Umanistica Digitale – ISSN: 2532–8816 – n. 24, 2026
F. Cristiano, E. Giovannetti – *Can Large Language Models Support Critical Discourse Analysis? A Pilot Experiment*
DOI: https://doi.org/10.60923/issn.2532-8816/23143

Can Large Language Models Support Critical Discourse Analysis? A Pilot Experiment

Francesca Cristiano

Cnr-Istituto di Linguistica Computazionale "A. Zampolli", Pisa, Italy
francesca.cristiano@ilc.cnr.it

Emiliano Giovannetti

Cnr-Istituto di Linguistica Computazionale "A. Zampolli", Pisa, Italy
emiliano.giovannetti@ilc.cnr.it

Abstract

This article presents a pilot experiment that explores the use of Large Language Models (LLMs) in the context of Critical Discourse Analysis (CDA). The study investigates the extent to which two LLMs can reproduce ideologically oriented discourse analyses. The proposed approach involves constructing a consensus-based gold standard from the annotations of three human raters, which was then used to evaluate the agreement between the automated analysis performed by the models and the human annotations. The case study examines a corpus of thirty opinion articles from ideologically diverse newspapers to investigate how the October 7 attack was portrayed in the media. The results indicate that LLMs perform well, particularly with respect to macro- and superstructural features, but struggle with microstructural phenomena such as euphemism detection, underscoring their potential role as supporting tools rather than substitutes for human analysis.

Keywords: critical discourse analysis, large language models, artificial intelligence, critical discourse studies, quantitative analysis

Questo articolo presenta un esperimento pilota che esplora l'uso dei modelli linguistici di grandi dimensioni (Large Language Models, LLMs) nel contesto dell'Analisi Critica del Discorso (Critical Discourse Analysis, CDA). Lo studio indaga fino a che punto due LLM possano riprodurre analisi del discorso orientate ideologicamente. L'approccio proposto prevede la costruzione di uno standard di riferimento basato sul consenso, ottenuto dalle annotazioni di tre valutatori umani, che è stato poi utilizzato per valutare il grado di allineamento tra l'analisi automatica prodotta dai modelli e le annotazioni umane. Il caso di studio esamina un corpus di trenta articoli

d'opinione provenienti da giornali ideologicamente diversi, per indagare come l'attacco del 7 ottobre sia stato rappresentato dai media. I risultati indicano che gli LLM si comportano bene, in particolare rispetto alle caratteristiche macro- e superstrutturali, ma possono avere difficoltà con fenomeni microstrutturali come il rilevamento dell'eufemismo, evidenziando così il loro potenziale ruolo come strumenti di supporto piuttosto che come sostituti dell'analisi umana.

Parole chiave: analisi critica del discorso, modelli linguistici di grandi dimensioni, intelligenza artificiale, studi critici del discorso, analisi quantitativa

1. Introduction

The development of public opinion, understood as a system of beliefs shared by a community, is influenced by the diffusion of the media. The latter are not mere means or technologies, but practices; they exist in specific times and spaces in history and operate in specific ways. Mass media thus are tools of construction and dissemination of meanings, and they are contexts in which sets of symbols are spread and consumed. Such meanings and symbols are able to influence and distort the representation of social reality (Kaneva & Hoover, 2009).

From a systemic perspective, the media also channels the audience's attention toward certain events defined as "newsworthy". People not only obtain information about public affairs from it, but they also learn to focus on a given topic based on the emphasis placed on it by the media (McCombs, 1972). Therefore, such themes gain prominence in media narratives and become the subject of public debate. By emphasizing particular facts over others, the media activates opinion dynamics based on current issues, gives visibility to actors and politicians and enables them to influence the direction of the community (Missier, 2021). Within the field of "traditional" forms of mass media, news plays a very important role. It is not only a narration of facts, but also a particular reconstruction of reality based on norms and values specific to a given society (van Dijk, 1983). The language used in the news discourse is, therefore, never neutral: it reflects socially shared ideologies that are not personal, but social, institutional or political (van Dijk, 1998).

Van Dijk (1995b: 248) defines ideologies as 'basic frameworks of social cognition, shared by members of social groups, constituted by relevant selections of sociocultural values, and organized by an ideological schema that represents the self-definition of a group. Besides their social function of sustaining the interests of groups, ideologies have the cognitive function of organizing the social representations (attitudes, knowledge) of the group, and thus indirectly monitor the group-related social practices, and hence also the text and talk of members.'

Critical Discourse Analysis (CDA), which is an analytical research aimed at exploring the mechanisms through which discourse—understood as any form of communicative event, and as a form of social practice (Fairclough, 1995)—contributes to the maintenance or subversion of power relations, is situated within this context (van Dijk, 1993b). CDA does not limit itself to the textual surface but delves into the deep structures of communication to reveal the processes of social reality construction through language.

CDA, therefore, allows us to interrogate the media not only as sources of information but as active agents in the construction of worldviews, capable of influencing collective beliefs, attitudes, and behaviors. It highlights how discourse is never neutral but it conveys interests, powers, and ideologies, by making clear what is implicit, naturalized, or concealed. Hence,

discourse is not only a product of a society, but also a powerful tool that contributes to the shaping of that given society (Fairclough et al., 2011).

Over the past few years, computational tools have been adopted to automate CDA, raising, however, numerous challenges. Although such technologies make it possible to analyze vast volumes of text, CDA remains deeply tied to the contextual and interpretive understanding of language. A key difficulty is semantic complexity: words shift in meaning depending on their position, register, and interaction with other elements of the text. In addition, there is the problem of context management: many ideologically relevant meanings emerge only within broad discursive structures or in relation to shared cultural events and references. Finally, implicature—those meanings suggested but not explicitly stated—represent a significant obstacle for computational models, which often struggle to distinguish between what is said and what is intentionally implied.

These difficulties highlight how the automation of CDA should not be understood as a replacement for human analysis but as a methodological support capable of amplifying exploratory skills and offering entry points to the critical reading of texts.

In this scenario, computational linguistics plays an increasingly central role in providing tools for the systematic and quantitative exploration of linguistic data, as it will be described in the following section.

