Dialogue Acts as a Lens on Human–LLM Interaction: Analyzing Conversational Norms in Model-Generated Responses

Arunima Maitra **Dorothea French** **Katharina von der Wense**

University of Colorado, Boulder
firstname.lastname@colorado.edu

Abstract

Large language models (LLMs) have revolutionized natural language generation across various applications. Although LLMs are highly capable in many domains, they sometimes produce responses that lack coherence or fail to align with conversational norms such as turn-taking, or providing relevant acknowledgments. Conversational LLMs are widely used, but evaluation often misses pragmatic aspects of dialogue. In this paper, we evaluate how LLM-generated dialogue compares to human conversation through the lens of dialogue acts, the functional building blocks of interaction. Using the Switchboard Dialogue Act (SwDA) corpus, we prompt two widely used open-source models, Llama 2 and Mistral, to generate responses under varying context lengths. We then automatically annotate the dialogue acts of both model and human responses with a BERT classifier and compare their distributions. Our experimental findings reveal that the distribution of dialogue acts generated by these models differs significantly from the distribution of dialogue acts in human conversation, indicating an area for improvement. Perplexity analysis further highlights that certain dialogue acts like 'Acknowledge (Backchannel)' are harder for models to predict. While preliminary, this study demonstrates the value of dialogue act analysis as a diagnostic tool for human–LLM interaction, highlighting both current limitations and directions for improvement.

1 Introduction

Large language model-based dialogue systems are hugely successful in open domain language generation tasks such as question-answering (Kann et al., 2022). Although these systems generally produce high-quality fluent dialogues and are able to hold conversations, their utterances sometimes fail to capture the nuances and emotions that are common in human–human interactions. Effective conversation depends on subtle patterns of dialogue acts – utterances that serve functions such as asking questions, signaling agreement, or providing acknowledgment. We aim to investigate the extent of interpersonal synergy exhibited by LLMs in comparison to human interactions. Synergy in terms of interactive and cooperative conversations refers to the way one agent responds to the other based on coordination and engagement between agents (Fusaroli and Tylén, 2016).

Dialogue acts are labels assigned to utterances that classify the intent of the speaker. In our study, we prompt the Llama 2 model developed by Meta (Touvron et al., 2023) and the Mistral-7B model developed by Mistral AI (Jiang et al., 2023), with context from the Switchboard Dialogue Acts (SwDA)[1] corpus of human telephonic conversations (Stolcke et al., 2000) to generate the next utterance. We then conduct dialogue act classification with a bert-base model (Raheja and Tetreault, 2019) and compare the dialogue acts of the LLM-generated responses to those of the gold-standard responses. Our study illustrates both the promise and the limits of current LLMs as conversational partners, and proposes dialogue act analysis as a human-centered diagnostic tool that complements surface-level metrics.

Dialogue acts, a basic unit of conversation and indication of quality and engagement (Deriu et al., 2021), allow us to measure the quality and type of utterances generated by models. By analyzing the specific types of dialogue act the Llama 2 and Mistral models struggle to generate, we gain a better understanding of the current limitations of LLMs. Dialogue acts have been especially studied in work with regard to classroom dialogue (Ganesh et al., 2021), thus models that better follow human conversation styles or use quality dialogue could mimic a teacher's discourse and guide each student individually based on their utterances, leading to more personalized feedback for each student (D'Mello

[1] https://catalog.ldc.upenn.edu/LDC97S62

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 317–325

and Graesser, 2013; Macina et al., 2023).

Badshah and Sajjad (2024) and Nadeau et al. (2024) suggest that Mistral outperforms Llama 2 in several aspects, including reduced hallucinations and enhanced engagement in back-and-forth conversations. Although Mistral tends to generate more engaging and informative dialogues, typical human conversations often diverge from this pattern. Human interactions frequently involve various cues and conventions, such as acknowledging others' opinions and providing affirmations. Our experimental results demonstrate that the Llama 2 model adapts its responses to more accurately reflect the nature of the ongoing conversation, aligning itself with the conversational style of the interlocutor.

