Teaching LLMs to unveil tendentious implicit contents of Italian political communication

Walter Paci, Lorenzo Gregori, Alessandro Panunzi
University of Florence

{walter.paci, lorenzo.gregori, alessandro.panunzi}@unifi.it

Abstract

This paper investigates the capacity of Large Language Models (LLMs) to interpret and explain non-bona fide true implicit content within the specific context of Italian political communication. While implicit meaning is a fundamental aspect of pragmatic competence, it is frequently exploited in political discourse through presuppositions and implicatures to convey tendentious messages without overt commitment. Building on previous research that highlighted the limitations of general-purpose models in zero-shot settings, we perform instruction-based fine-tuning on two state-of-the-art open-weight models: Llama3.1 and Qwen2.5. We utilize the IMPAQTS-PID benchmark, a dataset derived from the IMPAQTS corpus, to train these models to generate natural language explanations of manipulative implicit meanings. Our experimental results, validated through manual evaluation by independent annotators, demonstrate that fine-tuning significantly improves model performance compared to previous benchmarks, and that both models perform substantially better at explaining presuppositions than implicatures.

1. Introduction

In natural language use, the meaning conveyed by an utterance rarely coincides entirely with its explicit surface structure. Communication routinely depends on what is left unsaid, making the ability to infer implicit meanings a core aspect of linguistic competence (Grice, 1975; Sperber and Wilson, 1986). Although implicit communication often enhances conciseness and efficiency, it can also function as a rhetorical strategy, enabling speakers to convey messages without overtly committing to them. This is especially apparent in the use of *non-bona fide true* implicit content, a phenomenon that is particularly widespread in political discourse (Lombardi Vallauri, 2017). Political communication frequently exploits presuppositions, implicatures, and other types of implicit contents to deliver information without asserting them (Cominetti et al., 2024). Presuppositions allow speakers to frame contested propositions as already established or mutually accepted background knowledge (Stalnaker, 1978; Beaver, 2001), while implicatures enable the indirect transmission of evaluative or strategically loaded content (Grice, 1975; Levinson, 2000).

With the advent of Large Language Models (LLMs), increasingly sophisticated language processing abilities have emerged, including aspects of pragmatic competence (Ma et al., 2025). Recent research has begun to assess LLMs' capacity to interpret implicit content (Solidjonov, 2025; Cho and mook Kim, 2024). However, most existing studies concentrate on English-language datasets derived from written texts.

In this work, we analyzed the abilities of LLMs to interpret and explain implicit contents in Italian political communication. To this aim, we used a dataset derived from the IMPAQTS corpus as a data source, and performed the fine-tuning of two models, Llama3.1-8B-Instruct and Qwen2.5-7B-Instruct. Generated explanations are manually evaluated by independent annotators; moreover, BERTScore metric is evaluated on this task.

While our experimental setup focuses on a limited number of models, our objective is not to provide an exhaustive comparison across architectures or parameter scales, but to assess whether instruction-based fine-tuning on structured pragmatic data can improve the ability of LLMs to reconstruct implicit content. As such, the models considered in this study are intended as representative instances of current open-weight instruction-tuned systems and not a comprehensive sample of the model landscape.

More broadly, this work relates to the question of how pragmatic reasoning abilities interact with model characteristics such as scale and architecture. While we adopt a controlled fine-tuning setup, our results suggest that improvements in pragmatic interpretation may depend not only on model size, but critically on the availability of structured, task-specific supervision.

2. IMPAQTS corpus

The IMPAQTS (Implicit Manipulation in Politics – Quantitatively Assessing the Tendentiousness of Speeches) project is devoted to the systematic investigation of implicit manipulation in Italian political discourse. Its primary objective is the creation of a large-scale, multimodal corpus of political

Speech Type	Speeches	Words
Parliamentary speech	561 (39.99%)	889,769 (43.11%)
Rally	283 (20.17%)	480,983 (23.30%)
Party assembly	137 (9.76%)	229,379 (11.11%)
Statement in person	231 (16.46%)	299,404 (14.51%)
Broadcast statement	164 (11.69%)	126,427 (6.13%)
New media statement	27 (1.92%)	37,971 (1.84%)
Total	1403	2,063,933

Table 1: IMPAQTS corpus size

speeches annotated for instances of tendentious or questionable content conveyed implicitly. By focusing on the manipulative potential of linguistic implicitness, the project addresses a crucial dimension of political communication: the strategic use of presuppositions, implicatures, and other non-explicit meanings to present debatable content as taken for granted.

