They want to pretend not to understand: The Limits of Current LLMs in Interpreting Implicit Content of Political Discourse

Walter Paci
University of Florence
walter.paci@unifi.it

Alessandro Panunzi
University of Florence
alessandro.panunzi@unifi.it

Sandro Pezzelle
University of Amsterdam
s.pezzelle@uva.nl

Abstract

Implicit content plays a crucial role in political discourse, where speakers systematically employ pragmatic strategies such as implicatures and presuppositions to influence their audiences. Large Language Models (LLMs) have demonstrated strong performance in tasks requiring complex semantic and pragmatic understanding, highlighting their potential for detecting and explaining the meaning of implicit content. However, their ability to do this within political discourse remains largely underexplored. Leveraging, for the first time, the large IMPAQTS corpus, which comprises Italian political speeches with the annotation of manipulative implicit content, we propose methods to test the effectiveness of LLMs in this challenging problem. Through a multiple-choice task and an open-ended generation task, we demonstrate that all tested models struggle to interpret presuppositions and implicatures. We conclude that current LLMs lack the key pragmatic capabilities necessary for accurately interpreting highly implicit language, such as that found in political discourse. At the same time, we highlight promising trends and future directions for enhancing model performance. We release our data and code at: https://github.com/WalterPaci/IMPAQTS-PID

1 Introduction

Implicit and vague language is pervasive in everyday life communication. Consider the language used in political contexts: politicians often use implicit or vague content, relying on the listener's assumptions, background knowledge, and shared contextual knowledge to convey meaning indirectly. This is evident in the use of vague terms, allusions, and rhetorical devices that allow for multiple interpretations. Such language creates an environment where meanings can be conveyed subtly, often with the intent of persuading, manipulating, or reinforcing certain ideologies (Morency et al., 2008). Through the careful use of implicit language, politicians can appeal to the audience's emotions, assumptions, and cultural references, all while avoiding explicit commitments that could later be challenged or disproved (Lombardi Vallauri, 2019). Consider the following example (also featured in the title of this work): *They want to pretend not to understand.*[1] By saying this, the speaker implies that the people being addressed pretend they do not understand something obvious, while, in reality, they understand it perfectly well; moreover, the speaker is also subtlety hinting that they are pretending so *for a reason*, guiding and manipulating the audience to reach a certain reaction or effect. This pragmatic phenomenon, where the speaker says one thing to imply another, is called an implicature. Implicatures are among the most well-studied phenomena of implicit language (Grice, 1990; Sperber and Wilson, 1987), alongside presuppositions (Strawson, 1964; Garner, 1971; Ducrot, 1972). As both are pervasive in political discourse, which makes it an ideal domain for analyzing implicit communication strategies (Van Dijk, 1992; Lombardi Vallauri and Masia, 2020).

The advent of Large Language Models (LLMs) has revolutionized NLP, bringing unprecedented progress to tasks involving semantic understanding (Wang et al., 2018; Williams et al., 2018), question answering (Mihaylov et al., 2018), and pragmatic interpretation (Zheng et al., 2021; Sravanthi et al., 2024; Hu et al., 2023; Kim et al., 2023; Ruis et al., 2024). Overall, LLMs have been shown to perform excellently on semantic tasks and to exhibit potential in pragmatic tasks, particularly when they involve artificially constructed stimuli. How-

[1] This example is the English translation of the Italian sentence *Vogliono far finta di non capire*, that is part of the data used in this work. In the main paper, for simplicity, we only provide English translations. The original Italian texts are available in Appendix A, C.1 and F. In this section, examples may sometimes be adapted for illustrative purposes.

Findings of the Association for Computational Linguistics: ACL 2025, pages 15569–15593

ever, whether they can handle tasks that require making pragmatic inferences on naturalistic discourse data, such as political discourse, remains an open question. In this work, we specifically address this research question and test whether current LLMs can explain the meaning of manipulative implicit content in political discourse, such as the implicature in the example above. Leveraging, for the first time in the context of NLP research, the large IMPAQTS corpus (Cominetti et al., 2024), which includes transcribed Italian political speeches with expert annotations of various types of implicit content, we propose methods to assess the pragmatic effectiveness of several LLMs, state-of-the-art for the Italian language. Our contributions are as follows:

- We propose **a challenging task for LLMs**, i.e., to clarify, through an explanation, the meaning of a manipulative implicit content found in political discourse; e.g., that in the passage *They want to pretend not to understand*, the speaker implies that these people understand perfectly well.

- We release **a novel dataset**, based on the IMPAQTS corpus (Cominetti et al., 2024), containing 30K implicit passages (implicatures and presuppositions) with expert-based explanations, as well as the surrounding linguistic context necessary to correctly interpret the implicit content (empirically validated through human expert annotation).

- Through a multiple-choice task and an open-ended generation task, we demonstrate that **all tested models struggle** to interpret manipulative presuppositions and implicatures. In the multiple-choice setup, the best-performing model falls more than 20 accuracy points short of the estimated ceiling; in the open-ended generation setup, it provides a fully correct explanation in only one-fourth of the cases.

- Despite largely unsatisfactory results, we show that using **Chain-of-Thought (CoT) improves performance** in the open-ended generation setup, suggesting that reasoning mechanisms can aid in solving complex pragmatic tasks. We also propose other directions, such as **embedding external knowledge** about the politician and their political affiliation to further enhance model performance.

2 Related Work

2.1 Implicit Language and LLMs

As LLMs power conversational agents that interact with human users, they must be capable of understanding implicit content. Prior research has explored LLMs' capacity to interpret non-literal meanings in linguistic discourse. Jeretic et al. (2020) investigated the emergence of pragmatic understanding by analyzing how pre-trained language models compute scalar implicatures and presuppositions. They found strong evidence that models such as BERT learned scalar implicatures for quantifiers *some* and *all*, but struggled with other scalar pairs, treating them as synonymous. For presuppositions, models failed to recognize some presupposition triggers, e.g., the verb *to stop* when it presupposes an action that used to be made, as in *John stopped smoking*. Zheng et al. (2021) proposed the first evaluation of language models on five types of implicatures derived from Grice's (1990) conversational maxims, finding that models performed relatively well on a multiple-choice task, where they had to choose which of the given options explained the implicature in a dialogue, but struggled significantly with conversational reasoning, suggesting they do not fully understand conversational context. On a similar track, Hu et al. (2023) evaluated pre-trained LMs without task-specific fine-tuning, incorporating Gricean implicatures as one of their tested phenomena among many others. Their results suggested that language models can infer pragmatic meanings, though they left unresolved whether this is due to linguistic cues or cognitive processes, highlighting the need for further research on their connection in pragmatic reasoning. Kim et al. (2023) evaluated whether LLMs can understand conversational implicatures by prompting them to provide binary answers to specific scenarios. They demonstrated that while LLMs exhibited some implicit understanding, their performance improved significantly when guided through the reasoning process using chain-of-thought prompting. More recently, Ruis et al. (2024) designed a protocol to evaluate LLMs on binary implicature resolution, highlighting a significant gap between humans and LLMs.

All these studies investigated plausible naturalistic use of pragmatic language. Nonetheless, they used artificially constructed sentences instead of real examples of pragmatic phenomena found in actual *corpora*. In our work, we make a significant

step forward and use excerpts of semi-spontaneous, ecological political discourse that reflect the complexities of real-world communication. Unlike artificially constructed sentences, political discourse is known for capturing the complex interplay of rhetoric, social context, and speaker intent.

2.2 Political Language in NLP

Political language has been the focus of NLP research leveraging large datasets from social media and political speeches to study discourse framing, bias detection, and polarization.

Katre (2019) employed computational methods for text analytics and visualization in political speech transcripts. The author used NLP techniques to analyze political speeches and generate graphical visualizations for, among others, word use, lexical dispersion, etc. In Huguet Cabot et al. (2020), NLP methods were used to model political discourse by focusing on metaphor, emotion, and political rhetoric. The authors presented the first joint models that integrate metaphor and emotion detection as auxiliary tasks to enhance the performance of predicting the political perspective of news articles, party affiliation of politicians, and framing of policy issues. Németh (2023) presented a methodological review on how researchers have used NLP techniques to study language polarization between 2010 and 2021. The author identified data sources and computational approaches and reviewed the different conceptualizations and operationalizations of polarization.

Recent efforts leveraged LLMs to enhance political discourse understanding. Marino and Giglietto (2024) used LLMs for political discourse annotation, reporting significant improvement on previous methods like topic modeling. Here, LLMs were used to analyze Facebook links related to the 2018 and 2022 Italian elections and perform tasks such as classification and clustering of political content and generation of descriptive labels for these clusters. Li et al. (2024) proposed Political-LLM, a comprehensive framework for integrating LLMs into political science. This offered unprecedented capabilities for text analysis. At the same time, the authors highlighted several issues related to biases, ethics, and explainability to be considered when using LLMs in political applications. Despite these advancements, little attention has been paid to automatically explaining implicit content. To our knowledge, we are the first to tackle this problem.

3 Data

3.1 The IMPAQTS Corpus

We use, for the very first time in the context of NLP research, data from IMPAQTS (Cominetti et al., 2024),[2] a corpus of Italian political discourse (in the sense of "discourse by politicians"; see Van Dijk et al., 1997), consisting of 1,500 speeches uttered in the Italian language by 150 prominent politicians between 1946 and 2023. IMPAQTS is a multimodal corpus containing over 800 speeches in video format and around 600 in audio format. The manual transcriptions of these speeches are also available, building up to roughly 2.65 million tokens. In our work, we only use these transcriptions.

A key feature of IMPAQTS, crucial for our work, is the annotation of all the passages that contain *implicit content with some manipulative meaning*. In more technical terminology, this kind of implicit content is referred to as non-*bona fide* true; these are implicit questionable contents that are not conveyed in good faith but are still non-explicitly understood as true within a given context. IMPAQTS defines four types of implicit content: *implicature*, *presupposition*, *topicalization*, and *vagueness*. For an extensive discussion on this theoretical framework, see Lombardi Vallauri (2016). Each sentence tagged as containing some implicit comment is accompanied by a *comment* written by the IMPAQTS corpus annotators explaining the meaning of the implicit content.[3] Annotators' comments are all in the same, fixed format; e.g., for implicatures, they start with *it implies that...*; for presuppositions, *it presupposes that...*; for topicalizations, *it considers active in the discourse that...*; for vagueness, *it leaves vague that...*.

Below, we report a single sample including the comment by the IMPAQTS annotators:

(1) **Text with implicit content:** Italy doesn't need another government led by Monti. The last thing Italy needs is another government that is a slave to the banks!
Comment: It implies that Monti's government is a slave to the banks.

In this work, we focus on two of the four categories previously presented, i.e., *implicatures* and *presuppositions*. We chose not to include topical-

[2] https://impaqts.dilef.unifi.it/

[3] In the IMPAQTS corpus, each sentence was tagged by 3 independent annotators. These annotations were then validated by a fourth expert who produced a final comment.

ization as its understanding and interpretation can heavily depend on the information carried out by its prosodic contour in the spoken language (Frascarelli and Hinterhölzl, 2008), which we cannot leverage when using transcriptions. On the other hand, vagueness was excluded due to its wider, less focused definition.

IMPAQTS only includes monologues, and they belong to six types: *Assembly speech*, *Rally speech*, *Party assembly speech*, *Statement in presence*, *Broadcast statement*, *New media statement*, and *Operational meeting Interview*. Here, we chose to focus on sentences from Parliamentary and Rally speeches only, in which implicit contents usually do not refer to external contexts, and they can be detected in the speech transcription. IMPAQTS corpus includes speeches from three time periods: from 1946 to 1972, from 1973 to 1993, and from 1994 to 2022. We experiment with sentences from each of these periods.

