Tadej Todorović, Andrej Flogie, Daniel Hari
University of Maribor, Slovenia

2025, Vol. 22 (1), 19-34(228)
journals.uni-lj.si/elope
https://doi.org/10.4312/elope.22.1.19-34
UDC: [81'33:821.111.09-2]:004.89

Generative AI in Pragmatics: Assessing the Accuracy of Automated Speech Act Classification in Pinter's *The Birthday Party*

ABSTRACT

This study explores the feasibility of using generative AI (ChatGPT, Gemini, and DeepSeek) to automate speech act annotation in Harold Pinter's play *The Birthday Party*. Three chatbots – ChatGPT, Gemini, and DeepSeek – were tested under three scenarios varying in the amount of theoretical material provided. Each chatbot's output was compared to a manually annotated reference via a Python script measuring classification accuracy. Scenario 2 produced the highest accuracy overall (75–82%), while Scenario 1 underperformed, owing to incorrect reliance on external typologies, and Scenario 3 showed signs of overfitting. ChatGPT o1 emerged as the most accurate model, achieving 82% accuracy in Scenario 2. The findings suggest that GenAI chatbots can serve as valuable preliminary annotators when good prompt-engineering and well-curated theoretical material are provided. Future research could extend this methodology to more context-dependent texts, further refining prompt-engineering strategies and exploring larger linguistic corpora.

Keywords: pragmatics, speech act analysis, ChatGPT, DeepSeek, Gemini, Pinter

Generativna umetna inteligenca v pragmatiki: analiza natančnosti samodejne klasifikacije govornih dejanj v Pinterjevi drami *Zabava za rojstni dan*

IZVLEČEK

Študija raziskuje smiselnost rabe generativne umetne inteligence (ChatGPT, Gemini in DeepSeek) za avtomatizacijo anotacije govornih dejanj v Pinterjevi drami *Zabava za rojstni dan*. Trije klepetalni roboti – ChatGPT, Gemini in DeepSeek – so bili testirani v treh scenarijih, ki so se razlikovali glede na obseg predloženega teoretičnega gradiva. Rezultati vsakega klepetalnega robota so bili primerjani z ročno anotirano različico s pomočjo Python skripte, ki je izmerila natančnost klasifikacije. Scenarij 2 je na splošno dosegel najvišjo natančnost (75–82 %), medtem ko je bil scenarij 1 zaradi neustreznega zanašanja na tuje tipologije preslab, scenarij 3 pa je kazal znake preprileganja (angl. *overfitting*). ChatGPT o1 se je izkazal za najnatančnejši model, saj je v scenariju 2 dosegel 82-odstotno zanesljivost. Ugotovitve kažejo, da lahko klepetalni roboti GEN-UI služijo kot koristni predhodni anotatorji, če so na voljo dobro zasnovani pozivi in dobro pripravljeno teoretično gradivo. Prihodnje raziskave bi lahko to metodologijo razširile na besedila, ki so bolj odvisna od konteksta, nadalje izpopolnile strategije inženiringa pozivov in raziskale večje jezikovne korpuse.

Ključne besede: pragmatika, analiza govornih dejanj, ChatGPT, DeepSeek, Gemini, Pinter

1 Introduction

In this paper, we examine the potential of generative artificial intelligence (GenAI) for assisting and optimizing research in pragmatics. Specifically, we focus on research using speech act analysis, a powerful tool that enables quantitative and qualitative insight into various pragmatic topics of interest, mediation (Kádár et al. 2024; House et al. 2024), small talk (House and Kádár 2023), bargaining (Liu, House, and Kádár 2024), etc. Such research can yield unique and robust results in pragmatics; however, the initial data collection process is often time-consuming as it requires the researchers to manually annotate the data using some form of speech act typology. A tool such as GenAI that could either perform or, at the very least, facilitate this initial process would thus be especially beneficial for researchers in this field.

We thus performed a case study, testing the potential of selected GenAI tools (ChatGPT, Gemini, and DeepSeek) for annotating Harold Pinter's early play *The Birthday Party* (Pinter 1991) using a finite speech act typology developed by Edmondson, House, and Kádár (Edmondson and House 1981; Edmondson, House, and Kádár 2023). We decided to test the chatbots' capabilities using a literary work because historical documents, interviews, or other recordings usually require additional context for successful annotation, whereas a literary work is as close to a self-contained whole as possible. Next to identifying the most appropriate chatbot for the task, we also undertook the task of determining the best prompt (prompt-engineering) that researchers could use for this work. In doing so, we developed three scenarios for testing each chatbot: (1) instructing the chatbot to research the speech act typology online, providing it with a short annotated excerpt from *The Birthday Party*, and then instructing it to annotate the remainder of Act one; (2) providing the chatbot with a short description of the speech act typology (20 pages) and with a short annotated excerpt from *The Birthday Party*, and then instructing it to annotate the remainder of Act one; and (3) providing the chatbot with an exhaustive description of the speech act typology (80 pages) and with a short annotated excerpt from *The Birthday Party*, and then instructing it to annotate the remainder of Act one. Finally, we analysed the results by comparing the automatic annotations to the version of *The Birthday Party* manually annotated by human experts. In doing so, we endeavoured to answer the following research questions:

1. How successful are chatbots in providing an automatically annotated text in line with the given speech act typology?
2. Which scenario yields the best outcome (the highest fidelity to manual annotation)?
3. Are chatbots useful for performing such preliminary annotations or, at the very least, facilitating this process?

