Steering Pragmatic Interpretation in LLMs: A Diagnostic Evaluation of Few-Shot and Reasoning-Based Prompting for Indirect Speech Acts

Massimiliano Orsini[1], Dominique Brunato[2]
[1]University of Padua, Padua, Italy
[2] Istituto di Linguistica Computazionale "A. Zampolli" (CNR-ILC), ItaliaNLP Lab, Pisa, Italy
massimiliano.orsini10@gmail.com, dominique.brunato@ilc.cnr.it

Abstract

Pragmatic competence presents a persistent challenge for Large Language Models (LLMs), as it requires context-dependent inference beyond literal meaning. This study examines whether few-shot prompting can reliably steer LLMs toward appropriate interpretations of indirect speech acts under small-data conditions. Focusing on Italian, we evaluate three LLMs on a small dataset that captures pragmatic ambiguity through graded plausibility judgments. We compare a zero-shot baseline with multiple few-shot prompting configurations that vary in the number and composition of demonstrations, as well as in the presence of explicit pragmatic guidance. Results show that few-shot prompting does not yield robust or monotonic improvements overall. While performance improves substantially for conventionalized indirect speech acts, gains for non-conventionalized indirect speech acts are unstable and limited. In contrast, introducing explicit pragmatic reasoning along with demonstrations through guided chain-of-thought prompting appears more promising. Overall, these findings highlight the limits of example-based steering for pragmatic inference and suggest that explicitly modeling pragmatic reasoning may be a more effective direction in small-data settings.

Keywords: Indirect Speech Acts, Few-shot Prompting, Large Language Models Evaluation

1. Introduction

Pragmatic competence is a fundamental component of human communication. Beyond the literal meaning of sentences, speakers routinely convey intentions that depend on contextual inference, shared knowledge, and socially grounded expectations. Modeling such phenomena remains a major challenge for natural language processing systems, as it requires sensitivity not only to lexical and syntactic information, but also to discourse context and inferential reasoning.

Recent advances in large language models (LLMs) have led to substantial improvements across a wide range of linguistic tasks. Nevertheless, pragmatic phenomena remain especially hard, precisely because they rely on context-dependent and socially embedded knowledge rather than stable form–meaning mappings (Ma et al., 2025). In addition to modeling challenges, the study of pragmatic competence faces a methodological bottleneck. Constructing reliable resources for training and evaluation is non-trivial, and datasets targeting pragmatics are often limited in size or focused on specific, easily operationalizable phenomena. Moreover, evaluation is complicated by human label variation, which is particularly pronounced in pragmatic tasks (Jiang and Marneffe, 2022). These factors make it difficult to assess whether model behavior reflects genuine pragmatic understanding or superficial pattern matching.

Among the various pragmatic phenomena, indirect speech acts (ISAs) constitute a particularly well-studied and theoretically grounded case. Indirect speech acts are utterances whose intended communicative function diverges from their literal form (Searle, 1979). Correctly interpreting them requires recovering the speaker's intention by integrating linguistic form with contextual cues, rather than relying on surface meaning alone. A further crucial distinction concerns the difference between conventionalized and non-conventionalized ISAs. In conventionalized ISAs (C-ISAs), specific lexical or syntactic forms are strongly associated with particular communicative intentions (e.g. "*Can you...?*" for indirect requests) to the point that they induce a strong bias toward the indirect meaning in speakers, making it difficult to access to the literal interpretation even when the context requires it (Gibbs, 1983; Marocchini and Domaneschi, 2022). In contrast, non-conventionalized ISAs (NC-ISAs) derive their indirectness from context-specific reasoning rather than established linguistic conventions (Trott and Bergen, 2019; Bašnáková et al., 2013). Accordingly, while C-ISAs may benefit from distributional regularities learned during pretraining, NC-ISAs require more flexible contextual generalization and inference.

In our previous work, we provided empirical evidence supporting the theoretical distinction between conventionalized and non-conventionalized ISAs (Orsini and Brunato, 2025). That study focused on the construction of a small, manually curated benchmark for Italian and on the analysis of human plausibility judgments, with the goal of capturing pragmatic ambiguity across ISA

Proceedings of Learning Non-Literal Expressions with Small Data @ LREC 2026, pages 12–20
11 May 2026.

types. We also reported preliminary results on the performance of Italian LLMs, observing a clear advantage on conventionalized ISAs over non-conventionalized ones. However, model evaluation was not the primary focus of that work, leaving open questions regarding the robustness of model behavior under different evaluation settings and prompting conditions. Building on this previous work, we move beyond this initial analysis and focus on a more systematic evaluation of LLMs' pragmatic competence, specifically investigating the role of few-shot prompting as a lightweight steering mechanism.

