Direct and Indirect Interpretations of Speech Acts: Evidence from Human Judgments and Large Language Models

Massimiliano Orsini[1,*], Dominique Brunato[2]

[1]University of Padua, Padua, Italy
[2]ItaliaNLP Lab, Istituto di Linguistica Computazionale "A. Zampolli" (CNR-ILC), Pisa, Italy

Abstract

This paper introduces INDIR-IT (Indirectness for the Italian language), a linguistically informed, manually curated benchmark for evaluating large language models' (LLMs) understanding of indirect speech acts (ISAs) in Italian. By systematically contrasting conventionalized and non-conventionalized ISAs with literal interpretations, the corpus enables fine-grained assessment of pragmatic competence, an area still relatively underexplored compared to lexical and syntactic understanding. Preliminary results show that LLMs handle conventionalized ISAs relatively well, while performance on non-conventionalized ISAs remains more sensitive to model size and capacity. INDIR-IT offers a foundation for advancing research on pragmatic inference in both humans and LLMs.

Keywords

Indirectness, Speech acts, Italian benchmark, Large Language Models, Human evaluation

1. Introduction

Since Vaswani et al.'s seminal work [1], pre-trained large language models based on the transformer architecture (LLMs) have shown outstanding capabilities in understanding and generating natural language. However, these advances have also raised important concerns regarding interpretability. From a linguistic perspective, questions remain about the true nature and depth of the linguistic competence exhibited by these models [2, 3], and whether they can serve as computational evidence for usage-based theories of language [4]. In response, a growing body of research has focused on improving interpretability and systematically evaluating LLMs across diverse linguistic domains. This is often achieved through the development of standardized benchmarks, i.e. datasets paired with metrics designed to evaluate various models on specific tasks.

While substantial progress has been made in evaluating LLMs' syntactic, semantic, and general natural language understanding (NLU) abilities, pragmatic competences remains relatively underexplored despite its central role in human communication, where meaning depends on intentional language use, interactional context, and communicative effects [5]. This is due in part to the difficulty of operationalizing pragmatic phenomena, which encompass a wide range of abilities, such as resolving deixis, interpreting implicatures, understanding figurative language, adhering to conversational maxims, and deriving speaker intentions from indirect speech. These abilities are particularly relevant for designing more natural and humanlike dialogue systems.

In addition to the conceptual challenge, there is also a resource gap: most of the available resources are developed in English and often merely translated to fit another language. This practice risks neglecting language-specific pragmatic nuances and may compromise the validity and fidelity of evaluations conducted in non-English contexts.

This article intends to address both of these challenges by focusing on a central yet underrepresented pragmatic phenomenon: **indirectness**.

We outline a methodology for the construction of a dataset of indirect speech acts (ISAs) and a corresponding evaluation task in Italian. The dataset is designed with two complementary purposes: on the one hand, to measure the degree of competence of LLMs with regard to ISAs; and on the other, to provide insights into the interpretability of LLMs in processing indirectness in comparison with humans.

Contributions The contributions of this article can be briefly summarized in the following points:

- A methodology for developing a benchmark of ISAs that accounts for both their variety and degree of conventionality;
- INDIR-IT, a manually-curated Italian-language dataset and evaluation task constructed in accordance with this methodology[1];
- Preliminary results comparing human and LLM performance, providing initial insights into how current models handle ISA-related pragmatic competence.

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy
*Corresponding author.
massimiliano.orsini10@gmail.com (M. Orsini)

[1]The dataset is freely available at this link: https://huggingface.co/datasets/MaxiOr/ISA

In what follows, we first introduce key concepts from the linguistic literature on indirect speech acts and review existing NLP resources for evaluating model interpretability. We then present our novel dataset and describe the design of the associated evaluation task. Finally, we report and discuss the results of the human annotation study alongside preliminary evaluation outcomes across several LLMs.

2. Related Works

2.1. Indirect Speech Acts

Within the domain of pragmatics, the concept of speech acts is central, as they are defined as the minimal unit of communication [6]. In *How to Do Things with Words* [7], Austin makes a distinction between what is said (locution), what is intended (illocution) and the effect produced on the hearer (perlocution). This distinction is crucial for the pragmatic phenomenon known as indirectness, where the locution and the illocution of an utterance do not correspond to each other.

In Searle's framework [8], an indirect speech act is defined as the simultaneous performance of two speech acts: a primary act, which functions as the final intended meaning, and a secondary act that lends its locution to the primary act. This view, which is known as standard pragmatic view or literal force hypothesis (LFH) [9], establishes that the illocution of the secondary act, the literal force, is always functional for the retrieval process of the primary illocutionary force.

However, this literal-first processing assumption is far from universally accepted. An alternative proposal, the Direct Access View advanced by Gibbs [10], holds that listeners can often directly infer the intended meaning without fully processing the literal content, particularly when the context strongly supports a nonliteral reading. Several experimental studies support this view [11, 12, 13], especially in the case of conventionalized indirect speech acts, whose interpretation is often facilitated by lexicalized or syntactic triggers. Examples include indirect requests like "Can you V?" or indirect offers such as "Would you like to V?", which are often processed rapidly and effortlessly.

While conventionalized ISAs may often be identified via such surface cues, a large class of non-conventionalized ISAs remains highly context-dependent, as no fixed mapping exists between form and function. These acts require more complex inferential reasoning, often drawing on Theory of Mind (ToM) capacities [14, 15] and sensitivity to subtle discourse-level cues.

Importantly, despite decades of research, there is still no unified account of how indirect speech acts are processed. Competing models continue to propose differing mechanisms and processing orders, and much depends on contextual, cognitive, and conventional factors [16, 17]. This lack of consensus reflects not only the complexity of the phenomenon but also the variability observed even among human comprehenders.