2 Related Work

GenAI chatbots have been recognized as useful in many domains, including time-consuming tasks such as literature reviews, citation management, proofreading, summarizing, paraphrasing, etc. (Stokel-Walker 2023; Else 2023; Altmäe, Sola-Leyva, and Salumets 2023). They have been tested in various academic fields, such as machine translation of literary works (Mohar, Orthaber, and Onič 2020) and assistance in analysing historical documents (Hazemali

et al. 2024) with various degrees of success. In the field of pragmatics, there are several studies that examine chatbots themselves and their utterances, such as exploring Gricean Maxims to help inform the basic design of effective conversational interaction (Setlur and Tory 2022), using AI-generated conversations as human-like data for pragmatic analysis (Chen, Li, and Ye 2024), and investigating the politeness strategies of chatbots (Monteiro, Pereira, and Salgado 2023). In the field of speech act analysis, a recent study has examined whether chatbots are capable of assertion (Williams and Bayne 2024). Most of these studies focus on studying GenAI and generating new knowledge by examining their behaviour, which we believe to be a worthy endeavour; however, chatbots can also be useful as a facilitator of research in sometimes painstakingly slow processes, such as annotating large volumes of text, utterance by utterance, using a specific speech act typology. To our knowledge, no one has yet tried to use the capabilities of chatbots to aid in the analysis of such data.

3 Data and Methodology

3.1 Data

Our data includes a manual annotation of Harold Pinter's *The Birthday Party* (Pinter 1991), one of Pinter's most frequently performed works, with some critics ranking it among the greatest dramatic achievements of British theatre (Hribar 2004; Gavez 2016; Onič 2016). The play follows the events unfolding in a boarding house in an English seaside town, run by Meg and Petey Boles. It begins with an ostensibly mundane breakfast conversation between Meg and Petey and eventually transforms into a psychological play where two strangers, Goldberg and McCann, arrive at the boarding house, searching for Stanley, one of the "permanent" guests at the boarding house.

We chose this play for three reasons. First, we wanted to analyse chatbots' capabilities in analysing a literary work, which, compared to historical documents or diplomatic transcripts, is as close to a self-contained whole as possible. A speech act annotation of historical documents requires additional outside context, i.e., knowledge of the complex political situation during which the analysed discourse took place, so the researchers can attribute the correct speech acts to participants based on both their statements and their motivations in the context of the political situation (insofar as this information is known). For example, in the case of a speech act annotation of a mediation event between the EEC and the Slovenian and Croatian states, the authors first had to be familiarized with the political context that surrounded that mediation attempt (Kádár et al. 2024). Second, among the various types of literary works, plays are most appropriate for speech act annotation and subsequent analysis, as they feature (almost exclusively) direct speech, whereas other forms of literature, like novels or short stories, also include the narrative voice, which cannot be analysed in this way. Moreover, the resemblance to an ordinary everyday conversation that contains elements of naturally occurring discourse, such as hesitations, repetitions, self-corrections, or non-sequiturs, is highest in contemporary drama (Podbevšek and Žavbi 2021; Onič and Prajnč Kacijan 2020), as opposed to, for example, the language of Elizabethan drama, which is highly poetic. Third, *The Birthday Party* is one of the most exemplary absurdist plays; additionally, Pinter's use of dialogue is characterized as "standard English, but the conversation doesn't get anywhere"

(Schechner 1966, 176): i.e., Pinter's dramatic dialogue adheres to certain established everyday paradigms, typically English *small-talk* clichés, in which certain regularities apply which, whether written or unwritten, are quite firmly rooted in the English tradition (Onič 2016). This makes Pinter's plays and *The Birthday Party*, in particular, ideal material for a pragmatic analysis that utilizes this speech act typology.

3.2 Methodology

The purpose of the paper is to establish which AI tool is best suited for analysing and annotating a text, in our case a play, using a finite speech act typology developed by Edmondson, House, and Kádár (Edmondson and House 1981; Edmondson, House, and Kádár 2023). We chose this speech act typology because, compared to other speech act typologies, it is finite, which prevents the invention of new speech acts. This allows for comparison of different texts and ensures replicability of the research thus produced (Kádár et al. 2024). Furthermore, the typology has been widely and successfully used in pragmatics, as evidenced by the influential works of various authors in the field (House 1996; Edmondson, House, and Kádár 2023; Taguchi and Kádár 2025).

To determine the best tool for the job, we compare the results (the annotated play) of three AI tools available on the market to a manual annotation of *The Birthday Party*. The AI tool producing an AI-generated version with the least discrepancies compared to the manual annotation will be considered the most appropriate; our methodology is thus fundamentally contrastive. For the annotation, we chose the following AI tools: ChatGPT (OpenAI), Gemini (Google), and DeepSeek. While both ChatGPT and Gemini use a transformer-based architecture, they nevertheless utilize different training data: ChatGPT uses a massive dataset of both text and human-annotated examples, whereas Gemini uses a proprietary dataset, curated by Google. Additionally, the two have different strengths: advanced language understanding in Gemini's case and exceptional conversational ability in the case of ChatGPT (Rane, Choudhary, and Rane 2024). Both language understanding and conversational ability are variables relevant to pragmatics and might explain the differences in the final output. The final tool, DeepSeek, was added because of its lower cost of development and usage compared to ChatGPT and Gemini (DeepSeek-AI et al. 2025), even though it retains their capabilities and uses a new architecture – a more collective approach that uses a mixture of specialized neural networks that work in conjunction and not a massive, unified AI system (Moors 2025). Furthermore, since DeepSeek is open-source and relatively easier to run because of its lower resource usage, it would be much easier to run it locally and thus avoid various privacy and security concerns related to using generative AI tools. All the selected tools have different context windows: ChatGPT supports contextual lengths of up to 128000 tokens, Gemini up to one million tokens, and DeepSeek up to 163840 tokens. This allows for the processing of very long documents or complex conversations while preserving the full context, which is ideal for our study and, if our study is successful, for analysis of entire corpora of texts using this approach (e.g., hundreds of absurdist plays).

