***Journal of Artificial Intelligence and Technology*, 2026, 6, 284-296**
https://doi.org/10.37965/jait.2026.0860

ISTP
RESEARCH ARTICLE

Trusting ChatGPT? When a Subtle Variation in the Prompt Can Significantly Alter the Results

Jaime E. Cuellar,[1,2] Óscar Moreno Martínez,[1,2] Paula Sofía Torres Rodriguez,[1] Jaime Andrés Pavlich-Mariscal,[1,3] Andrés Felipe Micán-Castiblanco,[1,2] and Juan Guillermo Torres Hurtado[1,3]

[1]Pontificia Universidad Javeriana. Bogotá, Colombia
[2]Facultad de Comunicación y Lenguaje, Bogotá, Colombia
[3]Facultad de Ingienería, Bogotá, Colombia

(Received 07 July 2025; Revised 26 November 2025; Accepted 19 December 2025; Published online 22 February 2026)

***Abstract*:** How much can we trust highly complex predictive models like ChatGPT? This study tests if subtle changes in prompt structuring do not produce significant variations in the classification results of sentiment polarity analysis generated by the LLM GPT-4o mini. The model classified 100.000 comments in Spanish on four Latin American presidents as positive, negative, or neutral on 10 occasions, varying the prompts each time. The experimental methodology included exploratory and confirmatory analyses to identify significant discrepancies among classifications.

The results reveal that minor modifications to prompts, such as lexical, syntactic, modal, or even their lack of structure, impact the classifications. At times, the model produced undecided responses mixing categories, providing unsolicited explanations, or using languages other than Spanish. Statistical analysis using Chi-square tests confirmed significant differences in most comparisons between prompts, except in one case when linguistic structures were similar.

These findings challenge the robustness and trustworthiness of large language models (LLMs) for classification tasks, highlighting their vulnerability to variations in instructions. Moreover, it was evident that the lack of structured grammar in prompts increases the frequency of hallucinations. The discussion underscores that trust in LLMs is based not only on technical performance but also on the social and institutional relationships underpinning their use.

***Keywords*:** ChatGPT; large language models (LLMs); trust; robustness; sentiment analysis; Spanish

I. INTRODUCTION

Natural language processing (NLP) tools are widely utilized in various academic and business contexts today. These tools serve to analyze, interpret, or generate natural language, whether written or spoken [1]. Social sciences have employed them, for instance, to analyze electoral processes, social struggles, or political biases in social media [2–5]. These tools have also been central to academic debates in fields such as history, archival studies, and heritage preservation, given the accelerated digitization of archives, automatic transcription of oral sources, and cultural management/conservation efforts [6–9]. In business contexts, they have been used to analyze product and service reviews, lead social media marketing and advertising, and develop chatbots [10,11].

The latest advances in the NLP field are large language models (LLMs), defined as models with a large number of parameters and layers, trained on vast datasets from diverse sources [12]. Upon receiving instruction in the form of a prompt, LLMs can generate text-based responses that closely mimic the style and content that a human might produce. The most advanced LLMs are accessible via the Internet through an application programming interface (API), a communication interface with standardized syntax defining how two or more software components interact [13–15]. ChatGPT has emerged as one of the most popular LLMs. Within just two months of its launch at the end of 2022, it achieved over 100 million monthly active users [16].

As Zhang [17] proposed, LLMs might produce divergent results even when the instructions provided are semantically similar. These differences in outputs bring to light a discussion previously addressed in Science and Technology Studies (STS): how much can we trust highly complex predictive models like ChatGPT? At first glance, such trust should rely on the mathematical and statistical robustness of the model. Addressing this question, this article tests the hypothesis that subtle changes in the structuring of prompts do not lead to significant variations in the sentiment polarity classification results generated by the LLM GPT-4o mini. To this end, the robustness of the model was tested by having it classify a dataset of 100.000 YouTube comments using 10 different prompts to evaluate whether significant differences exist in its responses. In other words, this procedure served to measure consistency in the results of the prompts.

The remainder of the article is divided into four sections. The first, the literature review, examines research and conceptual frameworks on LLMs, sentiment analysis, prompt construction, trust, and robustness. The second, the methodology, outlines the process of collecting, structuring, processing, and analyzing comments, as well as the construction of the 10 prompts and the exploratory and confirmatory methods used to evaluate the robustness of the model. The third, the results, presents the findings from the previous steps and interprets them to determine whether significant differences exist. Finally, the fourth, the discussion,

Corresponding author: Jaime E. Cuellar (e-mail: jaimecuellar@javeriana.edu.co).

explores the implications of trusting LLMs with varying robustness, particularly in a context marked by their widespread adoption and the illusion of certainty they produce, a phenomenon that, as Nowotny [18] argues, lies at the heart of our human desire to regain control in an uncertain digital future.

II. LITERATURE REVIEW

As previously described, LLMs are models based on neural networks [12] that receive instructions in the form of prompts and can generate text-based responses. Their impact on various fields of knowledge has been immediate. In medicine, for instance, they have been used to provide personalized health information to patients and offer 24-hour customer service through chatbots [19]); in religion, they have supported the exegesis of the Quran and modern Islamic thought [20]; in economics, they have assisted in processing financial indicators [21]; and in communication, they have been employed to identify false information [22] and analyze comments on Chinese social media regarding events such as the Ukraine–Russia war [23].

Despite the recent surge in popularity of LLMs like ChatGPT, several limitations of the model have already been documented. Borji [24] establishes a typology of eleven failures in ChatGPT 3.5: errors in logical reasoning, factual inaccuracies, mathematical mistakes, difficulties in generating functional code fragments, biases that reproduce social stereotypes, linguistic errors, artificial or feigned emotional expressions, inability to connect with people through emotionally resonant content, lack of skill in offering original perspectives, lack of clarity in decision-making processes, and, finally, the dissemination of erroneous or propagandistic information.

These shortcomings are difficult to explain since LLM functions are treated as black boxes [10,25,26]. That is, while their input data, results, and even potential limitations are known, their internal processes remain opaque. Their predictions or analyses rely on a set of complex operations, from which only the surface is observable [10]. Not only are the analyzed models black boxes, but also their proprietary APIs also fall under this category, as their functionality is not documented in detail [27].

A. SENTIMENT ANALYSIS

Among the numerous tasks performed by LLMs, data classification stands out. The classification process enables the distribution of data into predefined categories. One example is sentiment analysis, which categorizes text into relevant groups to capture the opinions embedded within it [10], using computational tools and predefined or learned criteria. Within sentiment analysis, polarity analysis classifies texts into three categories: positive, negative, and neutral [10,11]. Its goal is to assign each element of a text set (tweets, YouTube comments, or other types) to the category that best fits its content.

Sentiment analysis has been widely used to investigate social phenomena in various contexts. For example, Chandra and Ritij [28] applied this technique to predict the outcome of the 2020 US presidential election between Joe Biden and Donald Trump. Similarly, Li *et al.* [29] used sentiment analysis as a basis for developing a metric to assess the impact of rumors on social media in terms of harm. In the health sector, Braig *et al.* [30] explored how sentiment analysis on Twitter (now X) can provide critical information for managing the COVID-19 pandemic. In education, Sultana *et al.* [31] utilized sentiment analysis on academic data to predict student performance. Lastly, Loomba *et al.* [32] applied this technique to Twitter data to analyze the perceptions of the Indian community regarding cryptocurrencies.

B. PROMPT CONSTRUCTION

A prompt is a written instruction provided to an LLM to perform a specific task or respond to a query, guiding its behavior and generating the desired outcomes. In this case, the task involved classifying YouTube video comments as positive, negative, or neutral. Since the same task can be requested using multiple instructions, it is essential to focus on how the prompt is structured to understand the most effective way to communicate with LLMs.

Prompt engineering is a relatively recent discipline that involves developing and optimizing prompts to effectively utilize LLMs, particularly for NLP tasks across various fields [33]. According to the Prompt Engineering Guide [34], developed by the group Democratizing Artificial Intelligence Research, Education, and Technologies, a prompt may include the following elements: a specific instruction or task for the model to perform, context or additional information to guide the model toward more accurate responses, input data or a question to address, and an output indicator specifying the expected type or format of the result.

