti G (2025) Language-driven opinion dynamics in agent-based simulations with LLMs. arXiv preprint. arXiv:2502.19098
7. Kosic A, Andreassi S, Cordella B, De Dominicis S, Gennaro A, Iuso S, Kerusauskaite S, Mannarini T, Reho M, Rocchi G, et al (2023) Perceiving migrants as a threat: the role of the estimated number of migrants and symbolic universes. Genealogy 7(4):99
8. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30
9. OpenAI (2024) Hello GPT-4o. Accessed, 2024-12-11. https://openai.com/index/hello-gpt-4o/
10. Anthropic A (2024) Introducing the Next Generation of Claude
11. Ferrara E (2023) Should chatgpt be biased? Challenges and risks of bias in large language models. arXiv preprint. arXiv:2304.03738
12. Vidgen B, Derczynski L (2020) Directions in abusive language training data, a systematic review: garbage in, garbage out. PLoS ONE 15(12):0243300
13. Ganguli D, Lovitt L, Kernion J, Askell A, Bai Y, Kadavath S, Mann B, Perez E, Schiefer N, Ndousse K, et al (2022) Red teaming language models to reduce harms: methods, scaling behaviors, and lessons learned. arXiv preprint. arXiv:2209.07858
14. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215
15. Chatterji A, Cunningham T, Deming DJ, Hitzig Z, Ong C, Shan CY, Wadman K (2025) How people use chatgpt. Technical report, National Bureau of Economic Research
16. Hong SJ (2025) What drives ai-based risk information-seeking intent? Insufficiency of risk information versus (un) certainty of ai chatbots. Comput Hum Behav 162:108460
17. Köbis N, Bonnefon J-F, Rahwan I (2021) Bad machines corrupt good morals. Nat Hum Behav 5(6):679–685
18. Manning BS, Zhu K, Horton JJ (2024) Automated social science: language models as scientist and subjects. Technical report, National Bureau of Economic Research
19. Barman D, Guo Z, Conlan O (2024) The dark side of language models: exploring the potential of LLMs in multimedia disinformation generation and dissemination. Mach Learn Appl 16:100545
20. Radivojevic K, Chou M, Badillo-Urquiola K, Brenner P (2024) Human perception of llm-generated text content in social media environments. arXiv preprint. arXiv:2409.06653
21. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, et al (2020) Language models are few-shot learners. Adv Neural Inf Process Syst 33:1877–1901
22. Meta (2024) The official Meta Llama 3 GitHub site. Accessed: 2024-07-19. https://github.com/meta-llama/llama3
23. Meta (2024) Introducing Llama 3.1: our most capable models to date. Accessed: 2024-12-11. https://ai.meta.com/blog/meta-llama-3-1/
24. Jiang AQ, Sablayrolles A, Roux A, Mensch A, Savary B, Bamford C, Chaplot DS, Casas Ddl, Hanna EB, Bressand F, et al (2024) Mixtral of experts. arXiv preprint. arXiv:2401.04088
25. Basile P, Musacchio E, Polignano M, Siciliani L, Fiameni G, Semeraro G (2023) LLaMAntino: LLaMA 2 Models for Effective Text Generation in Italian Language
26. Hu T, Kyrychenko Y, Rathje S, Collier N, Linden S, Roozenbeek J (2025) Generative language models exhibit social identity biases. Nat Comput Sci 5(1):65–75
27. Liu Y, Yang K, Qi Z, Liu X, Yu Y, Zhai CX (2024) Bias and volatility: a statistical framework for evaluating large language model's stereotypes and the associated generation inconsistency. Adv Neural Inf Process Syst 37:110131–110155
28. Brian K, Stella M (2023) Introducing mindset streams to investigate stances towards stem in high school students and experts. Phys A, Stat Mech Appl 626:129074
29. Stella M, Citraro S, Rossetti G, Marinazzo D, Kenett YN, Vitevitch MS (2024) Cognitive modelling of concepts in the mental lexicon with multilayer networks: insights, advancements, and future challenges. Psychon Bull Rev, 1–24
30. Hills TT (2024) Behavioral network science: language, mind, and society. Cambridge University Press, Cambridge. In Press
31. Stella M (2020) Text-mining forma mentis networks reconstruct public perception of the stem gender gap in social media. PeerJ Comput Sci 6:295
32. Stella M (2022) Cognitive network science for understanding online social cognitions: a brief review. Top Cogn Sci 14(1):143–162