The aim of this work is to set up a pilot experiment (described in Section 3) for assessing the interpretive use of the so-called Large Language Models (LLMs) and the tangible benefits they can bring to CDA. The study presents a framework for evaluating such tools in the specific context of a highly mediatized and controversial event such as the October 7, 2023 attack, in which divergent ideological narratives make the critical deconstruction of discourse particularly crucial. The analysis of the results obtained in this experiment is presented in Section 4. Finally, Section 5 of the article reports the conclusions of this study.

2. Related works

In recent years, language technologies have gained an increasingly important role within CDA, enabling the extension of linguistic observation to large amounts of data and complementing the hermeneutic intuition of the analyst with quantitative and computational tools. Among these, automatic terminology and collocation extraction technologies are particularly noteworthy, as they make it possible to identify recurring discursive patterns, anomalous frequencies, and semantic associations that reveal implicit ideological stances.

One of the most widely used tools in this field is Sketch Engine (Kilgarriff et al., 2014), a corpus linguistics platform that allows advanced searches across multilingual corpora, generates wordlists, identifies keywords, and produces word sketches—synthetic representations of the grammatical and semantic combinations of a word. Numerous studies have demonstrated the effectiveness of Sketch Engine for ideological and rhetorical analysis in media texts, particularly in contexts of conflict and geopolitical representation (Zhu & Liu, 2024). These studies show how the quantitative analysis of collocations and recurring terminology can highlight specific interpretive frames and contribute to the deconstruction of framing strategies.

At the same time, the advent of LLMs has opened new possibilities for assisted CDA. LLMs combine semantic understanding and text generation capabilities, which makes them promising tools for the automatic identification of themes, narrative polarizations, and complex discursive structures. Recent studies (Chew et al., 2023) have shown how LLMs can support deductive

coding, thematic analysis, and reconstruction of point of view in texts. While LLMs provide an effective complement to classical methods of critical analysis, there are still some methodological issues concerning the transparency, reproducibility, and reliability of their analyses.

Another recent study (Gillings et al., 2024) offers a theoretical reflection on the application of LLMs in Critical Discourse Studies, the umbrella term that encompasses Critical Discourse Analysis. The paper examines the potential impact of LLMs in three interrelated areas: LLMs as a form of linguistic data, LLMs as a support tool in the process of linguistic research, and the use of AI within a broader field that spans multiple social domains. The findings show that LLMs are highly efficient in implementing more mechanical tasks, while they face greater challenges in performing complex analyses that require deep reasoning and significant critical distance.

In another study, ChatGPT was used to conduct a quantitative analysis by replicating three previous corpus-based discourse analyses (Curry et al., 2024). The research highlights the strengths and weaknesses of LLMs. ChatGPT is able to semantically categorize words correctly, although it sometimes creates rather generic categories. However, the model performs imperfectly both in concordance analysis and in function-to-form analysis. The study also shows that the use of LLMs in Discourse Studies remains limited due to issues of reproducibility, replicability, ethics, and because of its non-deterministic nature.

A very recent theoretical essay addresses the limitations of LLMs in analyzing sensitive discourses such as hate speech, in particular, racial hate speech (DeJeu, 2025). The paper acknowledges the flexibility of LLMs in conducting a range of analytical functions across different types of texts and in their application to Critical Discourse Analysis. However, the specific role that LLMs can play in this field, and the way in which they should be applied—whether autonomously or in combination with human work—remains to be defined.

In light of this—and within a field still in an exploratory phase—the present study positions itself as an empirical-quantitative pilot experiment that applies LLMs directly to Critical Discourse Analysis. While earlier contributions have mainly provided theoretical discussions, exploratory insights, or replications of existing analyses, our work proposes a systematic approach based on a controlled corpus and a consensus-based human gold standard. This design makes it possible to empirically test the reliability of LLMs across different categories of discursive phenomena, highlighting both their potential and their limitations. In this sense, our contribution goes beyond anecdotal or illustrative uses of LLMs and offers a first step toward the development of reproducible methodologies for their integration into the toolkit of Critical Discourse Analysis.

3. The experiment

In this section, we introduce the experimental setup employed to test two LLMs in performing CDA on a journalistic corpus. The questions, based on van Dijk's framework, were posed to three human raters and to each model in five independent runs. Human and machine performance are not treated as equivalent in terms of knowledge or understanding. Instead, the extent to which the models' behavior matches or differs from human judgments based on theory under controlled conditions is examined. The human evaluations, harmonized through consensus, constitute the consensus-based gold standard, whose reliability was verified using inter-rater agreement measures and on which the models' accuracy was calculated. In this study, the term "accuracy" is used in an operational sense, referring to the degree of agreement with a

negotiated human consensus rather than to an objective notion of correctness, which would be incompatible with the interpretive nature of CDA.

The details are presented in the following subsections.

3.1 Corpus construction and case study selection

The corpus[1] consists of thirty articles comprising 28,948 words and 1,404 sentences. Rather than aiming at large-scale representativeness, the corpus was designed to foreground methodological questions related to interpretive evaluation and to be consistent with a pilot experiment.

For its construction, three newspapers were selected: The Jerusalem Post (TJP), The Electronic Intifada (TEI), and The Washington Post (TWP). In particular, The Electronic Intifada accounts for 12,194 words and 602 sentences; the articles from The Jerusalem Post include 8,688 words and 399 sentences; and those from The Washington Post contain 8,066 words and 403 sentences.

These outlets were chosen according to several criteria: they had to be English-language sources, they had to display different political and ideological orientations and target distinct audiences. This selection was made in order to ensure the presence of discursive variation within the corpus. The Washington Post was, therefore, included in the corpus to mitigate the risk of constructing a corpus composed exclusively of strongly polarized newspapers. While The Washington Post may be considered as the least polarized source among those selected, as it provides perspectives from both sides involved in the events, it may occasionally appear to display a slight pro-Israeli orientation. In other words, the corpus was not designed to ensure strict ideological symmetry across sources, but to capture a range of discursive positions relevant to the analytical task.