The IMPAQTS corpus comprises approximately 1,400 monologic speeches delivered by 150 Italian politicians between 1946 and 2023, totaling around 2.5 million tokens (see Table 1). The corpus includes monologues produced directly by politicians belonging to six types: parliamentary speeches, rally speeches, party assembly speeches, statements in presence, broadcast statements, and new media statements. The corpus is diachronically structured to cover nearly the entire history of the Italian Republic, subdivided into three major political periods (1946–1972; 1972–1994; 1994–2023). These divisions correspond to significant historical and institutional turning points and allow for the investigation of long-term changes in political rhetoric and implicit strategies.

The annotation framework of IMPAQTS is grounded in the distinction between *bona fide true* and *non-bona fide true* implicit contents. Only the last ones, that are potentially manipulative, are annotated and classified in the corpus. This criterion enables a principled differentiation between implicitness used for rhetorical manipulation and implicitness serving purely pragmatic or discourse-economy functions. The IMPAQTS annotation has been manually performed by expert linguists and contains, for each*non-bona fide true* implicit content, its classification and its explanation.

3. Previous Work

This work builds on a line of research investigating if LLMs can interpret and explain implicit content in Italian political communication, using data derived from the IMPAQTS corpus.

Paci (2025) evaluated pretrained LLMs on two tasks: a binary detection task on implicit content and a binary classification task where model had to classify a content as an implicature or a presupposition. Across nine multilingual models and multiple prompting strategies, models achieved unsatisfactory results in detection and remained close to chance in classification, suggesting that models did not exhibit any relevant kind of pragmatic reasoning capabilities when confronted with naturally occurring and strategically employed pragmatic language.

Subsequently, Paci et al. (2025) introduced the IMPAQTS-PID benchmark, containing over 30k implicit passages paired with expert explanations, and tested models in both multiple-choice and open-ended explanation settings. All tested models struggled to reconstruct manipulative implicit meanings: even the best-performing system remained far below the estimated performance ceiling, and fully correct explanations were produced only in a minority of cases. Chain-of-Thought prompting yielded slightly better results but they remain far from satisfactory. When models are required to freely generate an explanation of the implicit meaning rather than selecting among alternatives, performance drops sharply and fully correct explanations are produced only in a minority of cases. This suggests that the limitation does not merely concern classification or retrieval abilities, but the reconstruction of the intended communicative act itself.

Building on these findings, the IMPOLS shared task (Gregori et al., 2026) formalized the problem as a benchmark including detection, implicature/presupposition classification, and implicature type classification (models had to choose between conventional and (conversational) particularized or generalized implicatures), highlighting the difficulty of modeling context-dependent pragmatic inference in political speech.

Beyond political discourse, a growing body of work has investigated the ability of LLMs to perform pragmatic inference, particularly in relation to conversational implicatures and related phenomena. Recent studies have shown that, while models can capture certain scalar implicatures under controlled conditions, their performance remains unstable and highly sensitive to prompting strategies and contextual framing (Hu et al., 2023; Cho and mook Kim, 2024). Earlier benchmarks and analyses have sim-

ilarly highlighted that pragmatic inference in LLMs is often shallow, with models relying on surface-level heuristics rather than robust reasoning about speaker intentions (Min et al., 2022).

The present work extends these research lines by moving from evaluating general-purpose LLMs to explicitly training models to generate explanations of manipulative implicit content: we fine-tune Llama3.1-8B-Instruct and Qwen2.5-7B-Instruct on explanation pairs derived from IMPAQTS-PID and evaluate their outputs through human annotation and automatic similarity metrics.

