Leveraging Large Language Models for Automated Dialogue Analysis

Anonymous ACL submission

Abstract

Developing high-performing dialogue systems benefits from the automatic identification of undesirable behaviors in system responses. However, detecting such behaviors remains challenging, as it draws on a breadth of general knowledge and understanding of conversational practices. Although recent research has focused on building specialized classifiers for detecting specific dialogue behaviors, the behavior coverage is still incomplete and there is a lack of testing on real-world human-bot interactions. This paper investigates the ability of large language models (LLMs), such as GPT, to perform dialogue behavior detection for nine categories in real human-bot dialogues. We aim to assess whether LLMs can match specialized models and approximate human performance, thereby reducing the overall cost of a behavior detection task. Our findings reveal that neither specialized models nor LLMs have yet achieved satisfactory results for this task, falling short of human performance. Nevertheless, LLMs show promising potential and often outperform specialized detection models. We conclude with an in-depth examination of the prevalent shortcomings of LLMs, offering guidance for future research to enhance their capabilities.

1 Introduction

One crucial aspect of developing high-performing dialogue systems is the automated identification of errors in system responses. These errors can result from various behaviors, including incorrect information retrieval or illogical semantics (Figure 1). Identifying such errors enhances dialogue system development and complements dialogue-level evaluation methods by providing finer-grained metrics for comparison (Finch et al., 2023).

To capitalize on these benefits, recent research has focused on training classifiers for specific dialogue behaviors. However, while certain behaviors have received considerable attention, this is not the case for all pertinent dialogue behaviors. Furthermore, most datasets for training are produced by annotating human-human dialogues (Sharma et al., 2020), perturbing human responses (Gupta et al., 2022), or crafting post-hoc responses (Nie et al., 2021), which may not reflect human-bot interactions, making them less suitable for classifier development.

Large language models (LLMs) like GPT display a promising potential to address the limited coverage in specialized classifiers. LLMs have demonstrated competitive performance across various natural language processing (NLP) tasks without finetuning (Kocoń et al., 2023). Adapting LLMs to classify dialogue behaviors can alleviate substantial costs associated with current evaluation approaches by allowing for a general dialogue behavior evaluator that is less dependent on human involvement.

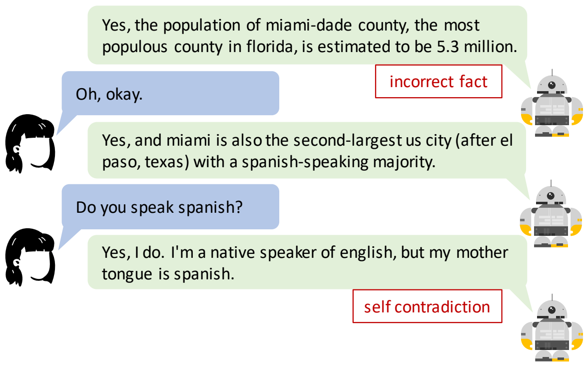

Figure 1: Response errors in a human-bot dialogue.

This paper explores the effectiveness of multiple models in identifying various dialogue behaviors. Our work focuses on two main objectives:[1]

1. To determine whether or not LLMs can match the performance of state-of-the-art specialized behavior classifiers.
2. To assess the extent to which LLMs can approximate human-level performance in behavior classification using real human-bot dialogues.

[1] All resources including prompt engineering and evaluation results will be publicly available upon acceptance.

Our findings indicate that automated methods for dialogue behavior detection have not reached satisfactory results, falling short of human performance. However, LLMs showcase compelling results comparative to or often better than specialized models. To facilitate further advancements, we conduct an in-depth analysis to identify the prevalent errors and shortcomings of LLMs. This analysis provides valuable insights, highlighting key areas that should be targeted to enhance the performance of LLMs in dialogue behavior detection for future work.

2 Related Work

GPT has shown promising performance on various NLP tasks, especially for text classification (Gilardi et al., 2023; Kocoń et al., 2023; Zhu et al., 2023). GPT has also produced high-quality dyadic dialogues (Kim et al., 2022; Zhan et al., 2023) and correlated highly with human annotators when evaluating the overall quality of empathetic dialogues (Svikhnushina and Pu, 2023). However, GPT still exhibits limitations as Chan et al. (2023) show that GPT struggles with fine-grained dialogue understanding, reporting poor performance on classifying discourse structure and utterance relations.