In recent years, guidelines [35,36] and recommendations for constructing prompts have been developed. Among others, they emphasize the specificity and clarity, structuring the type of input, specifying the desired result format, using delimiters, and breaking down complex sentences. Complementing this, advanced prompt engineering introduces various techniques for drafting prompts, including zero-shot prompting [37], few-shot prompting [38], chain-of-thought prompting [39], meta-prompting [17], self-consistency [40], generate-knowledge prompting [41], and tree of thoughts [42]. In summary, prompt engineering techniques have become increasingly sophisticated through these developments.

C. TRUST AND ROBUSTNESS

Trust in LLMs has become a critical issue, particularly as these models integrate into daily and professional activities of people. A key question in this context is: can we trust these models? This is a complex and challenging debate that spans technical and ethical issues and remains far from a definitive answer. For the social sciences, it is essential not only to address this topic from a critical and theoretical perspective but also to undertake empirical approaches that provide evidence and concrete perspectives. In this regard, Zanotti *et al.* [43] suggest that trust is not merely a desirable outcome in the use of LLMs but a necessary condition for their effective adoption and application across various domains.

On the one hand, various authors from engineering have explored the conditions for LLMs to be trustworthy. For example, Bolton *et al.* [44] argue that robustness in processing prompt variations, high capacity to recall relevant information, and the absence of hallucinations are fundamental criteria for the reliable use of these models. Similarly, Huang *et al.* [45] propose a set of dimensions necessary for building trust in LLMs, including veracity, security, fairness, robustness, privacy, machine ethics, accountability, and awareness. Additionally, Majeed and Hwang [46] highlight the importance of providing transparent information about the data sources used and assessing the credibility of the sources employed in training the model. From this perspective, it is understood that trust in LLMs requires compliance with predefined metrics and conditions.

Following this idea, one of the central elements identified in the literature as essential for generating trust in LLMs is robustness. Bolton *et al.* [44] define robustness as the ability of LLMs to withstand non-semantic variations in prompts, meaning they should be robust to modifications in prompts that do not substantially alter their meaning. This aspect is vital because it allows LLMs to maintain consistency in their responses even when inputs present slight variations. Huang *et al.* [45] expand on this idea by defining robustness as the ability of the model to sustain its performance under various circumstances, including changes in inputs, contextual variations, or the presence of noise or errors in the data. Koubaa *et al.* [47] interpret this concept as the ability to maintain consistent performance when faced with unexpected inputs or disruptions, ensuring the reliability and coherence of predictions across different scenarios. For a chatbot used to generate responses, this quality is evident in its capacity to handle a variety of language styles and topics, thereby ensuring a consistent user experience.

On the other hand, from the perspective of social sciences, Cook and Santana [48] argue that trust is a relational attribute that emerges between actors, rather than an individual or self-contained attribute. According to Cook, Hardin, and Levi [49], trust arises through encapsulated interests, when one party perceives that the other has incentives to act in their favor, motivated by a commitment to the relationship and an interest in maintaining a reputation for trustworthiness. Nonetheless, trusting LLMs adds a layer of complexity. Authors like Taddeo [50] and Grodzinsky *et al.* [51] discuss the concept of e-trust to study the trust relationship between humans and artificial agents. For Grodzinsky *et al.* [52], an artificial agent is a machine created by humans to operate without constant human intervention. The uniqueness of these agents lies in their ability to modify their internal state based on interactions with their environment over time. These variations could undermine trust in them.

In other words, trusting artificial agents differs from trusting other artifacts (such as a raft or a bicycle). LLMs, as artificial agents, do not require prolonged human intervention to function. Additionally, they can undergo training and eventually present variations in their responses to the same task. Taddeo [50] and Grodzinsky *et al.* [52] agree that opacity, software updates, and uncertain behavior resulting from learning pose unique challenges to trust in such agents.

Sociologists of science, such as Shapin [54], advocate for the relational definition of trust. In his case study on scientific knowledge in 17th-century England, Shapin realizes that the production of this knowledge was more closely tied to trust-based relationships among humans than to the notion of "scientific truth" in medicine. For Shapin, the word "trust" implies a relationship, while "truth" points to an essence; hence, he argues it is more appropriate to speak of trust rather than truth in this context. LLMs, in this sense, neither lie nor tell the truth; the question is whether to trust them. Rolin [55] extends this conception, suggesting that epistemic trust, trust in scientific knowledge, is not only directed at individual scientists or research teams but also at social practices and scientific institutions.

Additionally, STS have developed useful frameworks to broaden the discussion about trust in complex technological models. Nowotny [18] argues that trust in artificial intelligence does not depend solely on its technical robustness but also on factors such as explainability, transparency, and adaptability. The appeal of predictive algorithms and their seemingly precise responses lies in generating a false sense of certainty and control, which in turn fosters blind trust and even dependency. In other words, Nowotny [18] interprets trust in the robustness of a model as a manifestation of the human desire to eliminate uncertainty and regain control over the future. Humans pursue this longed-for certainty through predictive algorithmic operations, which constitute an illusion of security.

III. METHODOLOGY

This research employed an experimental quantitative methodology [56] that involved manipulating one or more variables to evaluate how these changes affected the outcomes. Specifically, it examined whether subtle adjustments in the structuring of prompts result in significant variations in the sentiment polarity analysis performed by the LLM GPT-4o mini. Our research based its results on both exploratory and statistical techniques such as principal component analysis (PCA), the Levenshtein distance, coincidence matrix, Chi-square test, and p-values. The four phases of the methodology were (1) data collection and structuring, (2) prompt design, (3) data processing, and (4) quantitative analysis. To build the corpus, 100.000 comments in Spanish were downloaded via the YouTube API concerning the presidents analyzed.

A. DATA COLLECTION AND STRUCTURING

In total, approximately 700.000 comments were initially collected from 24 YouTube channels corresponding to national and international news outlets that regularly publish content about the four Latin American presidents: Andrés Manuel López Obrador from Mexico, Nayib Bukele from El Salvador, Javier Milei from Argentina, and Gustavo Petro from Colombia. All selected channels produce content in Spanish and belong to hegemonic or widely recognized media organizations, such as major newspapers and television networks, as detailed in Annex 1. The search for videos within these channels followed a consistent keyword structure: "(Country) + (Name of the President) + '*presidente*'," and employed the YouTube API parameters: channel_id, query, published_after, published_before, and max_results, with a time frame set between January and June 2024. This combination of terms allowed for a focused selection of videos and comments specifically related to the four political leaders. Only videos containing more than one hundred comments were included. Duplicated comments were removed, and extremely long comments were excluded as statistical outliers. From the resulting cleaned dataset, a random sample of 100.000 comments was selected for analysis. There was no special preprocessing done to the text of the comments, that is, no lemmatization, stemming, or stopword removal. This was due to the fact that LLMs are able to process raw text [57], and any kind of preprocessing would have altered the original message. Comments were not filtered for spam or non-Spanish content, as these were methodologically relevant to observe how the model responded to noisy or multilingual inputs when prompted in Spanish. This procedure ensured a large, diverse, and representative corpus while maintaining transparency and methodological consistency across the four national contexts.

The database was constructed with the intent of testing the outcomes of sentiment analysis classification in Spanish. Therefore, it was crucial to target highly polarized comments. By collecting comments about sitting presidents, especially those perceived as left-leaning or right-leaning within the ideological spectrum, the aim was to capture manifest controversies, whether positive, negative, or neutral. Additionally, by incorporating media from local and international sources across four Latin American countries, the database was enriched with Spanish comments from different latitudes. The comments were left as raw data as the GPT-4o mini model was originally trained with unmodified full text.

B. PROMPT DESIGN

For this study, 10 prompts were created using the zero-shot prompting technique (Table I), as the goal was for the model to perform a task on new data without any additional training. Two prompts of reference were used, along with eight linguistic variations used by the research team (Table II) to manipulate specific dimensions (syntactic, lexical-semantic, and modal-pragmatic) as the goal was to test if subtle changes in prompt structuring produced (or not) significant variations in the classification results of sentiment polarity analysis generated by the LLM GPT-4o mini. The typology presented below is grounded on a linguistically motivated approach to prompt variation. Group A includes prompt-base 1, suggested by OpenAI [58] for sentiment analysis, and its four variations (prompts 3, 5, 7, and 9). Group B incorporates prompt-base 2, based on the one used by Zhang [51], for the same task and its four variations (prompts 4, 6, 8, and 10). The linguistic structural changes in the prompts are presented below.