Rather than aiming to improve performance per se, our goal is diagnostic: we ask whether few-shot prompting provides a reliable and stable way to steer pragmatic interpretation in a small-data regime, and whether any gains generalize beyond pattern-based cases.

Specifically, we address the following research questions:

- RQ1. To what extent can prompting-based interventions steer LLMs' pragmatic interpretation of ISAs beyond a zero-shot baseline?
- RQ2. Do prompting effects differ across conventionalized indirect speech acts, non-conventionalized indirect speech acts, and literal scenarios?
- RQ3. Does introducing explicit pragmatic knowledge and reasoning strategies in the prompt (e.g. via chain-of-thought instructions) lead to more stable and interpretable improvements in indirect speech act interpretation, compared to example-based prompting alone?

To answer these questions, we evaluate three Italian-capable LLMs under a range of zero-shot and few-shot prompting conditions, using both strict and relaxed accuracy measures derived from graded plausibility scores. Our results show that few-shot prompting yields non-monotonic and fragile effects, improving performance on conventionalized cases while failing to stabilize non-conventionalized interpretation and consistently reducing accuracy on literal scenarios.

2. Related Works

The automatic identification of Indirect Speech Acts (ISAs) has been explored within broader efforts to benchmark pragmatic competence in large language models. While evaluation resources for syntax and semantics are now abundant, pragmatic phenomena — and ISAs in particular — remain underrepresented in standard benchmark suites. This imbalance reflects both conceptual and methodological challenges: unlike morpho-syntactic or lexical-semantic tasks, ISA recognition requires modeling the mismatch between sentence form and communicative function, a phenomenon that is highly context-dependent and often inherently ambiguous.

Among existing resources, many attempts to scale ISA evaluation focus on relatively constrained and structurally regular subclasses of indirectness. Large-scale benchmarks such as BIG-Bench (Srivastava et al., 2023), as well as more targeted datasets including CIRCA (Louis et al., 2020) and GRICE (Zheng et al., 2021), which are part of the Pragmatic Understanding Benchmark (Sravanthi et al., 2024), typically tackle indirect speech acts through specific, recurring patterns, most notably indirect answers to polar questions or other highly conventionalized forms of indirect response. These settings allow for comparatively straightforward data generation and annotation, often supporting binary classification (direct vs. indirect) or limited interpretation spaces. While such resources have proven valuable for probing pragmatic inference at scale, their focus on restricted ISA types inevitably narrows the range of indirectness phenomena being modeled, underrepresenting not only more context-sensitive NC-ISAs but also less frequent lexical triggers of C-ISAs.

A different line of work seeks to enrich contextual information rather than expanding dataset size through structural constraints. Hu et al. (2023) propose scenario-based tasks for several pragmatic phenomena, in which models are presented with short contextual descriptions and must select the appropriate interpretation of an utterance among literal, indirect, and distractor options. This design increases contextual variability and more closely approximates natural discourse conditions, explicitly foregrounding the role of pragmatic inference. A closely related approach is adopted by Park et al. (2024), who evaluate LLMs' pragmatic abilities on a small Korean dataset targeting the four Gricean maxims, using a multiple-choice task in which models are given a context and an utterance to interpret among four possible interpretations. More recently, we introduce a resource explicitly focused on ISAs, named INDIR-IT Orsini and Brunato (2025), inspired by the scenario-based methodology of Hu et al. (2023) but narrowing its scope to ISA phenomena alone. Crucially, the dataset distinguishes between non-conventionalized ISAs (NC-ISA), which rely heavily on contextual inference, and conventionalized ISAs (C-ISA), where indirect meanings are more lexically or pragmatically routinized. This distinction allows for a finer-grained investigation of pragmatic competence, separating cases where indirectness emerges from general inferential reasoning from those supported by established communicative conventions. At the same time, the dataset

inherits the core trade-off of expert-designed resources: improved theoretical control and interpretability at the cost of limited scale.