Since both conventionalized and non-conventionalized ISAs play a central role in human interaction, mastering indirectness remains a major challenge for language models, which must grapple with these multiple layers of pragmatic reasoning to approach human-like communicative competence.

2.2. Pragmatics Understanding Benchmarks

Despite some criticism [18, 19, 20], benchmarks remain a central tool for evaluating the performance of (large) language models across a wide range of tasks. They offer a standardized framework to compare models' capabilities and have become an essential part of LLM development and assessment. While benchmarks for syntax, semantics, and general NLU are well developed— including recent efforts tailored to Italian [21, 22]—resources targeting pragmatic competence remain scarce, especially in languages other than English. This is particularly true for ISAs, a complex and context-dependent pragmatic phenomenon. One broad multilingual initiative that includes pragmatics-related tasks is BIG-Bench [23]. Although primarily aimed at probing the general capabilities of LLMs, it contains several tasks touching on pragmatics, including Implicature Recovery, which tests interpretation of indirect responses to polar questions (limited to binary yes/no inferences) and Intent Recognition, which evaluates models' ability to detect indirect requests.

Another recent contribution is the Pragmatic Understanding Benchmark (PUB) [24], which aggregates multiple tasks focused on different aspects of pragmatic competence, such as figurative language, presupposition, deixis, and indirectness. In PUB, three tasks specifically target indirectness, based on the CIRCA [25] and GRICE [26] datasets. CIRCA offers indirect responses to polar questions and includes both a classification task distinguishing between direct and indirect answers and an interpretation task for identifying the implied meaning. The GRICE dataset similarly focuses on indirect replies but extends the scope by including scalar implicatures.

Despite their usefulness, these datasets share several limitations. The context is minimal, often limited to a single question, which reduces the realism and ecological validity of the tasks. Additionally, the evaluation paradigm is typically binary or multiple choice, which may oversimplify the inherent ambiguity of non-conventionalized ISAs. The tasks often focus on a narrow range of ISA types, particularly indirect responses to yes/no questions,

Table 1
Examples of Scenarios included in INDIR-IT: ISA in bold, I = Indirect interpretation, L = Literal interpretation, D1–2 = Distractors.

Non-conventionalized Scenario	
Margherita non trova più il suo cellulare, così chiede a Fausto se sa dove si trova e lui le dice: **"Hai sentito lo squillo provenire dalla cucina prima?"**	
I: Fausto vuole far sapere a Margherita che il suo cellulare è in cucina. **L**: Fausto vuole sapere se Margherita ha fatto caso a un rumore proveniente dalla cucina. **D1**: Fausto intende dire che non ha la minima idea di dove si trovi il cellulare di Margherita. **D2**: Fausto vuole dire che ritiene improbabile che il cellulare sia in cucina.	
Conventionalized Scenario	**Literal Scenario**
Fausto e Margherita devono andare a mangiare fuori, ma Fausto è un po' stanco. Allora dice a Margherita: **"Puoi guidare?"**	Fausto e Margherita devono andare a mangiare fuori. Margherita però ha un po' di mal di testa, così Fausto le dice: **"Puoi guidare?"**
I: Fausto vuole che Margherita guidi per andare al ristorante. **L**: Fausto vuole assicurarsi che Margherita sia in condizioni di guidare. **D1**: Fausto vuole sapere se Margherita ha la patente. **D2**: Fausto intende dire che non ha voglia di andare a cena fuori.	**I**: Fausto vuole che Margherita guidi per andare al ristorante. **L**: Fausto vuole assicurarsi che Margherita sia in condizioni di guidare. **D1**: Fausto vuole sapere se Margherita ha la patente. **D2**: Fausto intende dire che non ha voglia di andare a cena fuori.

as these are generally easier to generate and annotate.

To address some of these limitations, Hu et al. [27] designed an indirectness understanding task embedded in short scenarios. Each item requires selecting the correct interpretation of an ISA from four options: the indirect meaning, the literal meaning, and two distractors. The task offers more variability in speech act combinations, though the dataset remains small (20 items total).

A more ambitious approach is proposed by Roque et al. [28], who suggest using ISA schemas, modeled after Winograd schemas [29]. These consist of paired contexts designed to favor either a literal or an indirect reading of the same utterance. While this method introduces richer contexts and greater variability, it remains easily scalable with minimal expert intervention only if it is applied to a limited set of ISA types.

3. Overview of INDIR-IT

3.1. Internal Partitioning

Inspired by Hu et al.'s work [27], the dataset presented in this paper consists of 100 scenarios. Each scenario includes a brief contextual description involving two characters, followed by an indirect speech act produced by one of the speakers. For each scenario, four candidate interpretations are provided: the indirect meaning, the literal meaning, and two lexical distractors, ranging from non-sequiturs to even another literal interpretation, albeit less plausible.

To investigate whether LLMs (and humans) process conventionalized and non-conventionalized ISAs differently, the dataset is split into two parts: 40 scenarios featuring non-conventionalized ISAs (NC-ISAs) and 30 pairs of conventionalized ISAs. Each pair includes the same utterance embedded in two distinct contexts: one favoring the indirect reading (C-ISAs) and one favoring the literal reading (Lit). This design, inspired in part by Roque et al. [28], allows us to probe models for context-sensitivity and bias in ISA interpretation.

In summary, the indirect interpretation is considered the target reading for both non-conventionalized and conventionalized scenarios, while the literal interpretation is expected to be preferred in literal scenarios.

Table 1 illustrates a representative example for each scenario included in the dataset[2].

3.1.1. Scenario design and coverage

In order to create a challenging and heterogeneous ISA dataset, the combinations of primary and secondary acts were designed to be as diverse as possible. However, some constraints limited this goal. First, not all primary acts can plausibly be expressed indirectly, as indirectness may conflict with their felicity conditions (e.g., declarations or promises). Second, not all secondary acts are equally suitable for every primary act, since the inferential paths required to recover the intended meaning of an ISA often follow conventionalized patterns.