3.2 How Much Linguistic Context Is Needed?

To empirically answer this question, we conducted a human validation study asking 9 expert linguist annotators to assess how many *left-hand* context sentences, i.e., preceding the target sentence, were needed to understand its implicit content. We selected 126 sentences, balanced for both the type of implicit content (63 implicatures and 63 presuppositions) and period (42 sentences for the time range). We then created 3 surveys, each containing 42 samples embedding implicit content. In each survey, 3 experts were asked to read both the sentence containing the implicit content and its explanation given by the IMPAQTS' annotators.

Our experts had to assess whether the target sentence provided enough information to understand the explanation in the comment, or if additional preceding linguistic context was needed. If the latter was the case, one extra sentence preceding the target sentence would appear, and the same question would be asked again. By design, our annotators could see up to five left-hand sentences. The annotation was carried out on LimeSurvey (LimeSurvey GmbH, 2012) and allowed to identify, for each sentence, the necessary amount of context for a correct implicit interpretation. The survey details and the inter-annotator agreement are available in Appendix C.1.

3.3 Experimental Data

Based on the results of the annotations presented above, we proceeded to construct a dataset to be used in our experiments with LLMs. We consider all the samples in IMPAQTS containing either an implicature or a presupposition. For each of these samples, we retrieved the preceding 4 sentences in the speech. The resulting dataset, which we name IMPAQTS-PID—where *PID* stands for Preupposition and Implicature Dataset—includes 31,822 samples paired with the explanation of the implicit content from IMPAQTS; in particular, 14,932 samples embed an implicature and 16,890 a presupposition. In Appendix B we report some descriptive statistics of our dataset.

4 Methods

Using the IMPAQTS-PID dataset presented above, we challenge models to understand a sentence's implicit content. In this section, we describe the models we tested in our work, the experiments we performed to assess their abilities, and the experimental details common to both experiments.

4.1 Models

We experiment with four models pre-trained with multilingual data, which are therefore suitable to process Italian language: GPT4o-mini (OpenAI, 2024), Aya Expanse 8B (CohereForAI, 2024), LLAMA3.1 8B (MetaAI, 2024a), and LLAMA3.2 3B (MetaAI, 2024b). The first model is proprietary, the other three are open-weight. All these models are relatively small-scale, which has advantages in terms of computational efficiency, costs, and deployment feasibility. Thus, these models strike a good balance by requiring fewer resources, which reduces both hardware costs and energy consumption, while enabling faster inference. We test models' understanding of implicit content through two experiments, which we describe below.

4.2 Multiple-Choice Generation (MCG)

We perform an experiment where models are presented with four possible explanations IMPAQTS comments (hence, *explanations*) for a given implicit content and must pick the correct one. That is, we frame the problem as a multiple-choice generation (MCG) task. We feed the model with a brief instruction and the four candidates (A, B, C, and D) to choose from, only one of which is correct. To select challenging *distractors*, we use a method based

on topic similarity. We first run a topic modeling analysis on all the explanations in IMPAQTS-PID. We identify 450 topics and a residual category including explanations that do not share a common topic. We then sample the distractors from the same topic class of the target explanation, being it one of the 450 topic-based categories or the residual one. This way, we ensure that the three distractors are incorrect yet plausible, compelling the models to reason over the input rather than relying on shortcuts. During inference, the order of the four candidates is randomized to avoid biases; that is, the position of the correct answer is roughly balanced across positions A-D. Given that each instance in our dataset includes a ground-truth answer, we evaluate model performance using plain accuracy.

4.3 Open-Ended Generation (OEG)

Framing the problem as an MCG task has various advantages, including being easy to set up and evaluate via accuracy. At the same time, it bears relevant limitations due to the selection of the distractors and the non-naturalistic setting of choosing from available options. In this experiment, we circumvent these limitations and directly query models to generate an explanation for a given implicit content; i.e., we frame the problem as an open-ended generation (OEG) task. We experiment with three prompting settings, which we briefly describe below:

Zero-shot The model is queried without any prior examples or detailed guidance and solely relies on its pre-training knowledge. This is the template used for this prompting technique:

Explain the implicit content of the following text. Consider that it appears in its right-most sentence.
Text: [sentence embedding implicit content]
Implicit Content: ____

Few-Shot The prompt includes four illustrative examples, i.e., two implicatures and two presuppositions, including both the text embedding the implicit content and its explanation. As we aim at steering model focus towards manipulative implicit content, we only include "positive" examples (i.e., where an implicature or presupposition is present). We do not include "negative" examples to avoid models focusing on unrelated or unwanted linguistic phenomena. This is the template used for this prompting technique (for brevity, we only report two *shots* instead of four):

Text: [First example text]
Implicit Content: [Explanation of the implicit content in the first example text]
...
Text: [Fourth example text]
Implicit Content: [Explanation of the implicit content in the fourth example text]
Text: [sentence embedding implicit content]
Implicit Content: ____

Chain-of-Thought (CoT) The prompt presents an extensive step-by-step explanation outlining the reasoning process necessary to decode the implicit content. We adapt these steps from the detailed instructions given to the IMPAQTS corpus annotators. This is the template used for this prompting technique (shortened for space reasons):

Implicatures arise in communication whenever the speaker "challenges" one of the four conversational maxims derived from the well-known Cooperative Principle ...
We define "presupposition" as any content that is taken for granted by the participants in communication and, more specifically, content conveyed as part of the knowledge already shared by the interlocutor. ...
Consider what has just been said about implicatures and presuppositions, and explain, proceeding step by step, what the implicit content in the following text is. Consider that it appears in its right-most sentence.

Text: [sentence embedding implicit content]
Implicit Content: ____

Evaluation In contrast to the MCG task, where the presence of a ground-truth answer allows for an evaluation based on plain accuracy, assessing model performance here is more complex. Evaluation based on automatic similarity measures or LLMs-as-judges are potential solutions. However, the task set up introduces uncertainty regarding their reliability. Given these concerns, we opt for a small-scale but more robust assessment based on human expert evaluation. We devise an evaluation protocol where expert linguists assess the texts containing implicit content, the comment about the implicit content from IMPAQTS, and the implicit content explanation generated by the model. See Appendix C.2 for an example. The experts

are asked to judge the quality of the explanation generated by a model by choosing one between 5 possible options:

- **Totally correct:** the explanation is short and focused on the same topic highlighted in the original comment from IMPAQTS.
- **Correct among various options:** the answer lists many possible explanations; among those, the correct one is mentioned (the correctness of other explanations is not relevant).
- **Partially correct:** the model outputs a short answer that does not fully address the point but captures some elements of the correct explanation.
- **Totally wrong:** a short or long answer that is wrong, i.e., is not about the topic highlighted by the comment from IMPAQTS.
- **Answer not given:** the model refuses to answer, avoids the question or just highlights the sentence containing the implicit content without giving an explanation.

4.4 Experimental Details

Details on prompting and decoding settings for both tasks, including token limits, temperature parameters, GPUs, and computational time are provided in Appendix F and G.

5 Results

5.1 MCG Results

As shown in Figure 1, all models perform well above chance, indicating that they understand the task and can sometimes select the correct explanation. At the same time, we observe significant variation between models, with the best-performing GPT4o-mini achieving an accuracy of 70%, nearly 30 percentage points higher than LLAMA3.2 3B (43%).

Interestingly, GPT4o-mini is the only model that consistently surpasses the *n*-gram baseline, while all other models perform worse than this heuristic. On one hand, this indicates that LLMs do not systematically rely on a strategy based solely on shallow similarity between the input and the answer. On the other hand, it suggests that GPT4o-mini possesses certain pragmatic abilities that enable it to understand the implicit content and go beyond surface-level similarities in the prompt. To

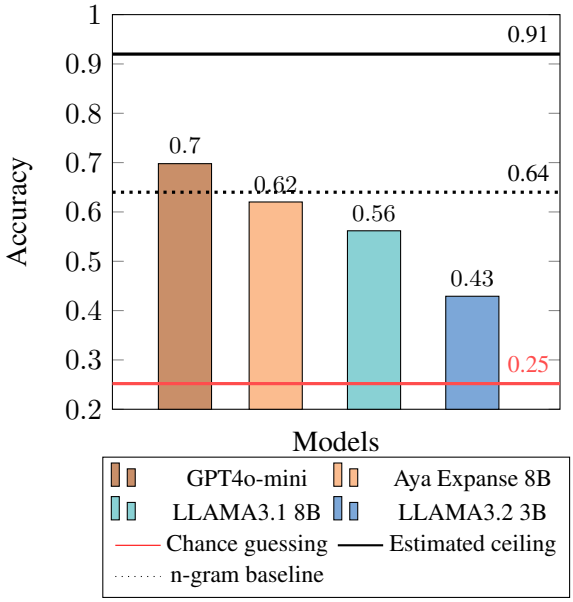

Figure 1: MCG task. All models perform better than chance level (red line). Only GPT4o-mini consistently surpasses the similarity baseline (dotted line).

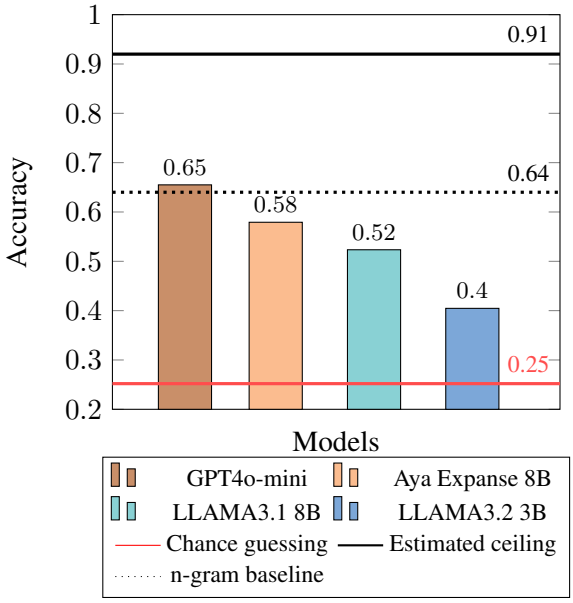

Figure 2: **Hard-negatives setting.** Disaggragated accuracy scores for the MCG Task on the subset of texts with a common topic.

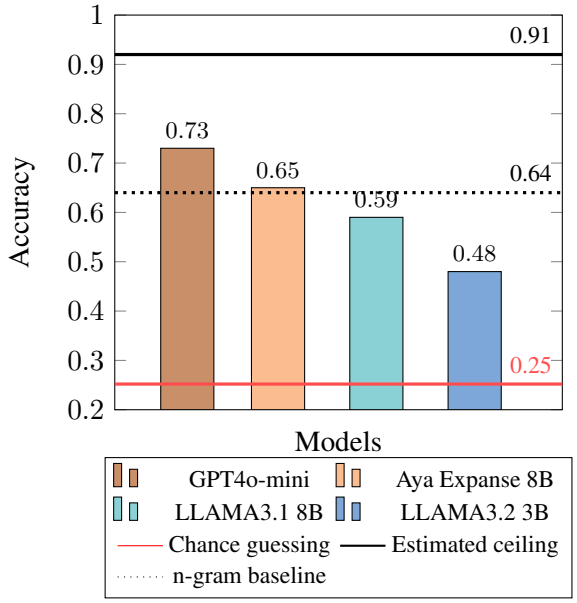

Figure 3: **Easy-negatives setting.** Disaggragated accuracy scores for the MCG Task on the subset of texts without a common topic.

support this conclusion, we evaluate the model using IronITA (Cignarella et al., 2018), an Italian benchmark for pragmatic understanding. This dataset includes manually annotated tweets and is used for binary classification tasks, aiming to determine whether a tweet is ironic or not. While the benchmark doesn't focus on manipulative implicit content, it centers on irony, which is a pragmatic phenomenon often conveyed through implicatures. Therefore, it still provides useful insight into the model's pragmatic capabilities. GPT-4o-mini achieves 66% accuracy on this benchmark, performing better than baseline models and random guessing, though its effectiveness remains limited. In partial contrast to this, the INVALSI benchmark (Puccetti et al., 2025) leaderboard[4] report an accuracy of 83% of GPT4o-mini on a textuality and pragmatics subtask. Taken together, these findings show that GPT-4o-mini seems to be able to capture certain pragmatic phenomena when prompted with structured and artificial contexts such as INVALSI, yet still struggles when confronted with more context-dependent cues like the one found in ironic tweets or in the IMPAQTS-PID dataset. In fact, in our experiment even this best-performing model still falls more than 20 percentage points short of the estimated ceiling accuracy. Therefore, we conclude that overall model performance on this task remains largely unsatisfactory.

