LLMs of Catan: Exploring Pragmatic Capabilities of Generative Chatbots Through Prediction and Classification of Dialogue Acts in Boardgames' Multi-party Dialogues

Andrea Martinenghi[1], **Gregor Donabauer**[2], **Simona Amenta**[1], **Sathya Bursic**[1], **Mathyas Giudici**[3], **Udo Kruschwitz**[2], **Franca Garzotto**[3], **Dimitri Ognibene**[1]
[1]Department of Psychology, University of Milano-Bicocca, Milan, Italy
[2]Information Science, University of Regensburg, Regensburg, Germany
[3]Polytechnic University of Milan, Milan, Italy
a.martinenghi1@campus.unimib.it, dimitri.ognibene@unimib.it

Abstract

Human language interactions involve complex processes beyond pure information exchange, for example, actions aimed at influencing beliefs and behaviors within a communicative context. In this paper, we propose to investigate the dialogue understanding capabilities of large language models (LLMs), particularly in multi-party settings, where challenges like speaker identification and turn-taking are common. Through experiments on the game-based STAC dataset, we explore zero and few-shot learning approaches for dialogue act classification in a multi-party game setting. Our intuition is that LLMs may excel in tasks framed through examples rather than formal descriptions, influenced by a range of pragmatic features like information presentation order in prompts and others. We also explore the models' predictive abilities regarding future dialogue acts and study integrating information on dialogue act sequences to improve predictions. Our findings suggest that ChatGPT can keep up with baseline models trained from scratch for classification of certain dialogue act types but also reveal biases and limitations associated with the approach. These insights can be valuable for the development of multi-party chatbots and we try to point out directions for future research towards nuanced understanding and adaptation in diverse conversational contexts.

Keywords: Multi-Party Dialogue, Game-Based Conversations, Dialogue Act Classification

1. Introduction

Human language-based interactions are not simply the mere delivery of messages and information but complex multilevel processes. In a pragmatic framework, every time two or more individuals are involved in a communicative interaction, they are performing actions: from sharing information to actually inducing or modifying another person's beliefs and behaviors (Austin, 1975).In this perspective therefore, an utterance is produced by a speaker within a specific communicative context and responds to certain communicative intentions (i.e. the objectives that one intends to achieve through one's utterance, for example to convince, explain, ask, swear, etc.). On the recipient's side, to comprehend and interpret communicative messages, a person engages in complex inferential processes aimed at understanding the communicative intention of the interlocutor.

Corpora from multi-party games offer unique opportunities to study these processes. First, games can deliver interactions in a natural environment, which game engines can record along with other happenings, making it possible to study precise connections between the players' utterances, the context, and their general strategies (Djalali et al., 2011). Second, chats based on games are ideal because they approximate spoken language without the need of transcription, and they manifest phenomena particular to multilogue, such as multiple conversation threads (Afantenos et al., 2015). Third, multi-party settings are particularly relevant, as humans tend to work in groups and teams, and both models and methods based on them provide unique challenges compared to two-party systems (Mahajan and Shaikh, 2021). Among these, speaker identification, turn-taking and tailoring the content of the response to each agent or person (Sibun, 1997).

This setting allows studying ChatGPT (Ouyang et al., 2022) and other generative chatbots-(GCBs)' understanding of dialogues and their ability to generalize to new contexts. In particular, most GCBs have been tuned for two-party dialogue. Their evident ability to participate in such interactions is matched by the difficulty of extracting any internal representation of the underlying skills or causes of occasional failures (Mahowald et al., 2024; Borji, 2023; Koyuturk et al., 2023). Having only an empirical appreciation of these skills and the end-to-end nature of the training of these systems together with the complexity of multi-party dialogue call for a nuanced and theory-based approach to study GCBs' capabilities and their ability to adapt to new contexts.

Previous studies tested the pragmatic skills of GCBs and their ability to engage in interactions

Games and NLP @LREC-COLING-2024, pages 107–118

and dialogues, especially in two-parties dialogues (but see (Wei et al., 2023)). While Barattieri di San Pietro et al. (2023) applied standardized psychological tests for pragmatic skills evaluation to GCBs and Ruis et al. (2024) work on conversational implicatures, we focus our study on the explicit recognition and prediction of speech acts (Searle, 1969) or dialogue acts (DAs) in multi-party settings (see section 2.1 for a review of this approach). Notwithstanding the observed dialogue skills, DAs are deemed to be challenging for GCBs (Gubelmann, 2024; Brown et al., 2020). The critique to GCBs' conversational and dialogue understanding is not new (Bender and Koller, 2020; Bender et al., 2021). Testing their performance in generalizing to the novel multiparty dialogue could contribute to this discussion. However, only a few works are present in the literature. Wei et al. (2023) implement a type of training under multi-party conditions which prevents studying the transfer of dialogue skills (and understanding) from two-party to multi-party. In (Chan et al., 2023), the focus is on sentence-level relationship parsing, which could not directly support language production and poses substantial complexities also to humans.

We investigate zero or few-shot learning approaches for classifying and predicting DAs in the game-based multi-party dataset STAC (Asher et al., 2016). Our study aims to explore the impact of example-based task formulation and pragmatic features on the performance of Game-Playing Chatbots (GCBs). We also examine the predictive capabilities of GCBs in forecasting future DAs and explore methods to incorporate information on the statistical distribution of DA sequences for improved predictions. Additionally, we analyze the coherence between DA and utterance wording prediction, considering the potential impact of disturbances on DA classification. Joint prediction of text and DA may enhance performance, but could also increase task complexity and affect results. Our study contributes to the understanding of how different dimensions of zero or few-shot learning approaches can enhance the classification and prediction of DAs in multi-party interactions.