[2]Appendix D provides the English translation for all the examples reported in the paper.

To address these challenges and expand coverage, scenarios were crafted to include longer contextual windows, allowing us to probe models on less frequently explored primary/secondary act pairings.

As a result, 26 distinct combinations were created for NC-ISAs, while 7 combinations were designed for C-ISAs, with indirect requests making up the majority. The difficulty of crafting different combinations for conventionalized ISAs might be due to the fact that indirectness is often adopted as a politeness strategy in order to decrease the imposing potential of such directive acts [8], and as consequence, indirect request might be those ISAs that mostly undergo conventionalization.

With regard to lexical triggers, the most represented is *'Puoi V?'*, functioning similarly to its English counterpart *'Can you?'*. However, the indirect meaning of conventionalized ISAs seems to be conveyed not only by a lexical entry but also by other factors such as modality, negation and grammatical person. This is clear by confronting *Puoi V?* and *'Posso V?'*, which conveys a different primary act, or *'Perché non V'* and *Perché V?'*, with the latter that does not trigger any conventionalized ISA at all. Since conventionality is only assumed beforehand, we cannot rule out this possibility for other forms of the same triggers that consequently are treated as trigger on their own. Each utterance in the dataset is labeled with both its primary and secondary act types: in literal scenarios, these labels are identical, as they are not supposed to convey any indirect meaning.

To clarify how these labels apply, we refer back to the examples in Table 1: in the non-conventionalized scenario, the primary act is labeled as a positive response, while the secondary act is a question, which reflects the indirect intention. In the conventionalized example, the utterance is a request (primary act) expressed through a question (secondary act). In the literal version of that scenario, both acts correspond to a question, with no indirectness involved.

The whole dataset, along with a complete list of all primary/secondary act combinations and triggers, is provided in the dataset card of the Hugging Face's repository.

3.2. Task Design

Based on the newly collected dataset, the task involves assigning a **plausibility score** ranging from 1 (not plausible) to 5 (very plausible) to each candidate interpretation of a given scenario. Rather than framing the task as a categorical classification, we opted for graded judgments in order to capture the intrinsic ambiguity of indirect speech acts, particularly in the case of NC-ISAs. In these cases, both the indirect and literal meanings may be conveyed simultaneously by the speaker, making it inappropriate to label any interpretation as definitively correct or incorrect. It is worth noting that similar caution may also apply to C-ISAs, at least until further empirical evidence confirms whether the Direct Access View systematically governs their interpretation in these contexts.

To ensure comparability between human and model evaluations, annotation instructions and model prompts were aligned as closely as possible. For models, the prompts include structural tags: COMPITO precedes the task instructions, STORIA introduces the scenario, and the question "Cosa intende dire Fausto?" ("What does Fausto mean?") follows immediately after the scenario. These tags help delineate task components while maintaining the consistency of the input. In both the prompts and human annotation interface, technical jargon is deliberately avoided. Interpretations are presented in random order and labeled with tags a, b, c, and d to prevent any biases related to order effects.

4. Human Annotation Procedure

The human annotation task was conducted with a total of 21 native Italian speakers recruited via the Prolific crowdsourcing platform[3]. To ensure annotation quality, only participants who reported Italian as their first language and who had no known language-related disorders were included. The final sample was balanced for gender (10 females and 11 males), with participants ranging in age from 21 to 63 years (mean age: 31).

To minimize the risk of participants inferring the purpose of the experiment and potentially biasing their responses, the raters were divided into three independent groups of seven annotators, with each group evaluating a different subset of the dataset.

In order to avoid exposing participants to both members of the conventionalized/literal pairs, these pairs were distributed across the sets so that each participant only saw one member of any given pair.

To limit the overall length of the task, each group was presented with a questionnaire containing 27 items. This distribution preserved the internal balance of the dataset while reducing the number of non-conventionalized scenarios included per set. Specifically, each questionnaire comprised 10 conventionalized scenarios, 10 literal scenarios, and 7 non-conventionalized scenarios, resulting in a total of 81 annotated items across the entire dataset.

4.1. Results

Results on the human annotation tasks are reported in Table 2 in terms of mean and standard deviation values for each interpretation.

Recall that in both non-conventionalized and conventionalized scenarios, the indirect interpretation was con-

[3]https://www.prolific.com/

Table 2
Results of the Human Annotation Task. Mean and standard deviation scores (in brackets) are reported for all interpretations across all conventionalized (C), literal (L) and non-conventionalized (NC) scenarios (S).

S	Ind	Lit	Dist1	Dist2
C	4.64 (0.36)	2.57 (1.10)	1.30 (0.36)	1.48 (0.43)
L	3.6 (1.17)	3.58 (1.07)	1.64 (0.74)	1.57 (0.56)
NC	4.22 (0.6)	3.33 (1.14)	1.67 (0.75)	1.59 (0.65)

sidered the target reading, while in literal scenarios the literal interpretation was expected to be preferred. Overall, human participants aligned with these expectations and exhibited clear, context-sensitive interpretive preferences across the three scenario types.

In conventionalized scenarios, the indirect interpretations received the highest ratings, consistent with expectations for conventionalized indirect speech acts. Literal interpretations in these scenarios were rated notably lower, indicating that participants were attuned to the pragmatics of the context.

In non-conventionalized scenarios, indirect readings remained the most favored, though literal interpretations showed a moderate increase in ratings, suggesting greater interpretive ambiguity when conventional cues are weaker.

In literal scenarios, participants rated both indirect and literal interpretations similarly, reflecting a balanced consideration of both meanings in contexts designed to support literal readings.

Across all scenarios, distractor interpretations consistently received low ratings, demonstrating participants' ability to identify and reject implausible alternatives.