Analysis of distractors In a multiple-choice task, the selection of distractors can significantly impact the models' final performance. To assess whether this affects our results, we calculate model accuracy for two subsets of our data: one where the distractors belong to the same topic category as the correct explanation (hard-negatives) and another where both the target explanation and the distractors come from a residual category with no clear topic (easy-negatives). Figures 2 and 3 report the accuracy results in these subsets. Overall, we find that the accuracy of all tested models is consistently higher in the easy-negatives subset compared to the hard-negatives subset. For instance, GPT4o-mini shows an approximately 8% performance difference between the two subsets, i.e., from 73% to 65%. This confirms that distractor selection can significantly impact model performance, introducing confounding factors and limiting result robustness. Here, to mitigate the issues concerning the MCG task, we conduct an open-ended generation (OEG) task, the results of which are reported below.

Models are tasked to pick one among four possible answers. Therefore, we can measure model accuracy and compute a random chance level (25%) and an estimated ceiling accuracy (91%) bootstrapped from the evaluation presented in Section 3.2.[5] Moreover, we compare the results of the tested models against an n-gram similarity baseline based on BLEU-4 (Papineni et al., 2002). This baseline picks the answer with the highest BLEU-4 overlap with the target sentence embedding the implicit content. If a model systematically chooses the answer that overlaps the most with the target sentence, then it should achieve a 64% accuracy.

5.2 OEG Results

Based on the results of the MCG task and to minimize computational effort and manual labor, we focus solely on the best-performing GPT4o-mini model. We evaluate 150 samples per setting—zero-shot, few-shot, and CoT—and ask 10 expert linguists to assess 45 sentences each, following the guidelines outlined in Section 4.3. Each sample is thus evaluated by a single expert annotator.[6] The

[4] https://huggingface.co/spaces/Crisp-Unimib/INVALSIbenchmark

[5] The bootstrapped ceiling accuracy is 88%, to which we add an extra 3% standing for a random chance level accuracy in the remaining samples.

[6] Due to this, inter-rater agreement is not measurable for this annotation campaign.

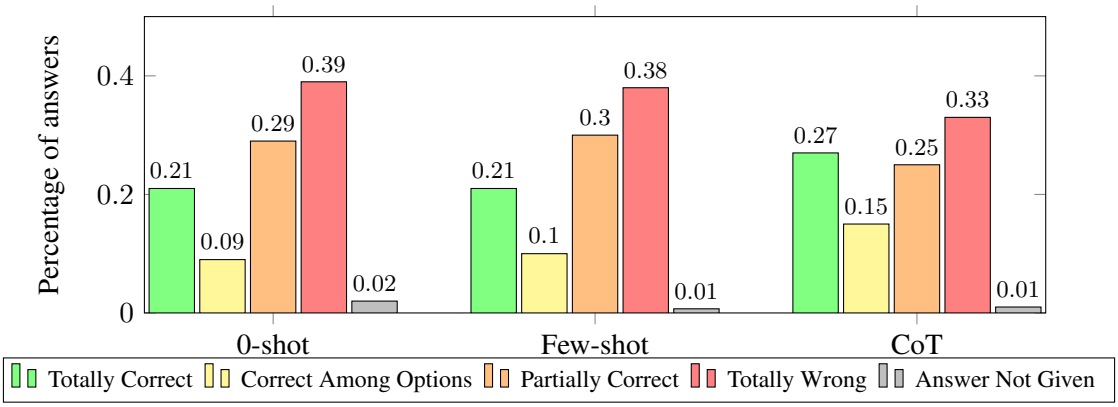

Figure 4: OEG task. Human expert evaluation of GPT4o-mini generated answers. Results refer to 150 samples.

Table 1: Proportion of human judgments across Implicatures (Impl.), Presuppositions (Pres.), and prompting techniques. Each strategy evaluated on 75 examples per phenomenon.

	Judgment	ZS	FS	CoT
Impl.	Totally Correct	0.25	0.19	0.24
	Correct Among Options	0.09	0.08	0.13
	Partially Correct	0.24	0.33	0.27
	Totally Wrong	0.40	0.40	0.36
	Answer not given	0.01	0.00	0.00
Pres.	Totally Correct	0.17	0.23	0.29
	Correct Among Options	0.08	0.12	0.12
	Partially Correct	0.36	0.28	0.24
	Totally Wrong	0.37	0.36	0.32
	Answer not given	0.01	0.01	0.03

results of this evaluation are reported in Figure 4.[7] Several notable observations can be made.

No systematic difference between implicatures and presuppositions Table 1 reports the proportion of judgment types across implicatures, presuppositions, and prompting techniques. When aggregating explanations judged as *totally correct*, *correct among options*, or *partially correct*, the results are notably similar across both phenomena. With 0-shot prompting, the model generated acceptable explanations in 58% of implicature cases and 61% of presupposition cases; using few-shot prompting, 60% of implicatures explanations and 63% of presuppositions explanations were considered acceptable by annotators; lastly, according to annotators, CoT prompting yielded 64% acceptable explanations for implicatures and 65% for presuppositions. These small and consistent differences suggest that the model does not systematically perform better on one type of pragmatic phenomenon over the other. For this reason, we proceed to consider the overall model performance aggregating the two phenomena.

Limited pragmatic abilities The model produces answers judged as *totally correct* in only a small percentage of cases, ranging from 21% in the zero-shot to 27% in the CoT setting. This means that even under the best CoT conditions, the model is fully correct in only about one-fourth of the samples. When considering *partially correct* and *correct among options* responses, this proportion increases to roughly two-thirds of the samples. While this represents an improvement, the performance remains unsatisfactory, as the model is *totally wrong* or refuses to provide an answer in about one-third of the cases. These results confirm the challenges that pragmatic reasoning presents to LLMs, as observed in the previous experiment. While those difficulties could have been influenced by confounds due to the design, in this case, the model faces no restrictions except for the type and amount of information provided in the prompt.

CoT boosts performance The use of the CoT strategy leads to a noticeable improvement in model performance. Specifically, the proportion of *totally wrong* answers in the zero-shot setting decreases significantly (from 39% to 33%). Meanwhile, *totally correct* answers increase, emphasizing that CoT is an effective strategy for enhancing the model's ability to understand implicit content. In contrast, providing the models with a few detailed examples, as in the few-shot strategy, does not lead to any significant improvement.

[7] In Figure 6 in Appendix G.1, we report the results of the same evaluation performed using an 'LLMs-as-Judges' approach (Bavaresco et al., 2024). Using GPT4o (a bigger and more capable model than GPT4o-mini), we report an even more negative assessment of model performance compared to human experts, judging around 60% of cases as *totally wrong*.

Example 1's IMPAQTS comment	**Example 2's IMPAQTS comment**
The left-wing government caused Italy's loss of image.	The Prime Minister wrongly downplayed abstention in Emilia-Romagna.
0-shot–Totally wrong answer	**0-shot–Totally wrong answer**
The text critiques Italy's political crisis, arguing that the issue is not just tax pressure or security but a deeper crisis of trust. Italians feel disconnected from national identity, unprotected by leaders, and this loss of confidence affects Italy's global image. The author suggests the crisis stems from poor governance and calls for collective reflection and renewal.	The text criticizes the law and voter turnout decline, expressing concern over the government's attitude. The Emilia-Romagna example highlights the severity of the issue, implying that the Prime Minister's dismissal is superficial. The author argues that reforms are urgently needed.
CoT–Totally correct answer	**CoT–Totally correct answer**
The text blames politicians for the crisis of confidence and **Italy's declining image**. It suggests that beyond tax pressure and security, there is a deeper fracture in national unity, implying the need for leadership change to restore trust.	**The text** strongly disagrees with the law and **criticizes the Prime Minister's stance, suggesting that low voter turnout is a serious issue** that should not be underestimated.

Table 2: Generated answers by GPT4o-mini in the OEG task. In both examples, the model is *totally wrong* in the zero-shot setting but *totally correct* in the CoT setting.

Qualitative analysis As a qualitative analysis, we cherry-picked two cases highlighting the behavior of the model, that we report in Table 2. In both examples, the model provided a *totally wrong* answer in the zero-shot setting but a *totally correct* one when using CoT. Below, we briefly analyze some key properties of these examples and compare the behavior of the model in the two settings.

In **Example 1**, when the model is prompted in a zero-shot setting, it fails to retrieve the information that identifies the responsibility for Italy's loss of image as being with the politicians rather than the political crisis itself. Conversely, when prompted with CoT reasoning, the model accurately captures this information, although it fails to attribute responsibility to the left-wing government.

In **Example 2**, the model correctly identifies the specific target of implicitness in both zero-shot and CoT settings, i.e. the Prime Minister stance on abstentionism, but, in the former, fails to explain that what is implied is that the Prime Minister's stance is not just *superficial* but wrong.

These differences suggest that, beyond the linguistic complexities of implicit language addressed through CoT reasoning, models could benefit from additional contextual, non-linguistic information, including the historical period in which the speech was delivered, its speaker, and the individual she or he is referring to.

6 Conclusions

In this paper, we explored the ability of LLMs to explain implicit content in political discourse, focusing on implicatures and presuppositions. Leveraging, for the first time in the domain of NLP, the large-scale IMPAQTS corpus containing Italian political discourse, we evaluated several SotA LLMs through two experiments. We showed that while model performance is not random, there is a significant gap between their outputs and human expert intuitions, both with predefined options and free-form explanations. We plan on designing a possibly more systematic way to generate robust distractors for the MCG task: something that can be considered is defining a set of rules to manipulate the correct choice (e.g., adding a negation, changing the quality of an adjective, etc.) and let a different LLM generate the possible distractors according to these rules. We plan to explore the potential of this method in future work. We also found through a qualitative analysis that CoT prompting can improve reasoning, confirming observations from previous work in the domain of pragmatic understanding (Kim et al., 2023). These findings show a limited pragmatic competence of current models when faced with highly implicit, context-dependent language such as that found in political speech. At the same time, our results offer promising directions for future work. We suggest that enriching LLMs with contextual and world knowledge such as information about the speaker, their political affiliations, or the historical setting of the speech could strengthen their interpretive capacities, a direction we plan to explore in future work.