To simultaneously determine the best procedure (prompt engineering) for producing optimal output (an annotation of *The Birthday Party* that deviates the least from the manual

annotation), we tested the three tools in three different scenarios. In the first scenario (1), we instructed the chatbot to research the specific speech act typology online, then we uploaded a sample of manual annotation (approximately 150 lines) for the chatbot to analyse, and we finally instructed the chatbot to produce its own annotation of the remainder of Act one – the play was provided in .docx format, with each utterance in the play numbered and on a new line; the manual annotation was annotated in the same manner. In the second scenario (2), we did not instruct the chatbot to research the typology but instead provided a short description of the typology (approximately 20 pages) and speech acts from a referential work (Edmondson, House, and Kádár 2023) alongside the same sample of manual annotation that we used in the first scenario. The remaining instructions for the chatbot were the same: to produce its own annotation of the remainder of Act one. The final, third scenario (3), differed from the second in that we uploaded a much more exhaustive and comprehensive description of the speech act typology (approximately 80 pages) from the same source (Edmondson, House, and Kádár 2023), while the other steps remained the same as in the second scenario.

We wanted to utilize the best available iterations of GenAI in use; however, different iterations have different capabilities: for instance, some allow uploading of texts and files, and some do not; some can research content online, while others cannot. To address such discrepancies, we had to slightly modify our prompts for each specific GenAI model used. Considering the rapid development of new iterations, we find it extremely relevant to mention the specific iterations used in our study and their capabilities at this time.

In testing ChatGPT, we intended to utilize the o1 iteration, which generates longer trains of thought before providing an answer (Wang et al. 2024). We decided against using ChatGPT o3, based on various technical and deployment-related parameters. Although ChatGPT o3 is a newer model, it is currently a smaller, more latency- and cost-optimized "mini" version, which makes it less appropriate for complex linguistic and pragmatic tasks. In recent comparative studies (Raffel et al. 2023), such smaller and/or more resource-efficient models have yielded lower performance in deeper discourse understanding and reduced capacity for long-term context retention in comparison to larger models like o1 or GPT 4o. Furthermore, ChatGPT o1 supports a larger number of parameters dedicated to more advanced forms of reasoning and linguistic understanding (Wang et al. 2024), which we believe is crucial for tasks such as speech act classification. However, ChatGPT o1 cannot currently research topics online, so we could not use it for Scenario 1. Instead, we used ChatGPT 4o for Scenario 1 and used the more powerful ChatGPT o1 for Scenario 2 and Scenario 3 (this is mentioned in the results). Nor can ChatGPT o1 read PDFs or documents, so the materials (manual annotation and the long and short theory) were provided in the prompt itself.

In testing Gemini, we intended to utilize the 2.0 PRO Experimental iteration, yet, similar to ChatGPT, different iterations offer different capabilities. We chose the Gemini 2.0 PRO Experimental because, according to Google's own documentation and independent analyses (Gemini Team et al. 2024; Chowdhery et al. 2022), it produces more advanced results in tasks related to logical reasoning and extended context retention – both of which are crucial for speech act analysis. However, because Gemini 2.0 PRO Experimental does not

currently support online research, we utilized Gemini 2.0 Flash for Scenario 1. In terms of technical specifications and reasoning capabilities, Gemini 2.0 PRO Experimental is also the closest model to ChatGPT o1, making a direct comparison of their outcomes the most methodologically sound approach. This enabled us to rule out effects stemming from significant differences in architecture or dataset size and focus instead on the models' actual ability to classify speech acts. However, it has the same limitations as ChatGPT o1 – it cannot research topics online or read PDFs or documents. Similarly to prompt-engineering for ChatGPT, we adopted the prompts for Scenario 2 and Scenario 2 for Gemini 2.0 PRO Experimental by providing the materials in the prompt itself. For Scenario 1, we used Gemini 2.0 Flash, which can research topics online and read documents, so the prompt was not further adapted.

In testing DeepSeek, we utilized the DeepSeek-R1 iteration, which is an open-source model that enables both researching online, and the upload of PDFs and DOCX files. or documents (DeepSeek-AI et al. 2025; Mercer, Spillard, and Martin 2025). Furthermore, DeepSeek-R1is, compared to ChatGPT o1 and Gemini 2.0 PRO Experimental, free to use and requires fewer resources to operate. Like ChatGPT o1 and Gemini 2.0 PRO Experimental, it offers enhanced reasoning capabilities and can process longer texts, making a direct comparison among the three models (ChatGPT o1, Gemini 2.0 PRO Experimental, and DeepSeek-R1) fully justified. Another advantage of DeepSeek-R1 is its open-source nature and relatively low computational demands, allowing for simpler local deployment and thus direct protection of sensitive data (Mercer, Spillard, and Martin 2025). According to published benchmarks (DeepSeek-AI et al. 2025), DeepSeek-R1 achieves statistically similar results to closed-source solutions on comparable text-intensive tasks, i.e., it should deliver at least an equivalent level of accuracy for speech act classification, while being free to use and requiring fewer resources to operate compared to ChatGPT o1 and Gemini 2.0 PRO Experimental.