The insights coming from our experiments will inform future development of multiparty chatbots based on similar few-shots approaches. The potential shown by this low-cost solution can also provide information on the challenges and opportunities for architectural (Wahlster, 2023) and learning-based approaches directing on the selection paradigm with different costs, e.g. full-training (Wei et al., 2023), fine-tuning (Ruis et al., 2024) or LoRa (Wang et al., 2023). In the spirit of *Games and NLP @ LREC-COLING* we provide all of our implementations as well as detailed results to the community to help reproducing our work[1].

2. Related Work

2.1. Dialogue Acts

When a person expresses an utterance, they are not only saying something: they are also *doing* something. This intuition that utterances possess both a descriptive and effective nature belongs to Austin (1975), who is considered the father of modern theory of speech acts. Austin (1975) formulated a theory of three kinds of acts: (1.a) *Locutionary acts*: acts *of* speaking, involved in the construction of speech; (1.b) *Illocutionary acts*: acts *in* speaking, concerning the meaning; (1.c) *Perlocutionary acts*: acts *by* speaking, relative to the consequences of speaking.

Following analysis and taxonomies of speech acts theory have focuses on Illocutionary acts and the role of intention versus that of convention (Horn and Ward, 2004). In the debate that followed, Grice (1957, 1975) was highly influential in suggesting that an utterance's power is to provide clues to the intention of the speaker. Searle (1969), while recognizing the importance of intentions in communication (which he confined to perlocutionary effects), pointed out that some communications succeed in virtue of the hearer's knowledge of certain rules governing the elements of the uttered sentence (illocutionary effects). Searle (1969)'s interpretation of the role of intention was aligned with Austin (1975)'s, and rejected by Strawson (2013), who argued that most commonplace speech acts succeed by producing the awareness that it was the speaker's intention to achieve a certain communicative goal.

Grice (1975)'s and Strawson (2013)'s recognition of intentions as such a central aspect of communication was continued by Bach and Harnish (1979), which was reflected in a greater consideration for the speaker's psychological state compared to Searle (1975). Their 4-classes taxonomy of Illocutionary acts, which along with Searle's taxonomy (1975) is among the most used in contemporary literature (e.g., Jurafsky and Martin (2024)), includes: (3.a) *Constatives*: committing the speaker to something's being the case (answering, claiming, confirming); (3.b) *Directives*: attempts by the speaker to get the addressee to do something (advising, asking, forbidding); (3.c) *Commissives*: committing the speaker to some future course of action (promising, planning, vowing); (3.d) *Acknowledgments*: express the speaker's attitude regarding the hearer with respect to some social action (apologizing, greeting, thanking).

[1] codebase at https://github.com/DimNeuroLab/llmGrounding

In pragmatics and in computational linguistics, such as in cuebased models, the term DA is often used as a synonymous of *speech act* (e.g., Jurafsky and Martin (2024); McTear (2022)), and a distinction is not fully clear. It was initially introduced into the field by Bunt (1981) "for referring to the functional units used by the speaker to change the context" (Bunt, 1994). Although sometimes equating DAs to speech acts, Jurafsky and Martin (2024) also outline a difference, describing DAs as the "combination of speech acts and grounding into a single representation of the interactive function of the turn or sequence" (Jurafsky and Martin, 2024). A third usage, that Horn and Ward (2004) suggest being the most used in cue-based literature, comes from Allen and Core (1997) to mean an act with internal structure related specifically to its dialogue function.

2.2. Multi-party Game Based Corpora

A growing body of corpora is based on games. In a survey on available corpora on multi-party dialogues (Mahajan and Shaikh, 2021), which with a sample size of over 300 publications is the only survey focusing on only multi-party corpora, games make up one of four categories of written corpora. *Settlers* (Afantenos et al., 2015) is the first published dataset with multiparty dialog discourse parsing (Asher et al., 2016), and prior to Molweni (Lin et al., 2020), the only one. Its content includes interesting features such as interleaved threads, creative language, and interactions between linguistic and extra-linguistic contexts (Asher et al., 2016).

Given the goal-oriented nature of games, that typically results in various sub-tasks, speech acts produced by players can be insightful in relation to their intentions, which are strictly related to the categorizaton of speech acts (e.g., Grice (1957, 1975)). *Settlers* has been used to study negotiation (Cadilhac et al., 2013). Other corpora, such as on *Avalon* (Stepputtis et al., 2023) and *Werewolf* (e.g., Lin et al. (2020)), have been used to investigate competitive-cooperative settings where private and competing beliefs and pieces of information are held by players, who are therefore encouraged to employ strategies that include deception.

2.3. Multi-party Chatbots

Although chatbots have a long history (Adamopoulou and Moussiades, 2020), multi-party chatbots are less studied, also due to the challenges of their design (Seering et al., 2020).

One of these challenges is the understanding of *who* is it talking to *whom* about *what*, which various methods try to tackle, such as response generators which incorporate Interlocutor-aware Contexts into a Recurring-Encoder-Decoder (Liu et al., 2019)

When developed for multi-party use, chatbots are often adapted from single-party systems. For example, a study by Wagner et al. (2022) used the Rasa framework to create a chatbot for goal-directed conversations in everyday scenarios.

Machine learning has led to the development of more advanced chatbots for multi-party environments, often involving role-playing to test different identities and features. In the LIGHT environment, humans and NPCs are assigned roles (e.g. wizard) and interact through conversations. R2C2 models are trained and tested using four methods to enhance turn-taking and coherence, addressing challenges in multi-party interactions (Wei et al., 2023). Role-playing identities are also created using ChatGPT and other LLM chatbots with dialogue engineering, as few-shot learning alone may not produce accurate and consistent behaviors (Wang et al., 2023).

2.4. Conversation State Extraction

State conversation extraction is considered playing a key role not only in understanding dialogues but also building dialogue systems (e.g., Gao et al., 2020). Unfortunately, its application on multi-party chats can be a puzzling task due to the presence of multi-threads and complex discourse relations. Various approaches have been tested, including on the STAC Corpus.