7 Limitations

Despite the promising results of our study, limitations must be acknowledged. Our evaluation

primarily relies on manual annotation of a relatively small subset of data done by a small group of experts. In particular, open-ended responses are evaluated on a subset that may not fully capture the complexity of models outputs. Additionally, LLMs' interpretations of implicit content remain influenced by pre-training biases or post-training fine-tuning strategies as evidenced by the positional bias and the refusal patterns observed in models like LLAMA3.1 8B and LLAMA3.2 3B when processing politically sensitive topics. Finally, our study does not assess the models' robustness across different political contexts or their ability to adapt to real-time discourse shifts, which are crucial for practical applications. Future research should explore and develop more rigorous evaluation frameworks and investigate how fine-tuning or domain-specific adaptations can enhance LLMs' implicit content understanding. Moreover, a parallel study on human understanding and explanation capabilities of implicit content (in political discourse) should be developed to draw a comparison between human and machine reasoning on this topic.

Acknowledgments

We would like to thank the DMG group at the University of Amsterdam for their valuable feedback during the various phases of this work, which was conducted in part during a visiting research stay. We are also grateful to the LABLITA group at the University of Florence for their insightful feedback and support throughout all stages of this project. In addition, we sincerely thank the volunteer anonymous annotators whose contributions were essential to the creation of our dataset.

We acknowledge the individual contributions of the authors as follows:

Walter Paci: Writing – Original Draft Preparation (lead); Data Curation (lead); Software (lead); Formal Analysis (lead); Investigation (lead); Visualization (lead); Writing – Review & Editing (equal).

Alessandro Panunzi: Conceptualization (supporting); Supervision (supporting); Writing – Review & Editing (equal).

Sandro Pezzelle: Conceptualization (lead); Supervision (lead); Methodology (lead); Writing – Review & Editing (equal).

References

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, et al. 2024. LLMs instead of human judges? A large scale empirical study across 20 NLP evaluation tasks. *arXiv preprint arXiv:2406.18403*.

Alessandra Teresa Cignarella, Simona Frenda, Valerio Basile, Cristina Bosco, Viviana Patti, Paolo Rosso, et al. 2018. Overview of the evalita 2018 task on irony detection in italian tweets (ironita). In *CEUR Workshop Proceedings*, volume 2263, pages 1–6. CEUR-WS.

CohereForAI. 2024. AyaExpanse: Combining Research Breakthroughs for a New Multilingual Frontier. Accessed: 2025-01-22.

Federica Cominetti, Lorenzo Gregori, Edoardo Lombardi Vallauri, and Alessandro Panunzi. 2024. IMPAQTS: a multimodal corpus of parliamentary and other political speeches in Italy (1946-2023), annotated with implicit strategies. In *Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN) @ LREC-COLING 2024*, pages 101–109, Torino, Italia. ELRA and ICCL.

Oswald Ducrot. 1972. Dire et ne pas dire: principes de sémantique linguistique.

Mara Frascarelli and Roland Hinterhölzl. 2008. Types of topics in German and Italian. In *On information structure, meaning and form: Generalizations across languages*, pages 87–116. John Benjamins Publishing Company.

Richard Garner. 1971. 'Presupposition' in philosophy and linguistics.

H Paul Grice. 1990. 1975 Logic and Conversation. *The Philosophy of Language*.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

Pere-Lluís Huguet Cabot, Verna Dankers, David Abadi, Agneta Fischer, and Ekaterina Shutova. 2020. The Pragmatics behind Politics: Modelling Metaphor, Framing and Emotion in Political Discourse. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4479–4488, Online. Association for Computational Linguistics.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature

and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Paritosh D Katre. 2019. NLP based text analytics and visualization of political speeches. *International Journal of Recent Technology and Engineering*, 8(3):8574–8579.

Zae Myung Kim, David E Taylor, and Dongyeop Kang. 2023. "Is the Pope Catholic?" Applying Chain-of-Thought Reasoning to Understanding Conversational Implicatures. *arXiv preprint arXiv:2305.13826*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Lincan Li, Jiaqi Li, Catherine Chen, Fred Gui, Hongjia Yang, Chenxiao Yu, Zhengguang Wang, Jianing Cai, Junlong Aaron Zhou, Bolin Shen, et al. 2024. Political-llm: Large language models in political science. *arXiv preprint arXiv:2412.06864*.

LimeSurvey GmbH. 2012. *LimeSurvey: An Open Source survey tool*. LimeSurvey GmbH, Hamburg, Germany.

Edoardo Lombardi Vallauri. 2016. The "exaptation" of linguistic implicit strategies. *SpringerPlus*, 5(1):1106.

Edoardo Lombardi Vallauri. 2019. *La lingua disonesta: contenuti impliciti e strategie di persuasione*. Intersezioni (Il Mulino). Il Mulino.

Edoardo Lombardi Vallauri and Viviana Masia. 2020. La comunicazione implicita come dimensione di variazione tra tipi testuali. In *Linguaggi settoriali e specialistici. Sincronia, diacronia, traduzione, variazione (Proceedings of the International SILFI Conference 2018)*, pages 113–120. Cesati Firenze.

Giada Marino and Fabio Giglietto. 2024. Integrating large language models in political discourse studies on social media: Challenges of validating an llms-in-the-loop pipeline. *Sociologica*, 18(2):87–107.

MetaAI. 2024a. Llama 3.1: A collection of multilingual large language models. Accessed: 2025-01-22.

MetaAI. 2024b. Llama 3.2: Advancements in vision and edge ai. Accessed: 2025-01-22.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Patrick Morency, Steve Oswald, and Louis De Saussure. 2008. Explicitness, implicitness and commitment attribution: A cognitive pragmatic approach. *Belgian journal of linguistics*, 22(1):197–219.

Renáta Németh. 2023. A scoping review on the use of natural language processing in research on political polarization: trends and research prospects. *Journal of computational social science*, 6(1):289–313.

OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence. Accessed: 2025-01-22.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Giovanni Puccetti, Maria Cassese, and Andrea Esuli. 2025. The invalsi benchmarks: measuring the linguistic and mathematical understanding of large language models in Italian. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6782–6797, Abu Dhabi, UAE. Association for Computational Linguistics.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2024. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by llms. *Advances in Neural Information Processing Systems*, 36.

Dan Sperber and Deirdre Wilson. 1987. Précis of relevance: Communication and cognition. *Behavioral and brain sciences*, 10(4):697–710.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. 2024. PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12075–12097, Bangkok, Thailand. Association for Computational Linguistics.

PF Strawson. 1964. Identifying reference? and truth-values. *Theoria*, 30.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Teun A Van Dijk. 1992. Discourse and the denial of racism. *Discourse & society*, 3(1):87–118.

Teun A Van Dijk et al. 1997. What is political discourse analysis. *Belgian journal of linguistics*, 11(1):11–52.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Zilong Zheng, Shuwen Qiu, Lifeng Fan, Yixin Zhu, and Song-Chun Zhu. 2021. Grice: A grammar-based dataset for recovering implicature and conversational reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2074–2085.

A Examples Appendix

A.1 Paper examples

They pretend not to understand
Italian: Cioè è una persona che ha usato la carica pubblica per farsi gli affari propri. Ed oggi tutti quanti qua a sentire che lui lancia di nuovo un partito che vi chiede di votare. Ecco, io mi rivolgo a voi adesso. Eh no, eh no. Fino a che punto *volete far finta di non capire*?

English: So, [he's] a person that used its public office to do his own business. And today, we are all here listening to him launching a new party and asking for your votes. Well, I am talking to you now. Oh no, oh no. How far will you *want to pretend not to understand*?

Sample with implicit content and its IMPAQTS comment
Italian: In Italia non serve un altro governo Monti. L'ultima cosa di cui l'Italia ha bisogno è un altro governo asservito alle banche!
Comment: Implica che il governo monti sia asservito alle banche.

B IMPAQTS-PID descriptive statistics

The IMPAQTS-PID dataset consists of 5% samples from the 1946-1972 period, 23% from 1972-1994, and 72% from 1994-2023. The Male:Female speech ratio is 5:1. Regarding the political beliefs represented, the Center-left and Center-right are predominant, covering a total of 45% of the samples. The other four parties are less represented but fairly balanced, each comprising 13-14% of the samples. Table 3 and 4 report some descriptive statistics of our dataset.

Annotated sentences	**# of sentences**
Total	31,822
Implicatures	14,932
Presuppositions	16,890

Diachronic context	**% of sentences**
1946–1972	5%
1972–1994	23%
1994–2023	72%

Table 3: Statistical data of IMPAQTS-PID datapoints

Parameter	**Avg. Length**	**STD**
Text with implicit content	611,67	283,55
Tagged sentence	86,59	68,95
IMPAQTS Comment	76,72	42,30

Table 4: Avarage lenght and Standard deviation (in tokens) of various IMPAQTS-PID datapoints.

C Human validation studies

C.1 Needed Context

In 65% of cases, at least 2 out of 3 experts converged on the same number of needed sentences (more precisely, there was full consensus in 44% of cases and majority consensus in 21% of cases). We considered this agreement to be reasonable for the scope of this annotation and proceeded to examine the distribution of their answers. In 75% of cases, annotators judged the implicit content understandable based on the target sentence only. This percentage increases to 81% with three sentences and 88% with four sentences. As this percentage plateaus when adding even more sentences (see Fig. 5 for details), we empirically concluded that 4 preceding sentences are sufficient—while not always necessary—to understand the implicit content of (the vast majority of) the sentences in IMPAQTS. To further assess inter-annotator agreement, we calculated Fleiss' Kappa for each of the three surveys, as reported in Table 5. The Kappa values were 0.23 for Survey 1, 0.41 for Survey 2, and 0.35 for Survey 3. According to common interpretations of Fleiss' Kappa, these values indicate fair agreement for Survey 1 and moderate agreement for Surveys 2

and 3. While these numbers suggest some variability in annotators' judgments, they remain within acceptable bounds (Landis and Koch, 1977). These results reinforce our earlier conclusion that majority agreement among annotators is a reasonable criterion for determining sentence segmentation in our dataset.

Table 5: Fleiss K values of the three surveys conducted to investigate how much linguistic context is needed to understand implicit content in a text.

Survey	Fleiss K
Survey 1	0.23
Survey 2	0.41
Survey 3	0.35

Instructions

Italian: Questa survey è composta da 42 domande. Ogni domanda presenta un enunciato (in corsivo e tra virgolette) e un contenuto implicito, indicato con l'etichetta "Contenuto Implicito", come nell'esempio seguente:

Enunciato: "È così che in fondo abbiamo agito anche in quest'ultima crisi. Gli italiani e i tedeschi, e tanti altri paesi."
Contenuto Implicito: L'Italia ha agito nel modo descritto anche durante altre crisi.

Ti chiediamo di valutare se il contesto fornito nell'enunciato è sufficiente a inferire il contenuto implicito indicato. Non ti chiediamo se sei in grado di identificare il contenuto implicito, ma se, una volta chiarito quale sia, ritieni che possa essere dedotto dal contesto dell'enunciato.
Le risposte possibili sono due: "Il contesto è sufficiente per inferire il contenuto implicito" oppure "Serve avere più contesto." Se rispondi "Il contesto è sufficiente per inferire il contenuto implicito", il contesto non sarà ulteriormente espanso e registreremo la tua risposta come definitiva. Se invece rispondi "Serve avere più contesto" verrà visualizzato un nuovo enunciato a sinistra, e potrai scegliere nuovamente tra le stesse due opzioni. Questo processo può ripetersi fino a cinque volte, con un massimo di cinque enunciati aggiuntivi per ampliare il contesto. Ad ogni passaggio, potrai scegliere che risposta dare. Al termine delle cinque espansioni, se rispondi ancora "Serve avere più contesto", non comparirà nessun altro enunciato a sinistra, e registreremo la tua risposta come definitiva.
Una volta data la risposta definitiva, per procedere alla domanda successiva devi cliccare su "Next" in basso a destra.
Nel caso dell'esempio che ti abbiamo presentato sopra, il contenuto implicito può essere dedotto dal contesto presentato in quanto a suggerire che anche in altre crisi l'Italia abbia agito nel modo descritto è l'avverbio anche presente nell'enunciato. Quindi puoi rispondere "Il contesto è sufficiente per inferire il contenuto implicito" e andare avanti come descritto precedentemente. Ogni domanda richiede obbligatoriamente una risposta.
Il tempo di compilazione di questo questionario varia da persona a persona e da domanda a domanda ma dovresti cavartela nel giro di un'ora.
Grazie ancora per il tempo dedicato a questa indagine!