For the quantitative comparison between the manually annotated classification (reference text) and the AI-generated classifications, we developed a Python script that automatically performs the following:

- Reads .docx files containing both the reference classification and the AI-generated classifications.
- Compares the corresponding speech act for each line/utterance in the dialogue.
- Measures the similarity between sentences (using SequenceMatcher and the Hungarian algorithm[1] for optimal line alignment).
- Aligns the reference and the AI-predicted speech act types and calculates the number of mismatches.
- Calculates the classification accuracy, i.e., the percentage of lines that were annotated correctly compared to the reference.

[1] Also known as the Kuhn-Munkres algorithm (Kuhn 1955) – this is a classic algorithm for solving the assignment problem in combinatorial optimization, where the best match between elements of two sets is sought to maximize the total similarity or minimize the total distance.

4 Results and Discussion

We have decided to segment the results into two sections. In the first part, we will present statistical data for all three scenarios, analysing the overall success rate of all three GenAI in this task and the general trends. In the second part, we will focus on a more detailed analysis of the mistakes and the general trends of those mistakes in the most successful chatbot, with specific examples from each speech act category.

Unsurprisingly, the chatbots were least successful in Scenario 1, where they were instructed to conduct online research on the speech act typology and, with the help of the manually annotated example, annotate the text. Our main concern was that the chatbots would not be able to differentiate between the various speech act typologies online and choose the one we prescribed. We hoped that providing the manually annotated example would ground the chatbots and steer them to the correct speech act typology; unfortunately, this was not the case. Both ChatGPT 4o and Gemini 2.0 Flash utilized speech acts that were outside the prescribed typology, with ChatGPT classifying 537 utterances and Gemini 2.0 Flash classifying fifty-six utterances with categories outside the classification. Surprisingly, DeepSeek-R1 managed to utilize the correct typology, yet its accuracy was still only 29%. We believe this result can be explained by the fact that, for Scenario 1, we were forced to use fewer capable iterations of ChatGPT and Gemini (ChatGPT 4o instead of ChatGPT o1 and Gemini 2.0 Flash instead of Gemini 2.0 PRO Experimental) because of prompt limitations: the more capable models do not yet have online research enabled. We conjecture that using the more capable models would improve usage of the correct typology, considering the similarities in the results in other scenarios.

That being said, the results of Scenario 1 indicate a clear winner, which was Gemini 2.0 Flash. Despite using a less capable model and despite assigning fifty-six utterances to speech act categories outside the prescribed technology, Gemini 2.0 Flash achieved an accuracy of 63%, which is more than twice as good as the other two chatbots, as shown in the table below.

Table 1. Accuracy of Chatbots in Scenario 1 (autonomous online research + manually annotated example).

	Total speech acts	Correct classifications	Mismatched classifications	Accuracy (%)
ChatGPT 4o	762	201	561	26%
Gemini 2.0 Flash	762	480	282	63%
DeepSeek-R1	762	221	541	29%

Furthermore, detailed examination of the results for Gemini 2.0 Flash in Scenario 1 indicates that it might have performed even better. It misclassified thirty-three instances of speech act Request as Command. Commands are not in our speech act typology, yet they do belong under Request, so it could be argued that Gemini 2.0 Flash still correctly recognized the pragmatic intent behind these misclassifications. On the other hand, it misclassified relatively "easy" categories: for example, it misclassified all four instances of Leave-Take, which is a ritualistic speech act that signifies the termination of an encounter between two speakers.

This is usually performed via tokens such as "Good night," "Bye," "See you," "Cheerio," etc. Nevertheless, Gemini 2.0 Flash classified three Leave-takes, "Ta-ta, Mrs. Boles," Ta-ta," and "Ta-ta, Stan," as a Greet, a speech act utilized for acknowledging the presence of the interlocutor ("Hello," "Hi," etc.), and one Leave-take, "See you later," as a Resolve (illocution used to express the speaker's actions). Furthermore, it classified an instance of How-are-you ("How are you keeping, Mrs Boles?"), another Ritualistic speech as, as a Request.

When using chatbots to facilitate the analysis of researchers, we would require them to at least identify the "easy" ritualistic speech acts, such as Leave-take and How-are-you, which are codified only by a few almost universal utterances. The fact that Gemini 2.0 Flash, as the most successful GenAI in Scenario 1, failed to do that, and in conjunction with the fact that it only achieved 63% total accuracy (still impressive, especially considering the much lower accuracy of ChatGPT 4o and DeepSeek-R1), means that using this scenario would not aid the researcher in their work.

Scenario 2 was much more successful, producing the most accurate classifications of all three scenarios in all chatbots. Chatbots were instructed to use only a short excerpt of the theory, which resulted in no "phantom" speech act classifications – all three chatbots used only the appropriate twenty-five speech acts from the finite speech act typology in their annotation. All three chatbots were comparable in their results, with ChatGPT o1 emerging on top, having achieved an impressive accuracy of 81%, followed closely by DeepSeek-R1 with 79% accuracy, and Gemini 2.0 PRO Experimental with 75% accuracy. The table below summarizes the success rate of each chatbot.

Table 2. Accuracy of Chatbots in Scenario 2 (short theory + manually annotated example).