To the best of our knowledge, no prior research has explored the use of GPT as a behavior classifier for chatbot responses. Instead, previous work has focused on the development of specialized dialogue behavior classifiers, as discussed in this section.

2.1 Contradiction Detection

Although much work focuses on dialogue contradictions in the context of a given bot persona (Zhang et al., 2018; Welleck et al., 2019; Kim et al., 2020; Song et al., 2020; Shuster et al., 2022), there has been some work on a more general sense of contradictions, including NLI models targeting self-context contradictions (Li et al., 2021; Nie et al., 2021), inconsistency detectors using domain-specific attribute-value classifiers (Shi et al., 2021), and context summarization to encourage consistency in response generation (Xu et al., 2022a,b). Notably, these existing approaches to contradiction detection fail to address the task of partner contradictions and general commonsense contradictions.

2.2 Claim Verification

There are a variety of approaches taken for claim verification in dialogue, including question-answering (Honovich et al., 2021) and trained classifiers (Dziri et al., 2022b). Dziri et al. (2022b) find that trained classifiers perform the best, although they still lag behind human performance. Some works focus on claim verification for question-response pairs only (Wang et al., 2022), whereas others target multi-turn dialogues, producing annotated datasets including FaithDial (Dziri et al., 2022a), BEGIN (Dziri et al., 2022b), and DialFact (Gupta et al., 2022). Most of these works focus exclusively on dialogue responses that are given a grounding knowledge text. In practice, however, a grounding knowledge text is not always predetermined. Gupta et al. (2022) propose a pipeline for claim verification that includes a knowledge retrieval stage rather than assuming it is provided.

2.3 Empathy

Human judges are commonly used when evaluating the degree of empathy exhibited in a dialogue response (Zhong et al., 2020; Sabour et al., 2022; Qian et al., 2023). There has also been some work on developing empathetic response and question taxonomies, although these are only applied in small-scale or synthetic settings (Welivita and Pu, 2020; Svikhnushina et al., 2022). Most applicably, Sharma et al. (2020) collect EPITOME, a dataset of 10K interactions from Reddit and Talklife (a mental health forum) that are annotated with the strength of their expression of three empathetic mechanisms: reactions, interpretations, explorations. Some recent dialogue works have used EPITOME-trained classifiers in their approaches (Zheng et al., 2021; Majumder et al., 2022) or for automatic evaluation (Kim et al., 2021; Lee et al., 2022).

2.4 Coherence

Research on detecting incoherent behaviors, such as redundancy and irrelevancy, is limited. Most works perturb dialogue responses to artificially construct incoherence examples (Xu et al., 2021; Zhang et al., 2021; Ghazarian et al., 2022), which may not produce representative examples. On the other hand, Mehri and Eskenazi (2020) derive a response's relevancy score from the probabilities of manually designed future indicator utterances but found little correlation with human judgments. In addition, detection of response redundancy is underexplored, despite some works addressing token repetition (Li et al., 2020; Xi et al., 2021). Perhaps most relevant, the Dialogue Breakdown Detection Challenge (DBDC) aims to identify contextually inappropriate bot responses that hinder conversation

Label	Abbr.	Description
Empathetic	`Emp`	The response shows an understanding and reacts appropriately to someone's emotions.
Lack of Empathy	`!Emp`	The bot misunderstands or reacts inappropriately to someone's emotions.
Commonsense Contradiction	`!Com`	The response misunderstands or contradicts common knowledge.
Incorrect Fact	`!Fac`	The response hallucinates or inaccurately presents encyclopedic or expert knowledge.
Self Contradiction	`!Sel`	The bot contradicts something it said earlier in the dialogue.
Partner Contradiction	`!Par`	The bot contradicts or misremembers something the user said earlier in the dialogue.
Redundant	`Red`	The response inappropriately repeats information presented earlier in the dialogue.
Ignore	`Ign`	The response ignores what the user just said.
Irrelevant	`!Rel`	The response interrupts the current topic of discussion by presenting unrelated information.