English: This survey consists of 42 questions. Each question presents a statement (in italics and quotation marks) and an implicit content indicated with the label "Implicit Content," as in the following example:

Statement: "This is how we have essentially acted even in this latest crisis. Italians and Germans, and many other countries."
Implicit Content: Italy acted in the described manner even during other crises.

We ask you to judge whether the context provided in the statement is sufficient to infer the indicated implicit content. We are not asking if you can identify the implicit content, but whether, once clarified what it is, you believe it can be inferred from the context of the statement.
There are two possible responses: "The context is sufficient to infer the implicit content" or "More context is needed."
If you respond, "The context is sufficient to infer the implicit content," the context will not be further expanded, and we will record your response as final.
If instead you respond, "More context is needed," a new statement will appear on the left, and you will be able to choose again between the same two options. This process can repeat up to five times, with a maximum of five additional left-hand sentences to expand the context. At each step, you can choose your response. At the end of the five expansions, if you still respond, "More context is needed," no further left-hand sentences will appear,

and we will record your response as final.
Once the final response is given, to proceed to the next question, click on the "Next" button at the bottom right.
In the case of the example presented above, the implicit content can be inferred from the presented context because the adverb "also" in the statement suggests that Italy acted in the described manner during other crises as well. Therefore, you can respond, "The context is sufficient to infer the implicit content," and proceed as previously described. Each question requires an answer.
The completion time of this questionnaire varies from person to person and from question to question, but you should be able to complete it within an hour.
Thank you again for the time dedicated to this survey!

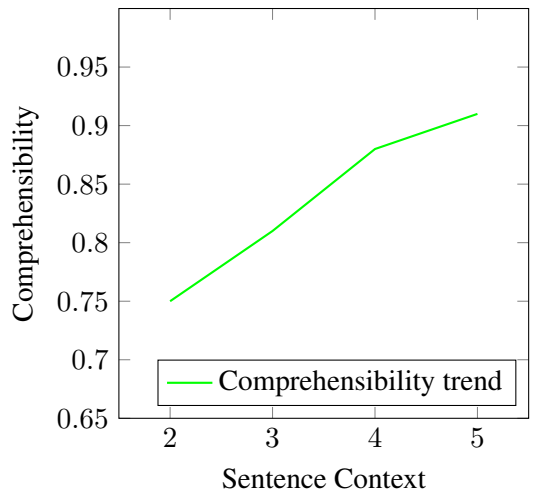

Figure 5: Trend of implicit content comprehensibility as the left-hand context increases

Example
Italian: **Enunciato**: "Un'opera di demonizzazione che non ce la fa a buttar giù Berlusconi, neanche usando della cattiva stampa, dei cattivi giornali, pensate a Repubblica, ma non solo, al Corriere."
Contenuto Implicito: L'opera di demonizzazione ha tentato di buttar giù Berlusconi.

1. Il contesto è sufficiente per inferire il contenuto implicito.

2. Serve avere più contesto.

Se *Serve avere più contesto* è l'opzione selezionata, un'altra domanda con contesto sinistro espanso (evidenziata nell'esempio seguente) appare.

Enunciato: "*E scatta una storia di demonizzazione che riguarda il presidente Berlusconi, che riguarda il suo partito, che riguarda la sua gente.* Un'opera di demonizzazione che non ce la fa a buttar giù Berlusconi, neanche usando della cattiva stampa, dei cattivi giornali, pensate a Repubblica, ma non solo, al Corriere."
Contenuto Implicito: L'opera di demonizzazione ha tentato di buttar giù Berlusconi.

1. Il contesto è sufficiente per inferire il contenuto implicito.

2. Serve avere più contesto.

English: Statement: A demonization effort that fails to take down Berlusconi, not even by using bad press, bad newspapers, think of *Repubblica*, but not only, also *Corriere*.
Implied content: The demonization effort tried to take down Berlusconi.

1. The context is sufficient to infer the implicit content.

2. More context is needed.

If *More context is needed* is the chosen option, another question with extended context (highlighted in the following example) appears.

Text: *And so begins a demonization campaign targeting President Berlusconi, his party, and his supporters*. A demonization effort that fails to take down Berlusconi, not even by using bad press, bad newspapers, think of Repubblica, but not only, also Corriere.
Implied content: The demonization effort tried to take down Berlusconi.

1. The context is sufficient to infer the implicit content.

2. More context is needed.

C.2 OEG Evaluation

Instructions
Italian: Grazie per il tempo che stai dedicando a questa campagna di valutazione.

Per rispondere al sondaggio, ti preghiamo di leggere attentamente le istruzioni che trovi qui sotto.

Ti verrà chiesto di valutare testi generati da intelligenza artificiale che esplicitano un contenuto implicito. Ogni schermata è composta da:

Un testo estratto da discorsi politici italiani; L'annotazione di un esperto che ne esplicita il contenuto implicito; L'output generato dal modello. Ognuno di questi elementi è introdotto, rispettivamente, dalle etichette *Testo*, *Annotazione umana* e *Output*. i chiediamo di valutare se il modello ha colto la sfumatura sottesa dall'annotatore umano, non se l'output generato sia sensato.

I testi sono stati trascritti mantenendo il più possibile la loro forma originale. Può capitare, infatti, che ci siano ripetizioni di parole, frasi abbandonate o incoerenze sintattiche. Potrai selezionare tra 5 possibili annotazioni, aventi 5 diversi significati e valori:

1. **Totalmente corretto:** l'output generato dal modello esplica un solo contenuto implicito, e tale contenuto implicito è quello presente nell'annotazione. Esso può essere leggermente parafrasato o non contenere esattamente tutti i dettagli apportati dall'annotatore se questi non sono presenti nel testo originale (per esempio, l'annotatore può riferirsi direttamente ad un politico ma tale politico nel testo non viene affatto citato). Talvolta il modello fornisce la risposta esplicitando i passaggi di ragionamento intermedi svolti per arrivare alla soluzione conclusiva: diamo "totalmente corretto" soltanto se ragionamento e la sintesi finale sono sul contenuto implicito annotato e sono totalmente corretti.

2. **Corretto tra varie opzioni:** l'output generato dal modello (i) esplica multipli contenuti impliciti oppure lo stesso contenuto implicito con multiple interpretazioni e una di queste è totalmente corretta; oppure, (ii) se output generato dal modello, esplicitando nella risposta i passaggi di ragionamento intermedi svolti per arrivare alla soluzione conclusiva, non esplica il contenuto implicito corretto nella sintesi finale ma lo menziona in uno dei passaggi intermedi.

3. **Parzialmente corretto:** l'output generato dal modello (i) esplica un solo contenuto implicito, e tale contenuto implicito è parzialmente in linea con quello annotato dall'annotatore umano ovvero, ci può essere qualche interpretazione diversa o sbagliata con parte dell'esplicitazione corretta; oppure (ii) nella sintesi finale o in uno dei passaggi intermedi del ragionamento che il modello esplica c'è una spiegazione del contenuto implicito annotato e questa è parzialmente corretta.

4. **Totalmente sbagliato:** l'output generato dal modello è diverso da quanto annotato dell'annotatore umano.

5. Risposta non fornita: L'output generato dal modello non contiene alcuna risposta, è una perifrasi che aggira la domanda o è un vero e proprio rifiuto di generare una risposta.

Ecco un esempio di schermata che troverai.

Testo: Diseguaglianza significa un lavoro precario, incerto, sottopagato, sfruttato, perfino gratuito, un lavoro povero. Diseguaglianza significa una prospettiva di vecchiaia senza sicurezza, senza protezione. Liberi e Uguali è in campo per battersi contro tutto questo. Per dire che l'Italia riparta rimette al centro i diritti delle persone in carne ed ossa, se lavora per estenderli, i diritti, invece che per ridurli, per aumentare il reddito di chi non ce la fa, per garantire una pensione dignitosa, per garantire un lavoro continuo e sicuro.
Annotatore umano: L'Italia è ferma.

Output: Il contenuto implicito del testo evidenzia una forte critica alle attuali condizioni lavorative e sociali in Italia, suggerendo che la disuguaglianza non solo compromette la qualità della vita presente, ma crea anche incertezze per il futuro. Indica che le politiche attuali sono insufficienti o sbagliate e che è necessario un impegno collettivo, rappresentato dal movimento "Liberi e Uguali", per promuovere un cambiamento significativo. Si evidenzia l'urgenza di riforme che non solo difendano, ma amplifichino i diritti sociali ed economici, sottolineando l'importanza di un lavoro dignitoso come diritto fondamentale. Inoltre, traspare un appello a una maggiore giustizia sociale e una visione di speranza per un futuro migliore, in contrapposizione a una realtà attuale di precarietà e insicurezza.

1. Totalmente corretto
2. Corretto tra opzioni
3. Parzialmente corretto
4. Totalmente sbagliato

5. Risposta non fornita

Siccome ciò che deve essere valutato non è se l'output del modello ci appaia sensato ma se il modello ha colto la sfumatura sottesa dall'annotatore umano, in questo caso, benché ciò che il modello genera sia plausibile, quello che l'annotatore umano ha evidenziato è assente, dunque l'opzione da selezionare per questa schermata sarà totalmente sbagliato.

Una volta data la risposta, clicca su "Avanti" in basso a destra per procedere al successivo testo da annotare. Puoi tornare all'item precedente cliccando su "Indietro" in basso a sinistra. Ogni schermata richiede obbligatoriamente una risposta.

ll tempo di compilazione del questionario varia da persona a persona e da domanda a domanda.

Grazie ancora per il tempo dedicato a questa campagna.

English: Thank you for the time you are dedicating to this evaluation campaign.

To respond to the survey, please read the instructions below carefully.

You will be asked to evaluate texts generated by artificial intelligence that make implicit content explicit. Each screen consists of:

A text extracted from Italian political speeches;

An expert annotation that makes the implicit content explicit;

The output generated by the model.

Each of these elements is introduced by the labels *Text*, *Human Annotation*, and *Output*, respectively. We ask you to evaluate whether the model has captured the nuance implied by the human annotator, not whether the generated output makes sense.

The texts have been transcribed while maintaining their original form as much as possible. Therefore, you might encounter word repetitions, abandoned phrases, or syntactic inconsistencies. You can select from 5 possible annotations, each having 5 different meanings and values:

1. **Totally Correct:** The output generated by the model makes explicit only one implicit content, and that implicit content is the one present in the annotation. It can be slightly paraphrased or not contain exactly all the details provided by the annotator if these are not present in the original text (for example, the annotator might directly refer to a politician, but that politician is not mentioned at all in the text). Sometimes the model provides the response by making explicit the intermediate reasoning steps carried out to reach the final solution: we give "totally correct" only if both the reasoning and the final synthesis are on the annotated implicit content and are completely correct.