Importantly, despite the different experimental paradigm, our findings offer additional support for the assumptions underlying Gibbs' Direct Access View of pragmatic comprehension [10]. Specifically, the consistently high ratings for indirect interpretations—even in contexts explicitly constructed to favour literal readings—suggest that comprehenders often bypass literal meanings when indirect interpretations are pragmatically accessible. This reinforces the notion that pragmatic inference does not obligatorily follow from a literal-first processing strategy, but rather may arise directly from contextual and discourse-level cues.

Additional support for this view emerges from the analysis of inter-annotator agreement, assessed using Krippendorff's α. For the entire annotated test set, we obtained a relatively moderate agreement of α = 0.642. Values are consistently higher in the conventionalized items (α 0.717) than in both the literal and the non-conventionalized ones (α 0.59 and α 0.6, respectively). Assuming lower agreement as an indication of a higher ambiguity level of an utterance, it appears that literal utterances in literal scenarios are perceived as ambiguous as indirect interpretations in non-conventionalized scenarios[4].

4.2. Qualitative Analysis

To have an in-depth understanding of the human annotation performance, we carried out a closer examination of specific scenarios that feature contrasting results. In particular, we analyzed two conventionalized/literal pairs (presented in Table 3), and two non-conventionalized scenarios (Table 4). For brevity, we report only their mean ratings. The full scenarios and associated interpretations are provided in Appendix D.3.

As mentioned in Section 3, different triggers may yield different outcomes, depending on their degree of conventionality. In the first conventionalized/literal pair in Table 3 featuring the trigger "*Perché non...?*" (Why not...?), the indirect interpretation was significantly rated higher in both scenarios. Conversely, in the second pair involving the trigger "Si può sapere...?" (Is it possible to know...?), the indirect interpretation was rated higher only in the conventionalized scenario, as expected. This asymmetry suggests that while both *Perché non...?* and *Si può sapere...?* may be considered conventionalized ISAs due to their frequent use in indirect communication, they likely differ in how strongly they activate the indirect reading across contexts.

Variation in conventionality is also evident in the non-conventionalized ISAs, depending on the inferential chain required to infer the indirect meaning, which results in different combinations of primary and secondary acts. As Searle [8] points out, the secondary act (i.e. the literal utterance of the sentence) often contains a reference to a preparatory condition of the primary act, which is considered one of the conditions that allow a speech act to be uttered felicitously. This holds for the first scenario in Table 4, where asking Margherita whether she has to work means asking for her availability to go out which can be loosely considered a preparatory condition for a subsequent proposal. Notably, this utterance may still be felicitous even if the speaker already knows the interlocutor's availability, highlighting its indirect character. In contrast, the second non-conventionalized scenario in Table 4 features a positive reply expressed through a promise that does not contain any references to a preparatory condition. We hypothesize that this is the reason why the literal interpretation received the highest mean score in this scenario.

[4]To further validate the reliability of the human annotations, Krippendorff's α was also computed separately for each of the three independent rater groups corresponding to the three questionnaires. The obtained values ranged from α = 0.485 to α = 0.754, indicating a consistent level of inter-annotator agreement across groups.

Table 3
Mean plausibility scores (1–5) assigned by annotators for conventionalized/literal pairs featuring the triggers "Perché non...?" and "Si può sapere...?". I = Indirect, L = Literal, D1/D2 = Distractors.

"Perché non...?"	**I**	**L**	**D1**	**D2**
Conventionalized	4.86	1.14	1.57	1.00
Literal	4.57	1.29	2.29	1.00
"Si può sapere...?"	**I**	**L**	**D1**	**D2**
Conventionalized	4.71	1.00	1.43	1.29
Literal	1.00	4.86	1.29	1.57

Table 4
Mean plausibility scores (1–5) assigned by annotators for two non-conventionalized scenarios. I = Indirect, L = Literal, D1/D2 = Distractors.

Scenario	**I**	**L**	**D1**	**D2**
Proposal as question	5.00	3.57	1.42	1.28
Positive reply as promise	3.14	4.86	1.14	1.00

5. Models Performance on INDIR-IT

This section presents a preliminary analysis of model performance on the INDIR-IT dataset. To this end, we evaluated three highly representative large language models, i.e. GPT-4o, Gemini 1.5 Flash, and Llama 3-8B Instruct, which differ in architecture, parameter size, and deployment setting. The primary goal here is not to exhaustively assess model performance on indirect speech acts, but rather to provide an initial demonstration of how the proposed dataset and methodology can be applied.

The models were tested in a zero-shot setting, using the same uncoupled literal/conventionalized pairs as in the human annotation task. In line with [27], zero-shot prompting was meant to assess models' implicit knowledge of indirectness as acquired during pretraining, rather than to optimize performance through fine-tuning or task-specific prompting strategies.

Figure 1 displays a general overview of the LLM models' performances, along with human reference. The detailed scores for all models are reported in Appendix B. Across scenarios, GPT-4 consistently showed the closest alignment with human preferences, particularly in identifying the most contextually appropriate interpretation.

More specifically, in conventional scenarios, all models approximated human preferences by assigning high ratings to indirect interpretations (GPT-4: M = 4.90; Gemini: M = 4.23; LLaMA: M = 4.90), with GPT-4 and LLaMA showing even stronger preferences than humans (M = 4.64). Models also gave higher scores to literal meanings (GPT-4: M = 2.87; LLaMA: M = 3.80) than humans did (M = 2.57), suggesting less sensitivity to suppressing literal readings when indirect meanings are expected.

In non-conventionalized scenarios, GPT-4 continued to strongly favor indirect interpretations (M = 4.76), more than humans (M = 4.22), while Gemini and LLaMA showed weaker alignment (Ms = 3.43 and 3.48, respectively). Literal ratings in NC scenarios were more comparable between humans and GPT-4 (3.33 vs. 3.24), but notably higher in LLaMA (M = 4.48), suggesting possible overgeneration of literal readings.

