Evaluating the abilities of LLMs and SpeechLMs in discovering implicit contents of Italian political speeches

Lorenzo Gregori, Walter Paci, Alessandro Panunzi
University of Florence
{lorenzo.gregori, walter.paci, alessandro.panunzi}@unifi.it

Abstract

This research investigates the pragmatic competence of Large Language Models (LLMs) in interpreting implicit meanings within Italian political discourse. Using the IMPAQTS-PIDMM dataset, which is a multimodal benchmark derived from the 2.5-million-token IMPAQTS corpus, the experiment evaluates how effectively models identify tendentious content such as presuppositions and implicatures. The study compares the performance of text-only LLMs against speech-based models (SpeechLMs) that process both audio and transcriptions to determine if acoustic cues enhance understanding. The results reveal that text-only models significantly outperform multimodal variants, with Qwen2.5-72B achieving the highest global accuracy of 0.863. Surprisingly, the inclusion of audio did not improve performance, as SpeechLMs like GPT-4o-mini-audio-preview and Qwen2-Audio-7B-Instruct obtained lower accuracy scores and a higher frequency of missed answers compared to their text-only equivalents. Across all tested architectures, models generally demonstrated a superior ability to process presuppositions over implicatures.

Keywords: political speech, implicit content, speechLM, IMPAQTS

1. Introduction

In natural language, the intended meaning of an utterance is seldom restricted to its explicit surface form. Effective communication frequently relies on what remains unsaid; thus, the ability to derive implicit meanings is a fundamental component of linguistic competence (Grice, 1975; Sperber and Wilson, 1986). While implicit communication often facilitates brevity and efficiency, it can also serve as a rhetorical tool to convey a message without asserting it explicitly. This is particularly evident in the use of *non-bona fide true* implicit content, which is prevalent in political discourse (Lombardi Vallauri, 2017). The strategic utility of implicit language in politics rests on several key tactics, with presuppositions and implicatures being the most frequent in political communication (Cominetti et al., 2024). Presuppositions are employed to present disputed propositions as part of the shared background knowledge (Stalnaker, 1978; Beaver, 2001), whereas implicatures are exploited to convey evaluative or strategic content indirectly (Grice, 1975; Levinson, 2000).

The rise of Large Language Models (LLMs) has led to the emergence of advanced capabilities in language processing, including pragmatic competence (Ma et al., 2025). Several recent studies have evaluated the pragmatic abilities of LLMs, specifically regarding implicit content understanding (Solidjonov, 2025; Cho and mook Kim, 2024). However, the majority of research in this field focuses on English-language data and relies solely on discourse transcriptions. In this experiment, we evaluate the performance of LLMs in understanding implicit content within Italian political speeches. We compare text-only models with speech-based LMs to determine whether, and to what extent, the inclusion of the original audio alongside its transcription enhances performance in this task.

This work exploits a dataset derived from the IMPAQTS corpus of Italian political speeches, that contains the annotation of tendentious implicit contents (sections 2 and 3). The experiment setup and the results are detailed in sections 4 and 5.

2. IMPAQTS corpus

The *Implicit Manipulation in Politics – Quantitatively Assessing the Tendentiousness of Speeches* (IMPAQTS) project represents a comprehensive effort to investigate the manipulative potential of implicit content within Italian political language. The primary outcome of this project is a large-scale, multimodal corpus annotated for tendentious content conveyed through implicit linguistic structures. The IMPAQTS corpus[1] (Cominetti et al., 2024) consists of approximately 1,400 monologic speeches delivered by 150 Italian politicians, totaling roughly 2.5 million tokens. The dataset adopts a speaker-oriented definition of political discourse, focusing exclusively on content produced by politicians rather than broader discussions of political topics.

The collection spans nearly the entire history of the Italian Republic (1946–2023) and is subdivided into three diachronic periods based on significant historical and legislative shifts[2]:

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

[2] Data density is higher in recent decades due to the greater availability of audiovisual records.

Proceedings of The 3rd Workshop on Natural Language Processing for Political Sciences (PoliticalNLP 2026) @ LREC 2026, pages 165–170
16 May 2026, Palma, Mallorca (Spain).