3. Experimental Setting

3.1. The Task

The task follows the original design of the INDIR-IT [1]. Each item consists of a short everyday-life scenario involving two characters. The model is asked to evaluate the communicative intention of one of the characters. For each scenario, the model is presented with four candidate interpretations of the speaker's intended meaning: one indirect meaning, one literal meaning, two lexical distractors.

The dataset consists of three types of scenarios:

- C-scenarios: containing C-ISAs
- L-scenarios: containing C-ISAs whose literal meaning is favored
- NC-scenarios: containing NC-ISAs

Exmples of these three types of scenarios can be found in Appendix A.

Rather than selecting a single label, models are instructed to assign a plausibility score from 1 to 5 to each interpretation, where higher values indicate greater plausibility. This formulation explicitly allows for graded judgments and inherent pragmatic ambiguity, rather than enforcing categorical decisions.

3.2. Prompting Conditions

For the diagnostic purpose of our study, we compare a *zero-shot* baseline against several *few-shot prompting conditions* designed to test whether limited exposure to ISAs can steer model behavior, and if this effect is consistent across different types of ISAs. The conditions include:

- zero-shot (0-shot): task instructions only;
- C-scenarios-only, 5 and 10 shots (C5, C10).
- NC-scenarios-only, 5 and 10 shots (NC5, NC10).
- L-scenarios-only, 5 shots (L5)
- Pairs of C-scenarios and L-scenarios sharing the same utterance, 5 shots (CL).
- Mixed C-scenarios and NC-scenarios, 5 and 10 shots (M5, M10).
- Mixed enhanced (ME): This condition mirrors the M10 prompt, but includes an additional instruction explicitly warning the model that the speaker's communicative intention may diverge from the literal meaning of the utterance.
- Guided chain-of-thoughts (CoT): where models were provided with a four-step heuristic inference process and 5 demonstrations where this process is applied. The steps require the model to first identify the literal meaning of the utterance among the four options, then to detect contextual cues that may confirm or reject this interpretation, and finally to select the appropriate option once a decision has been reached. The four steps are illustrated in Appendix B.

In all few-shot conditions, demonstration examples are selected following the design principles of the original dataset. In particular, examples are chosen so as to preserve, as far as possible, the distribution of combinations between primary acts (intended meanings, e.g. request) and secondary acts (surface forms, e.g. question) observed in the resource. This choice is meant to avoid introducing systematic biases in the demonstrations and to ensure that few-shot prompting reflects the diversity of pragmatic mappings present in the data.

The ratings for each interpretation are derived from the average of multiple human annotations collected as part of the INDIR-IT release. Since the task requires discrete scores on a 1–5 scale, these averaged values are rounded to the nearest integer before being included in the prompt.

In Table 1, we show the basic instructions common to all the prompting conditions.

3.3. Models

We evaluate three large language models that differ in training data and optimization strategies:

- LLaMA-3.1 (8B-Instruct)[2]: this model is the most compact variant in the Llama 3.1 series, featuring an instruction-tuned architecture optimized for multilingual tasks and high-efficiency inference (Dubey et al., 2024).
- ANITA[3]: an Italian-specific adaptation of Llama 3, this model was developed using QLoRA for parameter-efficient fine-tuning and Direct Preference Optimization (DPO) to align its outputs with regional linguistic standards (Polignano et al., 2026).

[1] The resource is available on HuggingFace at the following link: https://huggingface.co/datasets/MaxiOr/INDIRIT