	Total speech acts	**Correct classifications**	**Mismatched classifications**	**Accuracy (%)**
ChatGPT o1	762	623	139	82%
Gemini 2.0 PRO Experimental	762	575	187	75%
DeepSeek-R1	762	600	162	79%

ChatGPT o1, the most successful chatbot in this scenario and the most successful overall, misclassified only 139 speech acts. Of those, it struggled the most with *Opines*, which were classified as Tells (39), Remarks (4), Suggests (4), and Complains (2); *Resolves*, which it classified as Tells (12), Remarks (2), Requests (2), Willings (1), and Promises (1); *Complains*, misclassified as Requests (8), Remarks (4), Opines (3), and Tells (3); and *Remarks*, misclassified as Requests (7), Tells (2), Opines (2), and Complains (1). The entire table of misclassifications for ChatGPT o1 in Scenario 2 is presented below.

DeepSeek-R1 misclassified *Tells* as Discloses (15), Opines (7), Requests (2), Thanks (1), Remarks (1), and Complains; *Opines* as Tells (14), Complains (12), Requests (4), Remarks (3), Resolves (3), Suggests (1), Willings (1), and Discloses (1); *Complains* as Requests (12), Opines (10), Tells (4), and Discloses (1); and *Remarks* as Opines (11), Requests (8), Tells (5), and Resolves (2).

Table 3. Misclassifications of ChatGPT o1 in Scenario 2; the left column represents the misclassified reference category, the rows to the right show how ChatGPT o1 categorized them.

	Tell	Opine	Request	Remark	Suggest	Complain	Willing	Resolve	Promise	Total
Opine	39			5	4	2				**50**
Resolve	12		2	2			1		1	**18**
Complain	3	3	8	4						**18**
Remark	2	2	7			1	1			**13**
Request	4	2		2		1		2		**11**
Tell		4		3					1	**8**
Minimise	3	2	1							**6**
Suggest			5							**5**
Thanks	1	3								**4**
Disclose	2									**2**
Willing	1									**1**
Invite	1		1							**2**
Greet				1						**1**
Total										**139**

Gemini 2.0 PRO Experimental struggled the most with *Opines*, which were often misclassified as Tells (47), Discloses (3), Resolves (2), and Requests (1); *Complains*, which were misclassified as Opines (25), Tells (13), Requests (11), Remarks (2), and Discloses (1); *Remarks*, which were misclassified as Tells (17), Requests (11), and Opines (4); and *Resolves*, which were misclassified as Tells (9), Requests (4), and Willings (1).

We note that similar patterns emerge in all three scenarios, where the most misclassified speech acts were Opines, Tells, Complains, Resolves, and Remarks. This is unsurprising, considering the nature of such speech acts. The delineation between Opines and Tells, for example, is largely subjective (Edmondson, House, and Kádár 2023, 169), and the deciding factor is usually the person annotating the text, who "decides" on some criteria (note that this is not fatal for the methodology, as long as the criteria are applied consistently). This means that if the reference text was annotated differently (yet still consistently), the chatbots' success rate order might have been reversed. The accuracy would remain in the same range, as the differences between individual styles of annotation would average out. Overall, the results of Scenario 2 represent a (surprisingly) stellar result. At 75-82% accuracy, all three chatbots' performances could be used for conducting a preliminary classification of speech acts, which would facilitate the workload of researchers conducting speech act analysis. Before we examine ChatGPT o1's results in more detail, a presentation of Scenario 3 results is in order.

The results of Scenario 3 were, surprisingly, slightly worse than the results for Scenario 2. While the difference was marginal (a few percentage points, as indicated in the table below), it was detectable in all three instances.

Table 4. Accuracy of Chatbots in Scenario 3 (long theory + manually annotated example).

	Total speech acts	Correct classifications	Mismatched classifications	Accuracy (%)
ChatGPT o1	762	611	146	81%
Gemini 2.0 PRO Experimental	762	563	199	74%
DeepSeek-R1	762	579	183	76%

We conclude that Scenario 1 was not as effective because of the sheer amount of data the chatbots found online, which resulted in conflicting typologies being applied (as evidenced in the use of classifications that were not in the proposed finite typology), yet Scenario 3 suffered from a similar shortcoming. An explanation of this worse outcome might be readily available in studies examining the empirical relationship between dataset size, model size, and compute power in GenAI. These studies establish that while increasing dataset size improves model accuracy, the benefits taper off beyond a certain threshold (Kaplan et al. 2020; Chowdhery et al. 2022). At that point, models may begin to overfit. Overfitting is a phenomenon in machine learning that occurs when a model fits the training data too well or even exactly, which results in worse performance on any new or unseen data. To further improve the results, techniques such as regularization, pruning, and dropout to mitigate performance degradation might be required (Kaplan et al. 2020; Chowdhery et al. 2022). In Scenario 3, where we provided an exhaustive 80-page theoretical background, the model might have exhibited a tendency towards overfitting, making unnecessary distinctions and misclassifying instances. This mirrors findings in large-scale model training, where excessive data can paradoxically lead to poorer performance because of increased memorization. This aligns with other studies on machine learning, for example, findings from Kaplan et al. (2020), which highlight diminishing returns when dataset size surpasses a certain threshold.

Still, both Scenario 3 and Scenario 2 produced results in the range of approximately 75-82%, which makes them appropriate for research facilitation and, at the very least, preliminary annotation of data. Indeed, we would argue that the results are even better than purely statistical data shows. A more qualitative approach to the results of the best performing chatbot, ChatGPT o1 in Scenario 2, confirms that assertion. We can demonstrate this by examining and contextualizing the kinds of mistakes the chatbot made in individual categories.