4. Dataset

IMPAQTS-PID[1] (Paci et al., 2025) is a benchmark derived from the IMPAQTS corpus, that focuses specifically on implicatures and presuppositions, and includes more than 30,000 annotated instances accompanied by wide contextual windows to provide sufficient information for context interpretation. The benchmark is designed as a multiple choice task: each occurrence contains 4 possible answers about the content that is implicitly conveyed by the sentence. Only one answer is correct.

For the current experiment, we used IMPAQTS-PID as a dataset, discarding the multiple choices, and focusing on implicit content explanation. The dataset has been split into train and test set, obtaining 25,000 pairs of implicit contents and their explanation to exploit for training purposes (see Table 2).

	Train	Test	Total
Implicatures	11,802	3,123	14,925
Presuppositions	13,198	3,699	16,897
Total	25,000	6,822	31,822

Table 2: Dataset numbers

4.1. Dataset example

Below we report an example of a dataset instance, containing a transcription excerpt with an implicit content (within <s> and </s> tags), the implicit type, and the implicit explanation.

> **Transcription with context**: Ognuno di noi si è tagliato lo stipendio e c'è qualcuno che non ci crede. C'è qualcuno che non ci crede, ancora. Gli porti i bonifici e ti dice: "Io non ci credo". Se sono così inetti non ci parliamo. C'è un 44% invece di <s>persone che non votano, che sono disperate e che sono disilluse</s>.
>
> *Each of us has cut our salary, and there are still people who don't believe it. There are people who still don't believe it. You show them the bank transfers and they say, "I don't believe it." If they are that inept, then we have nothing to say to each other. Meanwhile, there is a 44% of <s>people who do not vote, who are desperate and disillusioned</s>.*
>
> **Implicit type**: Implicature
>
> **Implicit content explanation**: Chi non vota è disperato e disilluso.
>
> *Those who do not vote are desperate and disillusioned.*

[1] https://github.com/WalterPaci/IMPAQTS-PID

5. Experiment

5.1. Fine-tuning

In this work, we exploited the training dataset to specialize two LLMs, Llama3.1-8B-Instruct and Qwen2.5-7B-Instruct, through fine-tuning to perform implicit content explanation.

To contextualize the performance of the specialized models, we compare them against a strong general-purpose baseline. As reference, we adopt the zero-shot explanation setting introduced in Paci et al. (2025), where GPT-4o-mini is prompted to explicitly formulate the implicit content without task-specific training. This baseline represents the behavior of a general LLM relying solely on its pre-trained pragmatic competence, allowing us to assess whether improvements derive from instruction-based specialization rather than from model scale alone.

The selection of models reflects a controlled comparison between strong open-weight instruction-tuned systems (Llama3.1-8B-Instruct and Qwen2.5-7B-Instruct) and a general-purpose closed model baseline (GPT-4o-mini). While these models differ in architecture and training regimes, the goal of this comparison is to evaluate the impact of task-specific supervision relative to general-purpose capabilities.

5.2. Evaluation method

The evaluation was conducted through the manual annotation of 50 explanations generated by each language model: Llama3.1-8B-Instruct and Qwen2.5-7B-Instruct fine-tuned on IMPAQTS-PID dataset. The annotation was carried out independently by three annotators that were asked to evaluate the correctness of the generated explanations on a 0-4 scale, considering the following score meaning:

0. missing output;
1. wrong explanation;

2. the explanation contains correct elements, but it's wrong;

3. the output contains the correct explanation plus other elements;

4. the explanation is fully correct.

5.3. Results

Table 3 reports the number of explanations generated by the two fine-tuned models, manually evaluated and grouped by score, alongside an approximate distribution for the baseline model. While the results for the fine-tuned models are computed on our annotated sample with full access to instance-level labels, the baseline distribution is reconstructed from statistics reported in prior work (Paci et al., 2025).

However, the baseline results are not directly comparable at the same level of granularity: we cannot recover instance-level annotations, annotator agreement, or fully aligned evaluation conditions. For this reason, the baseline should be interpreted as an approximate but informative reference point, rather than as a strictly matched experimental condition.