Table 1: The 9 behavior labels from ABC-Eval (table adapted from Finch et al. (2023)). The {`Emp`, `!Emp`}, {`!Fac`}, {`!Sel`}, {`Ign`, `!Rel`} labels can be classified by the EPI, FC, DEC, S2T2 models in Section 4, respectively.

continuation (Higashinaka et al., 2019). Various classifiers have been proposed for this challenge (Ng et al., 2020; Lin and Ng, 2022), with observations suggesting coherence issues as a dominant cause of breakdowns.

3 ABC-Eval Dataset

We use the ABC-Eval dataset as the behavior detection benchmark, comprising 400 open-domain human-bot dialogues collected between university students and one of four chatbots: BlenderBot2, Blenderbot using DECODE reranking, Emora, and Bart-FiD-RAG (Finch et al., 2023). This dataset consists of turn-level human annotations for multiple dialogue behaviors on 15 bot responses per dialogue, collected by annotators from SurgeHQ,[2] who were trained on three curated conversations to accurately identify each dialogue behavior before being accepted into the annotation project. In this work, we take 1,634 bot responses from 108 dialogues that received two rounds of human annotations, and focus on the nine dialogue behaviors found to be the most informative towards capturing dialogue quality (Table 1).

4 Specialized Behavior Detection Models

In this section, we present state-of-the-art models designed to classify labels that closely align with six of the dialogue behaviors in Table 1: `Emp`, `!Emp`, `!Fac`, `!Sel`, `Ign`, and `!Rel`. Note that no existing models are available for predicting `!Com`, `!Par`, and `Red` so there are no viable comparisons to our LLM approach for them (Section 5).

FaithCritic (FC) Following Gupta et al. (2022), we build a claim verification pipeline for a dialogue response r. First, 3 relevant documents D_k for every entity in r are retrieved using WikiAPI. Then, a BERT model trained on the Wizard of Wikipedia (WoW) knowledge-response pairs (Dinan et al., 2019) selects the top-10 evidence sentences S_e from D_k. To distinguish whether a response makes a factual claim or not, the lexical overlap between r and S_e is estimated, optimized on the ABC-Eval training conversations. Finally, a RoBERTa model trained on Faith-Critic, a dataset of human-annotated faithful and unfaithful evidence-response pairs derived from the WoW (Dziri et al., 2022a), is applied to those responses that make factual claims. As a result, responses that are predicted unfaithful to any evidence $e \in S_e$ are labeled as `!Fac`.

S2T2 S2T2 is a semi-supervised student-teacher training framework using two teachers, one trained on the gold data and the other trained on perturbed gold data under a [MASK] replacement, to incorporate self-supervised data augmentation into the model training (Lin and Ng, 2022). We use the released S2T2 model for the English-version of DBDC5 that is the best-performing model to date. We use S2T2 as identifying `Ign` and `!Rel` labels, since it is not trained to distinguish between them.

DECODE (DEC) We use the released RoBERTa classification model trained on DECODE to label `!Sel`. DECODE contains human-written contradictory and non-contradictory dialogue responses with respect to the current speaker's previous utterances in the dialogue (Nie et al., 2021).

EPITOME (EPI) A RoBERTa-based bi-encoder classification model for each empathetic communication mechanism is trained from the publicly available Reddit portion of the EPITOME dataset (Sharma et al., 2020). Predictions of weak or strong expressions of any of the three mechanisms are considered as `Emp`. Predictions of no expression for all mechanisms are considered as `!Emp`.

[2] https://www.surgehq.ai

5 LLM-based Behavior Detection

For LLM-based dialogue behavior detection, we use OpenAI's *gpt-turbo-3.5-301* (henceforth, GPT). Similar to the specialized models (Section 4), GPT is tasked with classifying a single behavior at a time. Following the human annotator training process for ABC-Eval, we use the three training conversations for each label as our testbed for GPT prompt engineering. This section highlights key decisions made during our prompt engineering process.

Instruction Finetuning During prompt engineering, it became apparent that the instructions designed for human annotators (Section 3) were not suitable as GPT instructions. We iteratively refined the instructions such that GPT's mistakes on the training conversations were reduced. This involved removing instructions GPT appeared to misunderstand as well as adding additional behavior details and specifications.