1. 1946 – 1972: From the establishment of republican institutions to the end of the Fifth Legislature.
2. 1972 – 1994: Covering Legislatures VI through XI, marked by the shift toward a majoritarian electoral system.
3. 1994 – 2023: Spanning Legislatures XII to XIX, focusing on contemporary discourse.

IMPAQTS corpus contains speeches delivered in different contexts, with a vast majority (about 900k words) of parliamentary speeches (see Table 1).

Reflecting the performative nature of politics, the corpus integrates textual, audio, and video modalities: text transcriptions are aligned with the audio/video sources, making IMPAQTS a large valuable resource for multimodal analysis of political language.

3. IMPAQTS-PIDMM dataset

IMPAQTS-PIDMM dataset is a multimodal version of IMPAQTS-PID[3] (Paci et al., 2025), a large-scale benchmark derived from IMPAQTS, focused specifically on implicatures and presuppositions, comprising over 30,000 instances with contextual windows.

IMPAQTS-PID 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. The wrong options are built to be plausible, each one containing elements related to the context of the sentence. This ensures that the task is not trivial and requires a deep understanding of the occurrence content to be accomplished.

IMPAQTS-PID occurrences are enriched with a wide context window (up to 4 sentences preceding the one with the implicit content) to provide sufficient information for context interpretation.

IMPAQTS-PIDMM contains a small selection of IMPAQTS-PID occurrences (see Table 2), aligned with audio/video source. The dataset is specifically designed to test multimodal models on implicit content interpretation tasks.

3.1. Dataset example

Speech excerpt

[...] Quello che noi comunisti vogliamo oggi è che il 28 aprile ne rappresenti lo sbocco politico coerente. E questo non avverrà senza uno spostamento a sinistra, senza un rafforzamento nostro, che apra alle stesse forze politiche che sono state, e in parte sono ancora, vicine a noi la prospettiva di uno sbocco diverso dall'accordo con la Democrazia Cristiana alle condizioni che questo partito è disposto ad accettare.

[...] What we communists want today is for April 28th to represent a coherent political outcome. And this will not happen without a shift to the left, without a strengthening of our position, which offers those same political forces that have been—and in part still are—close to us the prospect of an alternative path; one different from an agreement with Christian Democracy on the terms that that party is willing to accept.

Possible implicit contents (4-choices)

- A. Il Governo ha intrapreso una politica di divisione del popolo e di fanatica esasperazione degli animi.
  The Government has undertaken a policy of dividing the people and of fanatical provocation of spirits.
- B. L'attuale sistema politico è sporco.
  The current political system is filthy.
- C. Prima era il tempo della rassegnazione.
  Before, it was the time of resignation.
- D. Attualmente le forze politiche vicine al PCI non hanno la prospettiva di uno sbocco diverso dall'accordo con la DC.
  Currently, the political forces close to the PCI (Italian Communist Party) have no prospect of an outcome other than an agreement with the DC (Christian Democracy)

Correct answer: D

4. Experiment

To test LLM on IMPAQTS-PIDMM, we needed SpeechLMs that jointly processes text and speech, leveraging acoustic and prosodic information. These models are not so common, given that the vast majority of SpeechLMs is trained on audio sources to perform speech recognition. Two models has been selected for the test: *Gpt-4o-mini-audio-preview*[4] and *Qwen2-Audio-7B-Instruct* (Chu et al., 2023). Both of them can process a mixed input with audio and text and perform acoustic processing of speech signal. In addition to this, a set of classic LLMs have been tested, including the text-only variant of the two multimodal models. Finally, four classes of language models are considered in this experiment: Aya Expanse, Llama, Qwen, and GPT. For each class, different LLM has been used, to extensively test different LLM versions and sizes.

For each element of the dataset, we provided to the models (a) an excerpt of the speech transcription containing a non bona fide true implicit content with a long left context, (b) the 4 possible answers about the implicit content explanation, and (c) the original source audio of the excerpt (for speechLM

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

[4] https://platform.openai.com/docs/models/gpt-4o-mini-audio-preview

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

	IMPAQTS-PID	IMPAQTS-PIDMM
Implicatures	14,932	539
Presuppositions	16,890	1,052
Total	31,822	1,591

Table 2: Dataset numbers

only). The models are queried with a simple zero-shot prompts[5] and we used accuracy as evaluation metric.

