AI-driven speech act annotation: accuracy and reproducibility across ChatGPT, LadderWeb and LLaMA

Nicola Brocca (Universität Innsbruck), *Elena Nuzzo* (Università Roma Tre) & *Joseph Wang-Kathrein* (Universität Innsbruck)

nicola.brocca(at)uibk.ac.at, elena.nuzzo(at)uniroma3.it, joseph.wang(at)uibk.ac.at

Abstract

This study evaluates three machine learning systems for annotating pragmatic categories, focusing on cancellations after accepting an invitation. The systems include the supervised model LadderWeb and the pre-trained models ChatGPT-4o and LLaMA-3.2. LadderWeb, built on Apache OpenNLP, was specifically designed for cancellation annotation. ChatGPT-4o was tested through a web interface to simulate non-expert use, while LLaMA-3.2 was run locally to ensure control, reproducibility, and data security. Both large language models were prompted using a few-shot learning approach (Brocca et al. to appear). System outputs were compared against a human baseline. GPT achieved the highest agreement across dimensions, with κ values ranging from substantial to almost perfect. LadderWeb also showed substantial agreement, whereas LLaMA performed considerably worse. Repeated testing after seven months revealed that GPT's results varied, though accuracy remained high, while LadderWeb and LLaMA produced self-consistent outputs. Notably, LLaMA improved when parameters were adjusted. These findings highlight the potential of pre-trained large language models such as ChatGPT-4o to support pragmatic corpus annotation, while also emphasizing their reproducibility challenges—an issue not observed with LadderWeb or LLaMA.

Keywords

LLM, pragmatic annotation, speech act, accuracy, reproducibility

1 Introduction

This article examines the accuracy and the consistency of a machine learning software called LadderWeb (Brocca et al. 2024), and two well-known large language models (LLMs) – ChatGPT-4o (OpenAI 2024) and LLaMA-3.2 (Meta AI 2025) – in annotating pragmatic categories.

Pragmatics has largely been "sidelined" during the "quantitative turn" in linguistics (Joseph 2008), primarily due to the non-discrete nature of its core categories (Scott-Phillips 2017). Unlike phonological or syntactic elements, there is "no simple and/or natural way to treat speaker meaning as a discrete object of enquiry" (Scott-Phillips 2017: 186). This difficulty is compounded by the challenge of operationalizing complex pragmatic notions such as conversational implicatures or presuppositions for corpus-based, quantitative research (O'Keeffe 2018). These challenges have resulted

AILing

Brocca, Nicola & Nuzzo, Elena & Wang-Kathrein, Joseph. 2026.
AI-driven speech act annotation
Special Issue: *Natural Language and AI*. Vol.3 No.1
DOI: 10.62408/ai-ling.v3i1.33

AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses ISSN: 2943-0070

in a shortage of openly available datasets, particularly for languages other than English. Pragmatically annotated corpora, such as the Switchboard Corpus (Calhoun et al. 2010) or DIADIta (De Felice and Strik Lievers 2024), tend to be relatively small in size compared to the part-of-speech (POS) tagged corpora. This limitation raises concerns about the generalizability of findings from such datasets and hinders cross-linguistic comparability.

Facilitating the annotation of pragmatic categories has been the *holy grail* in corpus-based pragmatic research for years: O'Keeffe et al. (2019: 60) noted, the "ideal scenario is to arrive at the point where we have robust pragmatic annotation tools." Similarly, Weisser (2018: 2) underscored the potential benefits of pragmatically annotated corpora, emphasizing their ability to not only enhance research on the interaction of linguistic levels but also support applied fields such as language teaching, textbook development, and the training of language professionals.

Recently, several studies have explored whether LLMs can accurately detect and annotate certain pragmatic categories when adequately prompted, yielding promising results (Bianco et al. 2025; Yu et al. 2024). However, their use also raises methodological concerns regarding both the accuracy of the annotation and its reproducibility (Brocca et al. to appear). This article evaluates the performance of different machine learning systems in annotation, considering both these aspects. Results are considered accurate when the annotations produced by the AI tools match a trusted reference, and reproducible when the same process, using identical inputs and conditions, produces the same outputs across multiple runs.

The article is organized to systematically address the key aspects of pragmatic annotation, beginning with the background in Section 2.1, which discusses the challenges and developments in the field. This section examines the theoretical and methodological obstacles encountered in creating robust frameworks for annotating pragmatic categories, as well as the progress made in overcoming these issues. It is followed by Section 2.2, which delves into technology-assisted approaches to pragmatic annotation. This section highlights the role of computational tools and machine learning technologies in advancing the efficiency and accuracy of annotation processes. In Section 3, the research questions guiding this study are outlined. The methodological framework is presented in Section 4, detailing the corpus and coding scheme employed in the study (Section 4.1). The subsequent subsections provide a thorough account of the study's technological components: Section 4.2 describes the training of LadderWeb, while Section 4.3 outlines the design of prompts for ChatGPT and LLaMA-3.2 to optimize their annotation performance. Section 4.4 concludes this section with an explanation of how the performance of the three systems was evaluated. The results of the study are presented in Section 5, quantitatively (Section 5.1) and qualitatively (Section 5.2), and discussed in Section 6, offering insights into the relative strengths and weaknesses of LadderWeb, ChatGPT-4o and LLaMA-3.2 in annotating

pragmatic categories. Finally, Section 7 concludes the article by summarizing the key outcomes, discusses their implications for corpus pragmatic research and proposes directions for future research in the field of automatic pragmatic annotation.

2 Background

2.1 Challenges in corpus pragmatics and pragmatic annotation

Most existing corpora are not well suited for pragmatics-focused research, as pragmatic functions cannot be automatically retrieved (Landert et al. 2023). Even employing a form-to-function approach –which involves identifying specific linguistic structures known or believed to express pragmatic functions– is challenging (De Felice and Strik-Lievers 2024). This is because pragma-discursive functions are typically realized through a wide variety of linguistic forms, making it unfeasible to compile an exhaustive list of relevant lexical items. Moreover, these functions often extend beyond individual words (Rühlemann and Aijmer 2015) and are highly context-dependent. For example, the sentence "we'll arrive by five o'clock" may function as a simple prediction in one context, but as a promise in another (Weisser 2015). Additionally, the categorization of certain pragmatic phenomena, such as individual speech acts, often varies across researchers, due to the lack of standardized reference frameworks, unlike those available at other levels of linguistic analysis. Therefore, annotating corpora with functional labels is necessary to enable effective analysis of pragmatic and discursive phenomena.

Due to its inherent challenges (lack of direct correspondence between form and function, and context-dependency), the annotation of pragmatic categories is still predominantly carried out manually, focusing on the specific pragmatic phenomenon being studied (Yu et al. 2024). Manual annotation allows researchers to identify all instances of a given pragmatic phenomenon, regardless of lexical variation or complexity. It is also context-sensitive, as human annotators can interpret meaning with nuance. However, manual annotation is resource-intensive, limiting its scalability and practical applicability. It requires extensive training of annotators and is susceptible to subjectivity, with inconsistencies and errors potentially arising from factors such as cognitive fatigue. To mitigate these issues, inter-coder agreement tests can be employed, though this further increases the overall workload.

All of this helps explain why, despite their growing number (e.g., Cavasso and Taboada 2021; Taylor 2016), corpora annotated with pragmatic information that support function-to-form searches remain rare (Rühlemann 2022; Weisser 2016; 2018) and are often small in size. This scarcity constrains the wider application of function-to-form approaches for analyzing pragmatic and discursive phenomena and raises concerns about the generalizability of resulting findings. Consequently, there is a clear

need for methods that can automate this annotation process. In this context, LLMs emerge as promising methodological candidates.

2.2 Technology-assisted pragmatic annotation

Recent years have seen significant advancements in the annotation of non-discrete categories in language research, driven in part by the development of semi-automatic annotation tools. These tools support both expert annotators and crowdsourced contributors, and have had a transformative impact on areas such as dialogue annotation (Weisser 2016; Zhao and Kawahara 2019), rhetorical analysis (Hamilton et al. 2024), and the annotation of relevance or stance (Gilardi et al. 2023). Traditionally, such annotation systems followed a function-to-form model (O'Keeffe 2018), relying on fully supervised learning approaches. These approaches used lexical and morphological features to predict linguistic functions with high accuracy, but required task-specific models trained on manually labeled datasets, along with extensive feature engineering to extract meaningful patterns from relatively small datasets.