The reference prompt for sentiment analysis provided a general instruction for the model to examine the sentiment expressed in the text and return a corresponding label, such as positive, negative, or neutral. In the official prompt suggested by OpenAI, number 1, a context is defined, the clarity of the task is reinforced, and a reminder is included to avoid returning any other text. In contrast, the literature-based prompt, number 2, further simplifies the instructions [59], directing the model to return only the sentiment label without adding additional context.

For prompts 3 and 4, a syntactic change was introduced by reordering sentences without altering their core meaning. That is, the order of the sentences in each prompt was modified. This adjustment aimed to guide the model to return only the label and placed the instruction at the beginning of the sentence, minimizing potential ambiguities in the instructions. Once again, it was emphasized that the desired outcome should be limited to a single word,

Table I. Prompts used for analysis

Group	Type of prompt	Identifier	Prompt
Group A (suggested by Open AI)	Reference prompt	1	As an AI with expertise in language and emotion analysis, your task is to analyze the sentiment of the following text. Please consider the overall tone of the discussion, the emotion conveyed by the language used, and the context in which words and phrases are used. Indicate whether the sentiment is generally positive, negative, or neutral, and return label without any other text.
	Syntactic change	3	Indicate whether the sentiment is generally positive, negative, or neutral, and return label without any other text. Please consider the overall tone of the discussion, the emotion conveyed by the language used, and the context in which words and phrases are used. As an AI with expertise in language and emotion analysis, your task is to analyze the sentiment of the following text.
	Lexico-semantic change	5	As an AI expert in natural language processing and sentiment analysis, your task is to analyze the sentiment of the following text. Please consider the overall tone of the comment, the sentiments conveyed by the text used, and the context in which words and sentences are used. Indicate whether the sentiment is generally positive, negative, or neutral, and return label without any other text.
	Modal-pragmatic change	7	As an AI with expertise in language and emotion analysis, you should analyze the sentiment of the following text. You might consider the overall tone of the discussion, the emotion conveyed by the language used, and the context in which words and phrases are used. You have to indicate whether the sentiment is generally positive, negative, or neutral, and return label without any other text.
	Unstructured prompt/ content word prompt	9	AI expertise language emotion analysis task analyze sentiment following text consider overall tone discussion emotion conveyed language used context which words phrases used indicate whether sentiment generally positive negative neutral return label without other text.
Group B (suggested by Zhang *et al.*, 2023)	Reference prompt	2	Please perform sentiment classification task. Given the text, assign a sentiment label from positive, negative, or neutral. Return label only without any other text.
	Syntactic change	4	Return label only without any other text. Given the text, assign a sentiment label from positive, negative, or neutral. Please perform sentiment classification task.
	Lexico-semantic change	6	Please perform sentiment classification task. Given the comment, assign a sentiment label from positive, negative, or neutral. Return label only without any other text.
	Modal-pragmatic change	8	You should perform sentiment classification task. Given the text, you might assign a sentiment label from positive, negative, or neutral. You have to return label only without any other text.
	Unstructured prompt/ content word prompt	10	Perform sentiment classification task given text assign sentiment label positive negative neutral return label only.

Source: Own elaboration.

Table II. Linguistic features of prompt variants

Prompt pair	Variation type	Linguistic feature manipulated	Change	Expected impact on results
1–2	Baseline vs. simplified	Instructional simplification	Omission of context and reminder of output constraint	Test whether minimalism reduces or maintains the consistency of the task
3–4	Syntactic	Sentence order	Instruction moved to the initial position	Reduce ambiguity
5–6	Lexical-semantic	Terminological variation	Language expertise → NLP (e.g.)	Evaluate semantic robustness and detect lexical sensitivity
7–8	Modal-pragmatic	Modality	Use of "should," "might," and "have to"	Test if the degree of obligation affects the task
9–10	Structural	Absence of grammatical words and punctuation	Telegraphic syntax	Examine tolerance to ungrammatical input

Source: Own elaboration.

such as positive, neutral, or negative, given the nature of the sentiment analysis task. Prompt 3 made an intra-sentence syntactic variation of prompt 1, while prompt 4 did so for prompt 2.

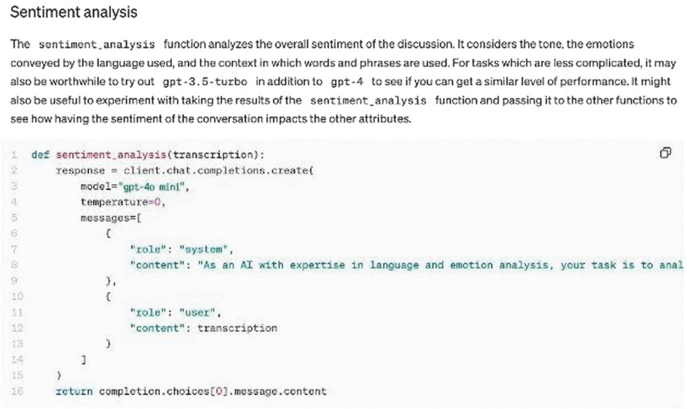
Prompts 5 and 6 underwent a lexical-semantic change. Instead of using the terms "language expertise" and "emotion analysis," the phrases "natural language processing" and "sentiment analysis" were used to test the robustness of the model concerning the recognition and processing of synonymous or semantically related terms when classifying a sentiment as positive, negative, or neutral. Prompt 5 made a lexical-semantic change of prompt 1, while prompt 6 did so for prompt 2.

Prompts 7 and 8 introduced a modal-pragmatic change. Variations in prompts including modals, such as "should," "might," or "have to," explored the effect of the level of obligation or emphasis in the instruction on the response of the model. In linguistic terms and within the semantic-pragmatic interface, modal verbs imposed varying degrees of directive force, potentially influencing how the model executed the task. For instance, a prompt using "have to" implies a stronger obligation and might lead the model to respond more accurately to the instruction, while a modal like "might" introduces a degree of epistemic flexibility, which could result in greater variability in responses. Prompt 7 made a modal variation of prompt 1, while prompt 8 did so for prompt 2.

Finally, prompts 9 and 10 were designed without grammatical words or punctuation to evaluate how the model handled semantically loaded but grammatically fragmented instructions. In this regard, the goal was to test the ability of the model to correctly interpret the underlying message using unstructured linguistic forms. While human language processing might eventually understand telegraphic language, it appeared that NLP models require precise syntactic structures or grammatically well-formed sentences to process specific tasks [60–62]. Prompt 9 made an unstructured variation of prompt 1, while prompt 10 did so for prompt 2.

C. DATA PROCESSING

Once the database was constructed and the prompts were designed, GPT-4o mini was instructed, via its API and Python, to classify the 100.000 aleatory comments as positive, negative, or neutral using the 10 prompts. The code utilized was developed based on OpenAI's documentation for sentiment analysis [63,64]. This code consists of several sections that can be modified to tailor the processing. In this case, given the choice of batch-based sentiment analysis code, the modifiable parameters included the model, temperature, and prompt, as shown in Fig. 1.



Sentiment analysis

The `sentiment_analysis` function analyzes the overall sentiment of the discussion. It considers the tone, the emotions conveyed by the language used, and the context in which words and phrases are used. For tasks which are less complicated, it may also be worthwhile to try out `gpt-3.5-turbo` in addition to `gpt-4` to see if you can get a similar level of performance. It might also be useful to experiment with taking the results of the `sentiment_analysis` function and passing it to the other functions to see how having the sentiment of the conversation impacts the other attributes.

```
def sentiment_analysis(transcription):
    response = client.chat.completions.create(
        model="gpt-4o mini",
        temperature=0,
        messages=[
            {
                "role": "system",
                "content": "As an AI with expertise in language and emotion analysis, your task is to anal
            },
            {
                "role": "user",
                "content": transcription
            }
        ]
    )
    return completion.choices[0].message.content
```

Fig. 1. Simplified data processing code. **Source:** OpenAI. 2023d. *Sentiment Analysis.* [Extract].

Sentiment analysis is used as the primary source in the code since this tool, widely adopted in recent years, facilitates data classification and allows for clearer and more practical comparisons between results. However, the goal of this article was not to evaluate the effectiveness of sentiment analysis but rather to use it as a means to reflect on the impact of prompt structuring on the classification process and its outcomes.