The results for the fine-tuned models are based on gold labels obtained by aggregating the scores assigned by three independent annotators. In all cases, at least two annotators agreed on the same score, allowing for straightforward label selection through majority voting.[2] A comparison of score distributions reveals a clear difference in output behavior between general-purpose and fine-tuned models. While the baseline exhibits a broad dispersion across all score categories, including intermediate scores (2–3), the fine-tuned models show a more concentrated distribution, with no instances of score 3 and a higher proportion of fully correct explanations. This is expected, given the fine-tuning strategy reported in Appendix 9.

The agreement among annotators is moderate, with a Krippendorff's α between 0.76 and 0.78 in the two models (Table 4).

The overall accuracy on the manually annotated test set is 0.54 for Qwen and 0.46 for Llama (see Table 5). For both the fine-tuned models and the baseline, these values are computed by considering only the explanations that received the highest score (4). Notably, the baseline achieves an overall accuracy of 0.22, confirming that fine-tuning substantially improves performance on this task.

Surprisingly, the generated explanations are much more valid when the implicit type is a presupposition, while a lower accuracy is observed for implicatures. This results can be interpreted in light of the different cognitive and linguistic properties of presuppositions and implicatures: presuppositions are typically associated with lexical or syntactic triggers (e.g., definite descriptions, factive verbs), which constrain the space of possible inferences and make the implicit content more directly recoverable from the surface form (Stalnaker, 1978; Beaver, 2001). In contrast, implicatures are highly context-dependent and require reconstructing the speaker's communicative intention, often relying on pragmatic reasoning about relevance, contrast, and discourse goals (Grice, 1975; Sperber and Wilson, 1986; Levinson, 2000).[3]

From a modeling perspective, this suggests that fine-tuned LLMs can internalize mappings between surface cues and implicit content when these mappings are structurally constrained, but still struggle when inference requires integrating broader discourse context and speaker intentions. This asymmetry mirrors the distinction between local, form-driven inference and global, discourse-level reasoning, highlighting a key limitation of current LLMs in pragmatic interpretation. In other words, fine-tuning appears to primarily enhance trigger-based reconstruction, while leaving intention-driven inference comparatively underdeveloped.

We also observe that the fine-tuned models consistently produce concise answers, as evidenced by the absence of outputs evaluated with score 0 (missing answer) and score 3 (correct but containing additional information), in contrast to the baseline distribution. This difference can be interpreted in light of known generation biases in general-purpose LLMs. Prior work has shown that such models tend to exhibit a verbosity bias, where longer responses are systematically preferred or produced independently of their actual informativeness (Zheng et al., 2023). Similarly, LLMs have been shown to favor additive strategies, introducing additional or loosely related content beyond what is strictly required by the task (Ji et al., 2023). In this perspective, the presence of intermediate outputs (score 3) in the baseline can be seen as a byproduct of over-generation, where partially correct explanations are embedded within broader, less focused responses. By contrast, instruction-based fine-tuning has been shown to improve alignment with task objectives and to constrain generation toward more targeted outputs (Ouyang et al., 2022), which is consistent with the more concise and focused behavior observed in our models.

To provide a more fine-grained comparison with

[2] Direct fine-grained level comparability of the baseline with our model is limited due to differences in annotation protocols: Paci et al. (2025) used a single expert evaluator to evaluate outputs.

[3] This interpretation is supported by our qualitative analysis conducted in Section 5.4.

Score	Llama3.1-8B			Qwen2.5-7B			Baseline (GPT-4o-mini)		
	Impl.	Pres.	Tot.	Impl.	Pres.	Tot.	Impl.	Pres.	Tot.
4	4	19	23	7	20	27	19	13	32
3	0	0	0	0	0	0	7	6	13
2	9	5	14	8	6	14	18	27	45
1	11	2	13	9	0	9	30	28	58
0	0	0	0	0	0	0	1	1	2

Table 3: Distribution of explanation quality scores for fine-tuned models and baseline. Baseline counts are reconstructed from Paci et al. (2025) using zero-shot results over 150 samples (75 implicatures, 75 presuppositions).