In contrast, modern pre-trained language models, such as ChatGPT-4o, are built on large-scale, unsupervised training over massive corpora of raw text, enabling them to learn broad, general-purpose language representations. These models can then be fine-tuned for specific tasks or used directly via prompt-based methods (Ouyang et al. 2022), minimizing dependence on task-specific annotated data and shifting the focus to effective prompt design (Brown et al. 2020; Wei et al. 2023). This shift enhances both scalability and flexibility in addressing a wide range of natural language processing (NLP) tasks (Liu et al. 2023; Wang 2023). Importantly, pre-trained LLMs have demonstrated exceptional performance across a variety of annotation challenges. ChatGPT, in particular, has outperformed crowdsourced workers on tasks such as relevance, stance, sentiment, and rhetorical move-step annotation (Gilardi et al. 2023; Kim and Lu 2024; Zhu et al. 2023). Notably, ChatGPT-4o achieves human-level accuracy in these tasks and surpasses its predecessor, ChatGPT-3.5, in terms of precision (Ostyakova et al. 2023).

Research further highlights the potential to balance high accuracy with reduced cost when using LLMs for annotation tasks. For instance, Hamilton et al. (2024) demonstrated that annotating propaganda techniques with GPT-4 reduced annotation time by a factor of ten compared to manual human efforts. This efficiency is likely to increase further as the technology continues to evolve. Additionally, Yu et al. (2024) present a compelling case for LLM-based annotation of pragma-discursive phenomena, such as the speech act of apologizing. Their study shows that with carefully designed prompting strategies, ChatGPT-4o achieves human-like accuracy, making the annotation process more efficient, scalable, and accessible. This advancement bridges

the gap between pragmatics and other linguistic domains in corpus-based research, opening new possibilities for pragmatic inquiry.

In a similar vein, Su and Ye (2025) demonstrate that LLMs are well-suited for tackling complex annotation tasks, highlighting their potential as valuable methodological tools. Their study compares GPT-4o and DeepSeek in the annotation of thanking speech acts and concludes that both models achieve high overall accuracy. Moreover, the findings suggest the feasibility of developing a collaborative framework between humans and LLMs in speech act research, reinforcing the potential of LLMs to automate fine-grained annotation in this domain.

Overall, the studies reviewed above demonstrate that LLMs like ChatGPT hold strong potential for annotating discourse-level features such as rhetorical moves and pragmatic units. Their performance often rivals that of human annotators, highlighting their viability as methodological tools that can reduce manual labor and offer scalable solutions for complex annotation tasks in pragmatics and discourse analysis. However, the use of commercial LLMs comes with notable limitations. These include the lack of transparency around their underlying technologies, the non-replicability of results due to continuous model updates, and inconsistencies introduced by personalized outputs across different user accounts.

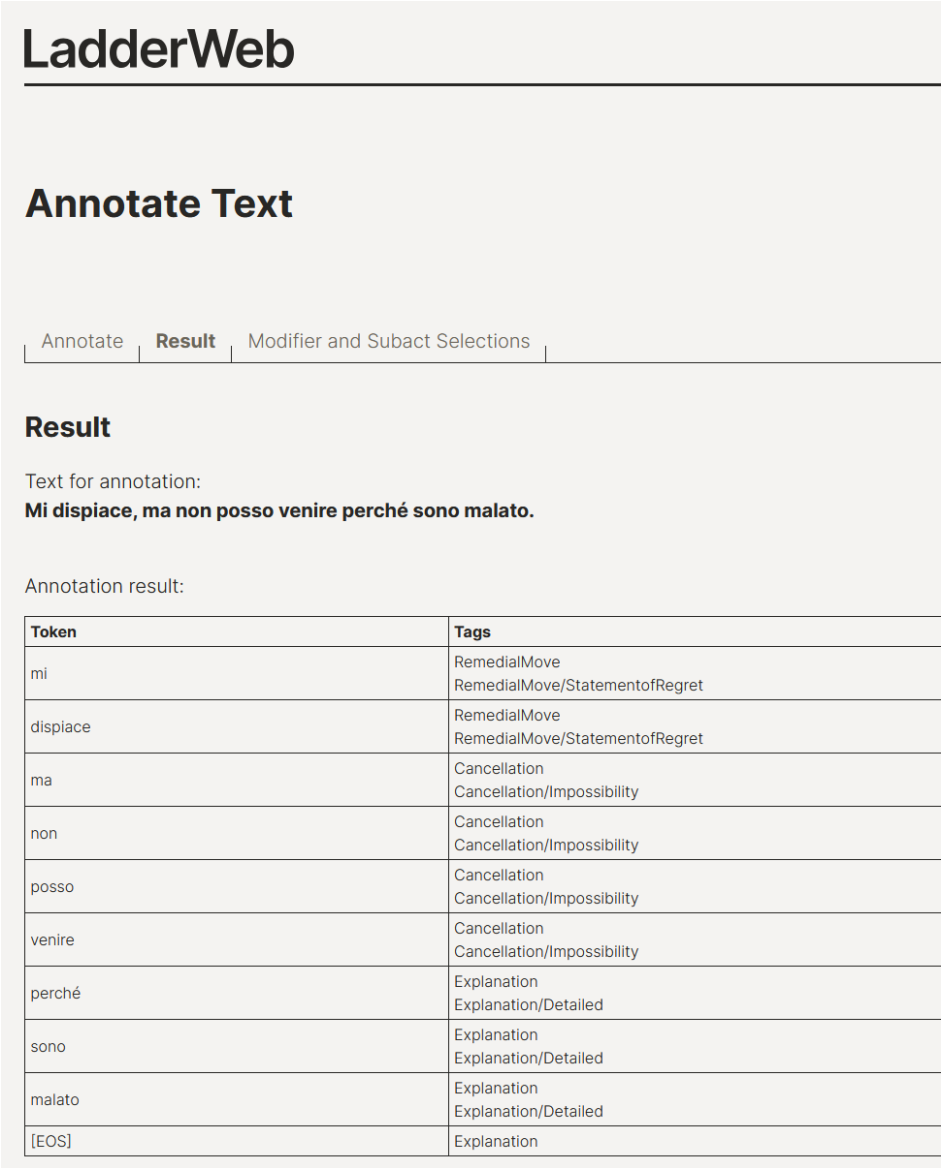
In contrast, open-source LLMs represent a significant move toward greater accessibility. Beyond lower costs, they offer key advantages over proprietary models—particularly in terms of transparency and reproducibility. Open-source models can be inspected, customized, and improved by a broad community, fostering diverse contributions that enhance both fairness and overall model quality. They also tend to provide stronger data privacy protections, as they are typically not designed to share user data with third parties. For these reasons, the academic community is increasingly advocating for the use of open-source LLMs. This shift not only broadens researchers' access to advanced tools but also supports a more open, collaborative, and reproducible research culture (Alizadeh et al. 2025).

3 Aims and research questions

Building on the promising performance of machine-learning tools in annotating pragmatic categories within the framework of speech act theory (Austin, 1970), this study aims to compare three AI technologies with differing characteristics: LadderWeb, ChatGPT-4o, and LLaMA-3.2.

LadderWeb is a web-based application developed specifically for annotating the speech act of cancellation. It is designed to support additional training sessions and allow for ongoing development, making it a flexible resource for this research. The authors began developing LadderWeb in 2022, training it with annotated data from the DISDIR corpus (refer to Sections 4.1 and 4.2). The application can be accessed through

its web interface at: https://ifd-ladderweb.uibk.ac.at/#/annotate. An example of annotation with LadderWeb can be seen in Fig. 1:



Token	Tags
mi	RemedialMove RemedialMove/StatementofRegret
dispiace	RemedialMove RemedialMove/StatementofRegret
ma	Cancellation Cancellation/Impossibility
non	Cancellation Cancellation/Impossibility
posso	Cancellation Cancellation/Impossibility
venire	Cancellation Cancellation/Impossibility
perché	Explanation Explanation/Detailed
sono	Explanation Explanation/Detailed
malato	Explanation Explanation/Detailed
[EOS]	Explanation

Figure 1: Screenshot of LadderWeb with an annotation of a cancellation.

ChatGPT-4o has been selected due to its status as one of the most advanced (with about 200 billion of parameters (Ben Abacha et al. 2024)) and widely used LLMs currently available, with extensive documentation and numerous positive reports regarding its effectiveness in annotating speech acts (Brocca et al. to appear; Yu et al. 2024; Su and Ye 2025). As a proprietary system, ChatGPT-4o offers a user-friendly interface and high performance. However, it presents certain ethical and methodological challenges,

notably the lack of transparency regarding its parameter configurations and update cycles, which may limit reproducibility and interpretability in research contexts.

LLaMA (Meta AI 2025) was selected as a well-documented, open-source LLM that can be run locally, providing full access to model weights, fostering transparency, reproducibility, and broad accessibility, and that has already shown strong performance on pragmatic annotations in Italian (Bianco et al. 2025). Version 3.2 with 3 billion parameters was chosen as it offers an optimal trade-off between capacity and memory requirements, making it feasible to run on a standard laptop (8 GB RAM). LLaMA-3.2 was employed through Ollama (Ollama 2025), an open-source platform designed to facilitate the local deployment of LLMs. Ollama enables users to download, run, and manage a variety of models directly on their own hardware, eliminating the need for reliance on cloud-based services. Ollama was paired with Open WebUI (https://github.com/open-webui/open-webui) – a browser-based graphical interface – to provide a more accessible interaction with LLaMA-3.2, engaging with the model through configurable chat environments without needing to work directly in the command line. In other words, Ollama was used as the backend engine for running the model, while WebUI was the front-end interface for interacting with it. When accessed through its WebUI, LLaMA-3.2 allows researchers to configure system-level parameters such as temperature, top-k, and top-p via a simple interface. The temperature parameter controls how predictable or varied the output is: low values lead the model to choose safer, more consistent tokens, whereas higher values allow for more variation. The top-k and top-p parameters further shape the pool of possible continuations[1]. Together, these parameters determine how many alternatives are considered and how freely the model can explore them.

