Exploring the Use of Large Language Models in Critical Discourse Analysis: A Consensus-Based Pilot Study

Emiliano Giovannetti, Francesca Cristiano

Cnr-Istituto di Linguistica Computazionale “A. Zampolli”
Via Moruzzi 1, 56124 Pisa, Italy
{emiliano.giovannetti, francesca.cristiano}@ilc.cnr.it

Abstract

Large Language Models (LLMs) are increasingly used in the social sciences and humanities (SSH) to support the analysis of complex textual data, raising methodological questions about evaluation and interpretive reliability. This paper explores the use of LLMs in Critical Discourse Analysis (CDA), considered here as a paradigmatic case of interpretive research in SSH, through a preliminary consensus-based evaluation framework. The study reports on a pilot experiment conducted on a small, theory-driven corpus of opinion articles addressing the October 7, 2023 attack and its aftermath. An LLM is asked to answer analytically motivated questions targeting different levels of discourse structure. Its responses are compared with annotations produced by multiple human analysts and aggregated through a consensus-based procedure. The results reveal an asymmetry in model performance: while LLMs align well with human consensus on macro- and superstructural features, they struggle with microstructural phenomena involving implicit meaning. These findings support the view of LLMs as epistemic support tools rather than replacements for human interpretation.

Keywords: critical discourse analysis, large language models, interpretive evaluation, consensus-based analysis, social sciences and humanities

1. Introduction

The growing availability of Large Language Models (LLMs) has generated considerable interest across the social sciences and humanities (SSH), where they are increasingly explored as tools for supporting the analysis of complex textual data (Underwood, 2025). LLMs produce semantically rich outputs and natural-language explanations, making them attractive for interpretive research domains. At the same time, their adoption raises methodological questions concerning evaluation, reliability, and the role of human judgment (Abdurahman et al., 2025).

To examine these questions in a concrete interpretive setting, this study focuses on Critical Discourse Analysis (CDA). CDA investigates how discourse contributes to the construction and reproduction of power relations and ideologies (Fairclough, 1995; van Dijk, 1998). Because it targets implicit meaning, evaluative framing, and ideologically loaded lexical choices, CDA represents a particularly demanding test case for assessing LLMs in SSH research. In recent years, the advent of LLMs has opened new possibilities for assisted CDA, while raising some methodological concerns. Although LLMs are highly efficient in implementing more mechanical tasks, they face greater challenges in performing complex analyses that require deep reasoning and significant critical distance (Gillings et al., 2025). LLMs appear flexible 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 they play in this field, and whether they should be applied autonomously or in combination with human work, remains still to be defined (DeJeu, 2025).

In the light of this, and within a field yet in an exploratory phase, the present study explores the use of LLMs in CDA through a consensus-based evaluation framework designed to assess their analytical behaviour in interpretive contexts. We report on an experiment conducted on a small, theory-driven corpus of opinion articles addressing a highly polarized and widely mediatized event, namely the October 7, 2023 attack and the subsequent escalation of the conflict. The choice of this case study reflects the high density of ideological positioning and discursive polarization characteristic of such contexts. Rather than asking whether LLMs can “perform” CDA autonomously, this study explores the conditions under which they may contribute to interpretive research practices. By comparing model outputs with consensus-based human annotations across different levels of discourse analysis, we aim to shed light on both the potential and the limitations of LLMs as epistemic support tools in the social sciences and humanities.

2. Theoretical and Methodological Framework

This study is grounded in Critical Discourse Analysis, with a particular reference to Teun A. van Dijk’s theoretical framework since he devotes significant attention to CDA applied to the study of the news. This analytical approach articulates discourse analysis across multiple, interconnected levels, namely macrostructure, superstructure, and microstructure, each associated with different degrees of explicitness and interpretive complexity (van Dijk, 1988).

The macrostructural level concerns the global meaning of a text. 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 superstructural level refers to the conventional organization of texts according to genre-specific categories, such as headlines, summaries, and commentary, which guide readers' expectations and interpretive trajectories. 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.

This distinction is methodologically relevant for LLM evaluation: macro- and superstructural features rely more on surface regularities, whereas microstructural phenomena require pragmatic inference and contextual sensitivity. The contrast provides a principled lens for identifying structural limits in model behaviour. Within interpretive SSH traditions, CDA exemplifies inquiry where multiple theoretically grounded readings may coexist (Wodak & Meyer, 2009). Here, disagreement is an epistemic resource rather than noise to be minimized.