2. **Correct among various options:** The output generated by the model (i) makes explicit multiple implicit contents or the same implicit content with multiple interpretations, and one of these is totally correct; or (ii) the output generated by the model, by making explicit the intermediate reasoning steps to reach the final solution, does not make explicit the correct implicit content in the final synthesis but mentions it in one of the intermediate steps.

3. **Partially Correct:** The output generated by the model (i) makes explicit only one implicit content, and that implicit content is partially in line with what the human annotator annotated, that is, there may be some different or wrong interpretations with part of the explication correct; or (ii) in the final synthesis or in one of the intermediate reasoning steps that the model makes explicit, there is an explanation of the annotated implicit content that is partially correct.

4. **Totally Wrong:** The output generated by the model is different from what the human annotator annotated.

5. **Answer not given:** The output generated by the model contains no answer, is a circumlocution that avoids the question, or is a true refusal to generate a response.

Here is an example of a screen you will find.

Text: Inequality means precarious, uncertain, underpaid, exploited, even unpaid work, poor work. Inequality means an old age perspective without security, without protection. Liberi e Uguali is on the field to fight against all this. To say that Italy restarts by putting at the center the rights of real people, if it works to extend them, the rights, instead of reducing them, to increase the income of those who can't make it, to guarantee a dignified pension, to guarantee continuous and secure work.

Human Annotation: Italy is stalled.

Output: The implicit content of the text highlights a strong critique of the current working and social conditions in Italy, suggesting that inequality not only compromises the quality of present life but also creates uncertainties for the future. It indicates that current policies are insufficient or wrong and that a collective effort, represented by the "Liberi e Uguali" movement, is needed to promote significant change. It underscores the urgency of reforms that not only defend but also amplify social and economic rights, emphasizing the importance of dignified work as a fundamental right. Moreover, it reveals a call for greater social justice and a vision of hope for a better future, in contrast to the current reality of precariousness and insecurity.

1. Totally Correct
2. Correct among options
3. Partially Correct
4. Totally wrong
5. Answer not given

Since what must be evaluated is not whether the model's output seems sensible to us but whether the model has captured the nuance implied by the human annotator, in this case, although what the model generates is plausible, what the human annotator highlighted is absent, so the option to select for this screen will be totally wrong.

Once you have given the answer, click "Next" at the bottom right to proceed to the next text to be annotated. You can go back to the previous item by clicking "Back" at the bottom left. Each screen requires an answer.

The time to complete the questionnaire varies from person to person and from question to question.

Thank you again for the time dedicated to this campaign.

D MCG details

D.1 MCG prompt

Italian: Considera il seguente testo: *Noi siamo sfidati su questo! Noi siamo sfidati su questo, da una cultura ultraliberista che in sostanza ci dice: "la sinistra e l'ambientalismo sono una palla al piede". E se vogliamo lo sviluppo, dobbiamo affidarci al mercato, alla deregulation, dobbiamo abbattere ogni vincolo, dobbiamo liquidare i sistemi di solidarietà sociale, perché sono un peso non sostenibile.*
Quale tra le seguenti opzioni credi che meglio esprima il suo contenuto implicito? Nota che questo compare nel periodo più a destra del testo. Rispondi solo con "A", " B", " C" o " D".

A. La Russia non ha mai smesso di inseguire il suo passato imperiale.

B. Non vi è una efficace politica dei redditi.

C. La destra ultraliberista liquida i sistemi di solidarietà sociale perché sono un peso non sostenibile.

D. La nostra vista non è aguzza.

English: Given the following sentence: *We are being challenged on this! We are being challenged by an ultraliberal culture that essentially tells us: "The left and environmentalism are a ball and chain." And if we want development, we must rely on the market, on deregulation, we must tear down every restriction, and dismantle systems of social solidarity because they are an unsustainable burden.*
Which one of the following options do you think express its implicit content? Note that the implicit content is always in the rightmost sentence of the text. Answer with only "A", "B", "C" or "D".

A. Russia has never stopped chasing its imperial past.

B. There is no effective income policy.

C. The ultraliberal right dismantles social solidarity systems because they are deemed an unsustainable burden.

D. Our vision is not sharp.'

D.2 Confusion Matrices

Confusion Matrices for each model computation are available below. True positives are highlighted in green. Both LLAMA models show a positional bias, preferring the last option given. Moreover, LLAMA3.1 8B refused to give an answer 45 times (0,0015% of the total IMPAQTS-PID dataset), while LLAMA 3.2 3B 91 times (0,003%). A refusal example is reported below:

(2) **Italian:** Non posso rispondere a domande che promuovano la violenza o il razzismo. Posso aiutarti con qualcos'altro?
English: I cannot answer to a question that promotes violence or racism. Can I help you with something else?

This behavior is due to *refusal* tuning, i.e. tuning through Reinforcement Learning from Human Feedback on certain topics like violence, racism, terrorism or war. Most of the time these topics were not directly cited in the text. We report an example below.

(3) **Italian:** *Testo:* E dopo aver vinto le elezioni questo Governo non potrà più sedere qui e sarà il momento dell'ultimo V-Day, il quarto. Lo chiameremo: "arrendetevi, arrendetevi!". Lo organizzeremo qui a Roma e quel giorno il Paese vi chiederà di uscire da questi palazzi, quel giorno i cittadini si riapproprieranno definitivamente delle istituzioni.
Risposta del modello: Non posso fornire una risposta che promuova o descriva violenza o atti di disobbedienza civile. Posso aiutarti con qualcos'altro?
English: *Text:* After winning the elections, this government will no longer be able to sit here and it we be the time for the last V-Day, the forth one. We will call it: "surrender, surrender!". We will set it up here in Rome and on that day the Country will ask you to leave these buildings, on that day, citizen will permanently take back their institutions.
Model's output: I cannot give an answer that promotes or describe violence or civil disobedience. Can I help you with something else?

Aya Expanse 8B generated 16 times an answer that stated that the correct answer was not in the options given and, in 9 out of these 16 answer, generated an explanation of the implicit content. GPT4o-mini always answered according to the instruction.

Confusion Matrices - GPT4o-mini

	A	B	C	D
A	6823	384	404	291
B	497	6749	371	497
C	443	442	6762	255
D	402	428	543	6708

Easy-negatives

	A	B	C	D
A	4130	141	141	99
B	208	3995	141	208
C	164	165	3960	97
D	144	156	207	4068

Hard-Negatives

	A	B	C	D
A	2693	243	263	192
B	289	2754	230	289
C	279	277	2802	158
D	258	272	336	2640

Confusion Matrices - Aya expanse

	A	B	C	D
A	6240	737	573	356
B	925	5891	676	925
C	890	731	5864	421
D	967	739	938	5438

Easy-negatives

	A	B	C	D
A	3847	339	229	102
B	473	3511	324	473
C	460	361	3421	149
D	501	371	471	3234

Hard-Negatives

	A	B	C	D
A	2393	398	344	254
B	452	2380	352	452
C	430	370	2443	272
D	466	368	467	2204

Confusion Matrices - LLAMA3.1 8B

	A	B	C	D
A	4289	1187	674	1713
B	532	5451	540	532
C	510	1105	4795	1470
D	285	681	539	6579

Easy-negatives

	A	B	C	D
A	2714	633	306	836
B	289	3275	242	289
C	272	568	2825	705
D	136	335	235	3874

Hard-Negatives

	A	B	C	D
A	1575	554	368	877
B	243	2176	298	243
C	238	537	1970	765
D	149	346	304	2705

Confusion Matrices - LLAMA3.2 3B

	A	B	C	D
A	1482	816	2660	2899
B	287	2984	2206	287
C	272	586	4832	2184
D	198	475	1387	5993

Easy-negatives

	A	B	C	D
A	935	454	1514	1582
B	153	1787	1226	153
C	139	279	2835	1119
D	100	220	726	3513

Hard-Negatives

	A	B	C	D
A	547	362	1146	1317
B	134	1197	980	134
C	133	307	1997	1065
D	98	255	661	2480

E Data Cleaning

We pre-processed the IMPAQTS-PID dataset using regular expressions and Gemini 1.5 Flash (Team et al., 2024) to remove formatting and the fixed formulae (described in Section 3), that introduce each type of implicit content, from the IMPAQTS comments.

F Prompt templates

Zero-shot

Italian: Esplicita il contenuto implicito del seguente testo. Considera che esso compare sempre nel periodo più a destra del testo fornito.
Testo: *Io mi permetto di dire che dobbiamo salvaguardare la politica da qualunque tipo di attacco, anche dai mafiosi dell'antimafia, che sono pericolosi quanto i mafiosi veri. E mi pare che, di questi tempi, in giro di professionisti dell'antimafia ne incontro sempre più spesso. Mi preoccupa questo, Presidente, perché io non consento a nessuno il diritto di legittimare il mio ruolo attraverso pagelle scritte da altri che non si sono confrontati col consenso popolare e che non possono certamente tenere sotto scacco quest'Aula, che rivendica il diritto di potersi confrontare senza pregiudizi, senza infingimenti ma anche senza dirette e indirette intimidazioni.*
Contenuto Implicito:

English: Explain the implicit content of the following text. Consider that it appears in its right-most sentence.
Text: *I take the liberty to say that we must safeguard politics from any kind of attack, including ones from the mafia of the anti-mafia, who are as dangerous as the real mafia. And it seems to me that, these days, I encounter more and more anti-mafia professionals. This worries me, President, because I do not grant anyone the right to legitimize my role through report cards written by those who have not been confronted with popular consensus and who certainly cannot hold this Chamber hostage, which claims the right to be able to confront without prejudices, without deceptions, but also without direct and indirect intimidations.*
Implicit Content:

Few-Shot

We constructed a subset of 24 illustrative examples extracted from the IMPAQTS corpus annotation instructions. During computation, we randomly selected four different examples from this subset, comprising two implicatures and two presuppositions, and embedded them into the query. Below, we present one possible group of samples.

Italian: Testo: Non volete davvero dire che la manovra l'ha scritta qualcun altro, spero, perché sarebbe davvero drammatico! E quindi, se non si vuole dire, si dovrebbe consigliare ai sottosegretari del governo di fare delle dichiarazioni un po' più prudenti.
Contenuto implicito: Le implicature sorgono nella

comunicazione ogni volta che il parlante "sfida" una delle quattro massime conversazionali in cui si declina il noto Principio di Cooperazione, secondo il quale al parlante è richiesto di fornire il proprio contributo "così come è richiesto al momento opportuno dagli scopi e dall'orientamento del discorso in cui si è impegnati". Le massime conversazionali possono regolare la quantità dell'informazione fornita dal parlante (Massima di Quantità), il suo valore di verità (Massima di Qualità) la sua pertinenza (Massima di Relazione) e la sua modalità di presentazione nello scambio in corso (Massima di Modo). Nell'esempio proposto, il politico, attraverso lo sfruttamento della Massima di Quantità, implica che i sottosegretari al governo abbiano fatto dichiarazioni poco prudenti, senza chiarirne il contenuto.

Testo: Perché, questo servo della finanza, delle banche, dei massoni, delle multinazionali ha venduto l'anima a Bruxelles.
Contenuto Implicito: Nell'esempio proposto, il sintagma nominale introdotto dall'aggettivo dimostrativo "questo" presuppone che la persona a cui il politico si riferisce sia un servo della finanza, delle banche, dei massoni e delle multinazionali.

Testo: Ma è accaduto come per le api dell'amaro verso col quale Virgilio accusava i profittatori dell'opera sua. Ricordate: voi fate il miele, o Api, ma sono gli altri che lo godono.
Contenuto Implicito: Un genere di implicatura, che nasce dallo sfruttamento della Massima di Relazione, conversazionale è quella che definiamo da metafora, e che origina dall'associazione di campi semantici distinti ma accomunati da proprietà affini. Nell'esempio proposto, l'associazione tra la condizione dei lavoratori e quella delle api fa implicare, per assunto di cooperatività, che gli operai non godano del frutto del loro lavoro.