The accuracy of the three tools is evaluated by addressing the following research questions:

[1] Top-k limits the model's choice to the k most probable continuations and selects one of them stochastically according to their relative probabilities. This constrains the search space and prevents the model from producing highly unlikely tokens while still allowing variation. For instance, if a model predicts the next word after "The cat sat on the" and the top three candidates are "mat", "floor" and "sofa", setting k = 3 means the model samples only from these three options. Top-p instead accumulates token probabilities until a predefined threshold p is reached and samples only from this dynamically defined set. This makes the selection sensitive to the local shape of the probability distribution and avoids excluding plausible tokens that fall outside a fixed k. For example, the model predicts the next word after "The cat sat on the" and assigns these probabilities: i. "mat" 0.40, ii. "floor" 0.25, iii. "sofa" 0.20, iv. "chair" 0.10, v. "carpet" 0.05. With a threshold of p = 0.8, the model begins summing probabilities from the highest downward. After adding "mat" (0.40), "floor" (0.25) and "sofa" (0.20), the cumulative probability reaches 0.85 and exceeds the threshold. These three words therefore constitute the sampling set. The model then selects one of them stochastically in proportion to their assigned probabilities, allowing variability while preserving contextual plausibility. Both methods introduce controlled randomness and help generate more diverse and fluent outputs than deterministic decoding (Lang 2025: 19-22).

RQ1: *How do pre-trained models and supervised learning annotation tools perform in annotating speech acts?*

This research question is addressed by comparing the annotation accuracy of the two pre-trained models, GPT-4o and LLaMA-3.2, and that of the customized tool LadderWeb. The comparison will include:
i. Quantitative analysis: Evaluating the accuracy of each tool's annotations against a gold standard created by expert annotators.
ii. Qualitative analysis: Investigating cases of discrepancy to uncover recurring patterns, contextual difficulties, or systematic errors in the annotation output of each tool.

Based on prior research (Liu et al. 2023; Wang 2023), pre-trained models are expected to achieve higher annotation accuracy due to their extensive exposure to large-scale datasets during training. In contrast, LadderWeb relies solely on manually annotated training data, which is inherently limited by the effort and scope of human annotation. However, fully harnessing the potential of pre-trained models requires the careful design and selection of prompts, a task that poses substantial methodological challenges. Moreover, the performance of pre-trained models can be negatively affected by biases or artifacts present in their training data.

When comparing the two pre-trained models, ChatGPT-4o and LLaMA-3.2, it is expected that ChatGPT-4o will demonstrate higher accuracy and more consistent adherence to the categories when appropriately prompted, given its broader training corpus, higher number of parameters, and instruction-following task alignment tuning. LLaMA-3.2, while also a highly capable model, may exhibit different strengths, such a more transparent controllability of the parameters.

RQ2: *How consistent and reliable are pre-trained models and supervised learning annotation tools?*

A consistent annotation tool must ensure that annotations yield the same results across time. This consistency may not hold for online LLMs such as ChatGPT, as the underlying model undergoes frequent and often undocumented updates. These updates may involve changes to the training data or adaptations based on user interactions, which can introduce variability in the outputs. In contrast, this issue does not affect models like LLaMA when used offline, since the model remains static and unaffected by external changes. The same applies to the annotation results produced by LadderWeb, as the tool is not adaptive and the training set is manually defined within the application. In addition, while ChatGPT accessed through the web interface automatically and opaquely selects decoding parameters, LLaMA run in a WebUI

interface allows users to configure these settings directly, ensuring greater transparency and control.

4 Data and procedures

4.1 Training data and annotation scheme

To maintain a clear focus and ensure the feasibility of the study, this research concentrates on the speech act of cancellation, intended as the act of withdrawing after having accepted an invitation. Although common in everyday interaction, cancellations remain relatively underexplored in speech act research. They are well-suited to corpus-based analysis, as they are frequently conveyed through instant messages, either written or spoken, typically in a single turn. At the same time, cancellations exhibit considerable variation in how they are expressed (see the analysed cancellations in the online supplementary material). This makes them ideal for applying a multi-layered coding scheme and offers a valuable opportunity to evaluate how well different models can implement it.

Cancellations have been examined in several languages, including through contrastive analyses (Brocca et al. 2023; Nuzzo and Cortés Velásquez 2020). This study, however, focuses exclusively on cancellations in Italian. This choice offers two key advantages: first, it allows us to build on existing annotated corpora, which serve as a valuable foundation for model training and benchmarking; second, it provides an opportunity to assess the capacity of pre-trained LLMs to handle a less-resourced language effectively (Brown et al. 2020: 14; Lai et al. 2023).

As mentioned earlier (Section 3), the development of both LadderWeb's training and GPT-4o's and LLaMA's fine-tuning relied on a pre-existing dataset of cancellation messages. This dataset is part of the DISDIR project—DIsdette e altre Strategie DI Rifiuto (translated as "Cancellations and Other Refusal Strategies")—a cross-cultural pragmatics study initiated in 2016 at Università Roma Tre, Italy (Nuzzo and Cortés Velásquez 2020; Cortés Velásquez and Nuzzo 2022). The data were collected through a questionnaire that included a mix of multiple-choice and open-ended Discourse Completion Task (DCT) items, as well as evaluative questions. The open-ended DCT was specifically designed to prompt participants to compose the kind of message they would send to cancel an invitation at the last minute after initially accepting it. The Italian responses obtained through this task make up the core of the corpus analyzed in this research.

The annotation of the dataset followed a data-driven methodology, with pragmatic categories emerging from the corpus itself, although shaped by insights from prior studies on refusals (e.g., Beebe et al. 1990). The overall annotation approach is grounded in Speech Act Theory, a framework that has long informed empirical work

in cross-cultural and interlanguage pragmatics, beginning with the CCSARP study led by Blum-Kulka et al. (1989). Within this tradition, speech acts are generally broken down into three components: the head act, internal modifiers, and supportive moves (or external modifiers).

However, as the annotation progressed, it became evident that the line between head acts and supportive moves was frequently blurred. In many cases, speakers did not explicitly state a head act; instead, the intended communicative force was conveyed through expressions typically classified as supportive moves, assuming a head act was present. To navigate this issue, some researchers have drawn a distinction between direct and indirect head act strategies. For example, in their study of refusals, Babai Shishavan and Sharifian (2016: 80) interpreted indirect strategies as functioning as head acts when a direct refusal was absent, but as supportive moves when they accompanied a direct refusal. Even so, this categorization remains interpretive and context-dependent.

Given these complexities, the decision was made to dispense with the conventional distinction between head acts and supportive moves. Instead, the analytical model adopted the concept of "sub-acts," defined as the smallest pragmatic units that together constitute a cancellation. The annotation scheme includes 12 types of sub-acts, which may be realized through different strategies and can be accompanied by four types of modifiers (details provided in the online Attachments). Emoticons and emojis were also taken into account during the annotation process and were typically treated as modifiers, with their contextual role assessed on a case-by-case basis. An Example of annotation is provided in (1):

(1)

	Sub-act	Strategy	Modifier
Mi sarebbe piaciuto molto essere presente questa sera	Willingness		*molto*= intensifier
ma ho avuto un contrattempo all'ultimo momento	Explanation	Generic	
e purtroppo non posso venire	Cancellation	Impossibility	*purtroppo*= evaluator

4.2 Training of LadderWeb

LadderWeb is a web application that manages annotations of short texts. Users can define annotation tags, annotate tokens in a text using these tags, retrieve the texts according to specific criteria, export the annotations, and train a model using Apache

OpenNLP. The trained model can then be applied in token classifiers that – given a text – annotate its tokens. As mentioned in Section 3, a supervised learning method was applied for training. A binary classifier is created for each tag and each language. During the training phase, each example is broken down into tokens, and each token is assigned either "+" or "-". If the token is marked with an annotation in the data sample, it receives "+"; otherwise it receives "-". The Apache OpenNLP library, specifically the Part of Speech (POS) Tagger (Apache Software 2025), is employed to create the binary classifiers. Apache OpenNLP provides a straightforward routine to build POS-taggers, which suits the requirements of LadderWeb. The POS-tagger first builds a statistical model from the training data. This training can be understood as constructing a function in a multi-dimensional space where each axis represents a feature of the token (for instance, the token itself, the preceding token, or the following token). The function should divide the space into two regions. Each token, having been annotated with or without a tag, is positioned in this space and marked with either "+" or "-". The aim is that the function produced during training separates the tokens marked with "+" from those marked with "-". When the trained model (i.e. the function) is used to annotate new texts, the tokens of the input text are mapped into the same multi-dimensional space. Depending on whether they fall on the "+" or "-" side of the function, the tokens are tagged or left untagged. The source code of LadderWeb is licensed under Creative Commons (CC-BY 4.0) and available on GitHub: https://github.com/kofnego-jw/ladderweb.