These considerations have direct implications for evaluating LLMs in SSH contexts. If interpretation is inherently plural and negotiated, rigid gold standards and accuracy-based metrics are insufficient to capture model behaviour meaningfully. Instead, evaluation must account for plausible human interpretations and the theoretical assumptions underlying them. For this reason, the present study adopts a consensus-based framework, aggregating multiple human annotations to define a shared interpretive reference. Consensus is not treated as absolute agreement, but as bounded convergence reflecting common ground and residual uncertainty. CDA thus functions as a methodological testbed for exploring how LLMs can be used in interpretive contexts, with evaluation serving as a structured means of assessing their analytical alignment.

3. The Experiment

The experiment reported in this paper is designed to explore how LLMs can be evaluated when applied to CDA. Rather than aiming at large-scale validation or performance benchmarking, the study adopts a small-scale, theory-driven design, intended to foreground methodological issues related to interpretation, consensus, and evaluation.

The corpus adopted in the experiment consists of thirty English-language opinion articles drawn from three newspapers with diverse political and ideological orientations: *The Jerusalem Post*, *The Electronic Intifada*, and *The Washington Post*. Opinion pieces were selected because they make ideological positioning particularly explicit and, therefore, constitute a suitable object for CDA. The articles cover the period from October 7, 2023 to January 7, 2024 and focus on the attack carried out by Hamas against Israel on October 7, 2023, together with its immediate political and military aftermath. This event was chosen as a case study due to its high degree of media visibility and discursive polarization, which results in sharply contrasting representations across different outlets.[1]

As previously stated, the analytical framework is grounded in van Dijk's model of discourse analysis and operationalized through a set of analytically motivated questions targeting different levels of textual organization.

We defined eight questions, distributed across the three main dimensions of the model. Q1–Q4 concern the macrostructure: Q1 asks raters to estimate the proportion of the article devoted to the events of 7 October 2023; Q2 focuses on the connotative framing of the attack, asking whether it is described in positive, negative, neutral terms, or not mentioned; Q3 and Q4 identify, respectively, the actors represented as the main agents of the action and those represented as its targets. Q5 addresses the superstructure and focuses on the function of the headline, distinguishing between informative, persuasive, and emotionally oriented titles. Q6–Q8 concern the microstructure: Q6 examines the presence of negatively connoted lexical items, Q7 the use of euphemisms as mitigating strategies in the representation of violent events, and Q8 investigates forms of linguistic dehumanization directed at specific groups. Taken together, these questions translate key CDA categories into a controlled annotation scheme that can be applied comparatively to both human raters and LLMs.

The resulting scheme includes both ordinal categories (Q1, Q6, Q7) and nominal categories (Q2–Q5, Q8), reflecting different degrees of interpretive gradience. The aim is not to exhaustively annotate all the possible discursive features, but to test how an LLM responds to questions that vary in terms of interpretive explicitness and contextual dependence. The full list of analytical questions, together with the prompts and the corpus of articles used in the experiment and all resulting data are publicly available in an online repository, in line with

[1] Given the political sensitivity of the topic, it is important to acknowledge that LLM outputs may also reflect alignment constraints and training-data biases embedded in the model. While a systematic bias analysis is beyond the scope of this pilot study, this factor should be considered when interpreting the results.

principles of transparency and reproducibility in SSH research.[2]

Three human analysts with expertise in CDA independently annotated the entire corpus by answering the same set of questions for each article. To reduce potential bias, the articles were anonymized and presented in random order, without reference to their source. In parallel, the same questions were submitted to GPT-4o via the ChatGPT web interface, conducting five independent runs for each article. Each run was performed in a separate temporary chat session to avoid cross-instance memory effects. This procedure was adopted to account for the non-deterministic nature of the model outputs and to assess the internal stability of the model's responses.

Given the interpretive nature of the task, the evaluation did not rely on a single authoritative annotation. Instead, both human and model-generated responses were aggregated using a consensus-based procedure, primarily based on majority agreement. In cases where no simple majority emerged, a predefined deterministic rule was applied to ensure consistency: for the ordinal categories (Q1, Q6, Q7), consensus was defined by selecting the median category, whereas for the nominal categories (Q2–Q5, Q8), a fixed tie-breaking criterion was used. To ensure the reliability of this reference standard, an inter-annotator agreement was first assessed among the human raters and, separately, across the multiple runs of the LLM.

Agreement was assessed using complementary metrics capturing both chance-corrected agreement (Fleiss' κ for human annotators) and distributional convergence; inter-run consistency of the LLM was quantified using the Mean Modal Agreement Ratio[3] (Artstein & Poesio, 2008), reported in Figure 1. Fleiss' κ indicated overall moderate agreement among human annotators, with lower values observed for microstructural categories, confirming their higher interpretive variability.