In this work, we ask whether dialogue act analysis can serve as a diagnostic lens for evaluating conversational coherence in large language models. Specifically, we investigate whether systematic differences in dialogue act distributions between human and model-generated utterances can reveal where LLMs diverge from human conversational norms. Beyond NLP evaluation, this question connects to cognitive science perspectives on pragmatic competence, highlighting whether LLMs reproduce or miss key interactional strategies, and where dialogue act mismatches may help explain why chatbot interactions sometimes feel less natural.

2 Background and Related Work

Recent work by Shaikh et al. (2024) highlights that large language models often fail to establish common ground in conversation, using significantly fewer grounding acts such as clarifications, acknowledgments, and follow-up questions compared to humans. Their study introduces a taxonomy of grounding behaviors and demonstrates that instruction-tuned models systematically underuse these acts, particularly in high-stakes domains like emotional support and teaching. While their analysis focuses on how LLMs manage grounding, our work addresses a complementary question: how well LLMs reproduce the broader functional structure of human conversation as captured by dialogue acts. By analyzing utterance-level dialogue act distributions, we contribute an orthogonal yet critical view of conversational alignment, revealing that models overproduce questions and opinionated statements, but underproduce backchannels and agreement. Together, these findings indicate that LLMs diverge from human norms not only in their ability to construct shared understanding but also in their broader interactional strategies.

2.1 Language Models and Prompting

LLMs excel in various language tasks, including text generation, summarization, and translation. Yi et al. (2024) notes that Meta's LLaMA-2 is optimized for interactive conversations, adapting to user input, while OpenAI's GPT-4 (Achiam et al., 2023) is more versatile. We use the LLaMA-2 13B-chat and Mistral-7B-Instruct models to compare dialogue act alignment, as they represent different optimization strategies and to explore how varying training regimes influence dialogue structure. While the models are not the most recent, the focus was on analyzing dialogue act patterns rather than raw performance, and on employing widely used open-source models to ensure accessibility, transparency, and replicability.

LLaMA-2 13B-chat is a chain-of-thought optimized model fine-tuned for dialogue using supervised learning and Reinforcement Learning from Human Feedback (RLHF) with human evaluations for coherence, helpfulness, and safety (Touvron et al., 2023). It employs 'ghost attention' to preserve system instructions across turns, making it strongly suited for coherent multi-turn conversations.

In contrast, Mistral-7B-Instruct is a lightweight, instruction-tuned version of the base model, fine-tuned on publicly available conversational and instruction datasets (Jiang et al., 2023). While it retains architectural efficiencies like Grouped-Query Attention (GQA) and Sliding-Window Attention (SWA), Mistral-7B-Instruct also benefits from instruction-following refinement. However, it does not appear to use RLHF or chat-specific alignment via continued conversational feedback.

Prompting methods are crucial for enhancing LLM performance and tailoring responses to user specifications (Henrickson and Meroño-Peñuela, 2023). The *system prompt* in these models instructs the model on how to respond, giving users some control over generated dialogues. In-context learning, a prompt engineering technique, provides task demonstrations to guide LLMs (Wu et al., 2024; Rubin et al., 2021; Dong et al., 2022). It can be zero-shot, one-shot, or few-shot, depending on the number of input-output examples provided. This method is particularly effective for models with

large context windows. We provide several lines of context and prompt the model to respond accordingly.

For example, Kosinski (2024) demonstrates that GPT-4 correctly completes 95% of a set of 40 traditional false-belief tests that are frequently used to assess Theory-of-Mind (ToM) in humans when given a large 32K context window size. By comparison, GPT-3 can only correctly solve 40% of the false-belief tasks because it is a smaller model (up to 1000 times smaller than GPT-4) with 2K context window size.

2.2 Dialogue Acts

Dialogue acts are the functional units of conversation, describing the communicative intent behind an utterance. Drawing from speech act theory (Searle et al., 1980) and conversation analysis, dialogue acts capture not only the literal meaning of an utterance but also the role it plays in interaction—for example, making a statement, asking a question, or providing feedback. In computational linguistics, the Switchboard Dialogue Act Corpus (SwDA) (Jurafsky, 1997) has become a widely used benchmark, defining a taxonomy of 44 dialogue act categories. A few representative examples are included in Table 1.