In literal scenarios, all models struggled to mirror the human balance between literal and indirect interpretations. LLaMA especially overvalued literal meanings, and GPT-4 gave similar scores to both interpretations. Distractor ratings remained low across models and humans, though LLaMA occasionally overvalued distractors.

Overall, the findings suggest that while LLMs can approximate human pragmatic reasoning, especially in highly conventional contexts, they still lack the fine-grained contextual sensitivity and interpretive flexibility exhibited by human participants.

5.1. Correlations between Humans and Models Ratings

To assess alignment between LLMs and human interpretations on INDIR-IT, we computed Pearson correlations between their ratings across the three scenarios and interpretation types for each. Table 5 presents a summary of these correlations, with an average score (AVG) reflecting overall agreement per scenario.

Among the evaluated LLMs, GPT-4 demonstrates the most robust and scenario-generalizable alignment with human interpretive preferences, particularly in contexts requiring nuanced reasoning (NC, L). Gemini exhibits moderate alignment, reliably scoring literal and distractor interpretations but falling short in indirect meaning resolution. In contrast, LLaMA demonstrates the weakest and most inconsistent agreement, especially in non-conventional scenarios.

In Table 6 we reported the results of the models on the same scenarios discussed in Section 4.2. As it can be seen, in the most challenging items, LLaMA often inverts the scores of the literal and indirect interpretations, assigning a higher score to the non-target option. Misalignment also frequently arises from disproportionately high scores assigned to distractors.

6. Discussion and Conclusion

This study introduced INDIR-IT, a novel dataset for the Italian language specifically designed to enable nuanced investigations into the processing of indirect speech acts

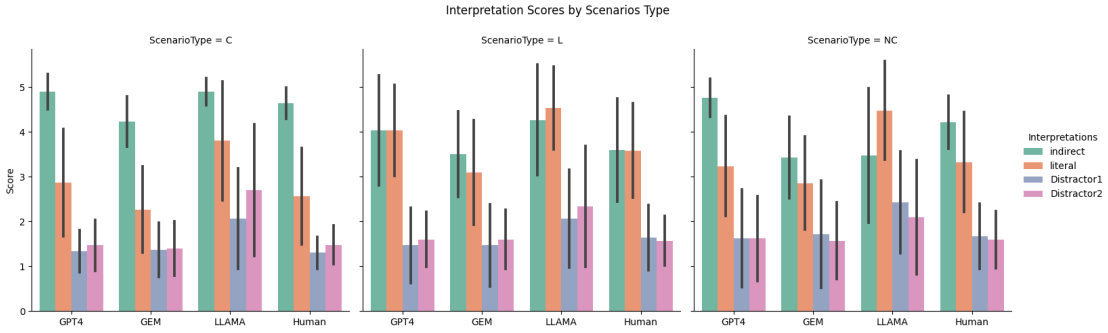

Figure 1: Model performance compared with the human annotation across each scenario type and each interpretation in terms of mean plausibility score with SD as error bars.

Table 5
Pearson correlation coefficient between human and models ratings for all interpretations across the three scenarios. Significant correlations (p value < 0.05) are bolded.

Model	S	Ind	Lit	D1	D2	AVG
GPT4	C	**0.49**	**0.78**	**0.57**	**0.45**	0.57
	L	**0.82**	**0.65**	**0.83**	**0.47**	0.70
	NC	**0.64**	**0.57**	**0.88**	**0.83**	0.73
Gemini	C	**0.40**	**0.61**	**0.62**	0.21	0.46
	L	**0.50**	**0.61**	**0.60**	**0.40**	0.53
	NC	0.29	**0.56**	**0.48**	**0.86**	0.55
Llama	C	-0.12	**0.36**	**0.51**	0.35	0.28
	L	**0.54**	**0.45**	0.55	0.28	0.46
	NC	**0.11**	-0.02	**0.65**	-0.20	0.14

Table 6
Scores assigned by the models on the scenarios discussed in the qualitative analysis (Section 4.2).

"Perché non...?"		**I**	**L**	**D1**	**D2**
C	GPT	5	1	2	1
	Gemini	4	1	3	2
	LLaMA	5	2	3	1
L	GPT	5	2	2	1
	Gemini	4	1	2	1
	LLaMA	3	5	4	1
"Si può sapere...?"		**I**	**L**	**D1**	**D2**
C	GPT	5	1	1	2
	Gemini	4	1	1	2
	LLaMA	5	4	1	5
L	GPT	1	5	1	2
	Gemini	1	5	1	2
	LLaMA	1	5	2	3
Proposal as question		**I**	**L**	**D1**	**D2**
NC	GPT	5	4	1	2
	Gemini	4	2	1	1
	LLaMA	3	5	2	1
Positive reply as Promise		**I**	**L**	**D1**	**D2**
NC	GPT	4	5	1	1
	Gemini	2	5	1	1
	LLaMA	2	5	2	3

(ISAs) by both humans and large language models (LLMs). Unlike previous benchmarks, this dataset systematically contrasts conventionalized and non-conventionalized scenarios, alongside literal interpretations, thereby providing a fine-grained tool for assessing pragmatic competence. This design makes it possible not only to evaluate overall model performance, but also to explore differences in how various forms of indirectness are handled, both by human annotators and by computational systems.

While the dataset and experimental task presented here constitute a preliminary implementation of this methodology, the results nonetheless offer several general insights into LLMs' pragmatic abilities, as well as into human performance. In terms of LLM performance, the findings consistently point to the role of model size in pragmatic competence. Larger models such as GPT-4o and Gemini Flash 1.5 display a markedly higher alignment with human judgments across all scenario types, while the smaller LLaMA 3 8B model struggles, particularly with non-conventionalized ISAs. The human annotation data also reveal meaningful patterns. As expected, indirect interpretations received higher and more consistent ratings in conventionalized scenarios, while literal and non-conventionalized scenarios elicited lower agreement levels, reflecting greater interpretive variability and ambiguity. Interestingly, this suggests that literal interpretations in literal scenarios are not necessarily fully transparent and may involve pragmatic inferencing comparable to that required for non-conventionalized ISAs. This is a finding that supports theoretical perspectives such as Gibbs' Direct Access View.