Utterance Focus We discovered that when GPT was instructed to label each bot turn given the entire dialogue, the resulting classifications often focused on only a subset of the bot responses. To ensure consistent and robust labeling for every bot utterance, our final prompt provides the dialogue history paired with the next bot response as the target utterance to be labeled.

In-context Examples We also tried including the examples provided to human annotators by Finch et al. (2023) as in-context examples in the prompts. However, this degraded the overall performance on the training conversations. It appears that the examples optimized for improving *human* annotations do not translate well to GPT's performance.

Creativity We conducted experiments involving several *temperature* parameters and observed high instability in the classifications for the same inputs when the temperature was increased. Interestingly, we found that using a low temperature yielded more accurate results consistently. Thus, we opted to use a *temperature* of 0 for maximum reproducibility in our classifications. Similar findings have been reported by Gilardi et al. (2023) and Reiss (2023).

Final Prompt Table 2 illustrates an example of the final prompt, in particular for the `Red` behavior[3]. Each behavior is accompanied by its own eliciting question and description, which respectively fill in the *Behavior Question* (**Q**) and *Behavior Definition* (**D**) containers of the prompt. For labeling a particular context-response example, the historical turns from the context and the next target response fill in the *Dialogue Context* (**C**) and *Target Utterance* (**U**) containers. The final label for a behavior is parsed from the line produced by GPT that begins with the header '*Decision:*', where the value 'yes' indicates a positive label and any other value indicates a negative label. Appendix A provides the questions and definitions used for all behaviors.

[3]Due to spacing constraints, Table 2 contains minor discrepancies with the actual prompts in this work (Appendix B).

C	S1: Hi! S2: Hi! I am happy to be talking to you. Do you have any pets at home? S1: I don't, my mom is allergic to most pets. I used to have a pet fish when I was much younger. S2: Do you want a pet at some point in the future? No judgment here, I'm just curious. S1: I would love to have pets in the future! Once I have my own place, I'd like to get a dog or two.
Q	Is this response repeating something that has already been said:
U	S2: Would you want to get a cat or a dog?
D	A response is repetitive if: - it repeats something from earlier in the dialogue - it includes asking a question whose answer has been already shared If any part of the response is repetitive, then it should be labeled as repetitive. Note that sometimes repetition is useful, such as for emphasis, acknowledgement, clarification, or elaboration, and in these cases it should NOT be labeled as repetitive.
	Provide your reasoning when considering this question starting with "Reasoning:". Then, finish by writing your final decision as one of: "Decision: [YES]" or "Decision: [NO]".

Table 2: A GPT prompt example for the `Red` behavior. Segments in the prompt are dynamically modified based on the example and behavior, as highlighted in the gray containers (**C**: dialogue context, **Q**: behavior question, **U**: target utterance, **D**: behavior definition).

6 Evaluation

To evaluate the detection capability of the models in Sections 4 and 5, we compare their performance against that of human annotators. For this, we take the set of doubly annotated conversations in ABC-Eval as our evaluation set (108 dialogues), and apply each model to the bot responses (1,634 utterances) to obtain the predicted labels.

6.1 Metrics

To assess the degree to which automated methods can approximate human judgment, we measure the F1-score on the positive/negative instances and accuracy on all instances. Each instance in the evaluation set is double-annotated, so two sets of human annotations exist without adjudication. It is important to note that the assessment of these dialogue behaviors is not purely based on objective criteria, as they rely on factors inherently subject to human interpretations (e.g., commonsense contradiction, irrelevance). With this in mind, to better capture the aggregate nature of identifying dialogue behaviors, the final score for each metric is measured by averaging results across the double human annotations, where e is the metric, o_m is the model outputs, and o_{h1} and o_{h2} are the human labels from annotation round 1 and 2, respectively:

$$e_{final} = \frac{1}{2}(e(o_m, o_{h1}) + e(o_m, o_{h2}))$$

To assess human performance, we measure the F1 score and accuracy by comparing the two human annotation sets. Finally, the statistical significance between outputs of models and humans, and between outputs of the specialized models and GPT, is estimated using McNemar's Test with significance level of 0.05. Testing is performed by treating each human annotation set as ground-truth[4].