Recent research on dialogue act classification treats it either as a text classification problem, where each utterance is classified in isolation (Lee and Dernoncourt, 2016), or as a sequence labeling problem (Kumar et al., 2018; Tran et al., 2017). According to Raheja and Tetreault (2019), some of the most promising models for dialogue act tagging are usually some sort of combination of the following models: conditional random fields (CRFs; Zhou et al., 2015), recurrent neural networks (RNNs; Chen et al., 2018), or BERT (Ribeiro et al., 2019). We classify dialogue acts using the Context-Aware Self-Attention Dialogue Act Classifier[2], which outperforms state-of-the-art methods by 1.6% on SwDA, the primary dataset for this task (Raheja and Tetreault, 2019). This model uses frozen BERT-base embeddings as input and employs a context-aware self-attention mechanism over dialogue turns, followed by a softmax classifier trained on the SwDA corpus. This design enables it to capture inter-turn dependencies critical for dialogue act identification.

[2] https://github.com/macabdul9/CASA-Dialogue-Act-Classifier.git

3 Experimental Setup

3.1 Dataset and Prompt

We use 1000 SwDA transcripts for the experiments, which are records of 2,400 two-sided telephonic conversations between two strangers with about 70 provided conversation topics, where each utterance is tagged with relevant dialogue acts. Since the dataset is a transcription of phone recordings, we removed the noises that did not contribute to the actual conversation. We prompt the Llama 2 and Mistral models with the following:

System Prompt: ***'You are a human, having a conversation with a stranger on telephone, about some topic from a predefined list. Given the context of the conversation, respond as best you can.'***

However, to assess the impact of different system prompt lengths and to finalize our choice of prompt, we experiment with both short and long variants:

Short System Prompt: *'You are a human having a conversation on telephone with another human you do not know, about some topic, from a given list. Given the context of the conversation, predict the next line as best you can.'*

Long System Prompt: *'You are having a conversation on telephone with someone you do not know. Given the context of the conversation, predict the next line as best you can. Respond with a single line. Your response should have dialogue act tags like- Statement-non-opinion, Acknowledge (Backchannel), Statement-opinion, Agree/Accept, Appreciation, Yes-No-Question, Non-verbal, Yes answers, Conventional-closing, Uninterpretable, Wh-Question, No answers, Response Acknowledgement like oh okay, Hedge, Declarative Yes-No-Question, Other, Backchannel in question form ,Quotation, Summarize/reformulate, Affirmative non-yes answers, Action-directive, Collaborative Completion, Repeat-phrase, Open-Question, Rhetorical-Questions, Hold before answer/agreement, and so on.'*

3.2 Methods

While the quality of generated responses seems to improve with in-context learning, the question remains how much prior knowledge is required for the dialogue systems to dynamically adjust their response strategies to align with human interactions (Brown et al., 2020).

Thus we initially conduct an experiment providing the LLMs the first 10 lines of utterances from the switchboard corpus as previous knowl-

Dialogue Act Tags	Example
Statement-non-opinion (sd)	Me, I'm in the legal department.
Acknowledge (Backchannel) (b)	Uh-huh.
Statement-opinion (sv)	I think it's great
Yes-No-Question (qy)	Do you have to have any special training?
Abandoned or Turn-Exit (%)	So, -

Table 1: Some Dialog Act Markup in Several Layers (DAMSL) tags

edge. The model is then prompted to generate the 11^{th} line. For evaluation, both the model-generated responses and the corresponding human responses from the SwDA corpus were automatically annotated with dialogue act labels using the BERT classifier. We then compared these labels along two dimensions: (1) distributional differences, by calculating the relative frequency of each dialogue act across model and human outputs; and (2) instance-level agreement, by measuring how often the dialogue act assigned to the model's response matched that of the gold standard human response in the same conversational context. In addition to these automatic comparisons, we manually inspected a random sample of 100 model–human pairs to qualitatively assess whether the classifier produced sensible labels and whether mismatches reflected genuine conversational differences rather than classifier errors. This qualitative check suggested that while the classifier rarely mislabels uncommon dialogue acts, the overall trends are robust.