Future work will aim to refine these preliminary results by expanding both the empirical scope and the range of model evaluations. In particular, INDIR-IT provides a foundation for more systematic investigations into how LLMs handle the interface between linguistic form, context, and pragmatic inference. Moreover, this methodology can be adopted to construct comparable datasets in other languages. A partial translation of INDIR-IT may also be feasible, but only for a subset of items, as certain lexical triggers are language-specific, and some non-conventionalized ISAs require culture-specific background knowledge in order for their intended meaning to be inferred.

7. Limitations

The limitations of this work concern both dataset construction and the experimental setup.

First, the selection of primary/secondary act combinations was not guided by their real distribution in Italian, as such labeled data are currently unavailable. While INDIR-IT includes a variety of combinations, it may not fully reflect natural frequencies. Future work could address this by expanding the dataset, possibly adopting hybrid methods that combine expert annotation with corpus extraction, as fully automatic approaches are not feasible given the contextual specificity required.

Second, inter-speaker variability poses challenges, especially in pragmatics. Since the task itself invites interpretive variation, a larger pool of annotators would help mitigate individual differences in pragmatic competence.

Third, model outputs are also sensitive to sampling variability. In this study, hyperparameters such as temperature, top-k, and top-p were not controlled. While allowing some randomness is appropriate given the inherent ambiguity of the task, future studies should standardize these parameters across models to ensure replicability and comparability.

Acknowledgments

This work has been supported by the project “XAI-CARE” funded by the European Union - Next Generation EU - NRRP M6C2 “Investment 2.1 Enhancement and strengthening of biomedical research in the NHS” (PNRR-MAD-2022-12376692_VADALA’ – CUP F83C22002470001) and the project "Language Of Dreams: the relationship between sleep mentation, neurophysiology, and neurological disorders" - PRIN 2022 2022BNE97C_SH4_PRIN2022.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, arXiv (Cornell University) (2017).

[2] A. Warstadt, S. R. Bowman, What artificial neural networks can tell us about human language acquisition, in: Algebraic structures in natural language, CRC Press, 2022, pp. 17–60.

[3] M. Baroni, On the proper role of linguistically-oriented deep net analysis in linguistic theorizing, ArXiv abs/2106.08694 (2021). URL: https://api.semanticscholar.org/CorpusID:235446467.

[4] R. Futrell, K. Mahowald, How linguistics learned to stop worrying and love the language models, arXiv preprint arXiv:2501.17047 (2025).

[5] D. Crystal, The Cambridge Encyclopedia of Language, Cambridge University Press, 2010. URL: https://books.google.it/books?id=J976wAEACAAJ.

[6] J. R. Searle, What is a speech act, 1996. URL: https://api.semanticscholar.org/CorpusID:142781882.

[7] J. L. Austin, How to Do Things with Words, Clarendon Press, Oxford [Eng.], 1962.

[8] J. R. Searle, Expression and Meaning: Studies in the Theory of Speech Acts, Cambridge University Press, Cambridge, 1979.

[9] S. C. Levinson, Pragmatics / Stephen C. Levinson, Cambridge textbooks in linguistics, Cambridge university, Cambridge, 1983.

[10] R. W. Gibbs Jr, A new look at literal meaning in understanding what is said and implicated, Journal of Pragmatics 34 (2002) 457–486.

[11] R. W. Gibbs, Do people always process the literal meanings of indirect requests?, Journal of experimental psychology. Learning, memory, and cognition 9 (1983) 524–533.

[12] E. Marocchini, F. Domaneschi, “can you read my mind?” conventionalized indirect requests and theory of mind abilities, Journal of Pragmatics 193 (2022) 201–221. URL: https://www.sciencedirect.com/science/article/pii/S0378216622000819. doi:https://doi.org/10.1016/j.pragma.2022.03.011.

[13] H. H. Clark, Responding to indirect speech acts, Cognitive psychology 11 (1979) 430–477.

[14] S. Trott, B. B. and, Individual differences in mentalizing capacity predict indirect request comprehension, Discourse Processes 56 (2019) 675–707. URL: https://doi.org/10.1080/0163853X.2018.1548219. doi:10.1080/0163853X.2018.1548219.

[15] J. Bašnáková, K. Weber, K. M. Petersson, J. van Berkum, P. Hagoort, Beyond the language given: The neural correlates of inferring speaker

meaning, Cerebral Cortex 24 (2013) 2572–2578. URL: https://doi.org/10.1093/cercor/bht112. doi:10.1093/cercor/bht112.

[16] P. Brown, S. C. Levinson, Politeness: Some Universals in Language Usage, Studies in Interactional Sociolinguistics, Cambridge University Press, Cambridge, 1987.

[17] R. W. Janney, H. Arndt, 1. Intracultural tact versus intercultural tact, De Gruyter Mouton, Berlin, Boston, 1992, pp. 21–42. URL: https://doi.org/10.1515/9783110886542-004. doi:doi:10.1515/9783110886542-004.

[18] S. R. Bowman, G. Dahl, What will it take to fix benchmarking in natural language understanding?, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 4843–4855. URL: https://aclanthology.org/2021.naacl-main.385/. doi:10.18653/v1/2021.naacl-main.385.