The corpus is composed of opinion articles, covering the period from 7 October 2023 to 7 January 2024. Opinion articles were chosen because they explicitly reflect the authors' political orientations and ideologies, and, implicitly, those of their respective newspapers. They are characterized by overt ideological framing and, thus, they are particularly suitable for a CDA-oriented analysis. All the articles were written by professional journalists, and are, therefore, similar in terms of quality. Despite being stylistically and rhetorically different, all the texts in the corpus can be examined according to the microstructural, macrostructural and superstructural categories as defined in van Dijk's framework.

The analysis focuses on the first three months following the attack because articles published during this period are more directly concerned with the events of 7 October 2023 and their immediate aftermath, thus providing a dense discursive context for the purposes of this CDA-oriented study. However, the inclusion of outlets that address the topic from different perspectives allows for the introduction of the Israeli military response, ensuring a certain degree of variability within the corpus. Furthermore, the outcome of the test is concerned with the ability of the LLMs to perform CDA-related tasks rather than with providing a comprehensive Critical Discourse Analysis of the corpus itself.

The articles were selected through a search conducted using two keywords, namely "7 October" and "October 7". These keywords were chosen to ensure that the corpus would exclusively include articles referring to the events of October 7, 2023, and their consequences. To avoid including articles containing information not related to these events, all the articles containing the aforementioned keywords were manually checked.

[1] https://github.com/klab-ilc-cnr/critical-discourse-analysis/blob/main/docs/Articles.md

3.2 Definition of the questions

The criterion adopted for the creation of the questions is based on the principle of CDA, with a particular reference to van Dijk's theoretical framework. Van Dijk devotes significant attention to CDA applied to the study of news.

Through his work, van Dijk developed a discourse analytical framework based on three main interconnected levels: social cognition, the structure of the text, and social analysis. Within this model, discourse is examined as a linguistic product, and as a socially and cognitively mediated practice that reflects and shapes the power relations of a given social group. Furthermore, according to his theory, the text is organised in three different levels, which are interdependent and support one another. These levels are known as macrostructure, superstructure, and microstructure (van Dijk, 1988).

The macrostructure concerns the global meaning of discourse. The main themes of a text (topics) are produced through well-defined rules and organized into a set of propositions. In journalistic texts, topics are not presented in a linear or sequential manner, but rather in an order according to which the most specific information precedes the less detailed one. Topics are an important aspect of news texts and they represent what the authors consider to be the most important information.

The superstructure relates to the overall structure of the text, which consists of a number of conventional categories organized in a specific order that varies according to the text genre. In news texts, the superstructure is characterized by an initial section called the summary, which is divided into headline and lead. The following category is that of the main event, which can sometimes contain that of the previous events and that of the consequences. Other news categories are background and verbal reactions. The former can be distinguished in history and context. Another optional category, which may be at the end of a news article, is the comment, containing conclusions, speculations and expectations frequently from the journalist regarding the events.

The microstructure concerns local linguistic elements, namely words and sentences that make up a text, as well as the strategies of local meaning and their underlying ideology. This level of analysis focuses on semantics, syntax, style, and rhetoric. Semantics concerns the meaning that wants to be emphasised in a news text, while syntax refers to the various ways in which syntactic categories are combined. Style, understood as lexical choice, is representative of the social characteristics of the author of a given text and depends on the sociocultural situation of the discourse. In news texts, as in any other genre, style is controlled by the communicative context. Finally, rhetoric refers to the set of verbal strategies used to capture the reader's attention and it plays a persuasive function.

The analytical questions we have formulated are based on van Dijk's theory, which has been adapted to the type of our study. Therefore, they aim to identify the discursive strategies and the ideological representations present in the examined texts through an analysis grounded in the three dimensions of the text listed above. All the questions, which will be introduced in the following section, can be consulted in a dedicated document.[2]

[2] https://github.com/klab-ilc-cnr/critical-discourse-analysis/blob/main/docs/Questions.md

3.2.1 Questions on the macrostructure

The first four of the eight questions we developed fall into the category of the macrostructure. The first, in particular, analyzes the space devoted, in percentage terms, to describing the specific events related to Hamas's attack on Israel on October 7, 2023. The raters were asked to indicate the approximate percentage by choosing one of the following ranges: "0–25%", "26–50%", "51–75%," and "76–100%".

The second question investigates the connotative value used to describe the attack. The raters were asked to select one of four options to indicate whether the attack was described positively, negatively, neutrally, or not mentioned at all.

Questions 3 and 4 are related and aim to identify the actors depicted as the main agents and those represented as the targets of the action. The response options include the following subjects: "Hamas", "Israeli army", "Israelis", "Palestine", "Palestinians", "none in particular", and "other".

3.2.2 Questions on the superstructure

Only one question, the fifth, falls into the category of the superstructure and it concerns the specific function of the headline—namely, whether it plays an informative role (e.g., "Attack on Gaza: over 1,000 casualties"), a persuasive role (e.g., "Unprecedented massacre: Gaza in flames"), or it is designed to elicit an emotional reaction (e.g., "Resistance or terrorism? The October 7 dilemma"). Within the category of the superstructure, the headline assumes a fundamental role, as it performs important textual and cognitive functions and often carries strong ideological implications.

3.2.3 Questions on the microstructure

Finally, three questions were included in the category of the microstructure. Question 6 examines the number of negatively connoted terms, which reinforce a specific narrative and/or convey prejudice (e.g., "Hamas terrorists", "Israeli settler-colonialism"). The response options are: "no terms", "between 1 and 5 terms", and "more than 5 terms".

Question 7 analyzes the number of euphemisms present in the text, which are often employed as a mitigating strategy used to describe violent events (e.g., "October Operation"). The raters (and the LLMs) were asked to choose among three options: "no euphemisms", "between 1 and 10 euphemisms", and "more than 10 euphemisms". Van Dijk (1993a: 186) defines euphemisms as follows: 'Our conceptual analysis of denial has already shown that denial may also be implied by various forms of mitigation, such as toning down, using euphemisms or other circumlocutions that minimise the act itself or the responsibility of the accused'.

Question 8 concerns linguistic dehumanization of a given group of people and examines the linguistic choices used to deny their humanity, deprive them of their human rights, and strip them of their moral status (e.g., "They act like a pack of wild dogs", "brutal people", "savage behaviour").

These linguistic representations have a strong manipulative impact and they negatively influence readers' perceptions of the targeted group. The groups presented among the possible answers are those previously mentioned as options for questions 3 and 4.