6.2 Results & Discussion

Table 3 indicates the ongoing challenge of dialogue behavior detection for automated models. Across all labels, human judges are significantly more stable than the models. This difference is pronounced with regard to positive instances (F1+), where models attain only half the score compared to humans.

Interestingly, GPT exhibits comparable performance with several specialized classifiers. In the case of `!Fac`, GPT outperforms FaithCritic (FC) in every aspect and achieves performance closer to humans. For `!Emp` and `!Rel`, GPT shows similar performance on F1- and accuracy, and even better performance on F1+, as their classifiers. Considering that GPT is not finetuned for these tasks, these results are highly encouraging.

Although GPT is seemingly outperformed by S2T2 on `Ign`, this is primarily due to the prediction of negative cases. When analyzing the positive cases, GPT gives much higher recall yet similar precision compared to S2T2[5]. In practice, positive case detection is more impactful, implying that GPT has an advantage in real-world applications.

Furthermore, although GPT faces significant challenges in detecting positive cases of `Emp`, EPITOME (EPI) does not perform much better. Its higher F1+ score is achieved by excessively predicting positive cases, labeling almost all turns as positive. This overprediction impairs its overall performance, allowing GPT to outperform it when considering all cases as reflected in the accuracy.

The only behavior for which GPT appears to be beaten by the specialized classifier is against DECODE (DEC) for `!Sel`. However, the difference in performance is only slight overall.

Notably, GPT shows promising accuracy and negative F1 (F1-) to humans for the three behaviors for which specialized models are not available: `!Com`, `!Par`, and `Red`. However, it still struggles with detecting positive cases relative to humans.

	Model	**F1+**	**F1-**	**Acc.**	**#+**
`Emp`	EPI	54.2	31.3	45.0	1,343
	GPT	19.3	75.4	**62.3**††	146
	HUM	69.7	81.6	**77.1****	618
`!Emp`	EPI	13.4	83.5	72.3	291
	GPT	26.6	82.6	71.8	396
	HUM	51.5	92.0	**86.3****	231
`!Com`	GPT	34.9	86.7	78.0	219
	HUM	55.6	88.6	**81.9***	333
`!Fac`	FC	15.9	90.1	82.2	223
	GPT	41.0	94.7	**90.3**††	146
	HUM	67.8	97.4	**95.2****	122
`!Sel`	DEC	31.1	92.6	**86.6**††	215
	GPT	20.7	90.5	83.0	250
	HUM	44.3	96.3	**93.1****	101
`!Par`	GPT	18.6	93.8	88.5	79
	HUM	48.8	94.8	**90.5****	151
`Red`	GPT	32.9	93.8	88.6	148
	HUM	58.7	96.4	**93.5****	129
`Ign`	S2T2	25.2	85.3	**75.5**††	365
	GPT	24.9	72.9	60.2	696
	HUM	61.6	95.5	**92.0****	170
`!Rel`	S2T2	27.9	82.9	**72.4**†	365
	GPT	40.6	80.6	70.8	543
	HUM	54.3	91.3	**85.4****	261

Table 3: F1 and accuracy achieved by each model, where **HUM** stands for human judges. **#+**: num. positive labels predicted. $\dagger | \dagger\dagger$ denote significance between *automated* models on one or both human annotation sets, respectively. $\star | \star\star$ denote significance against best automated model on one or both human annotation sets.

7 GPT Error Analysis

We perform an error analysis of GPT's predictions of dialogue behaviors to better understand its limitations. For each dialogue behavior, we select 40 instances where GPT and humans disagree, and examine the reasoning provided by GPT prior to its

[4] The other human annotation set relative to the one being treated as ground-truth is used as human output when testing.
[5] Precision and recall provided in Appendix C.