The LadderWeb model was trained after a dedicated annotation phase in which cancellation instances were manually added. The annotations were carried out by several non-expert raters under the supervision of two expert raters (Author 1 and Author 2). The taxonomy used is available in the online supplementary materials. Around 20% of the dataset was annotated through peer review, with non-expert raters cross-checking each other's work. These annotations were then examined and validated by one of the expert raters. For the remaining 80% of the data, cancellations were annotated individually by non-expert raters and subsequently reviewed by an expert rater to ensure reliability and accuracy. Any ambiguous cases were discussed in consensus meetings involving both expert raters. This process continued until a total of 1,100 cancellations had been annotated, after which the model was trained.

4.3 Instruction prompt for the pre-trained models

Since ChatGPT and LLaMA are pre-trained models, pragmatic annotation can be attempted using a single prompt that asks for the annotation of a target sentence. This approach, known as zero-shot prompting, is generally insufficient to ensure that the models adhere to a specific taxonomy during annotation (Brown et al. 2020). To overcome this limitation, it is necessary to include an instructional prompt before the

main annotation request. This method, referred to as a few-shot prompting (also followed by Hamilton et al. 2024; Yo et al. 2024), uses a sequence of chained example-based prompts to elicit more accurate and contextually grounded responses. As the instructional prompt for ChatGPT and LLaMA, we used a version first written by Brocca et al. (to appear) that had been tested to yield optimal results. The prompt was written in English, based on the assumption that GPT and LLaMA perform best in this language (Yin et al. 2024), while the annotation examples themselves were in Italian to match the language under investigation. During the instruction prompt, GPT and LLaMA were trained to learn the taxonomy of the speech act of cancellation (cf. 2.a), including definitions for each Sub-act, Strategy, and Modifier, accompanied by illustrative examples (2.b). The prompt also included ten in-context learning examples, that is, query–expected output pairs consisting of sample queries requesting the annotation of complete cancellation utterances together with their corresponding expected model outputs (see 2.c). These examples were carefully selected from the DISDIR corpus to ensure both representativeness and diversity. The whole script used for the fine-tuning phase is available in the online attachment at this link: https://doi.org/10.17605/OSF.IO/2HTNQ:

(2) a. Please learn the following contents. The speech act of cancellation in Italian may contain the following functional elements, also called Sub-acts: [labels and definitions from the taxonomy]
b. Example: [one example for each Sub-act, Strategy, and Modifier, as outlined in the taxonomy for cancellations]
c. My query: Annotate the following utterance: *Sono troppo stanca per uscire oggi, magari ci vediamo in settimana?* [‘I’m too tired to go out today, maybe we’ll see each other during the week.’]
Your output: *Sono troppo stanca per uscire oggi EXPLANATION - DETAILED, magari DOWNTONER ci vediamo in settimana OFFER OF REPAIR - UNCLEAR ALTERNATIVE [...]*

4.4 Testing the tools in pragmatic annotation tasks

Twelve cancellations were used to evaluate the annotation performance of the three tools under analysis. These cancellations, selected from the DISDIR corpus, included both typical and less common examples, covering a broad range of Sub-act, Strategy, and Modifier labels. Since LadderWeb requires each cancellation to be tokenized and each token to be tagged with three labels, Sub-act, Strategy, and Modifier (Section 4.2), the analysis was conducted at the token level. In total, the twelve cancellations comprised 214 tokens, resulting in 642 annotation items per model after tagging each token across the three categories (or marking the absence of a label). Notably, these

twelve cancellations were excluded from LadderWeb's training data and the instructional prompt used for the two pre-trained models.

For LadderWeb, each cancellation was submitted individually via the web interface. To initiate the annotation process in ChatGPT-4o and LLaMA-3.2, we first entered the instructional prompt into the browser-based interface. The models responded by confirming that they had processed the content and were ready to proceed. We then introduced the prompt (3), asking the models to annotate three cancellation utterances.

(3) Annotate the following utterances: [three cancellations inserted here]

After receiving the models' initial response, we repeated the prompt (2) three more times, each time supplying three new cancellations for annotation, until all twelve had been processed. We opted to submit small batches of cancellations across multiple iterations to avoid overloading the models' processing capacity and to reduce the risk of introducing annotation bias.

All prompts for ChatGPT-4o and LLaMA-3.2 were entered manually via the web interface rather than through API calls. While this approach is less efficient for large-scale annotation, it is more accessible to annotators without programming expertise and thus more feasible for widespread adoption. Annotations via ChatGPT-4o and LadderWeb were carried out in December 2024. About a month later, we repeated the annotation in LadderWeb, confirming that there was no difference from the first annotation. Annotations via LLaMA-3.2 and a second round of annotation via ChatGPT-4o took place in July 2025. For RQ2, we ran the annotation task twice with LLaMA-3.2. In the first run, we used default parameters to simulate usage by non-expert users and the usage in ChatGPT (e.g. temperature = 0.8, top-k = 40, top-p = 0.9). In the second run, we manually reduced the temperature to 0.05 in order to obtain less creative, more instruction-faithful results.

4.5 Evaluation of the models' performances

After collecting the annotations produced by the models, a benchmark was created through independent annotation by two expert raters to enable comparison. Each rater annotated the cancellations separately and they subsequently met to review their results and resolve any discrepancies. Although no formal statistical analysis of inter-rater agreement was performed at this stage, consensus was achieved through negotiation following the approach described by Löwen and Plonski (2016: 90). This negotiation process addressed the few cases of disagreement and ensured consistency in the final annotations. All human annotations were completed in December 2024. The resulting set of annotations is referred to as the gold standard, acknowledging that the annotation of pragmatic categories, such as speech acts, is intrinsically subjective.

To assess the models' performance, the agreement between each model and the benchmark was calculated for the annotation of Sub-act, Strategy, and Modifier with Cohen's κ (Hoek and Scholman 2017). Cohen's κ is a widely used measure that quantifies the degree of agreement between two raters classifying items into categorical codes while correcting for agreement that could occur by chance. The coefficient ranges from –1 (complete disagreement) to 0 (chance-level agreement) and 1 (perfect agreement), with commonly accepted benchmarks such as 0.41–0.60 indicating moderate agreement and 0.81–1 indicating almost perfect agreement (Artstein and Poesio 2008).

5 Results

5.1 Quantitative results

This section reports the quantitative results in the form of Cohen's κ values for the three tools in comparison with the gold standard across three layers, as illustrated in Table 1:

Table 1: Inter-rater agreement (Cohen's κ) for the annotations produced by the three models compared to the human benchmark.

	Sub-act	Strategy	Modifier
LadderWeb	0.746	0.716	0.774
ChatGPT-4o	0.95	0.861	0.868
LLaMA-3.2	0.594	0.443	0

The values indicate substantial agreement for LadderWeb, perfect agreement for ChatGPT-4o, and moderate agreement for LLaMA-3.2 in almost all categories. An intercategory comparison shows better agreement for Sub-acts than for Strategies. In the Modifiers category, there is a clear pitfall in the agreement of LLaMA-3.2 because the model did not label Modifiers. Although interacting with the LLM and asking more specifically for the annotation of this category would likely have resolved the issue, it would have compromised the comparability of the results.

To answer RQ2, we reran the same prompts roughly seven months after the initial annotation, using the same ChatGPT account and model version. In this second round, responses were noticeably slower. We also repeated the analysis with LLaMA-3.2, modifying only the temperature parameter, which we adjusted from 0.8 to 0.05.

Table 2: Inter-rater agreement (Cohen's κ) for the annotations produced by ChatGPT-4o on 21.07.2025 compared to the human benchmark, and by LLaMA-3.2 with lower temperature.

	Sub-act	Strategy	Modifier
ChatGPT-4o 2nd	0.9	0.822	0.584
LLaMA-3.2 T:0.05	0.678	0.552	0.392

As Table 2 shows, the annotations produced slightly different outputs. Concerning ChatGPT-4o, the observed inter-rater agreement was similar overall, with perfect agreement for the annotation of Sub-acts and Strategies. However, the inter-rater agreement for Modifiers dropped drastically, scoring 0.584. Concerning LLaMA-3.2, when the temperature is lowered to 0.05, performance improves somewhat, with $\kappa = 0.678$ for Sub-act, 0.552 for Strategy, and 0.392 for Modifier.