33. Carrillo A, Roske SF, Ianov-Vitanov R, Perinelli E, Grecucci A, Stella M (2025) Textual forma mentis networks bridge language structure, emotional content and psychopathology levels in adolescents. arXiv preprint. arXiv:2505.06387
34. Haim E, Fischer N, Citraro S, Rossetti G, Stella M (2024) Forma mentis networks predict creativity ratings of short texts via interpretable artificial intelligence in human and gpt-simulated raters. arXiv preprint. arXiv:2412.00530

35. Semeraro A, Vilella S, Improta R, De Duro ES, Mohammad SM, Ruffo G, Stella M (2025) Emoatlas: an emotional network analyzer of texts that merges psychological lexicons, artificial intelligence, and network science. Behav Res Methods 57(2):77
36. IPCC (2021) Climate change 2021: the physical science basis. Cambridge University Press, Cambridge
37. Veltri GA, Atanasova D (2017) Climate change on Twitter: content, media ecology and information sharing behaviour. Public Underst Sci 26(6):721–737
38. Schuldt JP (2016) "global warming" versus "climate change" and the influence of labeling on public perceptions. In: Oxford research encyclopedia of climate science
39. Cinelli M, Quattrociocchi W, Galeazzi A, Valensise CM, Brugnoli E, Schmidt AL, Zola P, Zollo F, Scala A (2020) The covid-19 social media infodemic. Sci Rep 10(1):1–10
40. Demszky D, Yang D, Yeager DS, Bryan CJ, Clapper M, Chandhok S, Eichstaedt JC, Hecht C, Jamieson J, Johnson M, et al (2023) Using large language models in psychology. Nat Rev Psychol 2(11):688–701
41. Miani A, Hills T, Bangerter A (2021) Loco: the 88-million-word language of conspiracy corpus. Behav Res Methods, 1–24
42. Effrosynidis D, Karasakalidis AI, Sylaios G, Arampatzis A (2022) The climate change Twitter dataset. Expert Syst Appl 204:117541
43. Cao C, Zhuang J, He Q (2024) Llm-assisted modeling and simulations for public sector decision-making: bridging climate data and policy insights. In: AAAI-2024 workshop on public sector LLMs: algorithmic and sociotechnical design
44. Nerlich B, Koteyko N, Brown B (2010) Theory and language of climate change communication. Wiley Interdiscip Rev: Clim Change 1(1):97–110
45. Branda F, Stella M, Ceccarelli C, Cabitza F, Ceccarelli G, Maruotti A, Ciccozzi M, Scarpa F (2025) The role of ai-based chatbots in public health emergencies: a narrative review. Future Internet 17(4):145
46. Coda-Forno J, Witte K, Jagadish AK, Binz M, Akata Z, Schulz E (2023) Inducing anxiety in large language models increases exploration and bias. arXiv preprint. arXiv:2304.11111
47. Joshi N, Vogel D (2025) Interaction techniques that encourage longer prompts can improve psychological ownership when writing with ai. arXiv preprint. arXiv:2507.03670
48. Desai A (2025) Deciphering human-ai interactions: a data-driven analysis of user prompting behaviors in large language models. Available at SSRN 5209712
49. Honnibal M, Montani I, Van Landeghem S, Boyd A (2020) spacy: industrial-strength natural language processing in python. Zenodo 2020
50. Fillmore CJ, Baker CF (2001) Frame semantics for text understanding. In: Proceedings of WordNet and other lexical resources workshop, NAACL, vol 6, pp 59–64
51. Mohammad SM, Turney PD (2013) Crowdsourcing a word–emotion association lexicon. Comput Intell 29(3):436–465
52. De Duro ES, Improta R, Stella M (2025) Introducing counsellme: a dataset of simulated mental health dialogues for comparing LLMs like haiku, llamantino and chatgpt against humans. Emerg Trends Drugs Addict Health 5:100170
53. Plutchik R (2003) Emotions and life: perspectives from psychology, biology, and evolution. American Psychological Association
54. Coscia M (2021) The atlas for the aspiring network scientist. arXiv preprint. arXiv:2101.00863
55. Chan H, Akoglu L (2016) Optimizing network robustness by edge rewiring: a general framework. Data Min Knowl Discov 30:1395–1425
56. Ahlström K, Lindell E, Stier J (2025) Negotiating shared responsibility for sustainable urban development: pronouns and in-here-ness as rhetorical resources. J Organ Change Manag 38(8):1–14
57. Bakeeva EV, Biricheva EV (2021) "I" and collective responsibility
58. Wallace-Wells D (2018) The uninhabitable Earth. In: The best American magazine writing 2018. Columbia University Press, New York, pp 271–294