Among these, a prominent strategy makes use of parsing algorithms. Li et al. (2023) propose a model that utilizes knowledge-enhanced features and symbolic knowledge graph relations to recognize emotions. Another model (Jia et al., 2020) leverages thread extraction based on dependency relations, along with a Thread-Encoder and Transformers, to enhance context understanding. While both systems require training, ChatGPT has shown superiority in zero-shot tasks compared to LLMs and fine-tuned models (Bang et al., 2023). However, its performance in predicting and classifying links between utterances has been limited (Chan et al., 2023).

In the specific task of DA classification, deep neural networks were compared on the STAC by İrsoy et al. (2019) along with their model of directed-acyclic-graph LSTM (DAG-LSTM) which exploits turn-taking and employs Tree-LSTM equations. Conditional Random Fields (CRFs) were used by a number of researchers including the corpus' authors (Cadilhac et al., 2013). With three types of features (lexical, syntactical and semantic), the model outperforms the frequency-based baseline. (See section 5 for a comparison of these models with ours).

State understanding in games may involve unique elements like strategies and secret identities of players. In Avalon, LLMs like ChatGPT were used to uncover players' secret roles by analyzing game dialogue through different state conceptualizations (Stepputtis et al., 2023). Each utterance was also labeled with a persuasion or lying strategy. In Werewolf, CNN/SVM models with manual rules were used to study players' behaviors based on their secret roles, aiming to train an agent to play like a human (Lin et al., 2020). In Settlers (STAC), CP-nets were utilized to predict players' strategic actions, specifically trades (Cadilhac et al., 2013).

3. Data

The STAC corpus (Asher et al., 2016) consists of multi-party chats annotated for discourse structure in the style of Segmented Discourse Representation Theory (SDRT) (Asher and Lascarides, 2003; Lascarides and Asher, 2009). It includes 45 online games sessions of *Settlers of Catan*, a popular boardgame[2].

Settlers of Catan is a competitive, strategic game where players need to exchange resources with each other, making bargaining a pivotal discourse element. It is played on a map made of hexes, which are associated to one of 5 resource types (Brick, Lumber, Ore, Grain, Wool, plus Nothing) and a number (2-12). At the beginning of the game, each player places 2 "Settlements" on an intersection that borders with 2-3 hexes, and can build more of those throughout the game. During a player's turn, he throws 2 dice, which sum indicates which hex will provide resources for the current turn to all players who own settlements on its intersections. Afterward, the player can negotiate with the others an exchange of resources, different combinations of which are needed to build more and better structures, and ultimately to win the game.

Afantenos et al. (2015) who provide annotations for the STAC corpus state that multi-party chats pose a series of challenges that cannot be addressed the same way of two-party chats. Among these, complex intersections of addresses between speakers that escape tree-like structures interpretations, and crossing dependencies. Therefore, they motivate that SDRT is particularly appropriate for the annotation of the STAC corpus because of three reasons: (a) it allows for distant attachments; (b) it is capable of dealing with fragments or non penitential utterances; (c) it can model non-tree like structures.

Their annotation process started with segmenting the turns into EDUs (Elementary Discourse Units), because within each turn the speaker may convey bits of information that carry different purposes (e.g., addressing a previous offer and proposing an offer to a third player). This part was initially done automatically, then corrected by hand. Each EDU was then annotated in a three layers fashion: (1) Type of speech act; (2) Dialogue act; (3) Relation type. These classes sum up with (4) the addressee. Layer (1), Type of speech act (or surface type/act), includes only the *Assertion*, *Question* and *Request* categories. Layer (2), DAs, has *Offer*, *Counteroffer*, *Accept*, *Refusal*, and *Other*, which labels units that either comment on strategic moves in the game or are not directly pertinent to bargaining. Layer (3) which contains 16 relation-based types (e.g., *Comment*, *Background*). For our study, we only consider layer (2).

Overall there are 13440 annotated EDUs. Along with the text and the annotated dialogue act, each segment's row carries information about the identity of the emitter, the emittee(s), the dialogue and others. In our study, we decided to utilize the whole dataset and to follow the original fragmentation of dialogues ($n = 1137$) to maintain consistency with the annotators' work. For some of our experimental runs, we treat each game ($n = 45$) as a single dialogue as we will explain later.

As we were not interested in predicting the DA *Other* (which could be any not further specified type of dialogue that is not related to the conversation about the game), we re-patched the original turns that had been segmented into EDUs, thus restoring the corpus into a sequence of turns. When this operation resulted in conflicting DAs (the only occurrence being *Other* plus a different dialogue act), we gave precedence to the other DA(either *Offer*, *Counteroffer*, *Accept* or *Refusal*). The number of turns that satisfy the experimental requirements amounts to 3939 (*Offer*: 981; *Counteroffer*: 647; *Accept*: 696; *Refusal*: 1615) when splitting by dialogues; 4552 (*Offer*: 1589; *Counteroffer*: 649; *Accept*: 697; *Refusal*: 1617) when splitting by games.

4. Methodology and Experiments

In general, we employ zero-shot and few-shot learning approaches when running classification or prediction tasks. In particular, we run our experiments using GPT-3.5 Turbo. We acknowledge that conducting comparisons among various LLMs could offer additional insights into how they perform on game-based multi-party dialogue corpora. However, our primary objective is to get an initial idea of these models' performance on the task. Therefore, we leave running such experiments as future work. Few-shot learning is a learning approach where the model is given at inference a small number of demonstrations of each new task it is asked to per-

[2]Rules available under: https://www.catan.com/sites/default/files/2021-06/catan_base_rules_2020_200707.pdf

form (Brown et al., 2020). Note that the weights of the model are not updated, thus, the model must use its prior knowledge to generalize from these examples to perform the task. As Brown et al. (2020) have shown, large language models excel at zero-shot, one-shot, and few-shot learning tasks, frequently matching performances of fine-tuned models. However, it has been shown that this type of few-shot learning can be unstable (Zhao et al., 2021; Ye and Durrett, 2022). The choice of prompt format, training examples, their number, or even their order all influence the performance and expose biases inherent in the model (Webson and Pavlick, 2022). Nonetheless, few-shot learning is being explored due to its speed, low cost and data efficiency in solving custom tasks (Ahmed and Devanbu, 2023; Wei et al., 2022).