Given the objective of this article, the only variable manipulated was the prompt in order to analyze changes in sentiment classification resulting from its variations. The temperature and model were kept constant. The temperature of the model was set to zero to ensure stability in the responses [65]. All other parameters were left at their default settings, and the GPT-4o mini model was used for all iterations.

With the code defined, the entire database was processed using each prompt as the instruction. This was done independently to ensure that the processing result of one prompt would not affect the result of the next. For each API request, only the tokens for the specific prompt and the text of each comment to be analyzed were sent. In total, one million data points were processed, resulting from analyzing 100.000 data points 10 times.

D. QUANTITATIVE ANALYSIS

To test the hypothesis of whether subtle changes in prompt structuring produce significant variations in results, two analyses were used: exploratory and confirmatory. On the exploratory side, PCA was used to analyze the similarity between prompts by encoding them into embeddings, along with a coincidence matrix between prompts that reflects the proportion of equivalent classifications among pairs of comments. The Levenshtein distance was also calculated. These analytical approaches aimed to perform a preliminary analysis of the classification of prompts and their similarity, which is why they are considered exploratory.

On the confirmatory side, the Chi-square test was conducted to determine whether the prompts produced significantly different results. This analysis aimed to answer the question of whether the observed differences in results were statistically significant.

1. EXPLORATORY ANALYSIS. For the first quantitative approach, embeddings were calculated for each prompt. Embeddings are high-dimensional vector representations generated by language models, capturing the semantic meaning of words or phrases and allowing their similarity to be measured based on distance in vector space [66]. Due to the high dimensionality of these vectors, PCA was applied to facilitate their visualization. This methodology was selected as it is part of the standard set of dimensionality reduction techniques commonly used for embedding visualization, such as SVD and t-SNE [67].

For the second quantitative analysis, a coincidence matrix was used to compare the extent to which prompts aligned in classifying comments. This matrix represents on a scale from 0 to 1 how closely a pair of prompts matched: the closer to 1, the more similar the classifications made by both prompts, whereas a trend toward 0 indicates greater discrepancies.

Finally, the third analytical approach involved calculating the Levenshtein distance, also known as edit distance, which is commonly used to determine the difference between two strings, A and B, in terms of the number of characters that must be added, removed, or substituted in string A to obtain string B, or vice versa.

2. CONFIRMATORY ANALYSIS. A Chi-square test was employed to evaluate whether the probability distribution functions (PDFs) of the classification results (positive, negative, or neutral) generated by each prompt were similar. This test assesses whether one prompt classifies each comment in a manner comparable to another prompt.

In this analysis, the observed frequencies (O_i) for event i represent the number of comments classified as positive, negative, or neutral by one of the prompts. On the other hand, the expected frequencies (E_i) are the predicted values for event i when using another prompt as a reference.

This procedure assumes that each classification result requested by a prompt is independent of another, allowing the Chi-square test to be applied to pairs of prompt results.
Where:

- i represents the classification events: positive, negative, or neutral.
- N is the total number of comments classified, which in this case was 100.000.

These expected frequencies allowed the calculation of the test statistic, defined as:

$$\chi^2 = \sum_{i=1}^{N} \frac{(o_i - E_i)^2}{E_i}$$

This value measures how different the classifications observed from the prompt under study are from the expected ones based on the reference prompt. Based on this value, the null hypothesis is either rejected or not rejected; specifically, the classification distributions produced by the prompt under study do not have statistically significant differences compared to those of the reference prompt.

With this in mind, the alternative hypothesis is proposed: the classification distributions produced by the prompt under study have statistically significant differences compared to those of the reference prompt.

We also conducted a human-based classification to establish a comparison against a ground truth. The main objective of this paper was to analyze variations in classification results generated by different LLM prompts. Although this type of comparison does not strictly require a predefined ground truth, we included a human evaluation layer to strengthen the analysis. Five participants with diverse educational and professional backgrounds manually labeled a sample of 1103 comments using the same sentiment categories applied by the LLM. The level of agreement between the human annotations and the LLM classifications was then measured using Cohen's Kappa coefficient [68], which ranges from −1 (total disagreement) to 1 (perfect agreement).

IV. RESULTS

After processing all the comments with each prompt, the outcome was a table containing the 100.000 comments and the way each prompt classified them. For instance, in Table III, a comment such as *"Nunca vi a el karma de actuar de manera tan instantánea" [sic]*, which translates as "I've never seen karma act so instantaneously," was classified as negative by two prompts, positive by five prompts, and neutral by three prompts. A minority of the processing results of the model were assigned to a fourth category, labeled undecided, even though this was not part of the instructions provided in the prompts. The undecided category encompassed any response that was not explicitly related to the sentiment classifications specified in the prompt: positive, negative, or neutral. For example, there were comments written in Russian to which the model also responded in Russian, and these responses did not correspond to the requested categories. There was also a comment that said "*Aver hazlo*" [sic], which translates as "Let's see, do it," to which the model replied, "Sure, please provide the text you want me to analyze." In other words, instead of classifying that comment, the model mistook it for an instruction, even though the instruction had already been given and was clearly delimited.

In Table IV, the percentage of each category for each prompt is shown. This table served as an initial approach to determine whether the prompts classified the comments in the same way. As can be observed, there are variations in these percentages. However, this table did not allow for an examination of variations in the classification of a particular comment. Moreover, this method did not verify whether the classifications were statistically different, which would be addressed in the confirmatory analysis.

As is evidenced in Table IV, the overall distribution of positive, negative, and neutral classifications varied across the 10 prompts even if prompts 1, 3, 5, and 7 (group A) and prompts 2, 4, 6, 8, and 10 (group B) seemed to have similar classifications within each group. The undecided results of prompt 9 were salient which suggested that grammatical and functional word cues (e.g. determiners, prepositions, and punctuation) were essential to preserve task comprehension and output consistency. Thus,

Table III. Example of classification process results

The comment's content	Prompt 1	Prompt 2	Prompt 3	Prompt 4	Prompt 5	Prompt 6	Prompt 7	Prompt 8	Prompt 9	Prompt 10
Que pedantes entrevistadores, ni siquiera pude acabar de ver la entrevista qué horror!!!!! Que no se dan cuenta con quién están hablando, una mujer muy inteligente brillante luz y esperanza para México!!!! Bendiciones para nuestra próxima presidenta!!!!! Porque este 2 de junio vamos a ganar, con Xóchitl	NEGATIVE	POSITIVE	NEGATIVE	POSITIVE	NEGATIVE	NEGATIVE	NEGATIVE	POSITIVE	NEGATIVE	NEGATIVE
Que factores de la Globalización determinaran el resultado y como se integra México a ese fenómeno sin verse afectado por las externalidades del neoliberalismo?	NEUTRAL	NEUTRAL	NEUTRAL	NEUTRAL	NEUTRAL	NEUTRAL	NEUTRAL	NEUTRAL	UNDECIDED	NEUTRAL
Que bueno !! Bukele será su talón de Aquiles	NEGATIVE	NEGATIVE	NEGATIVE	NEGATIVE	NEGATIVE	NEGATIVE	NEGATIVE	NEGATIVE	NEGATIVE	POSITIVE
Ay ahora sí exige ya que le queda.a la botarga 😂😂😂😂😂	NEGATIVE	NEGATIVE	NEGATIVE	NEGATIVE	NEGATIVE	NEGATIVE	NEGATIVE	NEGATIVE	NEGATIVE	POSITIVE
Nunca vi a el karma de actuar de manera tan instantánea	NEGATIVE	POSITIVE	NEUTRAL	POSITIVE	NEUTRAL	POSITIVE	NEGATIVE	NEUTRAL	NEUTRAL	POSITIVE

Source: Own elaboration.

Table IV. Percentage of sentiment classification by prompt

Prompt	NEGATIVE	POSITIVE	NEUTRAL	UNDECIDED
Prompt 1	70.84	18.33	10.81	0.02
Prompt 2	66.87	22.14	10.98	0
Prompt 3	67.02	18.78	14.19	0.01
Prompt 4	66.85	22.62	10.53	0
Prompt 5	69.32	17.6	13.07	0.01
Prompt 6	66.33	21.12	12.54	0
Prompt 7	71.27	18.1	10.62	0.01
Prompt 8	65.89	22.5	11.61	0
Prompt 9	69.58	19.86	9.47	1.1
Prompt 10	66.48	23.2	10.31	0.01

Source: Own elaboration.