Opines

Overall, ChatGPT o1 misclassified 50 Opines; however, thirty-nine of those were misclassified as Tells. The delineation between Opines and Tells is subjective, so different manual annotations might yield even higher accuracy for ChatGPT o1. In fact, from a research perspective, it would be useful to instruct the chat to mark any instance of Opine or Tell as Opine/Tell, and the researcher could then produce a more fine-grained verdict based on the needs of the project. In the case of *The Birthday Party*, distinguishing between Opines and Tells proves to be especially difficult, as characters often formulate their opinions as facts, for example, in the exchanges between Stanley and Meg.

Example 1.

STANLEY. The milk's off.	***Opine/Tell***
MEG. It's not.	***Opine/Tell***

Example 2.

MEG. Perhaps they couldn't find the place in the dark.	*Opine/Opine*
It's not easy to find in the dark.	***Opine/Tell***
STANLEY. They won't come.	*Opine/Opine*
Someone's taking the Michael.	*Opine /Complain*
Forget all about it.	*Request/Resolve*
It's a false alarm.	***Opine/Tell***
A fake alarm.	***Opine/Tell***

In example 1, ChatGPT o1 marked both utterances as Tell; at least for Stanley's utterance, this might be correct in certain cases. As annotators, we decided to mark this as Opine because it is a statement that Stanley and Meg dispute and because of the broader context – Stanley's badgering of Meg. However, Stanley formulates the utterance as a fact (Tell), so one could also adopt a different criterion and classify it as a Tell. Similarly in example 2, the line "It's not easy to find in the dark" could also be classified as a Tell if taken out of context, but we classified it as an Opine because Meg was continuing her speculation from the utterance before ("Perhaps they couldn't find the place in the dark"). Similarly for Stanley's "It's a false alarm": in most cases, this would be considered a Tell, but because we know from earlier that this is Stanley continuing his speculation, we label it as Opine. So, the misclassifications of chatbots are often related to the broader context and actual meaning of the text, which chatbots have not (yet) mastered.

Resolves

Most Resolves were also mislabelled as Tells, especially in instances when a Resolve followed a Request in Initiate-Satisfy pattern. Requests are often satisfied with either a Resolve or a Tell, and the chatbot had trouble differentiating between the "No" of Tell (Did you know? No.) and the "No" of Resolve (Come here. No.), as in Example 3 and 4.

Example 3.

GOLDBERG. Well, of course, you must have one.	
(He stands.) We'll have a party, eh?	
What do you say?	*Request/Request*
MEG. Oh yes!	***Resolve/Tell***

Example 4.

MEG. What do you mean?/	*Request/Request*
STANLEY. Come over here./	*Request/Request*
MEG. No.	***Resolve/Tell***

Complains

Complains were most often misclassified as Requests. As can be seen in examples 5 and 6 below, Complains in the text were often formulated as Requests, so we can see why the chatbots classified them as such. Only the broader context (mostly utterances before and after the Complain) determines that it is, in fact, a Complain. As in the case of Opines, chatbots were missing this additional context, a deficiency that might be rectified in further studies via smart prompt-engineering.

Example 5. (Lulu is scolding Stanley):

LULU: I mean, what do you do, just sit around the house like this all day long?

Complain/Request

Hasn't Mrs Boles got enough to do without having you under her feet all day long?

Complain/Request

Example 6.

MCCANN. Sure I trust you, Nat.

GOLDBERG. But why is it that before you do a job you're all over the place, and when you're doing the job you're as cool as a whistle?

Complain/Request

MCCANN. I don't know, Nat.

Remarks

Interestingly, Remarks were most often misclassified as Requests. Remarks are highly ritualistic speech acts, while Requests are substantive speech acts, so the discrepancy is worth addressing. In our manual annotation, we annotated utterances like Example 7 Remarks, as they were often followed by a more substantive Request and they function more as Remarks in the dialogue (Meg replies to the Requests, not Remarks); however, a different researcher might, like chatbots, interpret them as very mild Requests.

Example 7.

What's his name?/	*Request/Request*
MEG. Stanley Webber./	*Tell/Tell*
GOLDBERG. Oh yes?/	***Remark/Request***
Does he work here?/	*Request/Request*
MEG. He used to work./	*Tell/Tell*
He used to be a pianist./	*Tell/Tell*
In a concert party on the pier./	*Tell/Tell*
GOLDBERG. Oh yes?/	***Remark/Request***
On the pier, eh?/	***Remark/Request***
Does he play a nice piano?/	*Request/Request*
MEG. Oh, lovely./	*Opine/Opine*

Requests

Chatbots had the most difficulties recognizing requests that were not in a question form, which they classified as Tells, as we can see in Examples 8 and 9. However, all chatbots had remarkable overall results in terms of Requests. ChatGPT o1 correctly identified 242 out of 253 Requests, with an accuracy of 96%. This might also be because Requests are often in question forms, so they are relatively easy to recognize.

Example 8.

GOLDBERG. You know what I said when this job came up.	***Request/Tell***
I mean naturally they approached me to take care of it.	*Tell/Tell*
And you know who I asked for?	*Request/Request*
MCCANN. Who?	*Request/Request*

Example 9.