5. Results

Results are reported in Table 3, that contains the number of missed answers (i.e. the cases where the model did not answered with a A-D letter), the number of correct answers and the accuracy. Surprisingly, the two multimodal models performed poorly, if compared to the text-only models, highlighting a difficulty of these SpeechLMs to exploit the speech signal for this task.

Qwen2-Audio has been run twice, with text and audio, and with only text: the results show that the two runs obtained the same accuracy (about 0.61), but the variant with audio had a higher number of missed answers. Compared to Qwen2-Audio, Gpt-4o-mini-audio-preview obtained a similar number of missed answers and a lower accuracy (0.57), while the textual LLM (gpt-4o-mini) performed very well: 0.85 accuracy with no missed answers. In this case we run two different models (gpt-4o-mini and gpt-4o-mini-audio-preview), because the audio variant of GPT does not allow a text-only input. The best performance has been obtained by Qwen2.5-72B, that reached an accuracy of 0.86, with also 0 missed answers.

In general, we can observe that, despite the poor results obtained by multimodal models, implicit contents are well identified by textual models, that take into account the mere transcription: many of them reach an accuracy above 0.8. The number of parameters is an important feature for this task: bigger models have better performance. Among the classic LLM, Llama models obtained lower results, with its maximum accuracy of 0.76 (with 70B model), while Qwen models performed very well, reaching an accuracy above 0.8 even with small models (4B and 7B). GPT is the only one commercial LLM tested and it can't be properly compared with other models, given that we don't know how many parameters it has. Despite this its small version (gpt-4o-mini) reached 0.85 accuracy, that is close to the top.

Unfortunately, detailed information regarding the training data of these models is not publicly available. Consequently, it is not possible to formulate well-founded hypotheses about the factors that most strongly influence the performance gaps among different LLM families. For instance, variables such as the volume of Italian-language data included in training, as well as the extent to which political texts and public speeches are represented, may constitute relevant factors for this task.

Table 4 shows the accuracy differences between implicatures and presuppositions. In general, the models obtained a higher accuracy on presupposition, with some exceptions.

All the distractors used to build the multiple choice task are plausible answers. Anyway, some of that has the same topic of the right answer (topic constrained), while the others are unconstrained with respect to the topic. This result is obtained through topic modeling on the whole IMPAQTS-PID dataset (see (Paci et al., 2025) for further details). Table 5 shows the accuracy differences between questions with and without topic-constrained distractors. Data report a strong accuracy gap, highlighting much better performances when the topic unconstrained. This is expected, given that distractors with constrained topic have a higher semantic similarity with the right answer, making it harder to distinguish.

Finally, we analyzed the resulting discrepancies between the two speech models against the related textual models: gpt-4o-mini-audio-preview vs. gpt-

[5]Full prompts are reported in Appendix.

LLM	Missed	Correct	Accuracy
Gpt-4o-mini-audio-preview	53 (3.33%)	883	0.554
Qwen2-Audio-7B-Instruct	52 (3.27%)	937	0.589
Gpt-4o-mini	0 (0.00%)	1352	0.850
Qwen2-Audio-7B-Instruct (text only)	1 (0.06%)	964	0.606
aya-expanse-8b	0 (0.00%)	1223	0.769
aya-expanse-32b	0 (0.00%)	1305	0.820
Llama-3.1-8B-Instruct	0 (0.00%)	1003	0.630
Llama-3.2-3B-Instruct	0 (0.00%)	872	0.548
Llama-3.1-70B-Instruct-AWQ-INT4	14 (0.88%)	1203	0.756
Qwen2.5-7B-Instruct	0 (0.00%)	1320	0.830
Qwen2.5-32B-Instruct	15 (0.94%)	1284	0.807
Qwen3-4B-Instruct-2507	3 (0.19%)	1286	0.808
Qwen3-30B-A3B-Instruct-2507	0 (0.00%)	1257	0.790
Qwen2.5-72B-Instruct-AWQ	0 (0.00%)	1373	**0.863**

Table 3: Accuracy and number of missed answers of the tested LLMs on IMPAQTS-PIDMM dataset.