The following chart summarizes the values of interrater agreement across the machines and the annotation variables:

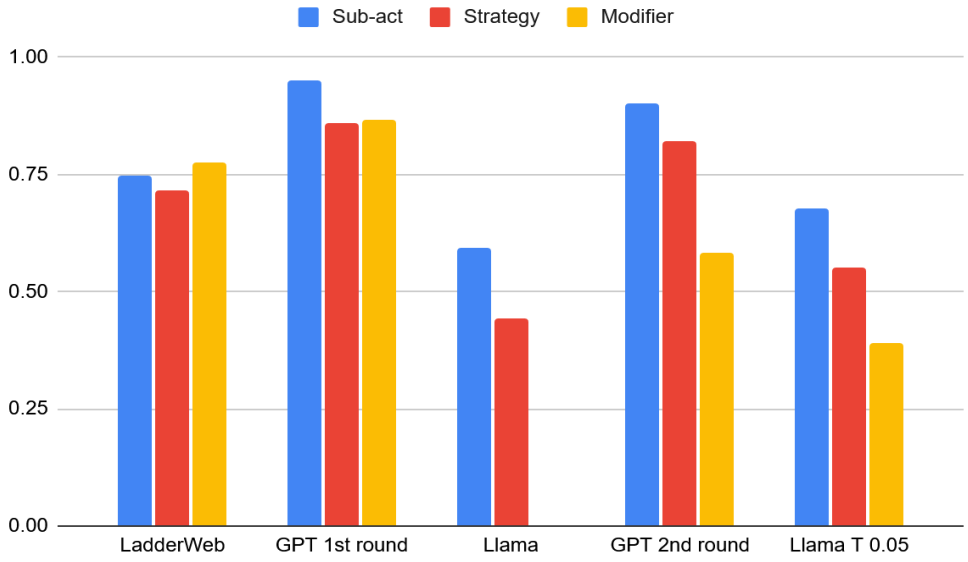


Figure 2: Value of the interrater agreement (K) across machines and annotation layers.

5.2 Qualitative Results

A closer examination of the misalignments with the human baseline, provides valuable insights into the machines' operational logic. This section presents a qualitative analysis of such misalignments across the three systems under study and the second round of annotation in ChatGPT-4o.

5.2.1 LadderWeb

LadderWeb left several Sub-acts unannotated. For example:

(4) *Sono desolata e spero che si presenteranno altre occasioni per passare una serata insieme*
'I am sorry, and I hope there will be other opportunities to spend an evening together' [our transl.][2]

In (4), the human benchmark identified two distinct Sub-acts: *Remedial Move* (*sono desolata*) and Offer of Repair (*spero che si presenteranno altre occasioni [...]*). However, LadderWeb failed to annotate either.

(5) *Non riesco stasera, ti chiamo domani*
'I can't tonight, I'll call you tomorrow'

In (5), the human annotation identified *Cancellation/Impossibility* and *Offer of Repair/No Alternative*. LadderWeb aligned with the first Sub-act (*Non riesco sta sera*) but left *ti chiamo domani* unannotated.

(6) *Divertiti anche per me, mi raccomando*
'Have fun for me too, please'

In (6), human raters annotated the entire utterance as Wishes. LadderWeb recognized *divertiti anche per me* but left *mi raccomando* unannotated. When it comes to Strategy annotation, further issues arise:

(7) *Però recuperiamo con una birra da me!*
'But we'll make up for it with a beer at my place!'

Expert raters annotated (7) as an Offer of Repair realized through the Strategy No Alternative. LadderWeb identified the Sub-act but left the strategy field blank. In some instances, LadderWeb fails to distinguish between the cut between Sub-acts and thus produces a conflated annotation, as in (8):

(8) *Ehi scusami so che ti creo un casino*
'Hey, sorry, I know that I'm causing you trouble'

[2] All translations are by the authors.

Human annotators labeled *scusami* as Remedial Move/Apology, and *so che ti creo un casino* as Appeal to Empathy. LadderWeb, however, applied both labels to *so* ('I know'), though the annotated segments were still aligned with the human version.

5.2.2 ChatGPT-4o, first round

In the case of ChatGPT-4o, misalignments manifest differently. Notably, there were no instances of missing annotations. Instead, discrepancies stemmed from differing interpretations, as illustrated below.

(9) *Non è successo niente, non ti preoccupare*
'Nothing happened, don't worry'

While the expert baseline classified this as an Explanation, ChatGPT-4o initially annotated it as an Appeal to Empathy. Regarding Strategy annotation:

(10) *Non mancherò la prossima occasione*
'I won't miss the next time'

Experts labeled this as an Offer of Repair using the Strategy Unclear Alternative. ChatGPT-4o, however, annotated it as an Offer of Repair with an Alternative. While not aligned with the baseline, the annotation remains within the bounds of plausible interpretation.

5.2.3 LLaMA-3.2 with default settings

Regarding LLaMA-3.2 with default settings, we observed more pronounced misalignments with the human baseline.

(11) *So che ti creo un casino*
'I know that I'm causing you trouble'
(12) *Non è successo nulla, stai tranquillo*
'Nothing happened, don't worry'

Both (11) and *non è successo nulla* in (12) were annotated as a Hedge. While these expressions do contribute to softening the tone of a cancellation and may function pragmatically as hedges, the instruction prompt explicitly indicated that Hedge is not a Sub-act but rather a modifier associated with specific lexical items. The use of Hedge here reflects a misinterpretation of the taxonomy's structure and its distinction between Sub-acts and Modifiers.

A similar confusion occurs with the expression *mi dispiace* ('I am sorry'), which was at times annotated as Appeal to Empathy. While the phrase does carry an emotional nuance, the instruction prompt clearly defined it as a Remedial Move / Statement of Regret. This tendency to freely interpret the labels becomes even more evident in the following cases (13 and 14):

(13) *Sono desolata*
'I am sorry'
(14) *Non mancherò*
'I won't miss it'

In (13), the model annotated the utterance as Explanation / Emotivo. The label Explanation diverges from the human annotation, which classified the segment as a Remedial Move / Statement of Regret. Furthermore, the strategy label Emotivo does not exist within the established taxonomy. In (14), the annotation Affirmazione (*sic*.) del Rifiuto ('Affirmation of Refusal') was assigned, which is also absent from the official label set and contains a spelling error (*Affirmazione* instead of *Affermazione*). Some misalignments were not only unexpected but also difficult to justify semantically:

(15) *Mi sarebbe piaciuto venire questa sera*
'I would have liked to come this evening'

The utterance in (15) was annotated as Greeting, possibly because of its position at the beginning of the message. However, the sub-act was annotated as Willingness and there is no justification for classifying it as a Greeting. This example underscores a clear misapplication of the label.

Finally, as we noted in Section 5.1, LLaMA-3.2 did not annotate Modifiers at all. This absence suggests that the model either ignored or was unable to process the part of the instruction referring to Modifier annotation. While it would be technically possible to prompt the model in a subsequent round to annotate Modifiers separately, doing so would have disrupted the controlled experimental setting. The goal was to assess the systems' capacity to follow complex annotation instructions in a single-pass scenario, as would be required in a realistic annotation pipeline. Therefore, this limitation highlights a crucial difference in instruction compliance among the models tested.

5.2.4 ChatGPT-4o, second round

Turning to the second-round annotations by ChatGPT-4o, we observed a high degree of internal consistency with its first-round annotations. For example:

(16) *Recuperiamo con una birra da me*
'Let's make up with a beer at my place'

This segment was annotated in both rounds as Offer of Repair/ Alternative. However, the human baseline considered it as Offer of Repair/ No Alternative, as the utterance does not provide a clear indication of when the alternative appointment will take place. While the misalignment is minor, it suggests that the distinction between Alternative and No Alternative remains subtle and potentially ambiguous for AI models. We also found some notable differences between the two rounds:

(17) *Non mancherò la prossima occasione*
'I won't miss the next opportunity'

In the first round, ChatGPT-4o annotated the Strategy of (17) as Offer of Repair/ Alternative, while in the second round it was labeled as Offer of Repair/ No Alternative. Human annotators opted for Offer of Repair/ Unclear Alternative. Interestingly, in the second round, ChatGPT-4o exhibited a more nuanced identification of Modifiers as shown in the following examples:

(18) *Non posso venire però recuperiamo*
'I can't come, but we'll make up for it'
(19) *Scusa se ho fatto un casino*
'Sorry if I made a mess'

In (18), *però* ('but') was annotated as a Downtoner, and in (19), *casino* ('mess') was marked as Evaluation. These Modifier annotations were not provided by human annotators but can be considered legitimate, indicating an improved sensitivity to pragmatic nuances in the second round.