[19] R. Aiyappa, J. An, H. Kwak, Y.-y. Ahn, Can we trust the evaluation on ChatGPT?, in: A. Ovalle, K.-W. Chang, N. Mehrabi, Y. Pruksachatkun, A. Galystan, J. Dhamala, A. Verma, T. Cao, A. Kumar, R. Gupta (Eds.), Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 47–54. URL: https://aclanthology.org/2023.trustnlp-1.5/. doi:10.18653/v1/2023.trustnlp-1.5.

[20] K. Zhou, Y. Zhu, Z. Chen, W. Chen, W. X. Zhao, X. Chen, Y. Lin, J.-R. Wen, J. Han, Don't make your llm an evaluation benchmark cheater, arXiv preprint arXiv:2311.01964 (2023).

[21] G. Attanasio, P. Basile, F. Borazio, D. Croce, M. Francis, J. Gili, E. Musacchio, M. Nissim, V. Patti, M. Rinaldi, D. Scalena, Calamita: Challenge the abilities of language models in italian, in: Italian Conference on Computational Linguistics, 2024. URL: https://api.semanticscholar.org/CorpusID:275357573.

[22] A. Seveso, D. Potertì, E. Federici, M. Mezzanzanica, F. Mercorio, et al., Italic: An italian culture-aware natural language benchmark, in: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), April 29-May 4, 2025, volume 1, 2025, pp. 1469–1478.

[23] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, A. Kluska, A. Lewkowycz, A. Agarwal, A. Power, A. Ray, A. Warstadt, A. W. Kocurek, A. Safaya, A. Tazarv, "...", Z. Wu, Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2023. URL: https://arxiv.org/abs/2206.04615. arXiv:2206.04615.

[24] S. L. Sravanthi, M. Doshi, T. P. Kalyan, R. Murthy, P. Bhattacharyya, R. Dabre, Pub: A pragmatics understanding benchmark for assessing llms' pragmatics capabilities (2024).

[25] A. Louis, D. Roth, F. Radlinski, "I'd rather just go to bed": Understanding indirect answers, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 7411–7425. URL: https://aclanthology.org/2020.emnlp-main.601. doi:10.18653/v1/2020.emnlp-main.601.

[26] Z. Zheng, S. Qiu, L. Fan, Y. Zhu, S.-C. Zhu, GRICE: A grammar-based dataset for recovering implicature and conversational rEasoning, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 2074–2085. URL: https://aclanthology.org/2021.findings-acl.182. doi:10.18653/v1/2021.findings-acl.182.

[27] J. Hu, S. Floyd, O. Jouravlev, E. Fedorenko, E. Gibson, A fine-grained comparison of pragmatic language understanding in humans and language models, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 4194–4213. URL: https://aclanthology.org/2023.acl-long.230. doi:10.18653/v1/2023.acl-long.230.

[28] A. Roque, A. Tsuetaki, V. Sarathy, M. Scheutz, Developing a corpus of indirect speech act schemas, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 220–228. URL: https://aclanthology.org/2020.lrec-1.28.

[29] H. J. Levesque, E. Davis, L. Morgenstern, The winograd schema challenge, in: Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning, KR'12, AAAI Press, 2012, p. 552–561.

A. Prompt

Below is the prompt fed to the models. In bold, the portions that are removed for the human annotation 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).
STORIA: Margherita non trova più il suo cellulare, così chiede a Fausto se sa dove si trova e lui le dice: "Hai sentito lo squillo provenire dalla cucina prima?"
Cosa intende dire Fausto?
a) Fausto vuole far sapere a Margherita che il suo cellulare è in cucina.
b) Fausto vuole sapere se Margherita ha fatto caso a un rumore proveniente dalla cucina.
c) Fausto intende dire che non ha la minima idea di dove si trovi il cellulare di Margherita.
d) Fausto vuole dire che a lui non importa se la loro conoscente sia sposata.

B. Models' Results

This section reports the models' results in terms of mean and standard deviation across each scenario and interpretation types. Row *Non-conventional 21* refers to the results obtained from the same 21 items administered to the annotators. Row *Non-conventional 40* refers to all non-conventionalized items of the dataset.

Table 7
GPT-4

Scenario type		I	L	D1	D2
Conventional	mean	4.90	2.87	1.33	1.46
	SD	0.40	1.22	0.48	0.57
Literal	mean	4.03	4.03	1.46	1.60
	SD	1.25	1.03	0.86	0.62
Non-conventional 21	mean	4.76	3.24	1.62	1.62
	SD	0.44	1.14	1.12	0.97
Non-conventional 40	mean	4.68	3.43	1.52	1.55
	SD	0.47	1.10	0.90	0.81

Table 8
Gemini 1.5 Flash

Scenario type		I	L	D1	D2
Conventional	mean	4.23	2.27	1.36	1.40
	SD	0.57	0.98	0.61	0.62
Literal	mean	3.50	3.10	1.46	1.60
	SD	0.97	1.18	0.94	0.67
Non-conventional 21	mean	3.44	2.89	1.71	1.57
	SD	0.98	1.13	1.23	0.87
Non-conventional 40	mean	3.63	2.83	1.62	1.42
	SD	0.90	0.96	1.12	0.78

Table 9
LLaMA-3 8B instruct

Scenario type		I	L	D1	D2
Conventional	mean	4.90	3.80	2.07	2.70
	SD	0.31	1.35	1.14	1.49
Literal	mean	4.27	4.53	2.07	2.33
	SD	1.26	0.94	1.11	1.37
Non-conventional 21	mean	3.39	4.39	2.43	2.09
	SD	1.61	1.20	1.16	1.30
Non-conventional 40	mean	3.80	4.48	2.47	2.05
	SD	1.49	0.99	1.26	1.17

C. Scenarios discussed in Section 4.2

C.1. *"Perché non...?"* Conventionalized/Literal Pair

CS: Margherita e Fausto stanno discutendo su cosa cucinare per cena. Fausto dice a Margherita:
LS: Margherita e Fausto stanno discutendo su cosa cucinare per cena. Fausto però era convinto che Margherita volesse fare la pizza, allora le dice:
ISA: "Perché non facciamo la pizza stasera?"
I: Fausto sta proponendo a Margherita di fare la pizza.
L: Fausto vuole capire perché non hanno più possibilità di fare la pizza.
D1: Fausto sta manifestando la sua frustrazione perché non hanno ancora preso una decisione.
D2: Fausto vuole far sapere a Margherita che lui non ha proprio voglia di pizza.