	Implicatures			Presuppositions		
LLM	Miss	Corr	Acc	Miss	Corr	Acc
Gpt-4o-mini-audio-preview	5 (0.93%)	273	0.506	48 (4.56%)	610	0.580
Qwen2-Audio-7B-Instruct	5 (0.93%)	290	0.538	47 (4.47%)	647	0.615
Gpt-4o-mini	0 (0.00%)	451	**0.837**	0 (0.00%)	901	0.856
Qwen2-Audio-7B-Instruct (text only)	0 (0.00%)	294	0.545	1 (0.00%)	670	0.637
aya-expanse-8b	0 (0.00%)	363	0.673	0 (0.00%)	860	0.817
aya-expanse-32b	0 (0.00%)	449	0.833	0 (0.00%)	856	0.814
Llama-3.1-8B-Instruct	0 (0.00%)	312	0.579	0 (0.00%)	691	0.657
Llama-3.2-3B-Instruct	0 (0.00%)	237	0.440	0 (0.00%)	635	0.604
Llama-3.1-70B-Instruct-AWQ-INT4	7 (1.30%)	426	0.790	7 (0.67%)	777	0.739
Qwen2.5-7B-Instruct	0 (0.00%)	415	0.770	0 (0.00%)	905	0.860
Qwen2.5-32B-Instruct	8 (1.48%)	436	0.809	7 (0.67%)	848	0.806
Qwen3-4B-Instruct-2507	2 (0.37%)	416	0.772	1 (0.10%)	870	0.827
Qwen3-30B-A3B-Instruct-2507	0 (0.00%)	423	0.785	0 (0.00%)	834	0.793
Qwen2.5-72B-Instruct-AWQ	0 (0.00%)	447	0.829	0 (0.00%)	926	**0.880**

Table 4: Accuracy and number of missed answers of the tested LLMs on implicatures and presuppositions.

4o-mini and the two run Qwen2-Audio-7B-Instruct (with and without the audio source). Data are reported in Tables 6 and 7. For each model family, we considered the implicit contents that have been correctly classified by the speech LM and misclassified by the text-only variant (*Speech LM ok*), and vice versa (*Text LM ok*). For Qwen, we found 20.30% of implicit contents misclassified by one of the two models, equally divided by items for which the multi-modal run succeded (about 10%) and items correctly classified by the text-only run. Table 6 reports also the numbers per class (presuppositions, implicatures, topic unconstrained and topic constrained), showing a substantial homogeneity[6]: the discrepancies in presuppositions and implicatures classification are approximately the same number and similar results are obtained for topic constrained/unconstrained items.

For gpt-4o-mini, we obtained 36.46% of item misclassified by one of the two models, with most of the wrong results obtained by the multi-modal model. This result was expected and follows the accuracy results measured above. Going deeper in class details, we can see that the speech LM performs better with presuppositions, while the text LM obtained better results with implicatures. This difference could suggest that the acoustic features have a stronger impact in processing presuppositions. This statement need to be confirmed by further analysis on more performative speech LMs.

6. Conclusions

This experiment evaluated the capacity of Large Language Models (LLMs) to identify implicit con-

[6]Percentage values are computed per class.