[2] HuggingFace handle: meta-llama/Llama-3.1-8B-Instruct

[3] HuggingFace handle: swap-uniba/LLaMAntino-3-ANITA-8B-Inst-DPO-ITA

Prompt instructions	English translation of the instructions
COMPITO: Leggerai delle storie brevi che descrivono una situazione ordinaria tra due personaggi: Fausto e Margherita. Ogni storia si conclude con una frase che Fausto rivolge a Margherita. Per ogni storia vengono fornite quattro possibili interpretazioni per spiegare l'intenzione comunicativa della frase di Fausto, in relazione alla situazione presentata. Ad ogni interpretazione, dovrai assegnare un punteggio da 1 a 5 in base alla sua plausibilità: (1 = non plausibile, 2 = poco plausibile, 3 = plausibile, 4 = più che plausibile, 5 = molto plausibile). Fornisci esclusivamente i punteggi finali per ciascuna interpretazione, senza spiegazioni o passaggi intermedi. --- SHOTS --- Non scrivere spiegazioni. Non scrivere testo aggiuntivo. Qualsiasi cosa fuori dal formato è un errore.	TASK: You will read short stories describing an ordinary situation between two characters: Fausto and Margherita. Each story ends with a sentence uttered by Fausto and addressed to Margherita. For each story, four possible interpretations are provided to explain the communicative intention of Fausto's sentence, in relation to the situation presented. For each interpretation, you must assign a score from 1 to 5 based on its plausibility: (1 = not plausible, 2 = slightly plausible, 3 = plausible, 4 = more than plausible, 5 = very plausible). Provide only the final scores for each interpretation, without explanations or intermediate steps. --- SHOTS --- Do not write explanations. Do not write additional text. Anything outside the required format is an error.

Table 1: Excerpt of the prompt instructions used in the experimental setup, along with their English translation.

- Minerva (7B)[4]: the instruct-tuned version of the largest entry in its specific family of LLMs and is natively trained in the Italian language (Orlando et al., 2024).

All models are accessed through the Hugging Face inference API and queried in a chat-based setting. We do not perform any parameter updates; all experiments are conducted in a prompting-only regime. No sampling strategy was employed at inference time and decoding was fully deterministic.

3.4. Evaluation Metrics

We calculated two accuracy measures[5]: **Strict accuracy**, where a prediction is counted as correct only if the target interpretation receives a strictly higher score than all other options. **Relaxed accuracy**, where a prediction is counted as correct if the target interpretation is tied for the highest score.

Relaxed accuracy is particularly appropriate for pragmatic tasks, where multiple interpretations may remain plausible even for human annotators. This metric allows us to distinguish between outright misinterpretations and cases where models recognize the intended meaning but fail to confidently separate it from alternatives.

[4]HuggingFace handle: sapienzanlp/Minerva-7B-instruct-v1.0

[5]Dataset coverage ranges from 90 items in the 10-shot conditions to 100 items in the 0-shot condition.

4. Results

Before analyzing the results, we first examined how the models adhered to the required output format. Llama3.1 consistently followed the format, whereas Anita generally did so, with only a few deviations occurring in prompts with 10-shot demonstrations. Minerva, by contrast, often produced only the interpretation it deemed correct, rather than providing all interpretations with their scores. In cases where the format was not respected, responses were still counted and evaluated if the intended answer could be inferred; otherwise, they were marked as incorrect.

As shown in Table 2, across all models, **few-shot prompting does not yield consistent improvements** over the zero-shot baseline. Both strict and relaxed accuracy exhibit **non-monotonic behavior** with respect to the number of demonstrations: in several conditions, performance improves with 5-shot prompting but stagnates or degrades when increasing to 10 shots. A notable exception is the NC10 condition, which yields comparatively better results for both ANITA and Minerva, as better shown in Figure 1.

These aggregate trends are largely reflected at the level of individual models, suggesting that few-shot effects are fragile and highly sensitive to configuration. In particular, this degradation may be partially attributed to increased prompt length, which has been shown to negatively affect model performance beyond a certain threshold (Levy et al., 2024)

Considering performance by scenario type, **in**

C	ALL	C	L	NC
0S	0.57 / 0.26	0.74 / 0.38	0.54 / 0.21	0.46 / 0.20
C5	0.51 / 0.31	0.73 / 0.56	0.39 / 0.22	0.46 / 0.23
C10	0.48 / 0.30	0.75 / 0.57	0.36 / 0.19	0.39 / 0.23
NC5	0.48 / 0.33	0.67 / 0.51	0.39 / 0.23	0.41 / 0.25
NC10	0.47 / 0.39	0.62 / 0.59	0.42 / 0.34	0.37 / 0.24
L5	0.47 / 0.30	0.61 / 0.44	0.37 / 0.23	0.43 / 0.23
CL	0.41 / 0.32	0.61 / 0.53	0.21 / 0.16	0.41 / 0.29
M5	0.48 / 0.32	0.71 / 0.52	0.36 / 0.21	0.41 / 0.25
M10	0.46 / 0.33	0.57 / 0.57	0.37 / 0.26	0.38 / 0.23
ME	0.46 / 0.34	0.67 / 0.56	0.36 / 0.27	0.40 / 0.26
CoT	0.55 / 0.39	0.69 / 0.55	0.62 / 0.43	0.40 / 0.24