We follow the same set of experiments with 30 lines of utterances instead of 10, and again with no prior context (zero-shot learning) where we prompt the model to converse with the user without having context knowledge about the utterances or dialogues of the assistant, to gain a better perspective of the extent to which prompt engineering influences the ability of dialogue assistants to engage in human-like conversations. It is essential to gather diverse system-generated responses to perform further analysis on the trends of the generated dialogues focusing on patterns, consistencies, and areas of improvement or divergence, as compared to human dialogues.

Our final experiments use the best performing context length of 30 previous utterances.

Dialogue Act Tags	Human	Llama 2	Mistral
Statement-non-opinion (sd)	35.29	18.19	17.64
Statement-opinion (sv)	23.52	36.37	35.29
Yes-No-Question (qy)	18.47	27.24	29.41
Acknowledge/Backchannel (b)	11.76	13.63	11.76

Table 2: Percentages of Dialogue Act tags of the selected utterances from the SwDA dataset, Llama 2 and Mistral

4 Results and Analysis

Table 2[3] shows the distribution of the top 4[4] dialogue act labels for the original utterances from the SwDA dataset next to the distribution of dialogue act tags among Llama 2's and Mistral's generated responses after conditioning the model on 30 lines of previous conversational context.

We see that humans use non-opinion statements significantly more compared to Llama 2 and Mistral. Both the models generate more opinion statements and questions, compared to humans. Although the models exhibit similar performance, Llama 2 demonstrates a greater tendency to acknowledge the provided prior context. However, solely measuring the overall distribution of dialogue acts might not be the most efficient method for identifying whether a dialogue act has been altered in the generated utterance compared to its original classification in the dataset.

Next, we consider the dialogue act tag for the generated versus the original sentence and provide a normalized confusion matrices, Figure 1 and Figure 2[5], indicating how many kept the same tags in the models' classification outcomes.

We see both LLMs predominantly respond with 'statement-opinion', as corroborated by Table 2.

[3]This table shows the distribution of dialogue acts from the gold standard 'next line' dialogues available from the transcripts after 30 lines of utterances, not the entire SwDA.

[4]Top 4 dialogue act tags are shown as they account for more than 85% of the dataset.

[5]Please refer to Table 2 for the list of dialogue act abbreviations.

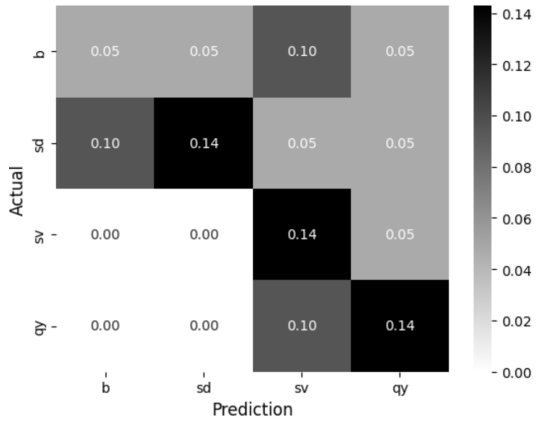

Figure 1: Confusion Matrix (Actual vs. Llama 2 13B Responses)

Llama 2 demonstrates a superior ability to align with the prompt and respond in a manner consistent with the given instructions, whereas Mistral encounters difficulties in following follow-up instructions related to a previous prompt. However, both models exhibit sub-optimal performance in accurately replicating outcomes compared to the ground truth data for our task. Further discussion on the models' accuracy is provided in the following sub sections.

4.1 System Prompt and Context Lengths

While running the experiments, we find that shorter and concise system prompts result in improved performance outcomes, whereas the utilization of broader prompts yields comparatively inferior results[6]. Longer system prompts, containing a significant amount of information, overwhelms the model and lead to incorrect associations. Whereas small and precise system prompts offer clearer guidance to the model, reducing ambiguity and potential confusion.