Before systematically running the experiments, we conducted a series of pre-tests on selected dialogue inputs and prompts to evaluate ChatGPT's behavior. We wanted to explore the model's ability to correctly classify the DAs of an input text and to investigate the rationale behind its choice. This iterative process was especially useful in elaborating more successful descriptions of the DAs as part of the prompts. Moreover, we needed to make sure that the produced output was consistent and suitable for an automatic analysis (parsing the response to a label). The interpretation of these results are illustrated in the qualitative analysis.

4.1. Tasks Evaluated

To comprehensively evaluate ChatGPT's capability in assessing multi-party dialogues as they naturally occur within multiplayer-games, we conducted a range of experiments with various objectives using the STAC dataset. Our objectives include the straightforward task of classifying relevant DAs to get the current status of the discussion, as well as predicting future DA types within an ongoing conversation. Figure 1 shows an example of the both tasks given previous textual turns. As *Other* could be any not further specified type of DA, in both tasks we will exclude samples that have this label during our experiments. However, we will keep previous DAs that are labeled with *Other* as context for few shot learning as they can provide helpful contextual cues.

Classification of Dialogue Act Types: Our first goal is to evaluate the classification performance of ChatGPT on game-relevant DAs covering the classes described previously (*Offer*, *Counteroffer*, *Accept* and *Refusal*).

Prediction of Dialogue Act Types: Our second approach extends beyond simple classification, focusing on the prediction of the subsequent DA following a prior conversation. Again, employing both zero-shot and few-shot learning methodologies across various scenarios (e.g. with different context length and number of samples during few-shot learning), we aim to predict relevant DA types.

4.2. Prompt Dimensions

We assess various dimensions that could potentially impact the quality of a prompt for both tasks, as it remains unclear which variables contribute to the model's performance and to what extent. Below, we offer a brief overview of the individual feature dimensions we vary in our experiments. Apart from different forms of these feature dimensions, the prompt always begins with an intro and ends with an output specification. The intro is always "I will give you a dialogue from a game of Settlers of Catan played by some players, you will need to predict the class of the next utterance." (in case of future DA type predicting) or a similar form (e.g. in case of current DA classification); The output specification is always "How could that dialogue continue? Very important: please respond with 1 possible continuation in this precise format: [class of utterance]" (again in case of predicting the future DA type), besides a few variations of the question (e.g. when running classification).

Game Description [GAME] *(name of the dimension as used in the columns of Table 2 and Table 3) in squared brackets***:** We evaluate two versions of this features as part of the prompt: (1) a summarized description of the game *Settlers*, and (2) no game description at all.

Number of Shots [SHOTS]: We experiment with four different versions: (0) no shots; (1) one or (2) two utterance(s) after the description of each DA class; (5) one utterance plus, after the description of the DA classes, four additional utterances for each DA class in random order (resulting in overall 5-shots);.

Context-Length of the Input Dialogue [CONTEXT]: For the length of context that we pass as part of the input dialogue we tested three variations: (1) one turn, (3) three turns and (5) five turns. For most of our runs we used condition (3).

Form of Player Names [PLAYERS]: Another variable we experiment with is the way player names are represented in the input dialogue. (1) in some prompts we report the original name of players, (2) in other prompts report an anonymized form of the player names (i.e., player_1, etc.) which remains consistent for active players within the same input dialogue.

Information about the Conditional Probability of the DAs [PROB]: This feature describes for a present DA which is/are the most likely dialogue act(s) to follow. We evaluate four different versions: (1) an indication of what is more likely to occur (i.e., "Offer often follows Other. Sometimes it follows Accept, less times it follows Refusal, Counteroffer

GAME	DIALOGUE	TURN	SPEAKER	TEXT	DA
13	405	75	Player 1	anyone has any wood?	Offer
13	405	76	Player 2	Nope, sorry.	Refusal
13	405	77	Player 3	haha no, seems to be a very clay-heavy game this	Refusal
13	406	80	Player 3	trading 1 ore for one sheep?	Offer

Figure 1: Dialogue and segmentation example with CONTEXT=3 and PLAYERS=YES. Context is highlighted in blue, target DA to classify/predict in red. For classification the turn text in green is added to the prompt, during prediction it is left out *(numbering gap between turns 77 and 80 is due to Server/Game Engine turns which are not included in the players-only dataset)*.

and Offer."), (2) an indication with the addition of an annotated two to three turns long exchange as example, (3) conditional probabilities for each possible combination, expressed as a percentage, and (N) no probabilities.

Domain of Dataset [DOMAIN]: Almost all of our experimental conditions followed (1) the partitioning of the dialogues based on the suggestion of the annotators. However, we also tested prompts using (2) games instead of dialogues as the splitting parameter for the dialogues.

Order of Features [ORDER]: Finally, we utilize the order in which the blocks are presented within the prompt as an additional parameter. We use three different variations: (1) Dialogue - Instructions - Classes; (2) Instructions - Classes - Dialogue; (3) Classes - Dialogue - Instructions.

5. Results

In line with common practice in NLP, we report on accuracy and macro F1-scores (Jurafsky and Martin, 2024, Ch. 4). We limited the results presented below to these metrics and a range of different prompt feature combinations to make the overview more concise. For more details about other runs as well as additional metrics, such as precision and recall, we refer to our codebase.