	Krippendorff's α	Kendall's τ
Llama	0.78	0.40
Qwen	0.76	0.40

Table 4: Inter-annotator agreement (Krippendorff's α) and annotators' score correlation with BERTScore (Kendall's τ)

	Llama	Qwen	Baseline
Impl.	0.17	0.29	0.25
Presupp.	0.73	0.77	0.17
Overall	0.46	0.54	0.22

Table 5: Accuracy of fine-tuned models and the baseline on explanation generation task.

the baseline, we leverage the disaggregated statistics distinguishing between implicatures and presuppositions: the baseline shows slightly higher performance on implicatures than on presuppositions, while fine-tuned models exhibit the opposite pattern, achieving substantially higher accuracy on presuppositions. This contrast suggests that task-specific supervision enables models to better exploit structural cues associated with presuppositions, while implicature interpretation remains more challenging due to its reliance on context-dependent pragmatic reasoning.

Finally, we measured BERTScore (Zhang et al., 2019) on our test set to assess whether this metric can reliably estimate explanation quality. BERTScore is a semantic similarity metric based on contextual embeddings that evaluates the alignment between generated and reference texts, and is commonly used in natural language generation as a proxy for semantic adequacy. To this aim, we computed the Kendall's τ correlation between the BERTScore and the gold labels, resulting in 0.40 for both fine-tuned models (see Table 4). This value indicates a moderate correlation with human judgments, suggesting that BERTScore alone is not sufficient to reliably select correct explanations.

5.4. Qualitative analysis

A manual inspection of the errors reveals that most failures do not originate from lexical or compositional misunderstanding, but from an incorrect reconstruction of the communicative act performed by the speaker. In particular, models tend to interpret utterances propositionally, while political discourse frequently encodes stance through presupposition, rhetorical structure, and indirect contrast. As a consequence, models systematically mis-handle (i) metadiscursive negation, (ii) generic polemical reference, and (iii) rhetorical accusations expressed through interrogative or self-praise constructions. The problem concerns pragmatic orientation rather than semantic content: models paraphrase the literal proposition instead of recovering the speaker's intended evaluative target.

One phenomenon that often occurs in both models' errors is the application of the grammatical negation to the embedded proposition instead of recognizing that the negation targets an attitude or stance toward the whole context.

> **Original**[4]**:** La considerazione non vale più per nessuno, né, per quanto ci riguarda, la nostalgia per il passato, né per quanto riguarda gli alti o i finti innovatori che assisterebbero il nostro paese.
> **IMPAQTS:** Vi sono finti innovatori che assistono il nostro paese.
> **Qwen:** In realtà gli alti e finti innovatori non assistono il Paese.
> **Llama:** In realtà gli alti e finti innovatori non assistono il nostro paese.
> '**Original:** *Consideration no longer applies to anyone, nor nostalgia for the past, nor to the so-called innovators assisting our country.*
> **IMPAQTS:** *There are fake innovators assisting the country.*
> **Qwen:** *In reality the so-called innovators do not assist the country.*

[4]We only report the relevant part of the following examples and not the whole context given to the model for space reasons.

Llama: *In reality the so-called innovators do not assist our country.*'

When the target of criticism is intentionally generic (e.g., an unspecified governing coalition), models replace it with a concrete entity, introducing hallucinated referents. In the following example, the model also states something that is not present nor relevant in the excerpt.

Original: mettiamo in guardia dal rischio di un esecutivo frutto di una nuova maggioranza tra diversi, una maggioranza improvvisata, che esiste solo in Parlamento e non nel Paese, una maggioranza che non rispecchia assolutamente la maggioranza degli elettori italiani.
IMPAQTS: Quella attuale è una maggioranza tra diversi, improvvisata, che esiste solo in Parlamento e non nel Paese, che non rispecchia assolutamente la maggioranza degli elettori italiani.
Llama: La maggioranza tra FI e i suoi alleati è improduttiva.
'**Original:** *We warn against a government based on a heterogeneous and improvised parliamentary majority not representative of voters.*
IMPAQTS: *The current majority is heterogeneous and unrepresentative.*
Llama: *The majority between FI and its allies is unproductive.*'

Models also tend to interpret self-praise constructions literally instead of contrastively; political self-attribution often presupposes a negative comparison with previous governments, like in the example that follows. Here, the speaker implies having done something previous governments failed to do, but models simply restate an explicit proposition.