3.3 The test and the data collection

In order to define a consensus-based gold standard derived from human annotations, three human analysts were consulted. This gold standard is essential to provide a measure of the

accuracy of the answers given by the LLMs. The human raters were selected on the basis of their background and experience in CDA. All the raters had a high level of proficiency in English, and were not aligned with any particular position regarding the Israeli-Palestinian conflict. They independently answered each question for each article in the corpus. The sources of the article were not disclosed, and the articles themselves were administered in a random order. While we acknowledge the fundamental role that an understanding of the context in which discourse is produced and spread plays in a critical discourse analysis, the empirical nature of our experiment required a controlled simplification. We have deliberately chosen not to mention the name of the newspaper in order to avoid any possible bias, which could have influenced the results of the experiment, and to offer a transparent comparison between the human and the artificial annotations. However, we recognise that even after having removed the name of the outlet, it could have been possible to infer the origin of an article. Nevertheless, eliminating the ideological hints from the text would not have been feasible, as CDA specifically focuses on these elements.

At the same time, the 8 questions and the 30 articles were submitted to the models. The LLMs used in the experiment were GPT-4o, which was accessed through the ChatGPT interface, and Llama-4 Maverick accessed via the Groq platform.

The choice of GPT-4o and Llama-4 Maverick was intended to compare two complementary families of general-purpose LLMs that differ not only in performance profile, but also in model ecosystem and deployment philosophy. GPT-4o was selected as a proprietary flagship model optimized for versatile, high-capability instruction following, whereas Llama-4 Maverick was selected as a native multimodal open-weight mixture-of-experts model. Their inclusion, therefore, makes it possible to test whether CDA-oriented analytical behaviour remains stable across both a closed commercial system and a more openly deployable model, rather than being tied to a single technological ecosystem.

Due to the infrequent indeterminacy of the answers given by the models, 5 independent runs per model were conducted for each individual question on each article. In addition, the questions were asked in a new session ("temporary chat" mode for ChatGPT, and a cleared chat for Llama) in order to prevent the models' internal memory from influencing subsequent responses.[3]

It should be noted that the evaluation does not rely on similarity scores derived from internal embedding representations of the models. Instead, the LLMs were prompted to produce categorical answers directly corresponding to the predefined response options for each analytical question. The comparison with the human annotations is, therefore, performed at the level of the final classifications rather than through intermediate semantic similarity metrics.

The responses produced by the human annotators were compiled into a single spreadsheet,[4] which constitutes the set of raw data used as the starting point. The spreadsheet contains 30 sheets, one per article. Each sheet lists the 8 questions by row, and the corresponding responses of the human raters by column (for a total of eight columns).

To calculate the accuracy of the LLMs, it was necessary to establish a consensus among the three human raters. To do so, the inter-rater agreement between them was first calculated, as described in detail in the following section. The actual calculation of the accuracy is presented and discussed in Section 4.

[3] https://github.com/klab-ilc-cnr/critical-discourse-analysis/blob/main/docs/Prompts.md

[4] https://github.com/klab-ilc-cnr/critical-discourse-analysis/blob/main/data/raw_data.xlsx

3.4 Inter-rater agreement

The evaluation of the inter-rater agreement (also known as inter-rater reliability) is a crucial step in ensuring the reliability of the assessments. Without an adequate level of agreement among the evaluators, the final consensus risks reflecting individual preferences or interpretations rather than a shared understanding of the criteria. This measure, therefore, makes it possible to verify that the decisions made by the raters are sufficiently consistent to serve as a solid basis for subsequent analyses.

The same principle is applied to the annotations generated across the five runs for each LLM, in order to assess the models' internal consistency and ensure a fair comparison with the human evaluations.

We would like to point out that, concerning LLMs, the inter-run agreement is used here as an indicator of output stability across repeated executions of the same analytical task. It should not be interpreted as evidence of analytical correctness. A high level of agreement across the runs merely indicates that the model tends to produce similar responses under the same conditions, whereas the validity of these responses must be assessed separately through their alignment with the consensus of human annotators.

To quantify the agreement between the human evaluators and between the outputs of the models, three complementary measures were adopted: Fleiss' Kappa (Fleiss, 1971), Krippendorff's Alpha (ordinal) (Krippendorff, 2013), and the Mean Modal Agreement Ratio (Artstein & Poesio, 2008). Multiple metrics were used to obtain a more robust assessment, capable of capturing agreement beyond chance (Fleiss' Kappa), consistency with respect to an intrinsic ordering of categories (Krippendorff's Alpha for ordinal data), and, finally, a more direct and intuitive measure of the frequency of agreement on the most selected category (Mean Modal Agreement Ratio).

Regarding human raters, we also calculated the raw agreement percentage (Holsti, 1969) and Cohen's Kappa (Cohen, 1960). The agreement percentage provides a straightforward measure of the proportion of identical annotations, offering an easily interpretable indicator of concordance. However, since this metric does not account for agreement occurring by chance, we complemented it with Cohen's Kappa, which adjusts for chance agreement in pairwise comparisons. These additional measures were included to give a more complete picture of the agreement patterns among the human annotators, particularly given the relatively small number of the raters and the potential variability in their interpretive judgments.

The eight questions were divided into two groups: those whose answers can be interpreted as ordered values, and those for which the categories have no intrinsic ordering. In the first case, Krippendorff's Alpha in its ordinal version was also computed, as it takes into account the distance between categories along the scale. To standardize calculations and enable ordering where applicable, each response option was first mapped to an integer. This coding made it possible both to compute agreement statistics using computational tools and to apply the metrics consistently to questions of different types.