C.2. *"Si può sapere...?"* Conventionalized/Literal Pair

CS: Margherita sta cucinando, quando Fausto nota che sta per mettere lo zucchero al posto del sale nell' acqua della pasta. Fausto allora le dice:

LS: Margherita sta cucinando. Fausto sente un buon odore provenire dalla cucina, così chiede a Margherita:
ISA "Si può sapere cosa stai facendo?"
I: Fausto biasima Margherita per la sua disattenzione.
L: Fausto vuole sapere cosa stia cucinando Margherita.
D1: Fausto si lamenta perché Margherita gli tiene troppe cose nascoste.
D2: Fausto si offre per aiutare Margherita a cucinare.

C.3. Proposal as Question

NCS: Fausto vuole andare a comprarsi un nuovo vestito, ma non si fida del suo stesso gusto in abbigliamento, allora dice a Margherita:
ISA: Sei a lavoro domani mattina?'
I: Fausto vorrebbe che Margherita andasse con lui per aiutarlo nell'acquisto del vestito.
L: Fausto vuole informarsi se Margherita lavora domani.
D1: Fausto vuole che Margherita rimanga a casa domani.
D2: Fausto vuole chiedere a Margherita di comprargli un nuovo vestito.

C.4. Positive Reply as Promise

NCS: Margherita chiede a Fausto se ci sia bisogno di ritirare dei contanti dal bancomat, visto che hanno programmato di fare un viaggio a breve. Fausto le risponde:
ISA: "Ci passo io domani".
I: Fausto intende dire che pensa che ci sia bisogno di contanti.
L: Fausto promette di passare domani a ritirare dei contanti.
D1: Fausto vuole che Margherita passi a ritirare i contanti.
D2: Fausto intende dire che pensa che non ci sia bisogno di contanti.

D. English Translation of all the Examples discussed in the Paper

D.1. Prompt

TASK: You will read short stories that describe an ordinary situation between two characters: Fausto and Margherita. Each story ends with a sentence that Fausto addresses 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 will have to 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)

STORY: Margherita can't find her cell phone anymore, so she asks Fausto if he knows where it is and he tells her: Did you hear the ring coming from the kitchen earlier?'
What does Fausto mean?
a) Fausto wants to let Margherita know that her cell phone is in the kitchen.
b) Fausto wants to know if Margherita heard a noise coming from the kitchen.
C) Fausto means to say that he doesn't have the slightest idea where Margherita's cell phone is.
d) Fausto wants to say that he thinks it is unlikely that the cell phone is in the kitchen.

D.2. Conventionalized/Literal Pair Presented in Table 1

C: Fausto and Margherita have planned to go out to eat, but Fausto feels a bit tired, so he says to Margherita: "Can you drive?"
L: Fausto and Margherita have planned to go out to eat, but Margherita has a bit of a headache, so Fausto says to her: "Can you drive?"
a) Fausto wants Margherita to drive to the restaurant.
b) Fausto wants to make sure that Margherita is able to drive.
c) Fausto wants to know if Margherita has a driver's license.
d) Fausto means that he doesn't feel like going out for dinner.

D.3. Scenarios discussed in Section 4.2

"PERCHE' NON?" PAIR - PROPOSAL AS QUESTION
CS: Margherita and Fausto are discussing what to cook for dinner. Fausto says to Margherita: "Why don't we make pizza tonight?"
L: Margherita and Fausto are discussing what to cook for dinner. However, Fausto was sure that Margherita wanted to make pizza, so he says to her: "Why don't we make pizza tonight?"
I: Fausto is suggesting making pizza to Margherita
L: Fausto wants to understand why they no longer have the possibility of making pizza.
D1: Fausto is expressing his frustration because they haven't made a decision yet.
D2: Fausto wants to let Margherita know that he really doesn't feel like eating pizza.

"IS IT POSSIBLE TO KNOW" PAIR - REPROACH AS QUESTION
C: Margherita is cooking, when Fausto notices that she is about to put sugar instead of salt in the pasta water. Fausto then says to her: "Is it possible to know what you are doing?"

L: Margherita is cooking. Fausto smells a good smell coming from the kitchen, so he asks Margherita: "Is it possible to know what you are doing?"
I: Fausto blames Margherita for her carelessness.
L: Fausto wants to know what Margherita is cooking.
D1: Fausto complains because Margherita keeps too many things hidden from him.
D2: Fausto offers to help Margherita cook.

NON CONVENTIONAL - PROPOSAL AS QUESTION
Fausto wants to buy himself a new suit, but he doesn't trust his own taste in clothing, so he says to Margherita: "Are you at work tomorrow morning?"
I: Fausto would like Margherita to go with him to help him buy a new suit.
L: Fausto wants to know if Margherita is working tomorrow.
D1: Fausto wants Margherita to stay home tomorrow.
D2: Fausto wants to ask Margherita to buy him a new suit.

NON CONVENTIONAL - POSITIVE REPLY AS PROMISE
Margherita asks Fausto if they need to withdraw some cash from the ATM, given that they have planned to take a trip soon. Fausto replies to her: "I'll stop by tomorrow."
I: Fausto means that he thinks there is a need for cash.
L: Fausto promises to come by tomorrow to pick up some cash.
D1: Fausto wants Margherita to come and collect the cash.
D2: Fausto means that he thinks there is no need for cash.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT (OpenAI) in order to: Text translation and Paraphrase and reword. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.