All the inter-agreement measures and the underlying data are made openly available in the repository cited above, in the interest of transparency and reproducibility. This approach reflects the assumption that, in interpretive SSH research, analytical validity emerges from negotiated convergence rather than from absolute correctness. The model's outputs were therefore evaluated by comparing their consensus labels with the consensus derived from human annotations, allowing us to assess whether the LLM's analyses fall within the range of plausible human interpretations.

Overall, this experimental design is intended to shift the focus from performance measurement to methodological reflection. By combining human disagreement, model stability, and consensus-based comparison, the experiment provides a framework for examining the conditions under which LLMs may act as epistemic support tools for interpretive analysis, rather than as autonomous analytical agents.

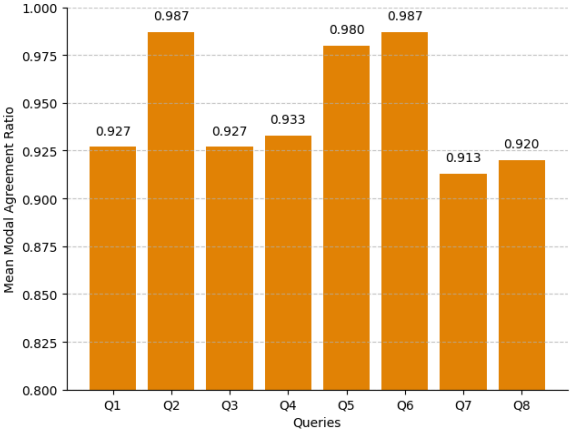

Figure 1: Inter-run agreement of the LLM across the eight queries, measured using the Mean Modal Agreement Ratio

4. Results and analysis

The results of the experiment reveal a differentiated pattern in the alignment between the LLM's outputs and the consensus-based human annotations, highlighting both the potential and the limitations of LLMs when applied to CDA (Fig. 2). Quantitatively, LLM consensus accuracy ranged from 0.67 to 0.97 across the eight analytical questions. Performance was the highest for superstructural and macrostructural categories (Q1–Q5), while lower scores were observed for microstructural phenomena (Q6–Q8), particularly euphemism detection. In particular, the headline function (Q5) reached 0.97 accuracy, while euphemism detection (Q7) dropped to 0.67, marking the largest divergence between model and human consensus. When a distribution-sensitive agreement measure was considered, alignment scores decreased consistently, indicating that strict majority comparison slightly overestimates convergence in cases of human disagreement.

---

2 https://github.com/klab-ilc-cnr/critical-discourse-analysis

3 Mean Modal Agreement Ratio measures how often annotators or model runs converge on the most frequent category. It provides an intuitive estimate of agreement and it is particularly suitable for interpretive tasks where some degree of disagreement is expected; in this study, it is used to assess the internal consistency of the LLM across multiple runs.

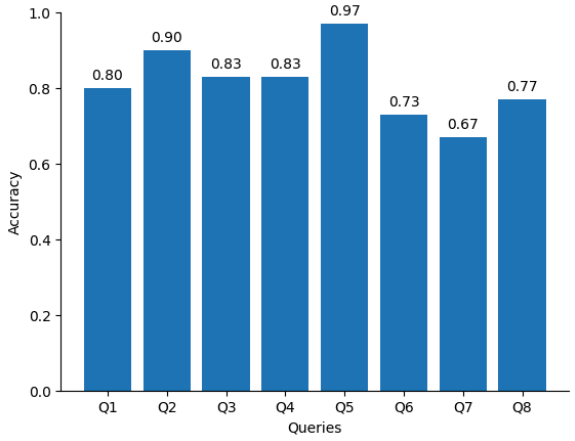

Figure 2: Accuracy of the LLM with respect to the human consensus across the eight queries

At the macrostructural level (Q1–Q4), the model reliably identifies the global themes, the relative emphasis on events, and the representation of social actors. An even higher convergence emerges for the superstructural features, particularly the headline function (Q5), which reaches 0.97 accuracy. This pattern aligns with van Dijk's (1985) notion of news schemata, where fixed organizational categories such as headlines follow predictable conventions that LLMs can effectively capture.

A different pattern emerges in microstructural analysis (Q6–Q8), where ideological meaning is often implicit. In euphemism detection (Q7), the model systematically misclassified highly explicit expressions such as "*mass slaughter*," "*carnage*," and "*terrible blow*" as euphemistic. From a CDA perspective, these terms are not mitigating but overtly explicit and emotionally charged. This suggests that the model equates euphemisms with lexical salience rather than pragmatic attenuation. Notably, these misclassifications were stable across the runs, indicating a consistent internal heuristic rather than random variation.