Table 2: Relaxed and strict accuracy (R / S) averaged across three LLMs (LLaMA 3.1, ANITA, Minerva), broken down by prompting condition (rows) and scenario type (columns).

the C-scenarios, all models show a substantial improvement over the zero-shot baseline under most conditions. Accuracy rises sharply and approaches ceiling values, with minimal differences between strict and relaxed evaluation.

This indicates that conventionalized cases are reliably learnable from surface-level patterns and benefit from limited exposure, independently of the specific few-shot configuration.

As for the **L-scenarios, no clear or consistent trend emerges across models**. The only notable effect is observed for Llama3.1, which shows a marked improvement in both relaxed and strict accuracy for literal interpretations under the CoT prompt, without any degradation in C or NC scenarios. For the other models, performance remains highly variable across prompting conditions and does not exhibit systematic improvements.

Lastly, in the **NC-scenarios**, for relaxed accuracy, **some few-shot configurations yield small local improvements** over zero-shot performance, but these gains are inconsistent and not cumulative. Increasing the number of demonstrations does not stabilize performance, and different configurations favor different models. On the other hand, for strict accuracy, all prompts improve, although the zero-shot baseline was already very low for the models (LLama3.1: 0.3, Anita: 0.1, Minerva: 0.2).

As further illustrated in Figure 1, relaxed and strict accuracy generally follow the same qualitative trends observed in the aggregated results. LLaMA 3.1 and Minerva exhibit comparable performance across both metrics, whereas ANITA shows a more pronounced gap between relaxed and strict accuracy. This reflects a higher frequency of tied scores among candidate interpretations, suggesting that ANITA more often distributes plausibility across multiple options rather than sharply favoring a single interpretation.

Overall, no single prompting factor emerges as a reliable determinant of performance across conditions. Instead, results suggest that task difficulty plays a central role: simpler items—such as those involving more frequent lexical triggers (e.g. *Puoi V?/Can you V?*) or primary/secondary act pairings (e.g stative act as positive/negative response) consistently achieve higher accuracy rates, regardless of the prompting strategy.

Regarding the CoT prompt, one possible explanation is that it facilitates the identification of literal interpretations, which may account for the observed improvements in the L-scenario condition.

It is also plausible that the CoT formulation shifts the nature of the task, making it closer to a classification problem rather than to a NLI task. This could explain why improvements remain limited in the NC condition, where indirect and literal interpretations are not in competition but structurally dependent, as the indirect interpretation is only licensed if the literal one is contextually valid.

5. Discussion and Conclusion

This study examined the impact of few-shot prompting on the interpretation of indirect speech acts in Italian, considering conventionalized, non-conventionalized, and literal scenarios.

Concerning the extent to which prompting-based interventions can steer the models' pragmatic interpretation beyond a zero-shot baseline, the results show that few-shot prompting generally does not provide robust or systematic improvements, with performance varying substantially across configurations and often degrading as more demonstrations are added.

As for the differences across conventionalized and non-conventionalized ISAs, as well as literal meaning of conventionalized ISAs, prompting effects clearly vary: only C-ISAs appear to benefit from in context learning from all the tested conditions, and guided chain-of-thought prompting also succeeds in mitigating the bias toward indirect meanings and improves performance on literal scenarios, while also preserving strong results on conventionalized cases. However, none of the strategies produces measurable gains on non-conventionalized indirect speech acts, which remain inherently difficult to construct and reason about.

Lastly, through the ME and CoT prompting conditions, we examined whether introducing explicit pragmatic knowledge and reasoning in the prompts could enhance model performance. Our results showed that the CoT prompt emerges as promising only for Llama 3.1, the most robust model in our evaluation: by explicitly structuring the inferential process, it improves performance without negatively affecting other scenarios. Importantly, this improvement in the L-scenarios should not be underestimated, as the literal interpretation of C-ISAs has been shown to be highly challenging even for human subjects (Gibbs, 1983). This suggests

Figure 1: Models' performance (relaxed and strict accuracy) across all conditions and scenarios.

that, at least for stronger models, explicit reasoning guidance-rather than additional examples—may be key to enhancing pragmatic interpretation in large language models.