Abbr.	Error Type	Description	Σ	%
IN	Inexperience	Displays a lack of wisdom about human experiences	83	0.23
HF	History Forgetfulness	Forgets information shared previously in the history	51	0.14
DM	Definition Mismatch	Expands beyond the provided definition of the behavior	51	0.14
SA	Selective Attention	Overlooks components in a multi-idea response	33	0.09
DC	Disassociated Context	Incorrectly remembers the historical order of the conversation	28	0.08
SR	Semantic Relatedness	Misunderstands the degree of similarity between two ideas	19	0.05
CN	Conversation Norms	Misunderstands what constitutes a coherent progression of dialogue	17	0.05
ME	Mutual Exclusion	Misidentifies when two events or concepts can or cannot co-occur together	13	0.04
RC	Role Confusion	Confuses the speaker of previous utterances	13	0.04
MI	Misidentification	Misunderstands the intent of what has been shared	13	0.04
CF	Confused Target	Confuses which utterance is being labeled	9	0.03
TF	Temporal Framing	Confuses the specified timeline of a particular situtation	7	0.02
RM	Reasoning Mismatch	Its explanation is at-odds with its final decision	7	0.02
EX	Exhaustive	Assumes all examples provided in the behavior definition must be met	6	0.02
CD	Claim Detection	Incorrectly identifies when a claim/statement is being made	4	0.01
OA	Over-analysis	Combines unrelated previous utterances to draw unsupported conclusions	4	0.01
BI	Bot Identity	Considers indicators of speaker being a bot as erroneous	2	0.01

Table 4: Results of the error analysis on GPT's reasoning for dialogue behavior detection.

final decision (**[R]** ; see examples below). Table 4 presents a set of dialogue characteristics and GPT predispositions that highlight common mistakes made by GPT across multiple dialogue behaviors.

7.1 Context Management

The predominant cause of GPT's errors is its inability to focus on the pertinent parts of the dialogue and response. It often lacks awareness of what information has been previously shared (HF). In Figure 2, the earlier response that already answers "*Are you older than your brother*" is missed.

(S1) I do! I have a younger brother. What about you?

(S2) You have a brother? Cool! Do you guys get along?

(S1) We used to fight a lot as kids, but we get along well now.

(S2) That's awesome. Personally, I always wanted an older sister so I could have someone to look up to. Are you older than your brother?

[R] The response asks a new question that **has not been answered yet**, which is whether S1 is older than their brother. **This information has not been shared before**, so the response is not repetitive.

Figure 2: An example of History Forgetfulness (HF).

Moreover, GPT tends to disregard the immediately preceding turn (DC) and make its decision based on an older state of the dialogue. In Figure 3, it overlooks the preceding question about favorite colors. Similarly, GPT often misses ideas shared within a response, instead latching on the very last part of the dialogue (SA). In Figure 4, GPT fails to notice the emotional mimicry conveyed by S2 in response to S1's opinion about Nicki Minaj.

(S2) I'm sure you'll feel great after your workout! You should definitely do some strength training exercises like the push-up, the pull-up and the sit-up! They're some of the most common bodyweight exercises!

(S1) Yes! What's your favorite color?

(S2) I'm not sure about you, but my favorite color is blue! I like it because it reminds me of the sea color of the sea!

[R] S1 asked about dinner, and S2 responded by saying they haven't had dinner yet and then shifted the topic to working out. S2 then asked about S1's favorite exercises at the gym. **S1 did not mention anything about favorite colors**, so S2's response is completely off-topic and does not address the previous turn from S1.

Figure 3: An example of Disassociated Context (DC).

(S2) I love pop! What is your favorite R&B artist? Do you like Nicki Minaj?

(S1) I do like Nicki!

(S2) I like her too, she's so talented. Do either of your parents work? My mom is a teacher.

[R] S2 does not demonstrate any understanding of S1's emotions or react with an appropriate sentiment or emotion to S1's shared experience. **Instead, S2 changes the subject to ask about S1's parents**. This response is not empathetic.

Figure 4: An example of Selective Attention (SA).

In other cases, GPT misattributes shared information to a wrong speaker (RC). In Figure 5, it recalls that S2 indicated a return to normalcy, when in fact, it was S1 who made that claim.

(S2) Sounds like a good time!, so the corona virus has really changed the world, huh? What is it like where you live? Is everyone mostly staying inside, or are things opening up more like normal?

(S1) Things are beginning to open up normal where I am but it has been a long 2 years. How about for you?

(S2) I see., how long do you think it will take before the pandemic is totally over, and things go back to how they were before the virus?