5.1. Classification of Dialogue Act Types

Baselines: For comparison of our classification results, we refer to Cadilhac et al. (2013) and Irsoy et al. (2019) who both performed this task with different approaches on the same dataset. While Cadilhac et al. (2013) adopt Conditional Random Fields (CRFs) to learn DAs, Irsoy et al. (2019) introduced a new architecture (DAG-LSTM) for contextual representations. However, we note that both also included the *Other* label in their evaluation which accounts for a high number of samples and thus leads to macro F1 performance that is not comparable. Therefore, adopting class-wise F1 scores (of the same classes we used in our setup) when comparing results is more fair.

When evaluated against Cadilhac et al. (2013) and Irsoy et al. (2019), our approach consistently produces higher class-wise F1 scores related to *Accept* and *Refusal*. However, it showed slightly worse performance in *Counteroffer* and *Offer* classifications, although demonstrating to be notably close in the *Offer* category. For detailed class-wise results, see Table 1.

As observable in Table 2 prompts with modified instructions, placed after the dialogue to classify, and differing only in the instruction formulation, resulted in very low classification Accuracy (0.560) and macro F1 (0.633) compared to the other variations. Another setup, with a single few-shot example, also exhibited low Accuracy of 0.571 and macro F1 score of 0.639.

For the exact wording used in our prompts, further details on the different experimental setups as well as additional results we refer to our codebase.

5.2. Prediction of Dialogue Act Types

As observable in Table 3, for runs where we selected "games" for prompt dimension [DOMAIN] we can observe higher metrics compared to the runs with "dialogue" for this feature. When compared with the baseline performance on dialogues (Accuracy = 0.266, F1 = 0.212), the improvements in Accuracy (0.345) and F1 (0.305) on games were carried both by increments in the same metrics for the *Offer* DA (Accuracy: 0.658 vs 0.619; F1 = 0.502 vs 0.400) both by the higher number of this class which accounts for the major difference between the two domains (1589 vs 981). This result underlines the impact that the distribution of the DAs can have on the major metrics, suggesting the need for a more complex interpretation of the results.

One combination of prompt dimensions (compare first row in Table 3) served as the basis for several variations of all other prompts, thus offering a point of reference for the interpretation of the impact of the feature dimensions. We show an example of this prompt in figure 2. When ranking by Accuracy, for all remaining features the results suggest that: (1 [GAME]) prompts result in better per-

	Offer	Counteroffer	Accept	Refusal
Cadilhac et al. (2013)	0.805	**0.585**	0.585	0.776
Irsoy et al. (2019)	**0.820**	0.517	0.643	**0.865**
OURs	0.719	0.384	**0.671**	**0.865**

Table 1: Dialog Act class-wise F1-score comparison with baselines. Best scores are highlighted in bold.

GAME	SHOTS	CONTEXT	PLAYERS	PROB	DOMAIN	ORDER	ACC	F1
NO	2	3	YES	(N)	DIAL.	(3)	0.560	0.633
NO	1	3	YES	(N)	DIAL.	(3)	0.572	0.639
NO	2	3	YES	(N)	GAMES	(1)	0.618	0.650
NO	1	1	YES	(N)	DIAL.	(2)	0.594	0.652
NO	1	3	NO	(N)	DIAL.	(3)	0.593	0.654
NO	1	5	YES	(N)	DIAL.	(2)	0.600	0.660
NO	1	3	NO	(2)	DIAL.	(2)	0.650	0.675
NO	2	3	YES	(N)	DIAL.	(1)	0.654	0.692
NO	1	3	YES	(N)	DIAL.	(1)	0.665	0.694
NO	2	3	NO	(N)	DIAL.	(1)	**0.691**	**0.716**

Table 2: Accuracy and macro F1-scores for dialogue act classification under different variations of the prompt features. Best scores are highlighted in bold.

formance when using a description of the game; (2 [SHOTS]) few-shot examples are useful; (3 [CONTEXT]) longer input dialogues result in better performance; (4 [PLAYER]) including anonymized player names instead of the original ones is useful; (5 [PROB]) the results suggest that more precise information about the conditional probabilities are better; (7 [ORDER]) the way in which the pieces of information within the prompts are ordered can be very impactful, with prompts that have the input dialogue first outperforming all the rest.

5.3. Qualitative Analysis

We select a series of turns and dialogues of various complexity to investigate ChatGPT's capability in understanding the given dialogue input and its reaction to the prompt variations. Most of the times, we ask the LLM to produce one or more possible continuation(s) and to specify its dialogue act, which allow us to evaluate the response's fitness through analysis of the speaker identity, the meaning of the utterance, the relationship with the given context, appropriateness of the chosen DA and the response's syntax. This allows us to shed more light on the reasons that informed the LLM's decision.

One of the things that ChatGPT seems to do best is associating the right DA to its own response. When mistakes appear, they tend to relate to the DA class *Other*, especially in the form of false negatives. Prominent cases are answers to difficult questions that do not constitute *Offers* or *Counteroffers*, which should be labeled as *Other* but are recognized as *Accept* or *Refusal* instead. Even more difficult are questions of these kinds that mention resources, such as when players discuss how to materially complete the trade with the UI. When asked how it would like the DAs to be described to avoid misunderstanding, ChatGPT can propose to include new categories, such as splitting *Other* in three classes: *Inquiry* for questions about resources and trades without making a formal offer, *Explanation* for clarifications related to ongoing actions, and *Other* for the remaining situations.

Errors in the form of the chosen DA (e.g., writing "class" where we expect the name of the class) tend to happen when the definitions or examples of the classes are defined or introduced in a way that is too ambiguous (e.g., asking to respond with "class, player: utterance").

ChatGPT's choice of speaker is not always great, with seemingly naive errors. Not rarely, it makes the last speaker continue the conversation, including by responding to their own question. Such poor performance in considering turn-taking and basic contextual information strikes in opposition to the excellent internal coherence demonstrated when prompted to produce entire stories or conversations from scratch. Evidence from this suggests us to specify that a *Counteroffer*/*Accept*/*Refusal* relates to another player's offer. Although helpful, this does not completely solve the issue, and it cannot be of assistance with another typical (semi)error, which is when the same player continues with a second offer.