We additionally experiment with various context lengths. First, we provide 10 lines of conversation between the conversation participants from the transcripts and ask the model to predict the next utterance. This short context length causes a high level of ambiguity in the generated response irrespective of the topic of conversation:

```
Assistant: Yeah, I know, it's kind
           of surprising, right?
Assistant: Oh wow, that's surprising.
Assistant: Wow, that's something.
```

Expanding the preceding context to include 30 lines of previous utterances results in a notable reduction of 'surprise' exhibited by the models. Moreover, the models demonstrate enhanced proficiency in maintaining coherence and relevance throughout the conversation, akin to human conversational comprehension:

```
Assistant: Yeah, it was a pretty
           chaotic time, you know?
Assistant: Yes, it's about time we give
           equal importance to all health
           issues, regardless of who they
           affect.
```

Finally, we conduct an experiment devoid of any preceding user and assistant interactions, instead supplying only the one utterance and instructing the models to continue the conversation. In this scenario, the models unsurprisingly exhibit difficulty in following the conversation, as depicted in the dataset, often introducing novel topics and information to sustain the interaction. This observed behavior suggests a limitation in the model's capacity to adapt its conversational style without contextual cues, resembling the behavior commonly observed in open-domain dialogue systems. Additionally, we note a unique reaction of the system to a subset of transcripts, ones in which the conversations exhibit overt one-sidedness or lack engagement. These especially dry transcripts are characterized by an average utterance length of less than 6 words per utterance for one or both participants, such as the example taken from the dataset below:

Speaker A: Are you still there?
Speaker B: Yes.
Speaker A: Okay,
Speaker B: it worked out fine.
Speaker A: Okay.

Subsequent conditioning of the model with such dialogue context results in the generation of utterances aimed at concluding the conversation, rather than perpetuating dialogue that contributes minimally to its progression or substance, previously shown in Abbasiantaeb et al. (2024).

4.2 Perplexity

The perplexity of a large language model is a measure of its prediction effectiveness on a certain dataset. It measures how likely a model finds a sequence of words by calculating the exponentiation of the average negative log-likelihood of the predicted tokens. A lower perplexity indicates better performance, as the model's predictions are closer

[6]Section 3.1 provides examples of long and short system prompts.

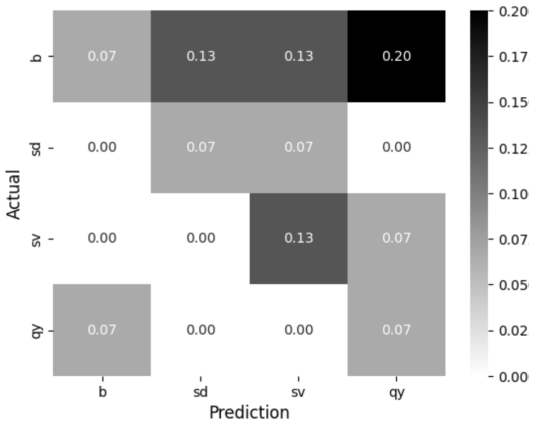

Figure 2: Confusion Matrix (Actual vs. Mistral 7B Responses)

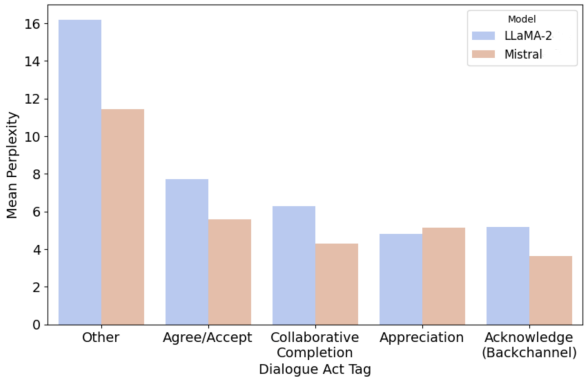

Figure 3: Top 5 Dialogue Acts with Highest Perplexity

to the actual sequence. We compute normalized perplexity (Roh et al., 2020) for both models on the gold response, using the first 30 lines of utterance from each conversation as context. Table 3 depicts the average normalized perplexity for the Llama 2 and Mistral models, where Mistral slightly outperforms Llama 2, indicating it produces more confident predictions.