Qs	Hum -FK	GPT -FK	Llama-FK	Hum-KA	GPT -KA	Llama -KA	Hum -MR	GPT -MR	Llama -MR
Q1	0.276	0.768	0.749	0.552	0.937	0.871	0.755	0.927	0.913
Q2	0.688	0.934	0.794	n/a	n/a	n/a	0.911	0.987	0.960
Q3	0.411	0.776	0.768	n/a	n/a	n/a	0.811	0.927	0.933
Q4	0.360	0.781	0.697	n/a	n/a	n/a	0.844	0.933	0.927
Q5	0.613	0.932	0.680	n/a	n/a	n/a	0.889	0.980	0.947
Q6	-0.020	-0.014	0.638	-0.071	-0.007	0.689	0.688	0.987	0.953
Q7	0.145	0.741	0.207	-1.566	-0.139	0.339	0.766	0.913	0.820
Q8	0.285	0.806	0.656	n/a	n/a	n/a	0.766	0.920	0.920

Table 1. Inter-rater agreement for human annotators and the two LLMs (GPT-4o and Llama-4 Maverick). Fleiss' Kappa (FK) and Mean Modal Agreement Ratio (MR) are reported for all questions, while Krippendorff's Alpha (KA, ordinal) is computed only for Q1, Q6, and Q7. "n/a" indicates non-ordinal questions.

Table 1 presents a comparative summary of the inter-rater agreement metrics for the human annotators and the LLMs across the eight questions. Krippendorff's Alpha was computed only for questions 1, 6, and 7, which involved ordinal response categories. For the remaining questions, which did not rely on ordinal data structures, the metric is marked as not applicable. Fleiss' Kappa and the Mean Modal Agreement Ratio were computed for all the questions, as they are applicable to both nominal and ordinal data.

Across all the questions and metrics, the LLMs display consistently higher inter-run agreement than human annotators, with particularly high values for Fleiss' Kappa and the Mean Modal Agreement Ratio, indicating strong internal consistency. For ordinal questions, Krippendorff's Alpha shows a more nuanced pattern: while Q1 yields positive values for both humans and LLMs, Q6 and Q7 display near-zero or negative values, indicating low or inverse agreement, especially among the human annotators.

These results suggest that the human annotators introduce a greater degree of interpretive variability, particularly in the ordinal classification tasks, whereas the LLMs maintain a more stable annotation pattern across the runs. It should be noted, however, that internal consistency does not imply external validity: a high level of agreement within the same type of annotator (human or machine) does not guarantee correctness with respect to an external gold standard. Nevertheless, the performance of the LLMs indicates potential as a reliable and internally coherent annotator in contexts where stability is a key requirement.

For a clearer view of the consistency observed across the various runs of the LLMs for each question, see Fig. 1. The bar chart reports the Mean Modal Agreement Ratio for all eight questions, comparing GPT-4o and Llama-4 Maverick. Both models display a generally high level of inter-run agreement across the tasks. However, Llama shows slightly lower agreement values for most questions, indicating a marginally higher variability between runs. The lowest agreement, and, therefore, the highest level of indeterminacy, occurs for Question 7 (on euphemism detection), which also shows relatively low agreement among the human raters.

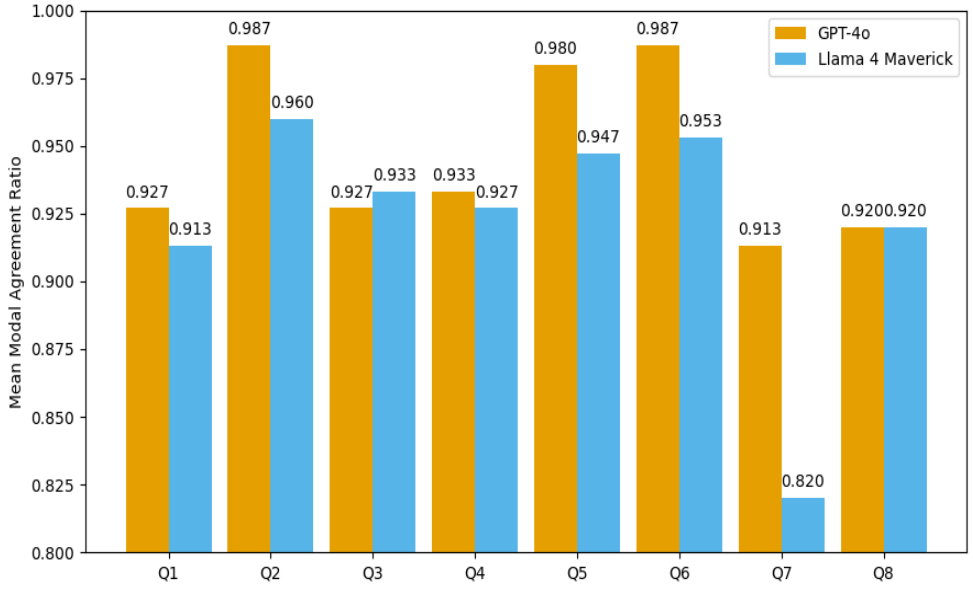

Fig. 1. Mean Modal Agreement Ratio for the two LLMs (GPT-4o and Llama 4-Maverick) across the five independent runs, reported separately for each question. Both models show a generally high degree of inter-run consistency. Slightly lower agreement is observed for Question 7 (euphemism detection), indicating greater variability for this microstructural phenomenon.

3.4.1 Consensus building and definition of the reference annotation

In order to compare the answers provided by the human raters with those provided by the LLMs, we first defined a consensus-based gold standard, which serves as a methodological reference

annotation rather than as an objectively correct interpretation of the texts. The basic criterion adopted was that of the "majority vote": the answer to be considered correct for a given question on a given article was determined either by unanimity (three identical responses from the three human raters) or by majority.

In cases where each human rater provided a different answer, the consensus was defined according to the type of question. For questions (Q1, Q6, Q7) whose answers can be considered ordinal, the average value was chosen as the consensus. For the other questions, in cases of divergence in the evaluations, priority was given to the response provided by "rater 1", based on their greater expertise in CDA. The cases in which adjudication was required were quite rare. Overall, this situation occurred in 9 out of 240 annotation decisions, which represents only 3.75% of the total. In particular, these instances were distributed across question 2 (articles 1 and 25), question 3 (articles 18 and 23), question 4 (articles 8, 16, 21 and 26), and question 8 (article 5).

With regard to the answers provided by the LLMs in their five runs, it was necessary to establish a similar consensus criterion based on the majority. In cases where no majority emerged—i.e., situations in which two pairs of runs produced different answers from each other and from a fifth run (for example "5, 5, 2, 2, 7")—the value with the lowest numerical code among the duplicated responses was selected (in this example, the answer "2"). No cases were recorded in which all the five runs produced distinct answers.