MEG. He hasn't mentioned it.	*Tell/Tell*
GOLDBERG (thoughtfully). Ah!	***Remark/Request***
Tell me.	***Request/Tell***
Are you going to have a party?	*Request/Request*
MEG. A party?	*Request/Request*

Other Speech Acts

Other speech act categories yielded less than 10 misclassifications across the entire text. Furthermore, the reasoning behind the mistakes is often like the above: the chatbots were unable to recognize the pertinent context. Some Tells, for example, were misclassified as Opines, usually because of emotive language in the utterances. Interestingly, none of the chatbots (excluding one count in case of Gemini 2.0 PRO experimental in Scenario 2), managed to recognize any of the Minimizes in the play, which were usually misclassified as Opines or Tells. Suggests were misclassified as Requests, which is not surprising, considering it is sometimes difficult to articulate the difference between the two. We use the criterion of benefit for the speaker for Requests and benefit for the hearer for Suggest, but more complex cases, which might benefit both the speaker and the hearer, complicate things and require additional *ad hoc* criteria. Thanks was misclassified as Opine or Tell, which is also due to a failure to grasp the necessary context of the text. On the bright side, Leave-Takes, Welcomes, and How-are-yous were classified with 100% accuracy by all chatbots in Scenarios 2 and 3. All chatbots also had a high accuracy in recognizing other ritualistic speech acts, such as Greets, yet only ChatGPT o1 (in Scenario 2 and 3) managed to recognize one instance of another ritualistic speech act, Extractor.

5 Conclusion

The purpose of this article was to determine the viability of using different GenAI chatbots for automated speech act annotation of texts for pragmatic purposes. We sought to answer the following questions:

1. How successful are chatbots in providing an automatically annotated text in line with the instructed speech act typology?
2. Which scenario yields the best outcome (the highest fidelity to manual annotation)?
3. Are chatbots useful for performing such preliminary annotations or, at the very least, facilitating this process?

In line with this, we can draw the following conclusions. Chatbots were (1) highly successful at annotating the text with the prescribed typology, yielding 75-82% accuracy; however, (2) the prompt-engineering that instructs the chatbots does matter: Scenario 1 offered only 26% (ChatGPT 4o), 29% (DeepSeek-R1), and 63% (Gemini 2.0 Flash) accuracy. Chatbots should therefore be provided with a much smaller reference frame (data provided) within which to operate. Furthermore, more is not always better: the more detailed theory in Scenario 3 yielded slightly worse results, though still useful. We believe this to be the result of overfitting, which is consistent with results from other studies on machine learning, which indicate diminishing returns when a dataset surpasses a certain threshold (Kaplan et al. 2020). Whether the accuracy could be further improved is subject to further studies, which should experiment with different prompts, as well as with the quantity and perhaps quality of the theory provided to the chatbots. Finally, we believe that the results warrant a tentative conclusion that (3) chatbots, using prompts such as Scenario 2, can be useful and can facilitate research in pragmatics by providing an automated preliminary annotation of the text. That being said, further research is needed, especially in terms of how the chatbots would perform in annotating texts that require further context, such as historical documents. One limitation of this study was that we tested the chatbots on a play, which typically includes the relevant context for the viewer/reader, whereas for annotation of a historical text, we require further historical context to properly classify utterances. Whether chatbots are capable of that is subject to further research. Furthermore, it would be useful to test the capabilities of chatbots in annotating and classifying texts in other domains of pragmatics and linguistics in general, such as gambits or ritual frame indicating expressions. Considering the relative similarity between the tasks, our approach could be beneficial in these areas as well.

References

Altmäe, Signe, Alberto Sola-Leyva, and Andres Salumets. 2023. "Artificial intelligence in scientific writing: A friend or a foe?" *Reproductive BioMedicine Online* 47 (1): 3–9. https://doi.org/10.1016/j.rbmo.2023.04.009.

Chen, Xi, Jun Li, and Yuting Ye. 2024. "A feasibility study for the application of AI-generated conversations in pragmatic analysis." *Journal of Pragmatics* 223:14–30. https://doi.org/10.1016/j.pragma.2024.01.003.

Chowdhery, Aakanksha, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, et al. 2022. "PaLM: Scaling language modeling with pathways." *arXiv*. https://doi.org/10.48550/arXiv.2204.02311.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, et al. 2025. "DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning." *arXiv*. https://doi.org/10.48550/ARXIV.2501.12948.

Edmondson, Willis, and Juliane House. 1981. *Let's Talk, and Talk about It: A Pedagogic Interactional Grammar of English*. Urban & Schwarzenberg.

Edmondson, Willis J., Juliane House, and Daniel Z. Kádár. 2023. *Expressions, Speech Acts and Discourse: A Pedagogic Interactional Grammar of English*. Cambridge University Press.

Else, Holly. 2023. "Abstracts written by ChatGPT fool scientists." *Nature* 613 (7944): 423. https://doi.org/10.1038/d41586-023-00056-7.

Gavez, Urša. 2016. "The reception of Harold Pinter's plays in Slovenia between 1999 and 2014." *ELOPE: English Language Overseas Perspectives and Enquiries* 13 (2): 51–61. https://doi.org/10.4312/elope.13.2.51-61.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, et al. 2024. "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context." *arXiv.* https://doi.org/10.48550/ARXIV.2403.05530.

Hazemali, David, Janez Osojnik, Tomaž Onič, Tadej Todorović, and Mladen Borovič. 2024. "Evaluating chatbot assistance in historical document analysis." *Moderna arhivistika* 7 (2): 53–83. https://doi.org/10.54356/ma/2024/biub3010.

House, Juliane. 1996. "Developing pragmatic fluency in English as a foreign language: Routines and metapragmatic awareness." *Studies in Second Language Acquisition* 18 (2): 225–52. https://doi.org/10.1017/S0272263100014893.

House, Juliane, and Dániel Z. Kádár. 2023. "Studying small talk from a pragmatic angle: An introduction." *Acta Linguistica Academica* 70 (4): 411–18. https://doi.org/10.1556/2062.2023.00704.