Similar issues arise in the analysis of dehumanization (Q8). In several cases, the model classified politically charged expressions such as "*Zionist enemy*" or references to "*Zionist aggression*" as instances of dehumanizing discourse. Human analysts, by contrast, did not reach the same conclusion, as these expressions articulate ideological positioning and moral condemnation without necessarily denying the humanity of the targeted group. In another instance, the model identified the Israeli military as the dehumanized subject on the basis of formulations describing it as inherently criminal or genocidal, even in the absence of animalistic metaphors or explicit denial of moral status. These examples point to a systematic tendency to conflate strong evaluative stance with dehumanization, overlooking the more specific linguistic mechanisms through which dehumanization operates in discourse.

Taken together, these cases illustrate a broader pattern in the model's behaviour: when confronted with microstructural phenomena that require pragmatic inference and contextual interpretation, the model tends to overextend analytical categories on the basis of lexical cues alone. Rather than assessing whether a given expression functions rhetorically as mitigation or dehumanization within its discursive context, the model appears to rely on surface features such as emotional intensity or ideological polarity. This behaviour does not occur sporadically, but recurs across texts and questions, indicating a structural limitation rather than isolated error.

From a methodological perspective, these findings gain further significance when considered alongside the contrast between the human interpretive variability and the model stability. The human annotations display non-negligible disagreement for precisely these microstructural categories, confirming that such phenomena are intrinsically open to interpretation. The model, by contrast, produces highly stable responses across multiple runs, suggesting internal consistency. In an interpretive SSH context, however, such stability may signal epistemic rigidity rather than analytical robustness, as consistent outputs can still reflect systematically simplified or biased readings.

Overall, the results highlight a fundamental asymmetry in the applicability of LLMs to interpretive analysis. While LLMs appear well suited to supporting exploratory work at the level of global structure and discursive organization, they struggle with forms of meaning that rely on implication, mitigation, and contextual negotiation—precisely the dimensions that CDA identifies as central to ideological critique (Fairclough, 1995; van Dijk, 1993). These limitations should not be interpreted merely as performance deficits, but as indicators of the epistemic boundary between computational pattern recognition and human critical interpretation.

5. Conclusions

This pilot study explored how LLMs can be integrated into CDA practice, employing a consensus-based evaluation framework to examine their interpretive behaviour. Instead of evaluating LLMs' ability to "perform" Critical Discourse Analysis independently, the study has focused on how their outputs relate to shared human interpretations and on what this relationship reveals about the epistemic role of LLMs in SSH research.

The experimental findings highlight a systematic asymmetry in the model behaviour. LLMs align closely with human consensus when addressing macro- and superstructural aspects of discourse, such as thematic framing, the actor representation, and the headline function. These dimensions rely on relatively explicit cues and

genre conventions, which appear well suited to the model's strengths in pattern recognition and contextual association. By contrast, the analysis of microstructural phenomena—where ideological meaning is often implicit, mitigated, or rhetorically mediated—reveals recurring limitations.

In conclusion, the findings brace a view of LLMs as epistemic support tools rather than replacements for human analysts. Their value lies in their ability to provide stable, reproducible perspectives that can assist exploratory analysis and highlight salient discursive patterns. At the same time, their limitations underscore the continued centrality of human judgment in the analysis of ideologically charged and rhetorically complex texts.

6. Bibliographical References

Abdurahman, S., Salkhordeh Ziabari, A., Moore, A. K., Bartels, D. M., and Dehghani, M. (2025). A primer for evaluating large language models in social-science research. *Advances in Methods and Practices in Psychological Science*, 8(2):1–25.

Artstein, R., and Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

DeJeu, E. B. (2025). Can (and should) LLMs perform critical discourse analysis? *Journal of Multicultural Discourses*, 19(3):188–195.

Fairclough, N. (1995). *Critical discourse analysis: the critical study of language*. Longman, London and New York.

Gillings, M., Kohn, T., and Mautner, G. (2025). The rise of large language models: challenges for Critical Discourse Studies. *Critical Discourse Studies*, 22(6):625–641.

Underwood, T. (2025). The impact of language models on the humanities and vice versa. *Nature Computational Science*, 5(9):695–697.

Van Dijk, T. A. (1985). Structures of news in the press. In T. A. Van Dijk, (Ed.), *Discourse and Communication*. Berlin, New York: DE GRUYTER, pp. 69–93.

Van Dijk, T. A. (1988). *News as discourse*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Van Dijk, T. A. (1993). Principles of critical discourse analysis. *Discourse & Society*, 4(2):249–283.

Van Dijk, T. A. (1998). Opinions and ideologies in the press. In A. Bell, and P. Garrett (Eds.) *Approaches to Media Discourse*. Oxford: Blackwell. pp. 21–63.

Wodak, R., and Meyer, M. (2009). *Methods of critical discourse analysis*. SAGE Publications, London, 2nd edition.