	Topic unconstrained			Topic constrained		
LLM	**Miss**	**Corr**	**Acc**	**Miss**	**Corr**	**Acc**
Gpt-4o-mini-audio-preview	32 (3.56%)	526	0.586	21 (3.03%)	357	0.515
Qwen2-Audio-7B-Instruct	32 (3.56%)	569	0.634	20 (2.89%)	368	0.531
Gpt-4o-mini	0 (0.00%)	807	0.899	0 (0.00%)	545	**0.786**
Qwen2-Audio-7B-Instruct (text only)	1 (0.11%)	580	0.646	0 (0.00%)	384	0.554
aya-expanse-8b	0 (0.00%)	715	0.796	0 (0.00%)	508	0.733
aya-expanse-32b	0 (0.00%)	794	0.884	0 (0.00%)	511	0.737
Llama-3.1-8B-Instruct	0 (0.00%)	585	0.651	0 (0.00%)	418	0.603
Llama-3.2-3B-Instruct	0 (0.00%)	517	0.576	0 (0.00%)	355	0.512
Llama-3.1-70B-Instruct-AWQ-INT4	6 (0.67%)	748	0.833	8 (1.15%)	455	0.657
Qwen2.5-7B-Instruct	0 (0.00%)	804	0.895	0 (0.00%)	516	0.745
Qwen2.5-32B-Instruct	8 (0.89%)	777	0.865	7 (1.01%)	507	0.732
Qwen3-4B-Instruct-2507	1 (0.11%)	774	0.862	2 (0.29%)	512	0.739
Qwen3-30B-A3B-Instruct-2507	0 (0.00%)	762	0.849	0 (0.00%)	495	0.714
Qwen2.5-72B-Instruct-AWQ	0 (0.00%)	831	**0.925**	0 (0.00%)	542	0.782

Table 5: Accuracy and number of missed answers of the tested LLMs on elements with topic-constrained and unconstrained distractors.

Qwen2-Audio	**Speech LM ok**	**Text LM ok**	**Total**
Presuppositions	112 (10.65%)	104 (9.89%)	216 (20.53%)
Implicatures	53 (9.83%)	54 (10.02%)	107 (19.85%)
Topic unconstrained	104 (11.58%)	90 (10.02%)	194 (21.60%)
Topic constrained	61 (8.80%)	68 (9.81%)	129 (18.61%)
Total discrepancies	165 (10.37%)	158 (9.93%)	323 (20.30%)

Table 6: Classification of correctness discrepancies in Qwen2 models between the speech LM and the related textual LM.

tent within Italian political discourse. A comparative analysis between textual and speech-based language models revealed that SpeechLMs achieved poor accuracy, whereas certain text-only models demonstrated high proficiency, reaching a peak global accuracy of 0.86. These findings suggest that textual transcriptions alone provide sufficient information for identifying implicit meaning in this context. Furthermore, the performance of textual models indicates that model scale is a critical factor, as larger models consistently outperformed smaller versions. Pragmatically, the models generally processed presuppositions more effectively than implicatures; moreover, acoustic features seem to play a role in the processing of presupposition. Finally, the difficulty of the task increased significantly when distractors were topic-constrained, sharing the same subject matter as the correct answer, which made the implicit content harder to distinguish.

7. 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.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

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.

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.

gpt-4o-mini	Speech LM ok	Text LM ok	Total
Presuppositions	59 (3.71%)	306 (29.09%)	365 (34.70%)
Implicatures	20 (1.90%)	195 (36.18%)	215 (39.89%)
Topic unconstrained	34 (6.31%)	285 (31.74%)	319 (35.52%)
Topic constrained	45 (5.01%)	216 (31.17%)	261 (37.66%)
Total discrepancies	79 (4.97%)	501 (31.49%)	580 (36.46%)

Table 7: Classification of correctness discrepancies in gpt-4o-mini models between the speech LM and the related textual LM.

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*.

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.

8. Appenidix

Prompt used for speech models

The following Italian prompt has been used to test speech models.

> Analizza il discorso.
>
> Quale tra le seguenti opzioni credi che meglio esprima il suo contenuto implicito? Nota che questo compare nella parte finale del discorso. Rispondi solo con "A", "B", "C" o "D".
>
> *OPTIONS*

English translation (not used in the experiment).

> *Analyze the speech.*
>
> *Which of the following options do you think best expresses its implicit content? Note that this appears in the final part of the speech. Answer only with "A", "B", "C", or "D".*
>
> *OPTIONS*

Prompt used for text-only models

The following Italian prompt has been used to test textual models.

> Considera il seguente testo:
>
> *TEXT*
>
> 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".
>
> *OPTIONS*

English translation (not used in the experiment).

> *Consider the following text:*
>
> *TEXT*
>
> *Which of the following options do you think best expresses its implicit content? Note that this appears in the rightmost sentence of the text. Respond only with "A", "B", "C", or "D".*
>
> *OPTIONS*