Original: Rivendico con orgoglio di aver fatto quello che gli italiani avevano dimenticato si potesse fare, ovverosia mantenere una promessa elettorale.
IMPAQTS: I governi precedenti non mantengono alcuna promessa elettorale, mentre Salvini sì.
Qwen: Il M5S ha mantenuto la promessa elettorale.
Llama: Il PD ha mantenuto la promessa elettorale di abolire i decreti sicurezza.
'**Original:** *I proudly claim to have done what Italians had forgotten was possible: keeping an electoral promise.*
IMPAQTS: *Previous governments do not keep promises, while the speaker does.*
Qwen: *M5S kept the promise.*
Llama: *PD kept the promise of abolishing the security decrees.*'

Another recurrent phenomenon is the strengthening of rhetorical questions into intentional claims: in the example that follows, an accusatory rhetorical question encoding criticism through presupposition is interpreted as an explicit intentional statement.

Original: si è domandato pochi giorni fa in presenza del ministro Napolitano perché, dopo due anni di governo della sinistra, non è giunta una sola voce, un invito forte da parte del governo a denunciare a chi subisce estorsioni.
IMPAQTS: Non è giunto un solo invito da parte del Governo a denunciare chi subisce estorsioni.
Llama: Il governo non è interessato a denunciare le estorsioni.
'**Original:** *He asked why, after two years of a left-wing government, no strong invitation to report extortion has been issued.*
IMPAQTS: *No invitation to report extortion was issued.*
Llama: *The government is not interested in reporting extortion.*'

These errors consistently arise when interpretation requires identifying the illocutionary function of the utterance before deriving its propositional content. Models instead derive meaning locally and only subsequently attempt inference, which leads to distortions in argumentative discourse. This suggests that, while fine-tuning improves the mapping between linguistic form and implicit content, current models still struggle with forms of inference that require global discourse interpretation rather than local trigger-based reconstruction, in line with recent findings on the limitations of LLMs in pragmatic reasoning (Hu et al., 2023; Cho and mook Kim, 2024; Ma et al., 2025). Importantly, this limitation cannot be attributed solely to model scale: previous work has shown that larger general-purpose models in zero-shot settings still fail to reliably reconstruct implicit content in political discourse (Paci et al., 2025), and that performance in complex reasoning tasks depends critically on the structure and relevance of the input signals rather than on scale alone (Min et al., 2022). This indicates that pragmatic competence does not robustly emerge just from scaling, but depends crucially on the availability of task-relevant training signals. In this perspective, our results suggest that even relatively simple instruction-based fine-tuning on structured pragmatic data can lead to substantial improvements, highlighting the importance of data construction and task formulation alongside model size.

6. Conclusions

In this study, we investigated the ability of Large Language Models to unveil and explain manipulative implicit communication in Italian political discourse. Our experiments demonstrate that while general-purpose models previously struggled with these tasks, specialized fine-tuning provides a significant performance boost.

Fine-tuning Qwen2.5 and Llama3.1 led to accuracy scores of 0.54 and 0.46, respectively, for fully correct explanations. Moreover, there is a big difference in how models handle different pragmatic phenomena; both models were highly effective at explaining presuppositions (reaching up to 0.77 accuracy) but found implicatures considerably more difficult to reconstruct.

We found that BERTScore has only a moderate correlation with human expert judgment and cannot be reliably used on its own to identify correct explanations.

These findings, together with previous evidence on the limitations of general-purpose models in pragmatic tasks, suggest that improvements in this domain depend not only on model scale but also on the availability of structured and task-specific training data, which enables models to better capture discourse-level and intention-driven aspects of meaning.

In future work, we will extend this line of research by exploring how task-specific supervision interacts with broader model characteristics, including scale and architectural differences, in order to better understand the conditions under which pragmatic reasoning abilities can emerge in LLMs.

7. Limitations

This study focuses on a limited number of models and does not aim to provide a comprehensive comparison across architectures, parameter scales, or training paradigms. In particular, we consider two open-weight instruction-tuned models (Llama3.1-8B-Instruct and Qwen2.5-7B-Instruct) and a single closed-model baseline (GPT-4o-mini), which are intended as representative instances rather than as an exhaustive sample of current LLMs.