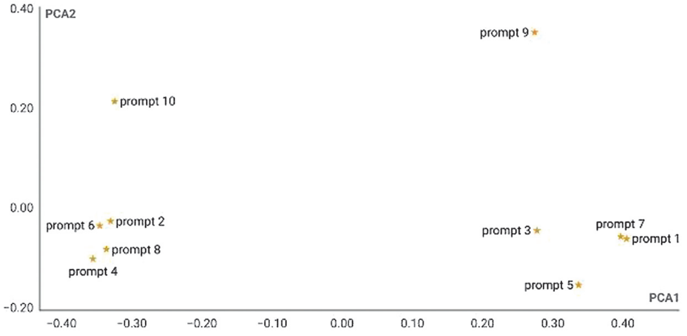


Fig. 2. Similarity of prompts in principal component space. **Source:** Own elaboration.

prompt engineering is not only a matter of wording but also of linguistic architecture. It seems to be that prompts that preserve grammatical integrity, explicit directives, and syntactic precision might foster stability and internal coherence. On the opposite, prompts that reduce structure and significantly alter lexical fields might increase ambiguity and misclassification.

Principal Component Analysis: To compare prompts that were more similar to each other (which, in theory, would be less likely to produce significantly different results), embeddings were generated for each prompt to evaluate their similarity.

As shown in Fig. 2, prompts from group A tended to cluster on one side of the graph, while those from group B were located on the opposite side, except for prompts 9 and 10, which exhibited markedly different behavior from the rest. This discrepancy may be due to the linguistic lack of structure of these prompts, making them the most distant or, in this case, the most different.

Additionally, prompts 1 and 7 from group A stood out for their proximity in the lower-right quadrant of the graph. This closeness could indicate that the model interprets both instructions in a similar way, likely due to their structural resemblance. Prompt 1 was the reference suggested by OpenAI, and prompt 7 was its modal variation. Their position suggested that these prompts might yield similar classification results. However, the counterparts in group B, prompts 2 and 8—another reference prompt and its modal variation, respectively—did not exhibit such proximity.

Other prompts that showed significant closeness are 2 and 6 from group B. Prompt 2 was the one suggested by Zhang *et al.* [17], and prompt 6 was its lexical-semantic variation. Again, the counterparts in group A, prompts 1 and 5 (reference prompt of OpenAI and its lexical-semantic variation), did not exhibit such proximity. These observations provided an initial framework for understanding, contrasting, and analyzing the results of the Chi-square test. This preliminary analysis helped identify patterns and possible similarities in how the model processed each of the prompts.

Coincidence Matrix: To quantify and make an initial approximation of the differences and similarities in classification among the prompts, the proportion of coincidences was calculated. This metric, defined as the number of exact matches between two prompts divided by the total comparisons made, allowed for a preliminary evaluation of the concordance level between the results obtained by different prompts in the sentiment classification of comments from the selected videos.

As shown in Table V, the results indicated coincidence values ranging from 0.92 to 0.98. Differences were more pronounced when comparing prompts based on the official prompt of OpenAI with those from Zhang *et al.* [51]. For example, comparing prompt 1 with prompt 2, the two reference prompts, produced a coincidence value of 0.93. In contrast, comparing prompt 1 with prompt 7 yielded a higher coincidence value of 0.98, as prompt 7 was nearly identical to prompt 1 but included modal verbs. The coincidence between this pair of prompts aligns with the distances observed in Fig. 2.

Similarly, the comparison between prompt 2 and prompt 6, which exhibited close proximity in Fig. 2, resulted in a coincidence value of 0.96. Although significant, this value is not as high as one might expect given their proximity. This is the same value observed between prompts 1 and 5, even though these did not exhibit as much proximity in Fig. 2.

Additionally, it was observed that the unstructured prompts, lacking grammatical words and punctuation, differed the most from the others. This underscored the importance of using well-formed sentences and proper punctuation when drafting prompts. Furthermore, all prompts exhibited some variation: the only instances with an exact coincidence value of 1.0 occurred when prompts were compared to themselves. The fact that coincidence values ranged from 0.92 to 0.98 raises questions about whether these differences were significant and their implications for trust and robustness in this model, as discussed in earlier sections.

Levenshtein Distance: As a complementary analysis, the Levenshtein distance between the tokens of each prompt was calculated, not at the character level but at the word level. Each word was assigned a unique number, and this encoding was used to convert the prompts into numerical arrays. The Levenshtein algorithm was then applied to these arrays. In practice, the Levenshtein distance values between prompts A and B represent the number of words that need to be added, removed, or substituted to transform prompt A into prompt B or vice versa.

As shown in Table VI, the number of words required to add, remove, or substitute to transform one prompt into another ranges from 0 to 68. Differences were more pronounced when comparing prompts based on the official prompt of OpenAI with those based on Zhang [51]. For example, comparing prompt 1 with prompt 2, the two reference prompts, yielded a distance of 55 words. In contrast, comparing prompt 1 with prompt 7 resulted in a distance of only eight words, as prompt 7 was nearly identical to prompt 1 except for the inclusion of modal verbs. The distance between this pair of prompts aligns with the proximities observed in Fig. 2 and the coincidences in Table V.

Similarly, the word distance between prompt 2 and prompt 6 was only one word, despite their coincidence value in Table V being 0.96. Interestingly, unstructured prompts, lacking grammatical words and punctuation, were not the most different from the rest, unlike in the previous analysis. Finally, the correlation between this matrix and the coincidence matrix was -0.71. This indicated a strong inverse correlation: the shorter the word distance, the higher the coincidence.

These analytical approaches provided a preliminary account of the classification results of the prompts and their differences but did not precisely determine whether these classification differences were significant. However, they enabled the formulation of the hypothesis to be tested using the Chi-square test.

Chi-square Test: As mentioned earlier, the Chi-square test was conducted, calculating both the test statistic (χ^2) and the p-values with a significance level of 5 %. Based on these values, it was determined whether the results of the prompts shared the same PDF as defined by Papoulis (2002).

For the quantitative analysis, the categorical results, such as positive, negative, neutral, and undecided, were transformed into the following numerical values: one, two, three, and four, respectively, for the Chi-square analysis. This transformation allowed the creation of a numerical vector for each prompt result to perform the calculations.

The p-values and test statistics are presented in Tables VII and VIII, respectively.

From Tables VII and VIII, it is concluded that only prompts 1 and 7 share the same PDF, as the null hypothesis was not rejected. In this sense, only for this pair of prompts could it be stated that the classification process did not have statistically significant differences. It probably occurred because modal verbs like "should" and "might" used in prompt 7 were interpreted by the

Table V. Coincidence matrix

	Prompt 1	Prompt 2	Prompt 3	Prompt 4	Prompt 5	Prompt 6	Prompt 7	Prompt 8	Prompt 9	Prompt 10
Prompt 1	1	0.93	0.95	0.93	0.96	0.93	0.98	0.93	0.94	0.92
Prompt 2	0.93	1	0.94	0.97	0.93	0.96	0.93	0.97	0.92	0.96
Prompt 3	0.95	0.94	1	0.94	0.96	0.94	0.95	0.94	0.92	0.93
Prompt 4	0.93	0.97	0.94	1	0.92	0.96	0.93	0.97	0.92	0.96
Prompt 5	0.96	0.93	0.96	0.92	1	0.93	0.96	0.92	0.93	0.92
Prompt 6	0.93	0.96	0.94	0.96	0.93	1	0.92	0.96	0.91	0.94
Prompt 7	0.98	0.93	0.95	0.93	0.96	0.92	1	0.92	0.94	0.92
Prompt 8	0.93	0.97	0.94	0.97	0.92	0.96	0.92	1	0.92	0.96
Prompt 9	0.94	0.92	0.92	0.92	0.93	0.91	0.94	0.92	1	0.92
Prompt 10	0.92	0.96	0.93	0.96	0.92	0.94	0.92	0.96	0.92	1

Source: Own elaboration.