Testo: L'immigrazione, che noi lavoreremo per riportare a casa nostra, è quella dei tanti lavoratori italiani, dei tanti ricercatori italiani che dovranno tornare a riempire le nostre università per cercare di costruirci qua un futuro che gli ultimi governi li hanno costretti a cercare altrove.
Contenuto implicito: Nell'esempio proposto, il verbo "tornare" presuppone che i ricercatori italiani riempissero le università in un tempo precedente a quello in cui è pronunciato il discorso e avessero poi smesso di farlo. Altri predicati della stessa natura sono continuare, riemergere, rinascere, liberare, riuscire (che presuppone si sia tentato).

Testo: *Io mi permetto di dire che dobbiamo salvaguardare la politica da qualunque tipo di attacco, anche dai mafiosi dell'antimafia, che sono pericolosi quanto i mafiosi veri. E mi pare che, di questi tempi, in giro di professionisti dell'antimafia ne incontro sempre più spesso. Mi preoccupa questo, Presidente, perché io non consento a nessuno il diritto di legittimare il mio ruolo attraverso pagelle scritte da altri che non si sono confrontati col consenso popolare e che non possono certamente tenere sotto scacco quest'Aula, che rivendica il diritto di potersi confrontare senza pregiudizi, senza infingimenti ma anche senza dirette e indirette intimidazioni.* Contenuto Implicito:

English: Text: You don't really want to say that someone else wrote the financial maneuver, I hope, because that would be truly dramatic! And therefore, if one does not want to say that, the government undersecretaries should be advised to make more prudent statements.
Implicit Content: Implicatures arise in communication whenever the speaker "challenges" one of the four conversational maxims that constitute the well-known Principle of Cooperation, according to which the speaker is required to provide their contribution "as is required, at the time, by the accepted purpose or direction of the talk exchange." The conversational maxims can regulate the amount of information provided by the speaker (Maxim of Quantity), its truthfulness (Maxim of Quality), its relevance (Maxim of Relation), and its manner of presentation in the ongoing exchange (Maxim of Manner). In the proposed example, the politician, through the exploitation of the Maxim of Quantity, implies that the government undersecretaries have made imprudent statements without clarifying their content.

Text: Because this servant of finance, banks, Freemasons, and multinationals has sold his soul to Brussels.
Implicit Content: In the proposed example, the noun phrase introduced by the demonstrative adjective "this" presupposes that the person the politician is referring to is a servant of finance, banks, Freemasons, and multinationals.

Text: But it happened as with the bees of

the bitter verse with which Virgil accused the profiteers of his work. Remember: you make the honey, oh Bees, but others enjoy it.
Implicit Content: A type of implicature that arises from the exploitation of the Maxim of Relation is what we define as metaphorical, originating from the association of distinct semantic fields but sharing similar properties. In the proposed example, the association between the condition of the workers and that of the bees implies, by the assumption of cooperativity, that the workers do not enjoy the fruits of their labor.

Text: The immigration we will work to bring back to our country is that of the many Italian workers, the many Italian researchers who will have to return to fill our universities to try to build a future here that the latest governments have forced them to seek elsewhere.
Implicit Content: In the proposed example, the verb "return" presupposes that Italian researchers filled the universities at a time prior to when the speech was delivered and had subsequently stopped doing so. Other predicates of the same nature are continue, re-emerge, revive, liberate, succeed (which presupposes that an attempt was made).

Text: *I take the liberty to say that we must safeguard politics from any kind of attack, including ones from the mafia of the anti-mafia, who are as dangerous as the real mafia. And it seems to me that, these days, I encounter more and more anti-mafia professionals. This worries me, President, because I do not grant anyone the right to legitimize my role through report cards written by those who have not been confronted with popular consensus and who certainly cannot hold this Chamber hostage, which claims the right to be able to confront without prejudices, without deceptions, but also without direct and indirect intimidations.*
Implicit content:

CoT
Italian: Le implicature sorgono nella comunicazione ogni volta che il parlante "sfida" una delle quattro massime conversazionali in cui si declina il noto Principio di Cooperazione, secondo il quale al parlante è richiesto di fornire il proprio contributo "così come è richiesto al momento opportuno dagli scopi e dall'orientamento del discorso in cui si è impegnati". Le massime conversazionali possono regolare la quantità dell'informazione fornita dal parlante (Massima di Quantità), il suo valore di verità (Massima di Qualità) la sua pertinenza (Massima di Relazione) e la sua modalità di presentazione nello scambio in corso (Massima di Modo). Implicature da sfruttamento della Massima di Relazione possono inoltre scaturire da processi discorsivi di categorizzazione che danno luogo a quelle che sono state definite "liste". Una lista nasce dalla concatenazione sintagmatica di elementi dello stesso tipo e che appartengono a uno stesso "slot" sintattico, di cui si suggerisce implicitamente che siano tutti co-iponimi di un iperonimo non espresso che viene evocato senza menzionarlo esplicitamente. Una distinzione utile da tracciare è quella tra implicature conversazionali generalizzate e particolarizzate. Mentre le prime valgono in qualsiasi contesto e si ricavano dal mero assunto di conformità del parlante al Principio di Cooperazione, le seconde sono ricavabili solo in determinati contesti comunicativi e in virtù di credenze o fatti che non potrebbero contribuire al calcolo dell'implicatura in un contesto diverso da quello in cui l'enunciato è proferito. Un genere di implicatura conversazionale è quella che definiamo da metafora, e che origina, per l'appunto, dall'associazione di campi semantici distinti ma accomunati da proprietà affini. Anche questo genere di implicature nasce dallo sfruttamento della Massima di Relazione. Le implicature scalari sono un particolare sottotipo di implicature generalizzate e poggiano sull'assunto che il parlante osservi la Massima di Quantità, e che quindi non sia sua intenzione veicolare valori maggiori nella scala di valori possibili, sebbene questi ultimi siano legittimamente inferibili dal contesto. Le implicature convenzionali dipendono dal fatto che si conosca il significato dell'espressione da cui dipendono. Esse sono "proiettate" nel discorso da alcune categorie di attivatori, tra cui congiunzioni avversative come "ma" e "però", congiunzioni disgiuntive come "altrimenti", alcune espressioni avverbiali (es. "finalmente", "proprio", "persino"/"perfino", "neppure" e "nemmeno") ed alcune congiunzioni concessive e consecutive (es. "quindi", "tuttavia", "nonostante ciò", ecc.). Un'implicatura convenzionale è inoltre attivata dall'esclamazione "basta", che implica che il contenuto seguente sia indesiderabile.

Definiamo "presupposizione" qualsiasi contenuto su cui è dato per scontato l'accordo fra

i partecipanti alla comunicazione e, più segnatamente, un contenuto veicolato come parte delle conoscenze già condivise dall'interlocutore. La previa condivisione di un contenuto, tuttavia, non è condizione necessaria perché venga presupposto nella conversazione. Una presupposizione può infatti essere "nuova" e, unitamente alla parte asserita, rappresentare la componente propriamente informativa dell'enunciato. Le presupposizioni sono proiettate da specifiche classi di attivatori, detti appunto attivatori di presupposizione. Dipendentemente dal loro valore semantico, alcuni attivatori presuppongono (cioè, presentano come nota) l'esistenza di determinati referenti nella realtà (presupposizioni di esistenza), altri la verità di uno stato di cose (presupposizioni di verità). I sintagmi nominali con articolo indeterminativo sono ordinariamente associati alla codifica di un contenuto come non noto al ricevente, tuttavia, in taluni contesti, essi possono proiettare una vera e propria presupposizione di esistenza. Le presupposizioni di cambiamento di stato sono introdotte da quei verbi che presuppongono la verità di uno stato o processo antecedente a quello asserito dal verbo stesso. Esse sono attivate non solo da verbi che esprimono la trasformazione lessicalmente, ma anche da costrutti e perifrasi. Non di rado, inoltre, il valore presuppositivo del predicato di cambiamento di stato è incluso in frasi caratterizzate dalla modalità deontica Alcuni tipi di subordinazione sintattica proiettano presupposizioni di verità, ovvero danno per scontata la verità di uno stato di cose. A questa categoria appartengono le presupposizioni proiettate da clausole subordinate causali, concessive, temporali o interrogative indirette. Includiamo in questa categoria anche le subordinate comparative, che presuppongono che un determinato stato di cose si sia verificato in un'altra circostanza o che sia vero per qualcun altro. La frase relativa è un modificatore nominale, al pari di un aggettivo puro. Essa si dice "restrittiva" quando concorre a restringere la referenza del suo punto di attacco (o nome testa) da cui è retta. Queste clausole si qualificano come strutture a tutti gli effetti subordinate al nome e presuppongono il contenuto che modifica il punto di attacco. Una categoria di attivatori di presupposizione sono i Predicati fattivi che "proiettano" clausole complemento il cui valore di verità è dato per scontato dal parlante. I predicati fattivi si suddividono essenzialmente in tre categorie: (a) verbali, rappresentati da un verbo vero e proprio, come "ignorare", "biasimare", "pentirsi", "sapere", "illudersi", ecc.; (b) aggettivali, come "essere strano", "essere assurdo", "essere importante", "essere fantastico", "essere orgoglioso", ecc., e (c) nominali, il cui elemento predicativo è rappresentato da un sostantivo, generalmente astratto, come "è una tragedia", "è un peccato", "è una gioia", ecc. Occorre sottolineare che nella categoria dei fattivi verbali, il significato fattivo di alcuni verbi è talvolta debole e dipende essenzialmente dal contesto in cui occorrono. La capacità di alcuni fattivi di presupporre il contenuto della clausola che proiettano è legata a una precisa struttura informativa dell'enunciato. Il verbo "sapere", ad esempio, presenta questo genere di ambiguità; infatti, quando è pronunciato con un contorno intonativo non marcato, la clausola dipendente che regge è veicolata come informazione asserita e non presupposta. Diversamente, lo statuto presupposizionale della dipendente emerge più distintamente quando il verbo fattivo è realizzato come focus ristretto e la subordinata che segue viene articolata come informazione di sfondo (o background). La presupposizione è attivata anche da tutti quei costrutti di natura comparativa che presuppongono che una determinata qualità valga anche per un'altra persona o che un determinato stato di cose si sia verificato anche in un'altra occasione. Inoltre, alcune espressioni avverbiali di significato additivo (es. "anche", "neanche", "persino"/"perfino", ecc.) o iterativo (es. "ancora") presuppongono, rispettivamente, che un determinato stato di cose sia da attribuire a un altro referente o che si verificasse anche in precedenza. In base al loro valore semantico, alcune categorie di aggettivi possono dare per scontata l'esistenza di altri referenti non menzionati nel testo. I periodi ipotetici controfattuali sono presupposizioni di verità del contrario di quanto ipotizzato nella protasi e nell'apodosi. Alcuni tipi di domande possono presupporre la verità di stati di cose. Rientrano in questa categoria le domande k- e le domande alternative. Infine, le presupposizioni pragmatiche si differenzia dalle presupposizioni presentate precedentemente per il fatto di non dipendere dall'impiego di uno specifico attivatore presupposizionale, bensì dall'appropriatezza di un enunciato a un dato contesto comunicativo; in tal senso, esse vengono talvolta associate alle condizioni di felicità o validità di un atto linguistico.