As a consequence, we do not systematically investigate how factors such as model architecture, parameter size, or language specialization (e.g., Italian-specific or multilingual models) influence the ability to reconstruct implicit content. Similarly, while the inclusion of a strong closed-model baseline provides a useful reference point, our analysis does not extend to a broader range of proprietary systems.

These limitations imply that our findings should be interpreted primarily as evidence of the effectiveness of instruction-based fine-tuning on structured pragmatic data. A more extensive evaluation across a wider variety of models, including larger-scale and closed-source systems, is necessary to fully assess the role of scale, architecture, and training data in pragmatic reasoning.

8. Bibliographical References

David I Beaver. 2001. *Presupposition and assertion in dynamic semantics*, volume 29. CSLI publications Stanford.

Ye-eun Cho and Seong mook Kim. 2024. Pragmatic inference of scalar implicature by llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 10–20.

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.

Lorenzo Gregori, Walter Paci, and Saccone Valentina. 2026. IMPOLS at EVALITA 2026: Overview of the IMPOLS Task. In *Proceedings of the Ninth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2026)*, Bari, Italy. CEUR.org.

H. P. Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press, New York.

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.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Stephen C Levinson. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. MIT press.

Edoardo Lombardi Vallauri. 2017. Implicits as evolved persuaders. In *Pragmemes and theories of language use*, pages 725–748. Springer.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. *arXiv preprint arXiv:2502.12378*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Walter Paci. 2025. "it's a further exercise in futility": implicit content detection and classification in italian political discourse. a pilot study. *AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses*, 2(2).

Walter Paci, Alessandro Panunzi, and Sandro Pezzelle. 2025. They want to pretend not to understand: The limits of current LLMs in interpreting implicit content of political discourse. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15569–15593, Vienna, Austria. Association for Computational Linguistics.

Dilyorjon Solidjonov. 2025. Pragmatic competence without embodiment? evaluating llm performance on implicature, presupposition, and speech acts.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and cognition*, volume 142. Harvard University Press Cambridge, MA.

Robert Stalnaker. 1978. Assertion. *Syntax and Semantics (New York Academic Press)*, 9:315–332.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

9. Appenidix

Models fine-tuning parameters

The following prompt has been used to perform instruction tuning of Llama and Qwen models.

> Il seguente estratto di un discorso politico contiene un contenuto implicito non bona fide vero tra i tag <s> e </s>. Esplicita questo contenuto implicito.
>
> *The following excerpt from a political speech contains a non-bona fide true implicit content marked between the tags <s> and </s>. Make this implicit content explicit.*

The following parameters have been used for Llama and Qwen models fine-tuning.

num_train_epochs=3

per_device_train_batch_size=2

gradient_accumulation_steps=8

learning_rate=2e-5

lr_scheduler_type="cosine"

warmup_ratio=0.1

Manual evaluation results

The following tables reports the judgments of each annotator per score. Both partial values on presuppositions and implicatures, and the overall values are reported.

Score	4	3	2	1	0
Total items					
annot 1	22	0	15	13	0
annot 2	19	0	15	16	0
annot 3	24	0	13	13	0
Implicatures					
annot 1	4	0	10	10	0
annot 2	3	0	7	14	0
annot 3	6	0	8	10	0
Presuppositions					
annot 1	18	0	5	3	0
annot 2	16	0	8	2	0
annot 3	18	0	5	3	0

Table 6: Number of explanation generated by fine-tuned Llama3.1-8B-Instruct evaluated by each annotator per score.

Score	4	3	2	1	0
Total items					
annot 1	26	0	13	11	0
annot 2	24	0	12	14	0
annot 3	26	0	15	9	0
Implicatures					
annot 1	6	0	8	10	0
annot 2	6	0	5	13	0
annot 3	7	0	8	9	0
Presuppositions					
annot 1	20	0	5	1	0
annot 2	18	0	7	1	0
annot 3	19	0	7	0	0

Table 7: Number of explanation generated by fine-tuned Qwen2.5-7B-Instruct evaluated by each annotator per score.