Table VI. Levenshtein distance matrix

	Prompt 1	Prompt 2	Prompt 3	Prompt 4	Prompt 5	Prompt 6	Prompt 7	Prompt 8	Prompt 9	Prompt 10
Prompt 1	0	55	39	57	15	55	8	57	34	61
Prompt 2	55	0	63	13	54	1	56	6	32	17
Prompt 3	39	63	0	60	49	63	40	63	55	68
Prompt 4	57	13	60	0	58	14	58	17	34	23
Prompt 5	15	54	49	58	0	53	23	56	43	61
Prompt 6	55	1	63	14	53	0	56	7	32	18
Prompt 7	8	56	40	58	23	56	0	57	38	63
Prompt 8	57	6	63	17	56	7	57	0	33	23
Prompt 9	34	32	55	34	43	32	38	33	0	28
Prompt 10	61	17	68	23	61	18	63	23	28	0

Source: Own elaboration.

Table VII. *p*-values

	Prompt 1	Prompt 2	Prompt 3	Prompt 4	Prompt 5	Prompt 6	Prompt 7	Prompt 8	Prompt 9	Prompt 10
Prompt 1	1	9.31E–104	1.05E–124	6.66E–124	2.77E–54	5.82E–104	0.118	1.34E + 138	2.10E–34	4.20E–157
Prompt 2		1	1.63E–149	7.26E–04	2.70E–162	6.14E–29	1.65E–123	9.65E–07	1.81E–62	6.79E–11
Prompt 3			1	1.60E–195	1.61E–27	2.42E–53	2.57E–143	8.79E–132	1.57E–221	2.00E–235
Prompt 4				1	1.62E–205	2.38E–50	3.39E–140	1.04E–13	1.83E–62	4.64E–03
Prompt 5					1	6.71E–87	5.16E–63	1.37E–168	2.17E–154	5.50E–251
Prompt 6						1	2.47E–124	2.10E–18	2.72E–118	3.47E–69
Prompt 7							1	8.71E–160	5.04E–36	1.75E–174
Prompt 8								1	1.53E–104	2.92E–20
Prompt 9									1	7.07E–79
Prompt 10										1

Source: Own elaboration.

Table VIII. Test statistics

	Prompt 1	Prompt 2	Prompt 3	Prompt 4	Prompt 5	Prompt 6	Prompt 7	Prompt 8	Prompt 9	Prompt 10
Prompt 1	0	474	571	567	247	475	4.27	635	155	720
Prompt 2		0	685	14.5	744	130	552	27.7	284	46.8
Prompt 3			0	897	123	242	657	604	1020	1.08E + 03
Prompt 4				0	943	229	642	60	284	10.7
Prompt 5					0	397	287	773	708	1.15E + 03
Prompt 6						0	569	81	541	315
Prompt 7							0	732	163	800
Prompt 8								0	478	90
Prompt 9									0	360
Prompt 10										0

Source: Own elaboration.

model as subtle semantic nuances or pragmatic softening in structurally complete instructions, which did not alter the core task directive and reflect the robustness of the model to linguistic changes that refer to contrasts in modality. This suggests that the model may be less sensitive to variations in modality when the instructions remain unambiguous and grammatically complete. More studies focusing on variations in the semantic-pragmatic interface might be needed to deeply discuss these particularities. Conversely, other types of linguistic variations as the ones described above affected the results given by the model in a sentiment analysis task.

However, the most noteworthy finding was that for all other cases, the null hypothesis was rejected: the p-values were extremely low (less than 0.05). It can be concluded that for almost all prompt pairs, there were statistically significant differences in the classification results for the same analyzed comments.

This leads to rejecting the hypothesis of this article and asserting that subtle changes in prompt structuring, such as lexical,

Table IX. Cohen's Kappa coefficients between prompts and human-based ground truth

Prompt	Cohen's Kappa coefficient
Prompt 01	0.54
Prompt 02	0.52
Prompt 03	0.54
Prompt 04	0.53
Prompt 05	0.56
Prompt 06	0.54
Prompt 07	0.54
Prompt 08	0.52
Prompt 09	0.52
Prompt 10	0.53

syntactic, or modal adjustments, or even their deconstruction, do indeed produce significant variations in the sentiment polarity classification results generated by the LLM GPT-4o mini. These findings introduced a reflection on trust and robustness in these models, a topic to be addressed in the discussion section.

A. COMPARISON AGAINST GROUND TRUTH

We also performed a classification of the comments using human raters over a sample of 1.103 comments. Five persons of different backgrounds labeled the comments with the same categories as the prompts. To determine the ground truth, we calculated the mode for each comment. If there was more than one mode, the comment was labeled as "undecided." To measure agreement between prompts and ground truth, we calculated the Cohen's Kappa coefficient. Table IX shows the coefficients for each prompt vs. the ground truth.

All of the coefficients were around 0.5, which meant moderate agreement. Differences between coefficients were minimal, with the prompt 5 having the highest (0.56) and prompts 2, 8, and 9 the lowest (0.52).

V. DISCUSSION

GPT-4o mini did not always provide responses limited to the categories positive, negative, or neutral. The 10 prompts instructed the model to classify comments exclusively into these three categories, yet in all variations of the instruction, at least one response was undecided. Occasionally, the model mixed categories, introduced new sentiments, provided explanations, or even used languages other than Spanish. These inconsistencies were related to the lexical, syntactic, and modal variations employed in the construction of the prompts.

All responses that did not fit the requested categories were identified as undecided. Prompt 9, the unstructured one from group A, yielded the highest number of undecided responses, exceeding 1.000, followed by prompt 10. This suggested that the lack of well-formed sentences and punctuation in prompts may lead to an increased number of hallucinations. Measuring the level of agreement between LLM prompts and human-based ground truth showed a moderate level of agreement. This suggests that although LLMs may be used for this kind of text classification, they still yield sub-optimal results.

Another key finding is related to the robustness of the LLM. The experiment demonstrated that, even for a sentiment analysis task with only three options, the model struggled to handle variations in prompts without affecting the classification process. This raised questions about trust: How can we trust classifications produced by LLMs with limited robustness? According to Bolton *et al.* [44], Huang *et al.* [44], and Koubaa *et al.* [47], trust is built on robustness. Yet, as shown, minor changes in the structure of the prompts significantly influenced the classification outcomes, revealing an underlying issue of consistency in the results.

This limitation in robustness is challenging to explain. As noted, these models operate as inscrutable black boxes [10,25,26]. In this case, GPT-4o mini never clarified the criteria used to define whether a comment was positive, negative, or neutral, nor whether these criteria were constant or variable. Consequently, trust in the model should not rely solely on its self-contained operability, presumed to be robust. As Nowotny [18] argues, uncertainty is an inherent characteristic of predictive algorithms.

The reflections above lead to the following question: Can we trust inherently uncertain models? If the answer is affirmative, it is because trust pertains both to adherence to pre-established metrics or criteria [44–46] and to a relational social attribute [48]. Trust in these models may stem from trust in the institution that creates and maintains them [49], prior social and technological experiences, or the incentives and dependencies they generate. These incentives relate to their ability to facilitate daily tasks: "Trust as reliance in computer artifacts means that we expect an object to do something to help us attain our goals" ([69,70] cited in [52]). Thus, the utility of these models and the trust they produce have led to dependency or, as Smith [71] terms it, "obligation."

As Nowotny [18] emphasizes, predictive systems cultivate trust precisely because they offer an illusion of control. This illusion is not merely epistemic but deeply social: we trust because we wish to tame uncertainty, and in doing so, we become increasingly dependent on these systems. In this sense, the kind of trust that LLMs inspire is less about truth and more about our attachment to stability amid algorithmic opacity. These perspectives highlight the importance of addressing both the potential and the limitations and biases inherent in LLMs [72], critical topics in the current context of their increasing adoption.

VI. CONCLUSIONS AND COMING WORK

The research demonstrated that LLMs, such as GPT-4o mini, were sensitive to subtle variations in prompt structuring. Lexical, syntactic, semantic changes, or their deconstruction resulted in significant changes in sentiment analysis outcomes. These differences challenged the robustness of the model, even when semantically similar instructions were used, exposing a latent problem of consistency in the outcomes.

Furthermore, the model exhibited undecided responses, classifying some comments outside the requested categories or providing unsolicited explanations. Unstructured instructions were particularly prone to generating hallucinations. This suggests that LLMs require clear prompts and grammatically well-formed sentences to minimize errors and improve the reliability of results.

In terms of trust, it was noted that trust did not rely exclusively on the technical performance of the models but also on the social and institutional relationships that legitimized their use. The growing dependency on these models, combined with the false perception of certainty they generate, may lead to uncritical acceptance of their results. This phenomenon emphasized the need for a more reflective adoption of LLMs in research and applications, while still acknowledging their potential for analyzing social phenomena. Following Nowotny's [18] insight, future research should interrogate not only how robust LLMs are, but why we so deeply *want* them to be robust, a desire that sustains the illusion of trust itself.