The understanding of contextual information related to the game is even more problematic: Frequently, ChatGPT makes a player offer a resource

GAME	SHOTS	CONTEXT	PLAYERS	PROB	DOMAIN	ORDER	ACC	F1
NO	2	3	YES	(N)	DIAL.	(1)	0.266	0.212
NO	2	1	YES	(N)	DIAL.	(1)	0.235	0.157
NO	2	3	YES	(1)	DIAL.	(1)	0.267	0.184
NO	2	3	YES	(N)	DIAL.	(2)	0.174	0.191
NO	0	3	YES	(N)	DIAL.	(1)	0.264	0.195
NO	2	3	YES	(2)	DIAL.	(1)	0.270	0.206
NO	2	3	NO	(N)	DIAL.	(1)	0.263	0.214
NO	2	3	YES	(3)	DIAL.	(1)	0.270	0.223
NO	2	5	YES	(N)	DIAL.	(1)	0.279	0.234
NO	5	3	YES	(N)	DIAL.	(1)	0.277	0.248
NO	2	3	YES	(2)	DIAL.	(2)	0.230	0.251
YES	2	3	YES	(N)	DIAL.	(1)	0.273	0.260
NO	2	3	YES	(N)	GAMES	(1)	0.318	0.262
NO	5	3	YES	(N)	GAMES	(1)	**0.350**	**0.305**

Table 3: Accuracy and macro F1-scores for next dialogue act prediction under different variations of the prompt features. Best scores are highlighted in bold.

that the same player had previously exhibited a need for or respond to offers that never happened. To better elucidate on its understanding of resources needs and offers, which as a key element of Settlers were also used as a major feature for classification by Cadilhac et al. (2013), we ask ChatGPT to illustrate the state of each player's needs and availability. The delivered representation within the same output tends to be black and white: for some players, it can be perfectly correct; for other players, a mixture of right and wrong conclusions. Some resources are misunderstood as wants when they are offered, and vice-versa.

Moreover, biases within the order of information can appear: Along with the preference for repeating the last turn's speaker in shorter dialogues, ChatGPT may exhibit primacy bias in favouring the first turn's speaker in long dialogues. At the same time, ChatGPT proves to be able to focus on a mid-dialogue offer making a player answer it, a common occurrence in multi-party conversations.

6. Limitations and Discussion

Firstly, in the prediction task, there appears to be a bias towards some DAs compared to others. *Offer* and *Counteroffer* show good recall and modest precision, indicating that ChatGPT often predicts them correctly but may also falsely predict them frequently. As *Accept* and *Counteroffer* are the first DAs in the description of DA classes, their position may contribute in explaining ChatGPT's bias in selecting them. On the contrary, *Accept* and *Refusal* show low recall and from modest to good precision. This suggests that while ChatGPT rarely predicts them, when it does, it is often correct. As a possible explanation, it may be helpful to note that predicting which DA follows an offer is expected to be challenging, as the range of possible responses is wide (*Accept*, *Refusal*, *Counteroffer* and *Other*), and affected by many exogenous factors, e.g. a player switching context or multiple line of conversation taking place simultaneously between different participants. At the same time, *Accept*, *Refusal* and *Counteroffer* are supposed to be cued by an offer

```
I will give you a dialogue from a game of Settlers of Catan
played by some players, you will need to predict the class
of the next utterance

The dialogue:
{dialogue}

It is very important that you consider what said by each
player, which represent their intentions, and the order in
which each player spoke. Build (but don't write) the
framework of which resources each player wants to trade for
giving and which to trade for receiving.

The admissible classes of utterances, with definition and
examples are:

"Offer: A proposal to trade resources between players, which
isn't related to another offer. Example1: Hey anyone have
any clay? Example2: Need wood or clay?",
"Counteroffer: A response to another player's offer,
proposing a different trade. Example1: I can do 1 of each
for 2 clay. Example2: (in response to an offer that
requested clay) What about sheep?",
"Accept: Agreeing to an offer or counteroffer made by
another player. Example1: I can wheat for clay. Example2:
(in response to an offer of ore) Sure",
"Refusal: Declining an offer or counteroffer made by another
player. Example1: (in response to an offer of wood) No, not
interested. Example2: (in response to an offer of ore for
sheep) Not as long as I keep losing ore from the robber",
"Other: Turns or statements that do not involve direct
trading, such as discussing game mechanics or making
observations about the current state of the game, including
questions that aren't offers or counteroffers. Example1:
What's up? Example2: (after a counteroffer) How do I accept
the trade?"

Please remember: If an utterance qualifies for "Other" but
also for one of the other 4 classes, it should then be
considered of the other class (not of the class "Other")

Very important: please respond with 1 possible continuation
in this precise format: [class of utterance]
```

Figure 2: Baseline prompt for prediction task.

or counteroffer, whereas the presence of *Offer* is more unpredictable, and this may help explaining *Offer*'s lower precision compared to *Accept* and *Refusal*.

Across different prompts and conditions, variations in performance metrics can be observed. Compared to the other DAs, *Offer* shows better stability; among the metrics, recall results are the most impacted, especially for *Counteroffer*. In particular, *Counteroffer*'s recall tends to be inversely correlated to *Offer*'s recall, implying competition in the prediction of these DAs.

ChatGPT's ability to classify the DAs proved good, with high precision and recall for all DAs but only modest for *Counteroffer*. Again, the lower performance of *Counteroffer* may be attributed to confusion with *Offer*. *Refusal* showed the highest precision, which could be attributed to a narrower and clearer realm of expressions when it comes to saying no to someone.