Model	Average Perplexity
Llama 2	2.96
Mistral	2.13

Table 3: Perplexity Evaluation Summary

In order to study the models' responses to various types of dialogue acts, we compute the average perplexity for each tag, and sort the tags based on their perplexity scores (highest and lowest), shown in Figure 3 and Figure 4 respectively.

These figures effectively show what types of responses– types of dialogue acts, the model finds the most or least confusing.

The 'Other' dialogue act type causes the most

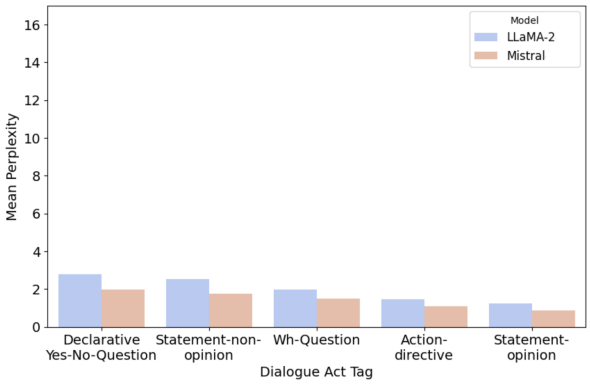

Figure 4: Bottom 5 Dialogue Acts with Lowest Perplexity

confusion for both models, i.e., both models struggle with less predictable dialogue types. This category includes utterances that do not clearly fit into any other dialogue act categories, encompassing statements such as correct-misspeaking, sympathetic comments, and greetings. The top five dialogue acts that contribute most to model confusion collectively account for less than 28% of the Switchboard dataset, with the majority stemming from the 'Acknowledge (Backchannel)' dialogue act (19%)[7]. This aligns with our findings that large language models are more likely to generate opinionated statements rather than simple agreement or acknowledgment, as the latter contributes minimally to advancing the conversation. Mistral outperforms Llama 2 in terms of lower perplexity for most dialogue act types.

The dialogue act which causes the least confusion among both the models is 'Statement-opinion', as both models tend to generate 'Statement-opinion' utterances (Burton et al., 2024), also inferred from Table 1. The five dialogue acts that lead to the lowest model confusion collectively comprise 52.4% of the Switchboard dataset, with the largest share attributed to the 'Statement-non-opinion' dialogue act (36%).

4.3 Classification Report for Llama 2 and Mistral

Table 4 and Table 5 show the precision, recall and F1 score for the top 4 dialogue act categories for the Llama 2 13B-chat model and Mistral-7B-Instruct model respectively.

The models achieved varying levels of performance across dialogue act categories. The overall

[7] https://web.stanford.edu/~jurafsky/ws97/manual.august1.html

DA Tags	Precision	Recall	F1
sd	0.75	0.43	0.55
sv	0.38	0.75	0.50
qy	0.50	0.60	0.55
b	0.33	0.20	0.25

Table 4: Precision, Recall and F1 for Llama 2

accuracy for the given categories is 0.48 and 0.33, respectively, for Llama 2 and Mistral, indicating significant room for improvement. In the context of human-like conversational carryover with prior knowledge, Llama 2 demonstrates a slight performance advantage over Mistral. Higher recall but lower precision in the Mistral model follows the previous result of lower perplexity but less aligned.

DA Tags	Precision	Recall	F1
sd	0.33	0.50	0.40
sv	0.40	0.67	0.50
qy	0.20	0.50	0.29
b	0.50	0.12	0.20

Table 5: Precision, Recall and F1 for Mistral 7B

5 Conclusions and Future Work

We analyze dialogues generated by Llama 2 and Mistral, using various levels of prompting and in-context learning, comparing them to the original human-human interactions from SwDA, utilizing dialogue acts to gauge similarity. In our research, we initially computed the percentages of the top four categories of dialogue act tags in both the original and corresponding LLM-predicted utterances. Our findings suggest that dialogue acts are not only a descriptive tool but also a potential predictor of when conversational systems fail to align with human norms. This analysis shows that in contrast to humans, who commonly use both opinionated and non-opinionated statements, language models exhibit a preference for generating opinion statements, potentially to add perceived value to the conversation. Additionally, these models tend to ask more questions, aiming to contribute more actively to the dialogue. Upon conducting further investigation using a confusion matrix, we discovered significant variations in the dialogue acts between the generated and original utterances, which were apparent in the differing proportions, as discussed previously. Llama 2's higher perplexity suggests that it might be more sensitive to context shifts and nuanced dialogue structures, resulting in more accurate classifications but higher uncertainty. Additionally, we found that context length significantly impacts response quality.