(20) *Un abbraccio*
'Hugs'

Example (20) was annotated as a Term of Endearment by the model, while human annotators labeled it as Farewell. Although both interpretations are contextually reasonable, the assigned label must align with the function defined in the taxonomy: according to the taxonomy, Term of Endearment is a Modifier, mostly associated with the Alerter, and not a Sub-act. Similarly, in example (12), *non è successo nulla* was annotated as Appeal to Empathy, consistent with the human baseline, but also received a Modifier label of Downtoner. According to our taxonomy, however, the label Downtoner should be applied to lexical elements that reduce the illocutionary force of

an utterance. In this case, the use of the term was not appropriate, reflecting a conceptual misunderstanding of the Modifier category.

5.2.5 LLaMA-3.2 with lower temperature

In the second annotation, when using LLaMA-3.2 with a lower temperature, the results showed a higher κ-index compared to the same model under standard parameters, although some misalignments still remained. Certain Alerters continued to be annotated due to their position at the beginning of sentences, as in example (15). Similarly, as observed with the model under default settings, Appeal to Empathy was overextended to encompass other categories. This likely reflects an interpretation driven by the semantic content of the label rather than by strict adherence to the instructions and examples. Nevertheless, the annotation did not show any major deviation from the instructions, particularly regarding the distinction between Sub-acts and Modifiers or the creation of new categories, which had occurred with LLaMA-3.2 at higher temperature. In the annotation of Modifiers, we also noted a clear improvement: most intensifiers were annotated consistently with the human annotators. However, signs of creative intervention continued to appear even at a temperature of 0.05: LLaMA-3.2 occasionally introduced unsolicited comments into its annotations. These comments were not required by the instruction prompt and could potentially confuse the annotation process, as illustrated in (21) and (22).

(21) The use of ‘mi dispiace’ is a polite way to express regret.

Example (21) contains an observation potentially aligned with the baseline, although in this case *mi dispiace* (‘I am sorry’) was erroneously annotated as Alerter / Greeting.

(22) ‘Non è successo nulla’ is a way to downplay the cancellation.

While (22) is correct in meaning, it is not pertinent to the task.

6 Discussion

Section 5.1 showed the interrater agreement measured with κ: the κ values indicate how closely each model – LadderWeb, ChatGPT-4o, and LLaMA-3.2 – aligns with the human baseline across layers, providing insight into both model-specific performance patterns and the complexity of the annotation tasks. We assumed that high κ corresponds to high accuracy.

As shown in Figure 1, Sub-acts yield the highest agreement scores across all machines. This confirms that Sub-acts are the most stable and reliably recognized

category, likely due to their high frequency and the smaller set of possible labels. In both rounds, ChatGPT-4o achieves near-ceiling performance on Sub-acts ($\kappa = 0.95$ and 0.9), while LadderWeb shows moderately high alignment ($\kappa = 0.75$). LLaMA-3.2, by contrast, performs significantly lower on Sub-acts ($\kappa = 0.59$ and 0.678).

Strategy annotations consistently yield lower agreement scores than Sub-acts. This supports the claim that Strategies, being subcategories of Sub-acts and therefore more numerous, are more sparsely represented in the training data of LadderWeb and in the instruction Prompt of ChatGPT-4o and LLaMA-3.2 and therefore more sensitive learning stability. While ChatGPT-4o still performs robustly in this layer, LLaMA-3.2's score drops markedly ($\kappa = 0.44$-0.55), reinforcing its difficulty with hierarchical or low-frequency categories.

Modifier annotation presents a particularly informative contrast. LadderWeb and ChatGPT-4o (first round) both reach relatively high agreement scores ($\kappa = 0.74$ and 0.87, respectively), which suggests that Modifier detection benefits from reliable lexical cues. However, ChatGPT-4o's performance in the second round drops ($\kappa = 0.58$), possibly overfitting to previously learned patterns. Most notably, LLaMA-3.2 with standard settings entirely fails to annotate Modifiers, possibly due to failing to process the instruction concerning Modifier tagging. This omission illustrates a fundamental limitation in instruction-following or representational capacity for LLaMA-3.2 in multi-layered annotation particularly in a category that is lexically dependent and therefore expected to be detected accurately by machine annotation.

Overall, the quantitative data can be summarized in the following several key findings: i) Sub-acts are consistently easier to detect due probably to their higher frequency and relatively clear-cut categorization; ii) Strategies and Modifiers exhibit greater variability possibly due to lower representation in the instructional prompt; iii) ChatGPT-4o benefits from its pre-training capacities, maintaining higher interrater agreement with the human base line across layers; and iv) LLaMA-3.2 shows the greatest instability across all layers. The better performance of LadderWeb over LLaMA-3.2 was unexpected in light of the previous literature (Liu et al. 2023; Wang 2023).

In Section 5.2, the qualitative analysis offers a valuable lens through which to interpret the quantitative results, providing insight into *why* the models behave as they do.

LadderWeb is a supervised learning model trained on a limited dataset. As such, it relies heavily on lexical overlap between input data and the training corpus. The analysis suggests that when lexical items diverge from its internal vocabulary used for the training (examples (4) to (8)), the system is often unable to assign an appropriate label. This limitation is exemplified by the system's failure to annotate expressions such as *sono desolata* or *non mancherò*, which, although pragmatically salient, fall outside of the system's learned vocabulary. A manual review of LadderWeb's training data verified this hypothesis: such items (*desolata* and *mancherò*) are not explicitly

represented in the training set and are therefore more prone to omission or misclassification. Like other supervised models (Hamilton et al. 2024), LadderWeb struggles to generalize beyond its dataset and exhibits a high sensitivity to lexical variation.

In contrast, the behavior of pre-trained LLMs, particularly LLaMA-3.2, reveals a distinct annotation strategy. From the analysis of the mismatching cases, these models appear to prioritize the semantic interpretation of a label over the strict taxonomy provided in the annotation instructions (Bianco et al. 2025). For instance, in example (12), LLaMA-3.2 labeled *non è successo nulla* as a Hedge in the Sub-act layer, despite the instruction that a Hedge is a Modifier, and provided a list of lexical items classifiable as Hedges.

Similarly, tagging (13) (*sono desolata*) as Explanation/Emotivo, or (14) (*non mancherò*) as Affirmation of Refusal, reflects a reliance on semantic plausibility rather than instruction compliance. These labels, while interpretively defensible, fall outside the specified taxonomy and thus indicate a partial disregard for prompt conditioning. This tendency appears more pronounced in LLaMA-3.2 (see (11) and (12)), but is also observable in ChatGPT-4o (see (9)). These cases suggest that ChatGPT-4o, despite its generally high performance, may at times default to label names that correspond closely to literal lexical meaning rather than examples and definitions specified in the task instructions. This behavior points to a latent tension between the model's trained semantic and the logic of the taxonomy.

One additional aspect deserves attention. While ChatGPT-4o's overall performance is very close to the gold standard established by human annotators, the second round of annotations revealed minor inconsistencies. This variability is relevant for future applications where LLMs might be deployed for automated pragmatic annotation of speech acts. In contrast, LadderWeb's output was stable across annotation rounds (see Section 4.4). Although its training demands are higher, requiring a curated and balanced dataset, the consistency of its output offers a level of replicability that is desirable in formal annotation pipelines. Finally, utilizing the downloadable LLaMA-3.2 through Ollama running locally on one’s own computer, allows for greater control over LLM. The replicability of the results does not depend on a company keeping the same model available; this, of course, contributes to transparency of our tests. LLaMA-3.2 also offers the possibility of employing and controlling techniques like RAG[3] or creating specialized models for specialized tasks through fine-tuning techniques.

[3] RAG stands for Retrieval-Augmented Generation. It is a method that enhances an LLM by giving it access to external information during generation. A retrieval system searches an uploaded database, or document collection for passages relevant to the user’s query. These retrieved passages are fed to the LLM as additional context. The LLM then generates an answer that combines its internal knowledge with the retrieved information. This approach gives users more control over the sources of the LLM output.

One final point concerns the variability of LLaMA-3.2 with respect to the temperature parameter. As expected, lowering the temperature increased accuracy, as the model produced more deterministic and consistent outputs, thereby reducing random variation in the annotations. Indeed, the annotation with a temperature of 0.05 appeared more closely aligned with the instructions than that with 0.8, since no additional invented categories appeared nor Strategy labels were introduced instead of the Sub-act labels. This suggests that reducing randomness helps LLaMA follow instructions more faithfully, leading to more reliable annotations. However, even at lower temperatures, the agreement remains notably below that of LadderWeb and GPT, indicating that temperature adjustment alone cannot fully compensate for the model's limitations in this task. Further experimentation with setting the temperature to zero or adjusting other decoding parameters, such as top-p and top-k, which also influence the generation of creative outputs, may potentially yield more accurate results.

Regarding data security and compliance with GDPR (European Union, 2016) standards, it is important to note that both LadderWeb and LLaMA-3.2, the last running locally, do not store user data. This makes them suitable for environments with strict data protection requirements. The same assurance cannot be offered for ChatGPT-4o, which operates as a commercial system and may involve data retention beyond the user's control.