Our range of test on different prompt variations reveal interesting insights: For instance, prompts with more examples per DA (and thus more shots for the few-shot learning) generally improve performance metrics, suggesting the usefulness of additional examples presented to the LLM. However, presenting the dialogue last negatively impacts performance across all metrics, indicating the importance of dialogue sequence in prediction accuracy. Manipulating context length of given previous turns also affects performance, with longer contexts generally improving recall and F1 scores, but with lower precision. Notably, shorter contexts result in lower F1 scores, particularly for the *Refusal* class.

Overall, these observations underscore ChatGPT's capabilities in certain DA predictions while highlighting areas for improvement, such as accurately predicting *Accept* and managing dialogue context effectively.

7. Conclusion

Previous literature assessed the ability of GCBs in solving an array of pragmatic tasks (e.g., implicatures, indirect speech acts, comprehension of fiugurative language, etc.; see (Hu et al., 2022; Barattieri di San Pietro et al., 2023; Ruis et al., 2024)), finding a performance comparable to that of humans. Does this mean then that GCBs engage in pragmatic processes in the same way as human do? Bender et al. (2021) have famously debated that GCBs do not possess human-like processes, defining them "stochastic parrots", lacking communicative intention and thus only mimicking language comprehension. Indeed, as pointed out by Hu et al. (2022), experiments showing that chatbots displaying human-like verbal behaviors should not necessarily lead to conclusions toward a similarity of processes of humans and AI. An interesting argument comes however from Lenci (2023), who brings a cognitive perspective into the debate arguing that even in humans "language understanding does not always consists in the construction of full-fledged, highly structured semantic representations or complex reasoning processes". Referring to the works of Ferreira et al. (2002); Karimi and Ferreira (2016), Lenci reminds us that humans make often use of shortcuts, heuristics and "good enough" representations in order to process language quickly and efficiently. From this stance we can gather that humans have both capabilities: they can, on the one hand, engage in deeper understading of the interlocutors' intentions forming and recalling theories of the mind and of shared knowledge, and on the other hand, rely on surface heuristics to reach easily their communicative objectives. When considering GCBs however, it appears that they strongly rely on the latter, without the ability to access the former processes. As suggested by Mahowald et al. (2024) large language models lack functional linguistic competence, that is the ability to rely on world knowledge to form models of the situation and the interlocutors and to engage in social pragmatic understanding of the communicative intentions. The lack of this functional ability (which is, to all intent and purposes the core of pragmatics), but, above all, the lack of the flexibility to engage in both levels of processing (the deep one and the surface one) might greatly impair the possibility of GCBs to generalize their verbal behaviors to more complex interaction involving multiple agents.

In conclusion, ChatGPT showed good ability in navigating through the DAs categorization, however our results show that it may over-rely on such "shortcuts" (Lenci, 2023) as it is less good at understanding the real state of the conversation. This is evident in joint speaker and utterance predictions that often deliver nonsensical outcomes about which ChatGPT is not aware. This indicates that applying off-the-shelf GCBs to multiparty dialogues may not be immediate and supports the adoption of expensive approaches, e.g. those involving full LLM training (Wei et al., 2023). However, the consistent results on the independent classification of DAs suggest that lower-cost but non-trivial solutions for multi-party GCBs should be explored.

8. Acknowledgements

We would like to thank the anonymous reviewers for their constructive feedback which has helped us improve the paper.

This work was supported by the project COURAGE funded by the Volkswagen Foundation, grant number 95564.

9. Bibliographical References

Eleni Adamopoulou and Lefteris Moussiades. 2020. Chatbots: History, technology, and applications. *Machine Learning with Applications*, 2:100006.

Stergos Afantenos, Eric Kow, Nicholas Asher, and Jérémy Perret. 2015. Discourse parsing for multi-party chat dialogues. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 928–937, Lisbon, Portugal. Association for Computational Linguistics.

Toufique Ahmed and Premkumar Devanbu. 2023. Few-shot training llms for project-specific code-summarization. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, ASE '22, New York, NY, USA. Association for Computing Machinery.

Keith Allan. 1998. Meaning and speech acts. *Retrieved June*, 28:2004.

James Allen and Mark Core. 1997. Draft of damsl: Dialog act markup in several layers.

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

John Langshaw Austin. 1975. *How to do things with words*, volume 88. Oxford university press.

Kent Bach and Robert M Harnish. 1979. Linguistic communication and speech acts.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Chiara Barattieri di San Pietro, Federico Frau, Veronica Mangiaterra, and Valentina Bambini. 2023. The pragmatic profile of chatgpt: Assessing the communicative skills of a conversational agent. *Sistemi intelligenti*, 35(2):379–400.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Emily M Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.

Ali Borji. 2023. A categorical archive of chatgpt failures. *arXiv preprint arXiv:2302.03494*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Harry Bunt. 1994. Context and dialogue control. *Think Quarterly*, 3(1):19–31.

Harry C Bunt. 1981. Rules for the interpretation, evaluation and generation of dialogue acts. *IPO annual progress report*, 16:99–107.

Anais Cadilhac, Nicholas Asher, Farah Benamara, and Alex Lascarides. 2013. Grounding strategic conversation: Using negotiation dialogues to predict trades in a win-lose game. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 357–368.

Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2023. Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations. *arXiv preprint arXiv:2304.14827*.

Alex Djalali, Sven Lauer, and Christopher Potts. 2011. Corpus evidence for preference-driven interpretation. In *Proceedings of the 18th Amsterdam Colloquim Conference on Logic, Language and Meaning*, AC'11, page 150–159, Berlin, Heidelberg. Springer-Verlag.

Fernanda Ferreira, Karl GD Bailey, and Vittoria Ferraro. 2002. Good-enough representations in language comprehension. *Current directions in psychological science*, 11(1):11–15.

Shuyang Gao, Sanchit Agarwal, Tagyoung Chung, Di Jin, and Dilek Hakkani-Tur. 2020. From machine reading comprehension to dialogue state tracking: Bridging the gap. Publisher: [object Object] Version Number: 1.