The data related to the conducted analyses and the corresponding consensus values were collected in three separate spreadsheets, one for the human raters,[5] one for the answers provided by GPT[6] and one by Llama.[7] In all three, the results for the eight questions were divided into eight distinct sheets. Each sheet also reported the various measures of inter-rater agreement.

4. Analysis of the results

The results of the accuracy analysis of the LLMs are summarized in Table 2, where two different measures are reported.

Question	*GPT-4o CA*	*GPT-4o SCA*	*Llama-4 CA*	*Llama-4 SCA*
Q1	0.80	0.66	0.60	0.56
Q2	0.90	0.85	0.90	0.75

[5] https://github.com/klab-ilc-cnr/critical-discourse-analysis/blob/main/data/HumanRatersData.xlsx

[6] https://github.com/klab-ilc-cnr/critical-discourse-analysis/blob/main/data/GPTData.xlsx

[7] https://github.com/klab-ilc-cnr/critical-discourse-analysis/blob/main/data/LlamaData.xlsx

Q3	0.83	0.78	0.70	0.66
Q4	0.83	0.60	0.40	0.32
Q5	0.97	0.84	0.70	0.65
Q6	0.73	0.54	0.73	0.60
Q7	0.67	0.58	0.50	0.44
Q8	0.73	0.58	0.60	0.51

Table 2. Accuracy of the two LLMs (GPT-4o and Llama-4 Maverick) across the eight analytical questions. The table reports both Consensus Accuracy (CA), based on majority-vote agreement with the human consensus, and Soft Consensus Accuracy (SCA), which accounts for the distributional variability of human raters and the runs of the LLMs.

The measure Consensus Accuracy (CA) was computed by comparing the consensus label produced by the runs of the LLMs with the consensus label obtained from the human raters. Let CiH denote the human consensus (majority vote among human raters) for instance i, and CiL denote the LLM consensus (majority vote among the runs of the LLMs) for the same instance. The metric is defined as:

$$LLM\ Consensus\ Accuracy = \frac{1}{N}\sum_{i=1}^{N}\delta(C_i^L, C_i^H)$$

where N is the total number of instances (in our case, articles) and $\delta(x,y) = 1$ if $x = y$ and 0 otherwise. This measure captures the proportion of cases in which the consensus decision of the LLMs coincides with the consensus decision of the human annotators.

Instead, the measure of Soft Consensus Accuracy (SCA) was introduced to account for the variability among the human raters and across the runs of the LLMs. Rather than reducing annotations to single categorical labels, both the human responses and the responses of the LLMs are represented as distributions over the set of categories. For each instance (article) i, let pi(k) denote the proportion of the human raters who are assigned class k, and qi(k) the proportion of the runs of the LLMs that are assigned the same class. The SCA is then defined as the average expected agreement between these two distributions across all N instances:

$$\mathrm{SCA} = \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{k=1}^{K} p_i(k)\, q_i(k)\right)$$

This measure captures the degree of probabilistic overlap between the human annotations and the annotations of the LLMs, thus reflecting not only full matches but also partial agreements in cases where the annotators were not unanimous.

Generally speaking, the results reported in Table 2 indicate that the LLMs perform well on several of the analytical questions, especially at the macro- and superstructural levels, although performance is less consistent on microstructural tasks. When considering the Consensus Accuracy (CA), values range from 0.40 to 0.97 depending on the question and the model. However, when rater variability is taken into account through the Soft Consensus Accuracy (SCA), the scores are systematically lower. This confirms that strict majority-vote agreement tends to overestimate the actual alignment between human and model judgments, since SCA captures the distributional variability among human raters and the runs of the models. While both models show generally high levels of inter-run consistency, GPT-4o tends overall to align more closely with the human consensus across most questions. However, this advantage is not uniform: the two models obtain the same CA on Q2 and Q6, and Llama-4 Maverick shows a slightly higher SCA on Q6. Overall, GPT-4o still appears to align more closely with the consensus judgments of human annotators.

This difference may reflect not only performance disparities (also in light of the substantial difference in parameter scale between the two models), but also differences in architecture, instruction tuning, and deployment philosophy between the two models.

The gap between CA and SCA is more evident in questions characterized by greater interpretive variability among the human raters, indicating that the performance of the models is partly constrained by the ambiguity of the task itself.

When the human annotators disagree substantially, the ability of the LLMs to reproduce a stable interpretation also decreases.

A closer analysis of the individual questions allows for more specific observations.

Let us start with the questions related to macrostructure. For Q1, which asks to estimate the proportion of the text devoted to the attack, both models show good accuracy, indicating that LLMs are generally able to distinguish the main thematic focus of a text. Nevertheless, GPT-4o displays a clearer alignment with human consensus. When considering the SCA values, which are influenced by the relatively moderate agreement among human raters, the measured accuracy decreases for both models but remains acceptable.

With regard to Q2, both models reach the same Consensus Accuracy, suggesting that they are equally effective at identifying the overall ideological orientation of the article. GPT-4o shows a higher Soft Consensus Accuracy, which indicates a somewhat closer alignment with the broader distribution of human judgments beyond simple majority-vote agreement.

Q3 concerns the identification of the main agent mentioned in the text. Again, both models show a good performance, suggesting that LLMs are capable of identifying the role played by the entities involved in the described events. However, GPT-4o again shows higher agreement with the human annotations.

Regarding Q4, GPT-4o identifies the main target of the action with reasonably high accuracy, whereas Llama-4 Maverick performs substantially worse than on Q3. Although human agreement on this question is not among the lowest in the dataset, the SCA values indicate that residual variability in human judgments still reduces the weighted accuracy measure.

The only superstructure question, Q5, concerns the nature of the article's title. For this question, the highest overall accuracy values are observed, particularly for GPT-4o. Both models perform well, but GPT-4o shows almost perfect alignment with the human consensus, suggesting that this task is relatively straightforward for current LLMs.

For Q6, which concerns the number of negatively connoted terms in the text, the two models show identical consensus accuracy. The SCA values are also relatively close, with Llama-4 Maverick showing a slightly higher soft agreement than GPT-4o. This suggests that, despite the relatively low agreement among the human raters, both models perform comparably on this task.