For future work, the experimental path continues. A key next step will be to test the hypothesis proposed in this study with other LLM models and languages, particularly English, in order to assess the consistency of the model across diverse linguistic contexts and thereby strengthen the validity of the findings.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflicts of interest to report regarding the present study.

REFERENCES

[1] A. Elliott, *The Routledge Social Science Handbook of AI*. New York: Routledge, 2022. [Online]. DOI: https://doi.org/10.4324/9780429198533.

[2] R. Saucedo, "GayTwitter: An investigation of biases toward queer users in AI and Natural Language Processing," *Saint Louis Univ. McNair Res. J.*, vol. 1, pp. 60–66, 2018.

[3] A. Khatua and N. Wolfgang, "Analyzing European Migrant-Related Twitter Deliberations," in Companion Proceedings of the Web Conference 2021. New York: Association for Computing Machinery, pp. 166–170, 2021. [Online]. DOI: https://doi.org/10.1145/3442442.3453459.

[4] O. Olabanjo *et al.*, "From Twitter to aso-rock: A sentiment analysis framework for understanding Nigeria 2023 Presidential Election," *Heliyon*, vol. 9, no. 5, p. 16085, 2023. [Online]. DOI: https://doi.org/10.1016/j.heliyon.2023.e16085.

[5] F. Karisma, "Acoso, soledad y desprestigio: un estudio sobre las formas, las rutas de atención y el impacto de las violencias digitales contra las candidatas al Congreso colombiano en 2022." 2023. [Online]. Available: https://web.karisma.org.co/acoso-soledad-y-desprestigio/

[6] C. Sporleder, "Natural Language Processing for cultural heritage domains: NLP for cultural heritage domains," *Lang. Linguist. Compass*, vol. 4, no. 9, pp. 750–768, 2010. [Online]. DOI: https://doi.org/10.1111/j.1749-818x.2010.00230.x.

[7] F. Pessanha and A. A. Salah, "A computational look at oral history archives," *J. Comput. Cult. Heritage*, vol. 15, no. 1, pp. 1–16, 2021. [Online]. DOI: https://doi.org/10.1145/3477605.

[8] J. Gray *et al.*, "Engaged research-led teaching: Composing collective inquiry with digital methods and data," *Digital Cult. Educ.*, vol. 14, no. 3, pp. 55–86, 2022.

[9] L. Marxen *et al.*, "Where did the news come from? Detection of news agency releases in historical newspapers," Master's thesis, École Polytechnique Fédérale de Lausanne, 2023. [Online]. DOI: https://doi.org/10.5281/zenodo.8333933.

[10] T. Ermakova *et al.*, "Commercial sentiment analysis solutions: A comparative study," in International Conference on Web Information Systems and Technologies. Fraunhofer-Institut für Offene Kommunikationssysteme FOKUS, 2021. DOI: https://doi.org/10.5220/0010709400003058.

[11] S. Kusal *et al.*, "AI based emotion detection for textual big data: Techniques and contribution," *Big Data Cogn. Comput.*, vol. 5, no. 3, p. 43, 2021. [Online]. DOI: https://doi.org/10.3390/bdcc5030043.

[12] T. Brown *et al.*, "Language models are few-shot learners," in Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20), Red Hook, NY, USA: Curran Associates Inc., Art. 159, pp. 1877–1901, 2020.

[13] D. Jacobson *et al.*, *Apis: A Strategy Guide: Creating Channels with Application Programming Interfaces*. Sebastopol, California, USA: O'Reilly Media, Inc., 2012.

[14] R. Mitchell,*Web Scraping with Python: Collecting More Data from the Modern Web*. Sebastopol, CA, USA: O'Reilly Media, Inc., 2018.

[15] J. Jünger, "A brief history of APIs," in Handbook of Computational Social Science, Volume 2: *Data Science, Statistical Modelling, and Machine Learning* Methods, U. Engel *et al.*, Ed. London: Routledge, pp. 17–32, 2021. [Online]. DOI: https://doi.org/10.4324/9781003025245.

[16] Data.ai, "ChatGPT's record-breaking growth: The fastest-growing consumer app in history." 2023. [Online]. Available: https://www.data.ai/

[17] Y. Zhang *et al.*, "Meta prompting for AGI systems," 2023. [Online]. DOI: https://doi.org/10.48550/arXiv.2311.11482.

[18] H. Nowotny, *In AI we trust: Power, illusion and control of predictive algorithms*. Hoboken, NJ, USA: John Wiley & Sons, 2021.

[19] S. Biswas, "ChatGPT and the future of medical writing," *Radiology*, vol. 307, no. 2, p. e223312C, 2023. [Online]. DOI: https://doi.org/10.1148/radiol.223312.

[20] A. El Ganadi *et al.*, "Bridging Islamic knowledge and AI: Inquiring ChatGPT on possible categorizations for an Islamic digital library (full paper)," *IAI4CH@AIIA**, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:265467558

[21] H. Li *et al.*, "Extracting financial data from unstructured sources: Leveraging large language models," *Soc. Sci. Res. Netw.*, vol. 39, no. 1, pp. 135–156, 2023. [Online]. DOI: https://doi.org/10.2139/ssrn.4567607.

[22] X. Chen *et al.*, "The Nexus between information disorder and terrorism: A mix of machine learning approach and content analysis on 39 terror attacks," *Dyn. Asymmetric Confl.*, vol. 15, no. 3, pp. 190–209, 2022. [Online]. DOI: https://doi.org/10.1080/17467586.2022.2055097.

[23] R. Rogers and X. Zhang, "The Russia–Ukraine war in Chinese social media: LLM analysis yields a bias toward neutrality," *Soc Media Soc.*, vol. 10, no. 2, 2024. [Online]. DOI: https://doi.org/10.1177/20563051241254379.

[24] A. Borji, "A categorical archive of ChatGPT failures," *ArXiv*, 2023. [Online]. DOi: https://doi.org/10.48550/arXiv.2302.03494.

[25] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: A review," *Artif. Intell. Rev.*, vol. 53, no. 6, pp. 4335–4385, 2020. [Online]. DOI: https://doi.org/10.1007/s10462-019-09794-5.

[26] B. Latour, *La esperanza de Pandora: ensayos sobre la realidad de los estudios de la ciencia*. Barcelona, Spain: Editorial Gedisa, 2021.

[27] G. Nogara *et al.*, "Toxic bias: Perspective API misreads German as more toxic," *ArXiv*, 2023. [Online]. DOI: https://doi.org/10.48550/arxiv.2312.12651.

[28] R. Chandra and S. Ritij, "Biden vs Trump: Modelling US general elections using BERT language model," *IEEE*, vol. 9, pp. 128494–128505, 2021. [Online]. DOI: https://doi.org/10.1109/ACCESS.2021.3111035.

[29] H. Li *et al.*, "Harmfulness metrics in digital twins of social network rumors detection in cloud computing environment," *J. Cloud Comput.*, vol. 13, no. 1, p. 36, 2024. [Online]. DOI: https://doi.org/10.1186/s13677-024-00596-x.

[30] N. Braig *et al.*, "Machine learning techniques for sentiment analysis of COVID-19-related twitter data," *IEEE*, vol. 11, pp. 14778–14803,

2023. [Online]. DOI: https://doi.org/10.1109/ACCESS.2023.3242234.

[31] J. Sultana *et al.*, “Prediction of sentiment analysis on educational data based on deep learning approach,” in Proceedings of the 2018 21st Saudi Computer Society National Computer Conference (NCC), pp. 1–5, 2018. [Online]. DOI: https://doi.org/10.1109/NCG.2018.8593108.

[32] S. Loomba *et al.*, “Sentiment analysis using dictionary-based lexicon approach: Analysis on the opinion of Indian community for the topic of cryptocurrency,” Ann. of Data Sci., vol. 11, no. 6, pp. 2019–2034, 2024. [Online]. DOI: https://doi.org/10.1007/s40745-023-00496-y.

[33] L. Giray, “Prompt engineering with ChatGPT: A guide for academic writers,” *Ann. Biomed. Eng.*, vol. 51, no. 12, pp. 2629–2633, 2023. [Online]. DOI: https://doi.org/10.1007/s10439-023-03272-4.