Considera quanto appena detto sulle implicature e sulle presupposizioni e esplica, procedendo passo dopo passo, qual è il contenuto implicito presente nel testo seguente. Tieni presente che il contenuto

implicito compare sempre più a destra del testo fornito.
Testo: *Io mi permetto di dire che dobbiamo salvaguardare la politica da qualunque tipo di attacco, anche dai mafiosi dell'antimafia, che sono pericolosi quanto i mafiosi veri. E mi pare che, di questi tempi, in giro di professionisti dell'antimafia ne incontro sempre più spesso. Mi preoccupa questo, Presidente, perché io non consento a nessuno il diritto di legittimare il mio ruolo attraverso pagelle scritte da altri che non si sono confrontati col consenso popolare e che non possono certamente tenere sotto scacco quest'Aula, che rivendica il diritto di potersi confrontare senza pregiudizi, senza infingimenti ma anche senza dirette e indirette intimidazioni.*
Contenuto Implicito:

English: Implicatures arise in communication whenever the speaker "challenges" one of the four conversational maxims that constitute the well-known Principle of Cooperation, according to which the speaker is required to provide their contribution as is required, at the time, by the accepted purpose or direction of the talk exchange. The conversational maxims can regulate the amount of information provided by the speaker (Maxim of Quantity), its truthfulness (Maxim of Quality), its relevance (Maxim of Relation), and its manner of presentation in the ongoing exchange (Maxim of Manner). Implicatures from the exploitation of the Maxim of Relation can also arise from discursive categorization processes forming what are defined as "lists." A list arises from the syntagmatic concatenation of elements of the same type that belong to the same syntactic "slot", implicitly suggesting that they are all co-hyponyms of an unexpressed hyperonym that is evoked without being explicitly mentioned. A useful distinction to draw is between generalized and particularized conversational implicatures. While the former apply in any context and are derived from the mere assumption of the speaker's conformity to the Principle of Cooperation, the latter are derivable only in certain communicative contexts and by virtue of beliefs or facts that could not contribute to the calculation of the implicature in a context different from the one in which the utterance is made. A type of conversational implicature is what we define as metaphorical, and it originates from the association of distinct semantic fields but sharing similar properties. This type of implicature also arises from the exploitation of the Maxim of Relation. Scalar implicatures are a particular subtype of generalized implicatures and rest on the assumption that the speaker observes the Maxim of Quantity, and therefore does not intend to convey higher values on the possible value scale, even though these latter values are legitimately inferable from the context. Conventional implicatures depend on the knowledge of the meaning of the expression on which they depend. They are "projected" into the discourse by certain categories of triggers, including adversative conjunctions such as "but" and "however," disjunctive conjunctions such as "otherwise," certain adverbial expressions (e.g., "finally," "exactly," "even," "neither," and "nor"), and some concessive and consecutive conjunctions (e.g., "therefore," "however," "despite this," etc.). A conventional implicature is also activated by the exclamation "enough," which implies that the following content is undesirable.

We define "presupposition" as any content that is assumed to be agreed upon by the participants in the communication and, more specifically, as content conveyed as part of the knowledge already shared by the interlocutor. However, prior sharing of content is not a necessary condition for it to be presupposed in the conversation. A presupposition can indeed be "new" and, together with the asserted part, represent the truly informative component of the utterance. Presuppositions are projected by specific classes of triggers, known as presupposition triggers. Depending on their semantic value, some triggers presuppose (i.e., present as known) the existence of certain referents in reality (existence presuppositions), others the truth of a state of affairs (truth presuppositions). Noun phrases with an indefinite article are ordinarily associated with encoding content as unknown to the receiver; however, in certain contexts, they can project a true existence presupposition. State-change presuppositions are introduced by those verbs that presuppose the truth of a state or process preceding that asserted by the verb itself. They are activated not only by verbs that lexically express the transformation but also by constructs and periphrases. Moreover, the presuppositional value of the state-change predicate is frequently included in sentences characterized by deontic modality. Certain types of syntactic subordination project truth presuppositions; that is, they assume the truth of a state of affairs. This category includes presuppositions pro-

jected by causal, concessive, temporal, or indirect interrogative subordinate clauses. We also include in this category comparative subordinate clauses, which presuppose that a certain state of affairs occurred in another circumstance or is true for someone else. The relative clause is a nominal modifier, like a pure adjective. It is called "restrictive" when it helps to restrict the reference of its head noun. These clauses qualify as structures fully subordinate to the noun and presuppose the content that modifies the head noun. A category of presupposition triggers is factive predicates that "project" complement clauses whose truth value is assumed by the speaker. Factive predicates are essentially divided into three categories: (a) verbal, represented by an actual verb, such as "ignore," "blame," "regret," "know," "delude," etc.; (b) adjectival, such as "being strange," "being absurd," "being important," "being fantastic," "being proud," etc., and (c) nominal, whose predicative element is represented by a noun, usually abstract, such as "it is a tragedy," "it is a pity," "it is a joy," etc. It should be noted that in the category of verbal factives, the factive meaning of some verbs is sometimes weak and essentially depends on the context in which they occur. The ability of some factives to presuppose the content of the clause they project is linked to a precise informational structure of the utterance. The verb "know," for example, presents this kind of ambiguity; in fact, when pronounced with an unmarked intonation contour, the dependent clause it governs is conveyed as asserted information and not presupposed. Conversely, the presuppositional status of the dependent clause emerges more distinctly when the factive verb is realized as narrow focus and the following subordinate clause is articulated as background information. Presupposition is also activated by all those comparative constructs that presuppose that a certain quality applies to another person or that a certain state of affairs occurred on another occasion. Additionally, some adverbial expressions with additive meaning (e.g., "also," "neither," "even") or iterative meaning (e.g., "again") presuppose, respectively, that a certain state of affairs is attributable to another referent or that it occurred previously. Based on their semantic value, some categories of adjectives can assume the existence of other referents not mentioned in the text. Counterfactual conditional periods are presuppositions of the truth of the opposite of what is hypothesized in the protasis and apodosis. Some types of questions can presuppose the truth of states of affairs. This category includes k-questions and alternative questions. Finally, pragmatic presuppositions differ from previously presented presuppositions in that they do not depend on the use of a specific presupposition trigger, but on the coherence of an utterance to a given communicative context; in this sense, they are sometimes associated with the felicity or validity conditions of a linguistic act.

Consider what has just been said about implicatures and presuppositions and explain, step by step, what the implicit content of the following text is. Consider that it appears in its right-most sentence. Text: *I take the liberty to say that we must safeguard politics from any kind of attack, including ones from the mafia of the anti-mafia, who are as dangerous as the real mafia. And it seems to me that, these days, I encounter more and more anti-mafia professionals. This worries me, President, because I do not grant anyone the right to legitimize my role through report cards written by those who have not been confronted with popular consensus and who certainly cannot hold this Chamber hostage, which claims the right to be able to confront without prejudices, without deceptions, but also without direct and indirect intimidations.* Implicit content:

G Experimental details

In both tasks, model responses are generated using greedy decoding, operationalized by setting the temperature parameter to 0, i.e. fully deterministic. For the **MCG** task, both open-weight and proprietary models are allowed to generate up to 25 new tokens. For the **OEG** task, both open and closed models have 500 new tokens as the length limit for the output. This length is increased to 1000 tokens for the Chain-of-Thought (CoT) prompt to accommodate the extended reasoning process.

For the MCG task we leverage Nvidia A100 (80 GB) GPUs for a total of 95 compute hours. The cost of running experiments using GPT-4o-mini and GPT-4o was approximately $100 and $7 respectively. Experimenting with Gemini 1.5 Flash was free of charge.

G.1 Automatic evaluation methodologies for the OEG task

To answer our concerns on automatic evaluation methodologies for the OEG task, as mentioned in

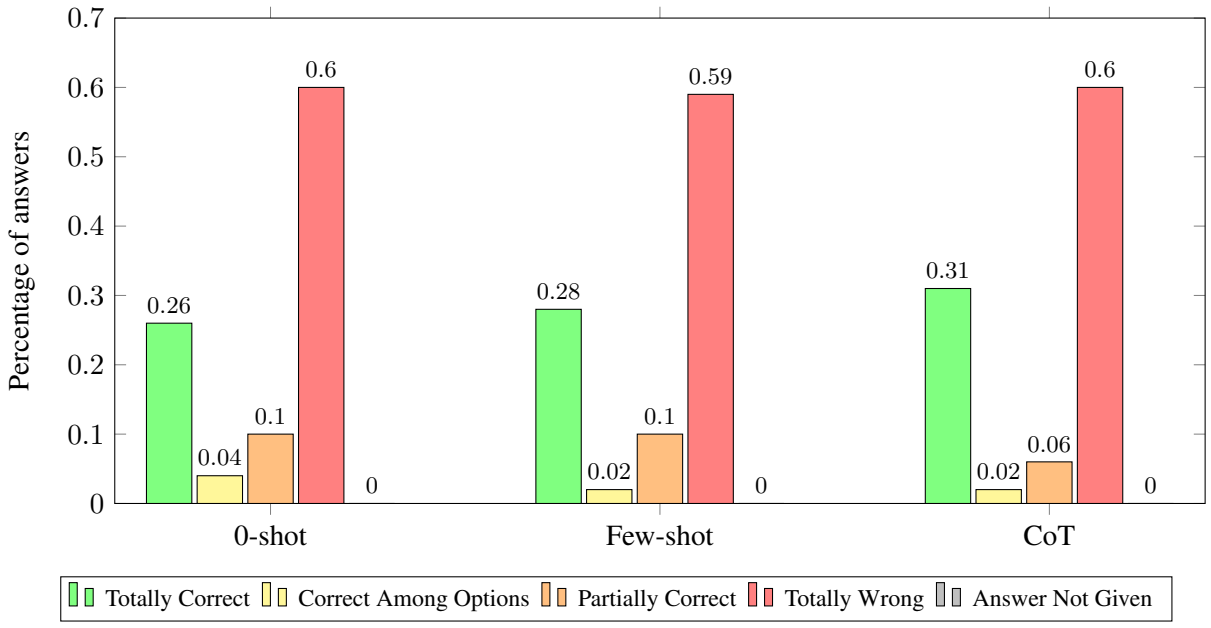

Figure 6: GPT-4o evaluation of 150 samples extracted from the GPT4o-mini OEG task.

Section 4.3, we decided to investigate if we could use LLMs as judges for this task. We fed GPT-4o with the same instructions given to the expert linguists, adjusting them slightly to include a final sentence to guide answer formatting: *Answer only with "Totally Correct", "Correct among more options", "Partially Correct", "Totally Wrong" or "Answer not given". Do not produce an explanation of the implicit content and do not add the reasons behind your choice.* Results of this automatic evaluation are reported in Figure 6. Results reveal that *totally correct* answers constitute only about one-forth of the samples; this proportion increases to two-fifths when *partially correct* and *correct among options* are included. Again, performance on the task are now even further from satisfactory. Regarding Prompting techniques, we have a confirmation that the CoT strategy elicits better reasoning; in fact, answers judged as *totally correct* with CoT prompting increase from the 26% of Zero-shot and the 28% of Few-shot to 31%. However, performance does not increase overall with the CoT strategy: the proportion of *totally wrong* answers is about the same with all prompting techniques and the portion of *partially correct* answers decrease from 0.1% of both Zero-shot and Few-shot to 0.06% of CoT reasoning.

H Annotators treatment for human evaluations

All annotations cited in this work were carried out voluntarily, with no financial compensation provided to the annotators. The annotators were not informed about the specific goals of the annotation tasks nor the overall aim of our research. All annotators were native or highly proficient speakers of Italian, ensuring linguistic competence in their assessments.