A similar pattern is observed for Q7, which asks the annotators to count the number of euphemisms in the text. Although human agreement for this question is higher than for Q6, both models show lower consensus accuracy than in the previous questions, with GPT-4o again outperforming Llama. These results suggest that detecting euphemisms remains a challenging task for LLMs. A closer inspection of the model responses indicates that while the models are often able to correctly identify euphemistic expressions, they sometimes classify explicit lexical choices as euphemisms, revealing limitations in the interpretation of pragmatic mitigation strategies. Finally, Q8 concerns expressions of dehumanization. Consensus accuracy is higher than for Q7 in both models, although the SCA values remain more modest, indicating that performance is still affected by suboptimal agreement among the human raters.

An analysis of the responses classified as incorrect (with respect to the human consensus) shows that both models occasionally interpret expressions such as "Zionist enemy" as dehumanizing. This suggests a tendency to overgeneralize the notion of dehumanization when ideologically marked expressions are present in the text. For this question, it was also necessary to include an explicit instruction in the prompt asking the models to consider the writer's opinion; without this clarification, the models tended to attribute dehumanizing expressions to the article itself even when they appeared only within reported speech.[8]

5. Conclusions

The results of this experiment offer several insights into the potentials and limitations of using Large Language Models in the context of Critical Discourse Analysis (CDA).

First, the models appear to perform better overall in tasks related to macrostructure (Q1–Q4), such as identifying the space devoted to the October 7 events, the ideological orientation of the description, and the main agents or targets of the action. In these cases, accuracy is often high, especially for GPT-4o, suggesting that LLMs are generally able to detect salient topics and their roles within the narrative. This finding aligns with McCombs' agenda-setting theory, since macrostructural topics are explicitly foregrounded in news discourse and, therefore, more easily captured by automated analysis.

Similarly, the only superstructure question (Q5, on the function of the headline) shows very high performance, which indicates that the models are particularly effective when dealing with conventional and easily recognizable textual functions. From a theoretical perspective, this

[8] A notable example appeared in Article 30, which reports a statement by Israeli Defense Minister Yoav Gallant, who referred to Palestinians by saying: "We are fighting human animals, and we will act accordingly".

resonates with van Dijk's notion of news schemata (van Dijk, 1985), where the fixed categories of headlines and leads follow predictable patterns that LLMs can reliably identify.

Among the microstructural tasks, the recognition of negative lexicalizations (Q6) yields moderate but comparable performance in the two models, whereas the identification of euphemisms (Q7) and dehumanizing expressions (Q8) remains more clearly problematic.

This difficulty reflects what Fairclough (1995) and van Dijk (1993b) highlight as the heart of CDA: ideology often works implicitly, through lexical choice, rhetorical figures, and subtle mitigation strategies. Precisely because such strategies aim to conceal or naturalize meaning, they represent the hardest layer for computational models to access.

First of all, with regard to the recognition of the euphemisms, the GPT-4o model, for instance, identified in the article "Failure to confront the genocide in Gaza will haunt humanity" (listed as article 24 in the dataset) of The Electronic Intifada, where the expression "crushing Hamas" has been isolated as a euphemism. This phenomenon was also observed, for example, in the article "October 7 was awful, but this is not a second Holocaust" (number 22) published in The Jerusalem Post, where the terms "mass slaughter", "carnage", and "terrible blow", used to describe Hamas's attack on Israel on October 7, 2023, were identified as euphemisms. The same three terms were classified as euphemisms by the Llama model as well.

As previously stated, the LLMs had difficulties in identifying dehumanizing expressions. In the already cited article number 24, for example, both models indicated "Israeli military" as the dehumanized subject. When specifically asked for an explanation, GPT-4o replied that it had identified "Zionist aggression", "Zionist arrogance" as dehumanising expressions, and "genocidal language of the entire Zionist political class", specifying that 'the article does not use overt animalistic language like "wild dogs" or "savage behavior", but it does dehumanize the Israeli military and political leadership through metaphors of shame, failure, and inherent criminality'. Llama, instead, reported that "Zionist aggression" implies a 'dehumanizing ideology driving the actions of the Israeli military'.

Such examples show how the models struggle exactly where CDA theorists insist that the ideological load of discourse is hidden: in the interplay between explicit meaning and implicit suggestion.

These findings highlight a systematic limitation of current LLMs. While they perform consistently in recognizing explicit and structurally salient features of discourse, they often fail to discriminate between literal and euphemistic formulations, misclassify expressions as dehumanizing when they are not, or overlook the distinction between the author's voice and reported speech unless this is explicitly indicated in the prompt. This echoes van Dijk's (1995a) emphasis on the importance of distinguishing between different voices in the text and on the cognitive mechanisms that sustain ideology through discourse. Such shortcomings reveal that, although reliable in detecting overt linguistic patterns, current models remain ill-equipped to address tasks requiring deeper interpretive nuance—such as capturing implicit rhetorical strategies or distinguishing between authorial stance and third-party voices—thus marking a critical frontier for future development.

In this study, these limitations are not directly related to the phenomenon of hallucinations. Hallucinations are typically defined as inaccurate outputs generated by AI systems that appear plausible but contain fabricated or unsupported information (Augenstein et al., 2024). In the cases discussed here, however, the models do not invent facts or introduce spurious content. Instead, the observed errors stem from interpretative inaccuracies: the models produce plausible but analytically fallacious explanations of discursive categories and strategies. In other words, the

difficulty does not lie in factual reliability, but in the interpretation and classification of complex discursive phenomena.

From a broader methodological perspective, this pilot experiment reinforces the idea that LLMs are not substitutes for human critical analysts, but support tools that can assist in the exploration of large corpora, highlight salient features, and generate preliminary considerations. The high level of internal consistency displayed by the model suggests that they can provide stable baselines for certain types of questions, while more nuanced interpretive tasks still demand expert human judgment.