Future research could expand this work in several directions. One promising avenue is to evaluate newer models, to assess whether advances in training and alignment reduce the dialogue act discrepancies we observed. Another is to extend dialogue act analysis beyond distributional comparisons, incorporating metrics for conversational flow, user engagement, and appropriateness in interactive settings. Finally, integrating human-in-the-loop evaluations, where human participants interact with models and provide feedback on coherence and naturalness, would help connect dialogue act diagnostics more directly to real-world conversational quality.

Taken together, these directions highlight the potential of dialogue act analysis to bridge natural language processing and human–computer interaction, supporting the development of conversational systems that are not only fluent but also socially and pragmatically aligned with conversational norms.

6 Limitations

A key limitation of the evaluation is that it compares model-generated dialogue acts to those annotated in the Switchboard corpus, implicitly treating the human responses as a single "gold standard." In natural conversation, however, multiple dialogue acts could be appropriate in the same context, e.g., a turn could plausibly be an Acknowledgment, a Yes–No Question, or an Opinion statement depending on the speaker's intent. As a result, this comparison may underestimate the flexibility of LLMs or exaggerate deviations from human norms. We therefore frame our findings as diagnostic rather than definitive, using distributional patterns, such as the models' tendency to overproduce questions and opinions, to highlight systematic behavioral differences. Future work could incorporate human judgments, multiple reference responses, or metrics for conversational diversity and context-sensitive appropriateness, providing a more nuanced assessment of how well LLMs emulate human interaction.

Our study is limited to English conversations, since the Switchboard Dialogue Act corpus is available only in English. While this choice ensures

comparability with prior work and leverages a widely studied benchmark, it also restricts the generalizability of our findings. Dialogue acts and conversational norms vary across languages and cultures; for example, the use of backchannels, politeness markers, or indirect questions can differ substantially. Future work should extend this analysis to multilingual corpora, which would allow us to evaluate whether the dialogue act patterns we identify are specific to English or reflect broader conversational tendencies in LLMs.

The SwDA utilized in our research provides a comprehensive array of dialogue acts; however, it lacks representation of certain emotional expressions commonly employed by humans. While dialogue act tags serve as valuable markers for categorizing communicative intents in dialogue, in the particular dataset, they inherently lack the capacity to encompass certain nuanced aspects of human expression, such as sarcasm. Consequently, the absence of explicit consideration for such emotional nuances within dialogue act frameworks represents a notable limitation, potentially leading to incomplete or inaccurate characterizations of human utterances in conversational AI systems.

Sometimes, there arise instances where the dialogue act labels assigned to the generated utterances align with those found in the reference data, yet substantial differences exist in the semantic content or pragmatic context of the dialogues. Such divergences underscore the inherent complexity of assessing dialogue quality solely through dialogue act matching, as they indicate potential limitations in capturing the richness and subtleties of human conversation beyond surface-level categorizations.

References

Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. 2024. Let the llms talk: Simulating human-to-human conversational qa via zero-shot llm-to-llm interactions. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 8–17.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sher Badshah and Hassan Sajjad. 2024. Quantifying the capabilities of llms across scale and precision. *arXiv preprint arXiv:2405.03146*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jason W Burton, Ezequiel Lopez-Lopez, Shahar Hechtlinger, Zoe Rahwan, Samuel Aeschbach, Michiel A Bakker, Joshua A Becker, Aleks Berditchevskaia, Julian Berger, Levin Brinkmann, and 1 others. 2024. How large language models can reshape collective intelligence. *Nature human behaviour*, 8(9):1643–1655.

Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. 2018. Dialogue act recognition via crf-attentive structured network. In *The 41st international acm sigir conference on research & development in information retrieval*, pages 225–234.

Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. 2021. Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, 54:755–810.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Sidney D'Mello and Art Graesser. 2013. Design of dialog-based intelligent tutoring systems to simulate human-to-human tutoring. In *Where humans meet machines: Innovative solutions for knotty natural-language problems*, pages 233–269. Springer.

Riccardo Fusaroli and Kristian Tylén. 2016. Investigating conversational dynamics: Interactive alignment, interpersonal synergy, and collective task performance. *Cognitive Science*, 40(1):145–171.

Ananya Ganesh, Martha Palmer, and Katharina Kann. 2021. What would a teacher do? predicting future talk moves. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4739–4751.

Leah Henrickson and Albert Meroño-Peñuela. 2023. Prompting meaning: a hermeneutic approach to optimising prompt engineering with chatgpt. *AI & SOCIETY*, pages 1–16.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Dan Jurafsky. 1997. Switchboard swbd-damsl shallow-discourse-function annotation coders manual. *www. dcs. shef. ac. uk/nlp/amities/files/bib/ics-tr-97-02. pdf*.

Katharina Kann, Abteen Ebrahimi, Joewie Koh, Shiran Dudy, and Alessandro Roncone. 2022. Open-domain dialogue generation: What we can do, cannot do, and should do next. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 148–165, Dublin, Ireland. Association for Computational Linguistics.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Preprint*, arXiv:2302.02083.

Harshit Kumar, Arvind Agarwal, Riddhiman Dasgupta, and Sachindra Joshi. 2018. Dialogue act sequence labeling using hierarchical encoder with crf. In *Proceedings of the aaai conference on artificial intelligence*, volume 32.

Ji Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. *arXiv preprint arXiv:1603.03827*.

Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. Mathdial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. *arXiv preprint arXiv:2305.14536*.

David Nadeau, Mike Kroutikov, Karen McNeil, and Simon Baribeau. 2024. Benchmarking llama2, mistral, gemma and gpt for factuality, toxicity, bias and propensity for hallucinations. *Preprint*, arXiv:2404.09785.

Vipul Raheja and Joel Tetreault. 2019. Dialogue act classification with context-aware self-attention. *arXiv preprint arXiv:1904.02594*.

Eugénio Ribeiro, Ricardo Ribeiro, and David Martins de Matos. 2019. Deep dialog act recognition using multiple token, segment, and context information representations. *Journal of Artificial Intelligence Research*, 66:861–899.

Jihyeon Roh, Sang-Hoon Oh, and Soo-Young Lee. 2020. Unigram-normalized perplexity as a language model performance measure with different vocabulary sizes. *CoRR*, abs/2011.13220.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2021. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*.

John R Searle, Ferenc Kiefer, Manfred Bierwisch, and 1 others. 1980. *Speech act theory and pragmatics*, volume 10. Springer.

Omar Shaikh, Kristina Gligorić, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2024. Grounding gaps in language model generations. *Preprint*, arXiv:2311.09144.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca A. Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *CoRR*, cs.CL/0006023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Quan Hung Tran, Ingrid Zukerman, and Gholamreza Haffari. 2017. A hierarchical neural model for learning sequences of dialogue acts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 428–437.

Qinyuan Wu, Mohammad Aflah Khan, Soumi Das, Vedant Nanda, Bishwamittra Ghosh, Camila Kolling, Till Speicher, Laurent Bindschaedler, Krishna P Gummadi, and Evimaria Terzi. 2024. Towards reliable latent knowledge estimation in llms: In-context learning vs. prompting based factual knowledge extraction. *arXiv preprint arXiv:2404.12957*.

Zihao Yi, Jiarui Ouyang, Yuwen Liu, Tianhao Liao, Zhe Xu, and Ying Shen. 2024. A survey on recent advances in llm-based multi-turn dialogue systems. *arXiv preprint arXiv:2402.18013*.

Yucan Zhou, Qinghua Hu, Jie Liu, and Yuan Jia. 2015. Combining heterogeneous deep neural networks with conditional random fields for chinese dialogue act recognition. *Neurocomputing*, 168:408–417.