7 Conclusions

The use of LLMs for pragmatic annotation is increasingly recognized as a promising approach. This study compared human annotations with those produced by the supervised learning model LadderWeb and the pre-trained LLMs ChatGPT-4o and LLaMA-3.2. ChatGPT-4o outperformed the other systems, achieving near-perfect agreement with the human baseline (RQ1). LLaMA-3.2 demonstrated only moderate agreement and frequently failed to annotate Modifiers, while LadderWeb showed intermediate performance.

However, a follow-up annotation conducted seven months later revealed that ChatGPT-4o produced different results, leading to a slight decline in inter-rater agreement with the human experts. This variability suggests that, although highly capable, ChatGPT-4o may be less stable over time than LadderWeb and LLaMA-3.2, making it a less reliable tool in longitudinal annotation projects and hindering replicability (RQ2). It is also worth noting that LadderWeb is trained specifically on the speech act of cancellation, whereas LLMs are designed as general-purpose models. Qualitative analysis indicates that pre-trained models do not adhere strictly to task-specific guidelines. Instead, they may draw on semantically related interpretations learned during training. This behavior can result in the assignment of technically incorrect or even non-existent labels (as the case of LLaMA-3.2 with high temperature)

according to the taxonomy, though these labels may still appear contextually plausible. Therefore, a possible solution would be to experiment with LLaMA-3.2 by adjusting different parameters, not only temperature, until the results reach the desired level of quality for the specific task of annotation.

In summary, ChatGPT-4o is the most accurate, though not entirely consistent. LadderWeb, while not fully accurate given the current limited training set, demonstrates intrinsic consistency. LLaMA-3.2, used with Ollama, has the potential for consistency but requires further experimentation to achieve sufficient accuracy. LadderWeb, ChatGPT-4o, and LLaMA-3.2 each offer distinct advantages for accelerating pragmatic annotation. The first two are particularly suitable for researchers or practitioners with limited technical backgrounds and their accuracy is particularly promising, confirming previous research (Brocca et al. to appear; Yu et al. 2024; Su and Ye 2025). While they are not yet a full substitute for human annotation, these tools can significantly improve inter-rater reliability when used in a complementary role alongside human annotators. In this context, they can bring research closer to the goals outlined by scholars in recent decades (O'Keeffe et al. 2019; Weisser 2018).

Future studies should examine the impact of various factors on the quality of AI-supported annotation. These include:

(i) the language of the data;
(ii) the prompt used, as suggested by Brocca et al. (to appear), who examined the influence of the number of examples in the instruction prompt;
(iii) the adopted taxonomy, since a less fine-grained taxonomy may yield higher human-machine agreement;
(iv) the type of data; written DCTs can be more standardised and may therefore produce higher levels of agreement than data collected in other ways, such as through oral role plays;
(v) the speech act under investigation, and
(vi) the pragmatic category to be annotated, since tasks may be more complex than speech-act annotation (e.g. Bianco et al. 2025 on humour annotation).

Examining these dimensions would clarify the conditions that foster reliable alignment between human judgements and model predictions, thereby enabling efficient annotation of pragmatic categories. This would, in turn, support the development of an accessible, user-friendly tool for researchers and promote greater consistency in corpus-based pragmatic analysis.

Attachments

Online attachments, including annotations and interrater agreement, the cancellation taxonomy, and the prompts used, are available here:
https://doi.org/10.17605/OSF.IO/2HTNQ

Attributions

Brocca, Nicola: Conceptualization, methodology, formal analysis, investigation, resources, data curation, writing-original, review, editing and funding acquisition. In particular, he wrote Section 1, Section 3, Section 4.2 (second paragraph), Section 4.3, Section 4.4, Section 4.5, Section 5, Section 6, Section 7.
Nuzzo, Elena: Conceptualization, methodology, resources, data curation, writing-original, review and editing. In particular, she wrote Section 2, Section 4.1.
Wang-Kathrein, Joseph: Software, resources, writing-original. He wrote Section 4.2 (first paragraph).

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this contribution.

Acknowledgement

We extend our gratitude to the following students for their contributions to the annotations essential for training LadderWeb: Caterina Colangeli, Nadine Mair, Maria Rudigier, Corrado Schininà, Tamara Walder, and Lukas Zehetgruber. The authors utilized ChatGPT-4o for language revisions.

Funding

This research was partially funded by the following:
- Digitization and Information Extraction for the Digital Humanities (DI4DH): A project initiated by the Digital Humanities Research Center at the University of Innsbruck (5th and 8th call).
- CLARIAH-AT (Funding Call 2022): Focused on the interoperability and reusability of digital humanities data and tools.

References

Alizadeh, Meysam & Kubli, Maël & Samei, Zeynab & Dehghani, Shirin & Zahedivafa, Mohammadmasiha & Bermeo, Juan D. & Korobeynikova, Maria & Gilardi, Fabrizio. 2025. Open-source LLMs for text annotation: a practical guide for model setting and fine-tuning. *Journal of Computational Social Science 8. 17.* https://doi.org/10.1007/s42001-024-00345-9

Artstein, Ron & Poesio, Massimo. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics* 34(4). 555–596. https://doi.org/10.1162/coli.07-034-R2

Austin, John Langshaw. 1970. *How to do Things with Words* (Reprint from corrected sheets of the 1963 reprint). Harvard University Press.

Apache Software Foundation. 2025. *Name Entity Recognition Tagger*. https://opennlp.apache.org/docs/1.9.1/manual/opennlp.html (last accessed on 10/12/2025).

Beebe, Leslie & Takahashi, Tomoko & Uliss-Weltz, Robin. 1990. Pragmatic transfer in ESL refusals. In Scarcella, Robin C. & Anderson, Elaine & Krashen, Stephen (eds), *Developing Communicative Competence in a Second Language*, 55–73. New York: Newbury House.

Ben Abacha, Asma & Yim, Wen-wai & Fu, Yujuan & Sun, Zhaoyi & Yetisgen, Meliha & Xia, Fei, & Lin, Thomas. 2024. MEDEC: A benchmark for medical error detection and correction in clinical notes. *arXiv preprint*.

Blum-Kulka, Shoshana & House, Juliane & Kasper, Gabriele. 1989. *Cross-cultural Pragmatics: Requests and Apologies*. Norwood, NJ: Ablex Publishing Corporation.

Bianco, Antonio & Brocca, Nicola & Garassino, Davide. 2025. Does ChatGPT get it? LLM-driven annotation of Humor. SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5500673 (last accessed on 10/12/2025).

Brocca, Nicola & Nuzzo, Elena & Wang-Kathrein, Joseph. to appear. Machine Learning for Pragmatic Annotation: Comparing Supervised and Pre-Trained Models in Speech Act Tagging.

Brocca, Nicola & Cortés Velásquez, Diego & Nuzzo, Elena & Rudigier, Maria. 2023. Linguistic politeness across Austria and Italy: Backing out of an invitation with an instant message. *Journal of Pragmatics* 209. 56–70. https://doi.org/10.1016/j.pragma.2023.02.018

Brocca, Nicola & Cortés Velásquez, Diego & Hirzinger-Unterrainer, Eva-Maria & Nuzzo, Elena & Rudigier, Maria & Spielthenner, Valentin & Wang-Kathrein,

Joseph. 2024. *LadderWeb*. University of Innsbruck. https://ifd-ladderweb.uibk.ac.at/#/annotate (last accessed on 10/12/2025).

Brown, Tom B. & Mann, Benjamin & Ryder, Nick & Subbiah, Melanie & Kaplan, Jared & Dhariwal, Prafulla & Neelakantan, Arvind & Shyam, Pranav & Sastry, Girish & Askell, Amanda & Agarwal, Sandhini & Herbert-Voss, Ariel & Krueger, Gretchen & Henighan, Tom & Child, Rewon & Ramesh, Aditya & Ziegler, Daniel M. & Wu, Jeffrey & Winter, Clemens, et al. & Amodei, Dario. 2020. Language models are few-shot learners. *arXiv preprint*. https://doi.org/10.5555/3495724.3495883

Calhoun, Sasha & Carletta, Jean & Brenier, Jason M. & Mayo, Neil & Jurafsky, Dan & Steedman, Mark & Beaver, David. 2010. The NXT-format Switchboard Corpus: A rich resource for investigating the syntax, semantics, pragmatics, and prosody of dialogue. *Language Resources and Evaluation* 44(4). 387–419. https://doi.org/10.1007/s10579-010-9120-1

Cavasso, Luca & Taboada, Maite. 2021. A corpus analysis of online news comments using the Appraisal framework. *Journal of Corpora and Discourse Studies* 4. 1–38. https://doi.org/10.18573/jcads.61

Cortés Velásquez, Diego & Nuzzo, Elena. 2022. Declining an invitation: The pragmatics of Italian and Colombian Spanish. In Gesuato, Sara & Salvato, Giuliana & Castello, Erik (eds), *Pragmatic Aspects of L2 Communication: From Awareness through Description to Assessment*, 143–163. Newcastle upon Tyne: Cambridge Scholars Publishing.