However, despite avoiding performance degradation on NC-scenarios, none of the prompting strategies leads to measurable improvements in this setting. This is particularly problematic, as non-conventionalized indirect speech acts are inherently harder to construct and therefore difficult to scale into large datasets. In this context, the ability to learn from small numbers of examples would be especially desirable. One possible direction for future work is to further enrich the guided inferential process, by expanding the number and specificity of reasoning steps and explicitly directing models to attend to fine-grained contextual cues, while simultaneously reducing the number of examples in order to mitigate degradation due to prompt length, and possibly, instructing the model to output its reasoning in order to enhance interpretability for further analysis.

6. Acknowledgements

This work has been (partially) supported by XAI-CARE - PNRR-MAD-2022-12376692 project, under the NRRP MUR program funded by the NextGenerationEU and by LLMs4EU "Large Language Models for the European Union" project, funded by the European Union through the Digital Europe Programme (DIGITAL-2024-AI-B-06-LANGUAGE - GA 101198470) under the grant agreement 101198470.

7. Bibliographical References

Jana Bašnáková, Kirsten Weber, Karl Magnus Petersson, Jos van Berkum, and Peter Hagoort. 2013. Beyond the language given: The neural correlates of inferring speaker meaning. *Cerebral Cortex*, 24(10):2572–2578.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and Zhiwei Zhao. 2024. The llama 3 herd of models.

Raymond W Gibbs. 1983. Do people always process the literal meanings of indirect requests? *Journal of experimental psychology. Learning, memory, and cognition*, 9(3):524–533.

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.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353, Bangkok, Thailand. Association for Computational Linguistics.

Annie Louis, Dan Roth, and Filip Radlinski. 2020. "I'd rather just go to bed": Understanding indirect answers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7411–7425, Online. Association for Computational Linguistics.

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. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8679–8696, Vienna, Austria. Association for Computational Linguistics.

Eleonora Marocchini and Filippo Domaneschi. 2022. "can you read my mind?" conventionalized indirect requests and theory of mind abilities. *Journal of Pragmatics*, 193:201–221.

Riccardo Orlando, Luca Moroni, Pere-Lluís Huguet Cabot, Simone Conia, Edoardo Barba, Sergio Orlandini, Giuseppe Fiameni, and Roberto Navigli. 2024. Minerva LLMs: The first family of large language models trained from scratch on Italian data. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 707–719, Pisa, Italy. CEUR Workshop Proceedings.

Massimiliano Orsini and Dominique Brunato. 2025. Direct and indirect interpretations of speech acts: evidence from human judgments and large language models. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 837–848.

Dojun Park, Jiwoo Lee, Hyeyun Jeong, Seohyun Park, and Sungeun Lee. 2024. Pragmatic competence evaluation of large language models for the korean language. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 256–266.

Marco Polignano, Pierpaolo Basile, and Giovanni Semeraro. 2026. Advanced natural-based interaction for the italian language: Llamantino-3-anita. *Scientific Reports*, 16.

John R. Searle. 1979. *Expression and Meaning: Studies in the Theory of Speech Acts*. Cambridge University Press, Cambridge.

Settaluri Lakshmi Sravanthi, Meet Doshi, Tankala Pavan Kalyan, Rudra Murthy, Pushpak Bhattacharyya, and Raj Dabre. 2024. Pub: A pragmatics understanding benchmark for assessing llms' pragmatics capabilities.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, "...", and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Sean Trott and Benjamin Bergen. 2019. Individual differences in mentalizing capacity predict indirect request comprehension. *Discourse Processes*, 56(8):675–707.

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, Online. Association for Computational Linguistics.

8. Appendices

A. Examples of Non-conventional Scenarios and Liter/Conventional Pairs

NC-scenario:

Margherita chiede a Fausto se sa se una loro conoscente sia sposata. Fausto le dice: "Le ho visto un anello al dito."

I - Fausto conferma a Margherita che la loro conoscente sia sposata.
L - Fausto informa Margherita che la loro conoscente possiede un anello.
D1 - Fausto vuole dire a Margherita che non sa se la loro conoscente sia sposata ma che di sicuro è molto ricca.