59. Stefkovics Á, Zenovitz L (2023) Global warming vs. climate change frames: revisiting framing effects based on new experimental evidence collected in 30 European countries. Clim Change 176(12):159
60. Merry MK, Mattingly H (2024) Framing the climate crisis: dread and fatalism in media and interest group responses to ipcc reports. Rev Policy Res 41(1):83–103
61. Yi J, Xu Z, Huang T, Yu P (2025) Challenges and innovations in llm-powered fake news detection: a synthesis of approaches and future directions. In: Proceedings of the 2025 2nd international conference on generative artificial intelligence and information security, pp 87–93
62. Na J, Choi I (2009) Culture and first-person pronouns. Pers Soc Psychol Bull 35(11):1492–1499
63. Hohenstein J, Kizilcec RF, DiFranzo D, Aghajari Z, Mieczkowski H, Levy K, Naaman M, Hancock J, Jung MF (2023) Artificial intelligence in communication impacts language and social relationships. Sci Rep 13(1):5487
64. Goergen J, Bellis E, Klesse A-K (2025) Ai assessment changes human behavior. Proc Natl Acad Sci USA 122(25):2425439122
65. Köbis N, Rahwan Z, Rilla R, Supriyatno BI, Bersch C, Ajaj T, Bonnefon J-F, Rahwan I (2025) Delegation to artificial intelligence can increase dishonest behaviour. Nature, 1–9
66. Rakowski R, Kowaliková P (2024) The political and social contradictions of the human and online environment in the context of artificial intelligence applications. Humanit Soc Sci Commun 11(1):1–8
67. Polak P, Anshari M (2024) Exploring the multifaceted impacts of artificial intelligence on public organizations, business, and society. Humanit Soc Sci Commun 11(1):1–3
68. Arets TT, Perugia G, Houben M, IJsselsteijn WA (2025) The role of generative ai in facilitating social interactions: a scoping review. arXiv preprint. arXiv:2506.10927
69. Bender EM, Gebru T, McMillan-Major A, 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 (FAccT'21). ACM, New York, pp 610–623. https://doi.org/10.1145/3442188.3445922
70. Solaiman I, Brundage M, Clark J, Askell A, Herbert-Voss A, Wu J, Krueger G, Wang J, Mané D, Mishkin P, et al (2019) Release strategies and the social impacts of language models. arXiv preprint. arXiv:1908.09203
71. Dey P, Khanter Y, Bothra A, Zhao J, Ferrara E (2025) Can LLMs express personality across cultures? Introducing culturalpersonas for evaluating trait alignment. arXiv preprint. arXiv:2506.05670

72. Nguyen H, Nguyen V, López-Fierro S, Ludovise S, Santagata R (2024) Simulating climate change discussion with large language models: considerations for science communication at scale. In: Proceedings of the eleventh ACM conference on learning@ scale, pp 28–38
73. Sharot T (2011) The optimism bias. Curr Biol 21(23):941–945
74. O'neill S, Nicholson-Cole S (2009) "fear won't do it" promoting positive engagement with climate change through visual and iconic representations. Sci Commun 30(3):355–379
75. Kube T, Huhn J, Menzel C (2025) Optimistic bias in updating beliefs about climate change longitudinally predicts low pro-environmental behaviour. Br J Soc Psychol 64(3):12905
76. Meer TG, Brosius A, Hameleers M (2023) The role of media use and misinformation perceptions in optimistic bias and third-person perceptions in times of high media dependency: evidence from four countries in the first stage of the covid-19 pandemic. Mass Commun Soc 26(3):438–462
77. Chen C, Shu K (2024) Combating misinformation in the age of LLMs: opportunities and challenges. AI Mag 45(3):354–368
78. Rossetti G, Stella M, Cazabet R, Abramski K, Cau E, Citraro S, Failla A, Improta R, Morini V, Pansanella V (2024) Y social: an llm-powered social media digital twin. arXiv preprint. arXiv:2408.00818
79. Guo Y (2018) A survey on methods and theories of quantized neural networks. arXiv preprint. arXiv:1808.04752
80. Nagel M, Fournarakis M, Amjad RA, Bondarenko Y, Van Baalen M, Blankevoort T (2021) A white paper on neural network quantization. arXiv preprint. arXiv:2106.08295
81. Li S, Ning X, Wang L, Liu T, Shi X, Yan S, Dai G, Yang H, Wang Y (2024) Evaluating quantized large language models. arXiv preprint. arXiv:2402.18158
82. OpenAI: API reference - chat. https://platform.openai.com/docs/api-reference/chat. Accessed 2024-05-30

Publisher's note