[34] Dair.ai, “Prompt engineering guide,” 2023. [Online]. Available: https://dair.ai/

[35] S. Ozdemir, *Quick Start Guide to Large Language Models: Strategies and Best Practices for Using ChatGPT and Other LLMs*. Boston, MA, USA: Addison-Wesley Professional, 2023.

[36] G. Marvin *et al.*, “Prompt engineering in large language models,” in International Conference on Data Intelligence and Cognitive Informatics. Springer Nature, pp. 387–402, 2023. [Online]. DOI: https://doi.org/10.1007/978-981-99-7962-2_30.

[37] J. Wei *et al.*, “Finetuned language models are zero-shot learners,” ArXiv, 2021. [Online]. DOI: https://doi.org/10.48550/arXiv.2109.01652.

[38] L. Reynolds and K. McDonell, “Prompt programming for large language models: Beyond the few-shot paradigm,” Extended Abstracts of the 2021, CHI Conference on Human Factors in Computing Systems, pp. 1–7, 2021.

[39] J. Wei *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 24824–248837, 2022. [Online]. DOI: https://doi.org/10.48550/arXiv.2201.11903.

[40] X. Wang *et al.*, “Self-consistency improves chain of thought reasoning in language models,” *ArXiv*, 2022. [Online]. DOI: https://doi.org/10.48550/arXiv.2203.11171.

[41] J. Liu *et al.*, “Generated knowledge prompting for commonsense reasoning,” *ArXiv*, 2021. [Online]. DOI: https://doi.org/10.48550/arXiv.2110.08387.

[42] J. Long, “Large language model guided tree-of-thought,” *ArXiv*, 2023. [Online]. DOI: https://doi.org/10.48550/arXiv.2305.08291.

[43] G. Zanotti *et al.*, “Keep trusting! A plea for the notion of trustworthy AI,” *AI Soc.*, vol. 39, no. 6, pp. 2691–2702, 2024. [Online]. DOI: https://doi.org/10.1007/s00146-023-01789-9.

[44] W. J. Bolton *et al.*, “RAmBLA: A framework for evaluating the reliability of LLMs as assistants in the biomedical domain,” *ArXiv*, Cornell University, 2024. [Online]. DOI: https://doi.org/10.48550/arxiv.2403.14578.

[45] Y. Huang *et al.*, “TrustLLM: Trustworthiness in large language models,” *Arxiv*, 2024. [Online]. DOI: https://doi.org/10.48550/arxiv.2401.05561.

[46] A. Majeed and S. Huang, “Reliability issues of LLMs: ChatGPT a case study,” *IEEE Reliab. Mag.*, vol. 1, no. 4, pp. 1–11, 2024. [Online]. DOI: https://doi.org/10.1109/MRL.2024.3420849.

[47] A. Koubaa *et al.*, “Exploring ChatGPT capabilities and limitations: A survey,” *IEEE Access*, vol. 11, pp. 118698–118721, 2023. [Online]. DOI: https://doi.org/10.1109/ACCESS.2023.3326474.

[48] K. S. Cook and J. J. Santana, “Trust: perspectives in sociology,” in The Routledge *H*andbook of *T*rust and *P*hilosophy, J. Simon, Ed. New York: Routledge, pp. 189–204, 2020. [Online]. DOI: https://doi.org/10.4324/9781315542294.

[49] K. S. Cook, R. Hardin, and M. Levi, *Cooperation Without Trust?* New York, NY, USA: Russell Sage Foundation, 2005.

[50] M. Taddeo, “Defining trust and e-trust: From old theories to new problems,” *Int. J. Technol. Hum. Interact.*, vol. 5, no. 2, pp. 23–35, 2009. [Online]. DOI: https://doi.org/10.4018/jthi.2009040102.

[51] W. Zhang *et al.*, “Sentiment analysis in the era of large language models: A reality check,” *ArXiv*, 2023. [Online]. DOI: https://doi.org/10.48550/arXiv.2305.15005.

[52] F. Grodzinsky *et al.*, “Toward a model of trust and e-trust processes using object-oriented methodologies,” in *ETHICOMP 2010 Proceedings*. A. Bisset, et al. Ed. Tarragona, Spain: Universitat Rovira i Virgili, pp. 14–16, 2010.

[53] F. Grodzinsky *et al.*, “Trust in artificial agents,” in The Routledge *H*andbook of *T*rust and *P*hilosophy, J. Simon, Ed. New York: Routledge, pp. 298–312, 2020. [Online]. DOI: https://doi.org/10.4324/9781315542294.

[54] S. Shapin, *A Social History of truth: Civility and Science in Seventeenth-Century England*. Chicago, IL, USA: University of Chicago Press, 1995.

[55] K. Rolin, “Trust in artificial agents,” in The Routledge *H*andbook of *T*rust and *P*hilosophy, J. Simon, Ed. New York: Routledge, pp. 298–312, 2020. [Online]. DOI: https://doi.org/10.4324/9781315542294.

[56] J. W. Creswell and J. D. Creswell, *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Thousand Oaks, CA, USA: Sage publications, 2018.

[57] H. Naveed *et al.*, “A comprehensive overview of large language models,” *ACM Trans. Intell. Syst. Technol.*, vol. 16, no. 5, pp. 106:1–106:72, 2025. [Online]. DOI: https://doi.org/10.1145/3744746.

[58] OpenAI, “Meeting minutes,” *OpenAI Platform*, 2023. [Online]. Available: https://platform.openai.com/docs/tutorials/meeting-minutes

[59] L. Krause and P. T. J. M. Vossen, “The Gricean Maxims in NLP: A survey,” in Proceedings of the 17th International Natural Language Generation Conference, pp. 470–485, 2024.

[60] R. IV. Logan *et al.*, “Cutting down on prompts and parameters: simple few-shot learning with language models,” *ArXiv*, 2021. [Online]. DOI: https://doi.org/10.48550/arXiv.2106.13353.

[61] J. D. Zamfirescu-Pereira *et al.*, “Why Johnny can’t prompt: How non-ai experts try (and fail) to design llm prompts,” in Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM, Art. 21, 2023. [Online]. DOI: https://doi.org/10.1145/3544548.3581388.

[62] A. Leidinger *et al.*, “The language of prompting: What linguistic properties make a prompt successful?” *ArXiv*, 2023. [Online]. DOI: https://doi.org/10.48550/arXiv.2311.01967.

[63] K. Guzman-Gil, “Batch processing with the batch API,” *OpenAI Cookbook*, 2024. [Online]. Available: https://cookbook.openai.com/examples/batch_processing

[64] OpenAI, “Batch API,” *OpenAI Platform*, 2023. [Online]. Available: https://platform.openai.com/docs/guides/batch/rate-limits

[65] OpenAI, “Create thread and run,” *OpenAI Platform*, 2023. [Online]. Available: https://platform.openai.com/docs/api-reference/runs/createThreadAndRun

[66] OpenAI, “Documentación sobre embeddings y análisis de similitud,” *OpenAI Platform*, 2023. [Online]. Available: https://platform.openai.com/docs/guides/embeddings

[67] OpenAI, “Introducing text and code embeddings,” OpenAI, 2022. [Online]. Available: https://openai.com/index/introducing-text-and-code-embeddings/

[68] J. Cohen, “A coefficient of agreement for nominal scales,” *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960. [Online]. DOI: https://doi.org/10.1177/001316446002000104.

[69] S. C. Goldberg, “Trust and reliance 1,” in The Routledge *H*andbook of *Trust and Philosophy*, J. Simon, Ed. New York: Routledge, pp. 97–108, 2020. [Online]. DOI: https://doi.org/10.4324/9781315542294.

[70] M. Coeckelbergh, “Can we trust robots?” *Eth. Inf. Technol.*, vol. 14, pp. 53–60, 2012. [Online]. DOI: https://doi.org/10.1007/s10676-011-9279-1.

[71] J. E. H. Smith, *The Internet is not What You Think it is: A History, A Philosophy, A Warning*. Princeton, NJ, USA: Princeton University Press, 2022.

[72] A. Páez, “Negligent algorithmic discrimination,” *Law Contemp. Prob.*, vol. 84, no. 3, pp. 19–33, 2021. [Online]. DOI: https://doi.org/10.2139/ssrn.3765778.