H Paul Grice. 1957. Meaning. *The philosophical review*, 66(3):377–388.

H Paul Grice. 1975. Logic and conversation. *Syntax and Semantics*, 3:43–58.

Reto Gubelmann. 2024. Large language models, agency, and why speech acts are beyond them (for now)–a kantian-cum-pragmatist case. *Philosophy & Technology*, 37(1):32.

Laurence R Horn and Gregory L Ward. 2004. *The handbook of pragmatics*. Wiley Online Library.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2022. A fine-grained comparison of pragmatic language understanding in humans and language models. *arXiv preprint arXiv:2212.06801*.

Ozan Irsoy, Rakesh Gosangi, Haimin Zhang, Mu-Hsin Wei, Peter Lund, Duccio Pappadopulo, Brendan Fahy, Neophytos Nephytou, and Camilo Ortiz. 2019. Dialogue act classification in group chats with dag-lstms. *CoRR*, abs/1908.01821.

Qi Jia, Yizhu Liu, Siyu Ren, Kenny Zhu, and Haifeng Tang. 2020. Multi-turn response selection using dialogue dependency relations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1911–1920. Association for Computational Linguistics.

Daniel Jurafsky and James Martin. 2024. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 3rd edition (draft), 3rd February 2024. https://web.stanford.edu/~jurafsky/slp3/.

Hossein Karimi and Fernanda Ferreira. 2016. Good-enough linguistic representations and online cognitive equilibrium in language processing. *Quarterly journal of experimental psychology*, 69(5):1013–1040.

Cansu Koyuturk, Mona Yavari, Emily Theophilou, Sathya Bursic, Gregor Donabauer, Alessia Telari, Alessia Testa, Raffaele Boiano, Alessandro Gabbiadini, Davinia Hernandez-Leo, et al. 2023. Developing effective educational chatbots with chatgpt prompts: Insights from preliminary tests in a case study on social media literacy. In *31st International Conference on Computers in Education*.

János Kramár, Tom Eccles, Ian Gemp, Andrea Tacchetti, Kevin R McKee, Mateusz Malinowski, Thore Graepel, and Yoram Bachrach. 2022. Negotiation and honesty in artificial intelligence methods for the board game of diplomacy. *Nature Communications*, 13(1):7214.

Alex Lascarides and Nicholas Asher. 2009. Agreement, Disputes and Commitments in Dialogue. *Journal of Semantics*, 26(2):109–158.

Alessandro Lenci. 2023. Understanding natural language understanding systems. a critical analysis. *arXiv preprint arXiv:2303.04229*.

Wei Li, Luyao Zhu, Rui Mao, and Erik Cambria. 2023. SKIER: A symbolic knowledge integrated model for conversational emotion recognition. 37(11):13121–13129.

Youchao Lin, Miho Kasamatsu, Tengyang Chen, Takuya Fujita, Huanjin Deng, and Takehito Utsuro. 2020. Automatic annotation of werewolf game corpus with players revealing oneselves as seer/medium and divination/medium results. In *Workshop on Games and Natural Language Processing*, pages 85–93, Marseille, France. European Language Resources Association.

Cao Liu, Kang Liu, Shizhu He, Zaiqing Nie, and Jun Zhao. 2019. Incorporating interlocutor-aware context into response generation on multi-party chatbots.

Khyati Mahajan and Samira Shaikh. 2021. On the need for thoughtful data collection for multi-party dialogue: A survey of available corpora and collection methods. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 338–352, Singapore and Online. Association for Computational Linguistics.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*.

Michael McTear. 2022. *Conversational ai: Dialogue systems, conversational agents, and chatbots*. Springer Nature.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2024. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by llms. *Advances in Neural Information Processing Systems*, 36.

John R Searle. 1969. *Speech acts: An essay in the philosophy of language*, volume 626. Cambridge university press.

John R Searle. 1975. A taxonomy of illocutionary acts.

Joseph Seering, Michal Luria, Connie Ye, Geoff Kaufman, and Jessica Hammer. 2020. It takes a village: Integrating an adaptive chatbot into an online gaming community. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA. Association for Computing Machinery.

Penelope Sibun. 1997. Beyond dialogue: the six w's of multi-party interaction. In *Working Notes of AAAI97 Spring Symposium On Mixed-Initiative Interaction, Stanford, CA*, pages 145–150.

Simon Stepputtis, Joseph Campbell, Yaqi Xie, Zhengyang Qi, Wenxin Zhang, Ruiyi Wang, Sanketh Rangreji, Charles Lewis, and Katia Sycara. 2023. Long-horizon dialogue understanding for role identification in the game of avalon with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11193–11208, Singapore. Association for Computational Linguistics.

Peter F Strawson. 2013. Intention and conventionin speech acts. In *Symposium on JL Austin (Routledge Revivals)*, pages 380–400. Routledge.

Nicolas Wagner, Matthias Kraus, Tibor Tonn, and Wolfgang Minker. 2022. Comparing moderation strategies in group chats with multi-user chatbots. In *Proceedings of the 4th Conference on Conversational User Interfaces*, CUI '22, New York, NY, USA. Association for Computing Machinery.

Wolfgang Wahlster. 2023. Understanding computational dialogue understanding. *Philosophical Transactions of the Royal Society A*, 381(2251):20220049.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. 2023. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Jimmy Wei, Kurt Shuster, Arthur Szlam, Jason Weston, Jack Urbanek, and Mojtaba Komeili. 2023. Multi-party chat: Conversational agents in group settings with humans and models. *arXiv preprint arXiv:2304.13835*.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. *Advances in neural information processing systems*, 35:30378–30392.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

Ozan İrsoy, Rakesh Gosangi, Haimin Zhang, Mu-Hsin Wei, Peter Lund, Duccio Pappadopulo, Brendan Fahy, Neophytos Nephytou, and Camilo Ortiz. 2019. Dialogue act classification in group chats with dag-lstms.