House, Juliane, Dániel Z. Kádár, Tadej Todorović, Matjaž Klemenčič, David Hazemali, Tomaž Onič, and Katja Plemenitaš. 2024. "Capturing power in diplomatic language use: The case of a closed-door mediatory negotiation and its aftermath during the breakup of the former Yugoslavia." *Journal of Language and Politics*. https://doi.org/10.1075/jlp.24036.hou.

Hribar, Darja. 2004. "Harold Pinter in Slovene translation." *ELOPE: English Language Overseas Perspectives and Enquiries* 1 (1–2): 195–208. https://doi.org/10.4312/elope.1.1-2.195-208.

Kádár, Dániel Z., Juliane House, Tadej Todorović, Tomaž Onič, David Hazemali, Katja Plemenitaš, and Donathan Brown. 2024. "The language of diplomatic mediation – A case study of an emergency meeting in the wake of the Yugoslav wars." *Language & Communication* 96: 54–66. https://doi.org/10.1016/j.langcom.2024.02.004.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. "Scaling laws for neural language models." *arXiv*. https://doi.org/10.48550/arXiv.2001.08361.

Kuhn, Harold William. 1955. "The Hungarian method for the assignment problem." *Naval Research Logistics Quarterly* 2 (1–2): 83–97.

Liu, Shiyu, Juliane House, and Dániel Z. Kádár. 2024. "Bargaining in Chinese livestream sales events." *Discourse, Context & Media* 60:100787. https://doi.org/10.1016/j.dcm.2024.100787.

Mercer, Sarah, Samuel Spillard, and Daniel P. Martin. 2025. "Brief analysis of DeepSeek R1 and its implications for generative AI." *arXiv.* https://doi.org/10.48550/ARXIV.2502.02523.

Mohar, Tjaša, Sara Orthaber, and Tomaž Onič. 2020. "Machine translated Atwood: Utopia or dystopia?" *ELOPE: English Language Overseas Perspectives and Enquiries* 17 (1): 125–41. https://doi.org/10.4312/elope.17.1.125-141.

Monteiro, Mateus De Souza, Vinícius Carvalho Pereira, and Luciana Cardoso De Castro Salgado. 2023. "Investigating Politeness strategies in chatbots through the lens of conversation analysis." In *Proceedings of the XXII Brazilian Symposium on Human Factors in Computing Systems*, Maceió, Brazil, 1–12. Association for Computing Machinery. https://doi.org/10.1145/3638067.3638068.

Moors, Sarah. 2025. "DeepSeek changes everything we thought we knew about building smart machines." *Digital Health Insights*, January 29. https://www.dhinsights.org/news/deepseek-changes-everything-we-thought-we-knew-about-building-smart-machines..

Onič, Tomaž. 2016. "Slogovne značilnosti … [premolk] … Pinterjevega dialoga." *Primerjalna književnost* 39 (2). https://ojs-gr.zrc-sazu.si/primerjalna_knjizevnost/article/view/6367.

Onič, Tomaž, and Nastja Prajnč Kacijan. 2020. "Repetition as a means of verbal and psychological violence in interrogation scenes from contemporary drama." *Ars & Humanitas* 14 (1): 13–26. https://doi.org/10.4312/ars.14.1.13-26.

Pinter, Harold. 1991. *Plays One*. Faber & Faber.

Podbevšek, Katarina, and Nina Žavbi. 2021. "Jezikovna norma v luči odrske govorne estetike." *Jezik in Slovstvo* 66 (2–3): 145–56. https://doi.org/10.4312/jis.66.2-3.145-156.
Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. "Exploring the limits of transfer learning with a unified text-to-text transformer." *arXiv.* https://doi.org/10.48550/arXiv.1910.10683.
Rane, Nitin, Saurabh Choudhary, and Jayesh Rane. 2024. "Gemini versus ChatGPT: Applications, performance, architecture, capabilities, and implementation." *SSRN Scholarly Paper*. Social Science Research Network. https://doi.org/10.2139/ssrn.4723687.
Schechner, Richard. 1966. "Puzzling Pinter." *The Tulane Drama Review* 11 (2): 176–84. https://doi.org/10.2307/1125196.
Setlur, Vidya, and Melanie Tory. 2022. "How do you converse with an analytical chatbot? Revisiting Gricean maxims for designing analytical conversational behavior." *arXiv*. https://doi.org/10.48550/ARXIV.2203.08420.
Stokel-Walker, Chris. 2023. "ChatGPT listed as author on research papers: Many scientists disapprove." *Nature* 613 (7945): 620–21. https://doi.org/10.1038/d41586-023-00107-z.
Taguchi, Naoko, and Dániel Z. Kádár. 2025. "Pragmatics: An overview." In *The Encyclopedia of Applied Linguistics*, edited by Carol A. Chapelle, 1st ed., 1–8. Wiley. https://doi.org/10.1002/9781405198431.wbeal1338.pub2.
Wang, Kevin, Junbo Li, Neel P. Bhatt, Yihan Xi, Qiang Liu, Ufuk Topcu, and Zhangyang Wang. 2024. "On the planning abilities of OpenAI's O1 models: Feasibility, optimality, and generalizability." *arXiv*. https://doi.org/10.48550/ARXIV.2409.19924.
Williams, Iwan, and Tim Bayne. 2024. "Chatting with bots: AI, speech acts, and the edge of assertion." *Inquiry*: 1–24. https://doi.org/10.1080/0020174X.2024.2434874.