D2 - Fausto vuole dire che a lui non importa se la loro conoscente sia sposata.

English translation:

Margherita asks Fausto if he knows if an acquaintance of theirs is married. Fausto tells her: "I saw a ring on her finger."

I - Fausto confirms to Margherita that their acquaintance is married.
L - Fausto informs Margherita that their acquaintance owns a ring.
D1 - Fausto wants to tell Margherita that he doesn't know if their acquaintance is married but that she is certainly very wealthy.
D2 - Fausto wants to tell her that he doesn't care if their acquaintance is married.

C/L-pair:

C-scenario:

Margherita e Fausto stanno andando in macchina al supermercato. Fausto a un certo punto chiede a Margherita: "Puoi svoltare a destra?"

L-scenario:

Margherita e Fausto stanno andando in macchina al supermercato. Fausto, che sa che in questo periodo ci sono molte cantieri aperti in strada, a un certo punto chiede a Margherita: "Puoi svoltare a destra?"

Options:

I - Fausto vuole che Margherita svolti a destra.
L - Fausto vuole sapere se si può svoltare a destra.
D1 - Fausto vuole sapere se Margherita è in grado di usare lo sterzo.
D2 - Fausto vuole che Margherita lo riaccompagni a casa.

English translation:

C-scenario:

Margherita and Fausto are going to the supermarket by car. At one point, Fausto asks Margherita: "Can you turn right?"

L-scenario:

Margherita and Fausto are going to the supermarket by car. Fausto, who knows there are many open roadworks at the moment, asks Margherita: "Can you turn right?"

Options:

I - Fausto wants Margherita to turn right.
L - Fausto wants to know if it is possible to turn right.
D1 - Fausto wants to know if Margherita is able to use the steering wheel.
D2 - Fausto wants Margherita to take him back home.

B. Inferential Steps for the Guided CoT

Instructions:

1. Individua quale delle quattro interpretazioni corrisponde al significato letterale della frase.
2. Individua l'indizio contestuale rilevante per trovare l'intenzione finale.
3. Valuta se l'interpretazione letterale coincide anche con l'intenzione comunicativa finale di Fausto.
4. Quale interpretazione rappresenta l'intenzione finale di Fausto?

STORIA:

Mentre Margherita legge un libro sul divano, Fausto cerca di appendere un quadro al muro, ma sta avendo un po' di difficoltà. Fausto dice: "Mi piacerebbe che qualcuno mi aiutasse."

Cosa intende dire Fausto?

a) Fausto esprime a Margherita il desiderio che qualcuno venga ad aiutarlo ad appendere il quadro.
b) Fausto vuole che Margherita gli dia una mano ad appendere il quadro.
c) Fausto vuole distrarre Margherita dalla lettura.
d) Fausto avvisa Margherita che non riuscirà ad appendere il quadro a meno che qualcuno non lo aiuti.

PASSAGGI:

1. Interpretazione letterale?
a) Fausto esprime a Margherita il desiderio che qualcuno venga ad aiutarlo.
2. Indizio contestuale?
Margherita ha la possibilità di aiutare Fausto.
3. Interpretazione letterale è l'intenzione finale? No.

4. Intenzione finale? b) Fausto vuole che Margherita gli dia una mano ad appendere il quadro.

English translation:

1. Identify which of the four interpretations corresponds to the literal meaning of the sentence.
2. Identify the relevant contextual cue to determine the final intention.
3. Assess whether the literal interpretation also matches Fausto's final communicative intention.
4. Which interpretation represents Fausto's final intention?

STORY:
While Margherita is reading a book on the couch, Fausto is trying to hang a picture on the wall, but he is having some difficulty. Fausto says: "I would like someone to help me."

What does Fausto mean?
a) Fausto expresses to Margherita the desire that someone comes to help him hang the picture.
b) Fausto wants Margherita to help him hang the picture.
c) Fausto wants to distract Margherita from reading.
d) Fausto informs Margherita that he will not be able to hang the picture unless someone helps him.

STEPS:
1. Literal interpretation? a) Fausto expresses to Margherita the desire that someone helps him.
2. Contextual cue? Margherita is in a position to help Fausto.
3. Does the literal interpretation match the final intention? No.
4. Final intention? b) Fausto wants Margherita to help him hang the picture.