Looking ahead, future studies could move beyond opinion pieces and include a broader range of journalistic genres, as well as additional newspapers, in order to see whether model performance changes across contexts and formats of discourse. Another promising direction concerns the systematic comparison of different LLMs. In this study we introduced two models, allowing for a first comparative assessment of their analytical performance. While both models show generally high levels of inter-run consistency, GPT-4o tends to align more closely with the human consensus. Future work could extend this comparative approach by including additional models—such as Claude or other open-weight LLMs—in order to better understand how differences in model architecture and training scale may affect performance in discourse-oriented analytical tasks. It would also be valuable to combine LLM-based analysis with corpus-linguistics tools such as Sketch Engine, which may provide complementary quantitative insights into lexical and collocational patterns. From a methodological perspective, greater attention should be paid to prompt design, particularly in tasks where distinguishing reported speech from the writer's own stance is crucial. Finally, it would be worth applying this framework to other sensitive domains of discourse, including racism, gender discrimination, and hate speech, to examine whether LLMs are able to engage with the subtle ideological and power dynamics that CDA seeks to reveal.

Data availability statement

The data that support the findings of this study are openly available on GitHub at https://github.com/klab-ilc-cnr/critical-discourse-analysis

Acknowledgement

Scientific publication produced under the agreement between the National Research Council (CNR) – Institute of Computational Linguistics "A. Zampolli" and the RUT Foundation".

References

Artstein, Ron, and Massimo Poesio. 2008. "Inter-Coder Agreement for Computational Linguistics." *Computational Linguistics* 34 (4): 555–596. https://doi.org/10.1162/coli.07-034-R2.

Augenstein, Isabelle, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele,

Shivam Sharma, and Giovanni Zagni. 2024. "Factuality challenges in the era of large language models and opportunities for fact-checking." *Nature Machine Intelligence*, *6*(8), 852–863. https://doi.org/10.1038/s42256-024-00881-z

Chew, Ryan, Jonathan Bollenbacher, Michael Wenger, Jacob Speer, and Alice Kim. 2023. "LLM-Assisted Content Analysis: Using Large Language Models to Support Deductive Coding (Version 1)." *arXiv*. https://doi.org/10.48550/ARXIV.2306.14924.

Cohen, Jacob. 1960. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 20 (1): 37–46. https://doi.org/10.1177/001316446002000104.

Curry, Neil, Paul Baker, and Gavin Brookes. 2024. "Generative AI for Corpus Approaches to Discourse Studies: A Critical Evaluation of ChatGPT." *Applied Corpus Linguistics* 4 (1): 100082. https://doi.org/10.1016/j.acorp.2023.100082.

DeJeu, Elizabeth B. 2025. "Can (and Should) LLMs Perform Critical Discourse Analysis?" *Journal of Multicultural Discourses* 19 (3): 188–195. https://doi.org/10.1080/17447143.2025.2492145.

Fairclough, Norman. 1995. *Critical Discourse Analysis*. London: Longman.

Fairclough, Norman, Jane Mulderrig, and Ruth Wodak. 2011. "Critical Discourse Analysis." In *Discourse Studies: A Multidisciplinary Introduction,* edited by Teun A. van Dijk, 357–378. London: SAGE Publications Ltd. https://doi.org/10.4135/9781446289068.n17.

Fleiss, Joseph L. 1971. "Measuring Nominal Scale Agreement among Many Raters." *Psychological Bulletin* 76 (5): 378–382. https://doi.org/10.1037/h0031619.

Gillings, Michael, Thomas Kohn, and Gerlinde Mautner. 2024. "The Rise of Large Language Models: Challenges for Critical Discourse Studies." *Critical Discourse Studies,* 1–17. https://doi.org/10.1080/17405904.2024.2373733.

Holsti, Ole R. 1969. *Content Analysis for the Social Sciences and Humanities*. Reading, MA: Addison-Wesley.

Kaneva, Nadia, and Stewart M. Hoover. 2009. *Fundamentalisms and the Media*. New York: Continuum.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. "The Sketch Engine." *Lexicography* 1 (1): 7–36. https://doi.org/10.1007/s40607-014-0009-9.

Krippendorff, Klaus. 2013. *Content Analysis: An Introduction to Its Methodology*. 3rd ed. Thousand Oaks, CA: SAGE Publications, Inc.

McCombs, Maxwell E., and Donald L. Shaw. 1972. "The Agenda-Setting Function of Mass Media." *Public Opinion Quarterly* 36 (2): 176–187. https://doi.org/10.1086/267990.

Missier, Chiara A. 2021. "Framing Fundamentalism in the Digital Media Space." *International Communication Studies* 30 (1): 20–41.

Van Dijk, Teun A. 1983. "Discourse Analysis: Its Development and Application to the Structure of News." *Journal of Communication* 33 (2): 20–43. https://doi.org/10.1111/j.1460-2466.1983.tb02386.x.

Van Dijk, Teun A. 1985. "Structures of News in the Press." In *Discourse and Communication,* edited by Teun A. van Dijk, 69–93. Berlin: De Gruyter. https://doi.org/10.1515/9783110852141.69.

Van Dijk, Teun A. 1988. *News as Discourse.* Hillsdale, NJ: L. Erlbaum Associates.

Van Dijk, Teun A. 1993a. *Elite Discourse and Racism.* Thousand Oaks, CA: SAGE Publications, Inc.

Van Dijk, Teun A. 1993b. "Principles of Critical Discourse Analysis." *Discourse & Society* 4 (2): 249–283. https://doi.org/10.1177/0957926593004002006.

Van Dijk, Teun A. 1995a. "Discourse Analysis as Ideology Analysis." In *Language and Peace,* edited by Christina Schäffner e Anita Wenden, 17–33. Aldershot: Dartmouth.

Van Dijk, Teun A. 1995b. "Discourse Semantics and Ideology." *Discourse & Society* 6 (2): 243–289. https://doi.org/10.1177/0957926595006002006.

Van Dijk, Teun A. 1998. "Opinions and Ideologies in the Press." In *Approaches to Media Discourse,* edited by Allan Bell e Peter Garrett, 21–63. Oxford: Blackwell.

Zhu, Jing, and Yang Liu. 2024. "The Representation of Israel–Hamas Conflict in the American and Chinese News Coverage: A Corpus-Assisted Comparative Critical Discourse Analysis." *International Journal of Social Sciences and Public Administration* 4 (2): 360–378. https://doi.org/10.62051/ijsspa.v4n2.49.