De Felice, Irene & Strik Lievers, Francesca. 2024. Building a pragmatically annotated diachronic corpus: The DIADIta project. In *Proceedings of the 10th Italian Conference on Computational Linguistics – CLiC-it 2024*, 1–7. Aachen: CEUR-WS.

Gilardi, Fabrizio & Alizadeh, Meysam & Kubli, Maël. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences of the United States of America* 120(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

European Union. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). *Official Journal of the European Union* L 119. 1–88. https://eur-lex.europa.eu/eli/reg/2016/679/oj (last accessed on 10/12/2025).

Félix-Brasdefer, J. César. 2003. Declining an invitation: A cross-cultural study of pragmatic strategies in American English and Latin American Spanish. *Multilingua* 22(3). 225–255. https://doi.org/10.1515/mult.2003.012

Félix-Brasdefer, J. César. 2008. Sociopragmatic variation: Dispreferred responses in Mexican and Dominican Spanish. *Journal of Politeness Research* 4(1). 81–110. https://doi.org/10.1515/PR.2008.004

Hamilton, Kyra & Longo, Luca & Bozic, Bojan. 2024. GPT-assisted annotation of rhetorical and linguistic features for interpretable propaganda technique detection in news text. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24)*, 1431–1440. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3589335.3651909

Hoek, Jet & Scholman, Merel. 2017. Evaluating discourse annotation: Some recent insights and new approaches. In *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-13).*

Joseph, Brian. 2008. The editor's department: Last scene of all… *Language* 84. 686–690. https://doi.org/10.1353/lan.0.0063

Jurafsky, Dan & Martin, James H. 2024. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models* (3rd ed.). Online manuscript released August 20, 2024. https://web.stanford.edu/~jurafsky/slp3/ (last accessed on 10/12/2025).

Kim, Minjin & Lu, Xiaofei. 2024. Exploring the potential of using ChatGPT for rhetorical move-step analysis: The impact of prompt refinement, few-shot learning, and fine-tuning. *Journal of English for Academic Purposes* 71. https://doi.org/10.1016/j.jeap.2024.101422

Lai, Viet Dac & Ngo, Nghia Trung & Veyseh, Amir Pouran Ben & Man, Hieu & Dernoncourt, Franck & Bui, Trung & Nguyen, Thien Huu. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint*. https://doi.org/10.48550/arXiv.2304.05613

Landert, Daniela & Dayter, Daria & Messerli, Thomas C. & Locher, Miriam A. 2023. *Corpus Pragmatics.* Cambridge: Cambridge University Press.

Lang, Fabian. 2025. *Künstliche Intelligenz in Seminar- und Abschlussarbeiten: Ein Praxisleitfaden für Studierende mit Handlungsempfehlungen, Prompt-Beispielen und kritischer Einordnung*. Heidelberg: Springer. https://doi.org/10.1007/978-3-662-71542-0

Liu, Pengfei & Yuan, Weizhe & Fu, Jinlan & Jiang, Zhengbao & Hayashi, Hiroaki & Neubig, Graham. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys* 55(9). Article 195, 1–35. https://doi.org/10.1145/3560815

Löwen, Shawn & Plonsky, Luke. 2016. *An A–Z of Applied Linguistics Research Methods.* London: Palgrave.

Meta AI. 2025. LLaMA 3.2 [Computer software]. Meta. https://www.llama.com/models/llama-3/ (last accessed on 10/12/2025).

Nuzzo, Elena & Cortés Velásquez, Diego. 2020. Canceling last minute in Italian and Colombian Spanish: A cross-cultural account of pragmalinguistic strategies. *Corpus Pragmatics* 4. 333–358. https://doi.org/10.1007/s41701-020-00084-y

O'Keeffe, Anne. 2018. Corpus-based function-to-form approaches. In Jucker, Andreas & Schneider, Klaus & Bublitz, Wolfram (eds), *Methods in Pragmatics*, 587–618. Berlin & Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110424928-023

O'Keeffe, Anne & Clancy, Brian & Adolphs, Svenja. 2019. *Introducing Pragmatics in Use.* London: Routledge.

Ollama. 2025. *Ollama: Run and manage large language models locally.* https://ollama.com/ (last accessed on 10/12/2025).

OpenAI. 2024. ChatGPT-4o [Large language model]. Retrieved 12 December 2024 from https://chat.openai.com/(last accessed on 10/12/2025).

Ostyakova, Lidiia & Smilga, Veronika & Petukhova, Kseniia & Molchanova, Maria & Kornev, Daniel. 2023. ChatGPT vs. crowdsourcing vs. experts: Annotating open-domain conversations with speech functions. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 242–254. Prague, Czechia: Association for Computational Linguistics.

Ouyang, Long & Wu, Jeff & Jiang, Xu & Almeida, Diogo & Wainwright, Carroll & Mishkin, Pamela & Zhang, Chong & Agarwal, Sandhini & Slama, Katarina & Ray, Alex & Schulman, John & Hilton, Jacob & Kelton, Fraser & Miller, Luke & Simens, Maddie & Askell, Amanda & Welinder, Peter & Christiano, Paul & Leike, Jan & Lowe, Ryan. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35. 27730–27744.

Rühlemann, Christoph & Aijmer, Karin. 2015. Corpus pragmatics: Laying the foundations. In Rühlemann, Christoph & Aijmer, Karin (eds), *Corpus Pragmatics: A Handbook*, 1–28. Cambridge: Cambridge University Press.

Rühlemann, Christoph. 2022. What can a corpus tell us about pragmatics. In O'Keeffe, Anne & McCarthy, Michael J. (eds), *The Routledge Handbook of Corpus Linguistics*, 263–280. New York: Routledge.

Scott-Phillips, Thomas C. 2017. Pragmatics and the aims of language evolution. *Psychonomic Bulletin & Review* 24(1). 186–189. https://doi.org/10.3758/s13423-016-1061-2

Shishavan, Homa Babai & Sharifian, Farzad. 2016. The refusal speech act in a cross-cultural perspective: A study of Iranian English-language learners and Anglo-Australian speakers. *Language & Communication* 47. 75–88. https://doi.org/10.1016/j.langcom.2016.01.001

Su, Hang & Ye, Jun. 2025. Large language models for automating fine-grained speech act annotation: A critical evaluation of GPT-4o and DeepSeek. *Corpus Pragmatics* 9. 463–482. https://doi.org/10.1007/s41701-025-00200-w

Taylor, Charlotte. 2016. *Mock Politeness in English and Italian.* Amsterdam: John Benjamins.

Wang, Haifeng & Li, Jiwei & Wu, Hua & Hovy, Eduard & Sun, Yu. 2023. Pre-trained language models and their applications. *Engineering* 25. 51–65. https://doi.org/10.1016/j.eng.2022.04.024

Wei, Xiang & Cui, Xingyo & Cheng, Ning & Wang, Xiaobin & Zhang, Xin & Huang, Shen & Xie, Pengjun & Xu, Jinan & Chen, Yufeng & Zhang, Meishan & Jiang, Yong & Han, Wenjuan. 2023. ChatIE: Zero-shot information extraction via chatting with ChatGPT. *arXiv preprint*. https://doi.org/10.48550/arXiv.2302.10205

Weisser, Martin. 2015. Speech act annotation. In Rühlemann, Christoph & Aijmer, Karin (eds), *Corpus Pragmatics: A Handbook*, 84–110. Cambridge: Cambridge University Press.

Weisser, Martin. 2016. DART – The dialogue annotation and research tool. *Corpus Linguistics and Linguistic Theory* 12(2). 355–388.

Weisser, Martin. 2018. *How to do Corpus Pragmatics on Pragmatically Annotated Data: Speech Acts and Beyond.* Amsterdam: John Benjamins.

Yin, Ziqi & Wang, Hao & Horio, Kaito & Kawahara, Daisuke & Sekine, Satoshi. 2024. Should we respect LLMs? A cross-lingual study on the influence of prompt politeness on LLM performance. *arXiv preprint*. https://arxiv.org/abs/2402.14531 (last accessed on 9/12/2025).

Yu, Danni & Li, Luyang & Su, Hang & Fuoli, Matteo. 2024. Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology. *International Journal of Corpus Linguistics* 29(4). 534–561. https://doi.org/10.1075/ijcl.23087.yu

Zhao, Tianyu & Kawahara, Tatsuya. 2019. Joint dialog act segmentation and recognition in human conversations using attention to dialog context. *Computer Speech & Language* 57. 108–127. https://doi.org/10.1016/j.csl.2019.03.001

Zhu, Yiming & Zhang, Peixian & Haq, Ehsan-Ul & Hui, Pan & Tyson, Gareth. 2023. Can ChatGPT reproduce human-generated labels? A study of social computing tasks. *arXiv preprint*. https://doi.org/10.48550/arXiv.2304.10145