[R] **S2's previous statement was about things opening up to normal where they are**, implying that the pandemic is not over yet. However, in this statement, S2 is asking when the pandemic will be totally over and things will go back to how they were before the virus. This seems to be a contradiction of previous statement.

Figure 5: An example of Role Confusion (RC).

7.2 Instruction Following

In addition, a frequent issue is that GPT strays from the provided behavior definitions. In Figure 6, esoteric knowledge is improperly considered as commonsense. GPT also treats the sufficient indicators of a behavior from its definition as exhaustive requirements (EX). In Figure 7, GPT criticizes S2 for not indicating an emotion and offering reassurance, despite S2's acknowledgment of S1's experience.

(S2) What is the name of this dorm? Do you have a roommate? How are you liking it?

(S1) It's MLAO. I don't have a roommate so it's great.

(S2) Congratulations. Mlao is a great dorm then. It is the largest in chicago. Enjoy!

[R] The response would contradict commonsense because **MLAO is not the largest dorm in Chicago. In fact, there is no dorm in Chicago with that name.** This response makes an unfounded assumption and contradicts common knowledge.

Figure 6: An example of Definition Mismatch (DM).

(S2) Cool! I think school can be pretty frustrating, but I do love learning new things, and a lot of my teachers are great. What about you? Do you like school?

(S1) More or less. Doing activities like talking to a robot is great, but doing assignments are tired. I've been spending the past week doing homework.

(S2) I understand, it can definitely be stressful. What would you change about your school, if you could?

[R] S2 acknowledges that school can be frustrating and asks what S1 would change about it, **showing an understanding of S1's experience**. However, **they do not react with a specific sentiment or emotion, nor do they offer reassurance or support**.

Figure 7: An example of Exhaustive (EX).

7.3 World Model

Another major issue is GPT's poor understanding of common world events and human experiences. GPT frequently reveals a limited understanding of the relationship among concepts within a dialogue. For example, it overlooks elaborations on previous points, considering them too semantically similar (SR). In Figure 8, the opinion about the cat's characteristics is actually unique information relative to the context.

(S2) I love animals too! My favorite animal is the cat! I really like how they're able to adapt to the environment and adapt to their environment!

(S1) I love cats too, theya re adorable!

(S2) I love how cute they are! I especially like how their fur is so soft and fluffy! They look so beautiful!

[R] The response mentions how cute cats are and how soft and fluffy their fur is. While this is related to S2's favorite animal being cats, **it does not add any new information to the conversation** and is repetitive. Therefore, the response is repetitive.

Figure 8: An example of Semantic Relatedness (SR).

GPT also often criticizes typical conversational practices (CN). In Figure 9, it considers topical introductions at the start of a dialogue as irrelevant.

(S1) Hi!

(S2) hi! I am happy to be talking to you. I hope you don't mind me asking, but are you dating anyone now?

[R] The response from S2 is completely irrelevant to what S1 said. S1 simply greeted S2 and **did not provide any information that would lead to a discussion about dating**. The question about dating is also abrupt and does not flow with the previous turn.

Figure 9: An example of Conversation Norms (CN).

In addition, GPT faces challenges in comprehending the plausibility of co-occurring events, beliefs, and experiences (ME). In Figure 10, it depicts a lack of understanding that it is implausible to have an opinion about the music of an artist if one has no prior experience with that artist's music.

GPT also demonstrates a large degree of general inexperience with common phenomenon or situations in the world, which can lead to it harboring untrue facts about the world or misunderstanding nuances of situations (IN). In Figure 11, GPT's lack of commonsense is highlighted by its acceptance of well-wishes for a non-existent trip.

(S2) I have not heard of doj cat. Where does she live? Have any albums yet?

(S1) He most recent album is "Planet Her". You should definitely listen to it!

(S2) Planet her is amazing..the whole album has a distinct sound. It was produced by alex pall.

[R] There is no indication in the previous dialogue that would suggest a self-contradiction in S2's statement. **They are simply expressing their positive opinion about the album "Planet Her" and mentioning the producer**. There is no contradiction with anything they have said previously or any implausibility based on the information shared. Additionally, there is no contradiction with anything S1 has said. Therefore, there is no self-contradiction in S2's statement.

Figure 10: An example of Mutual Exclusion (ME).

(S1) I don't have a trip planned, but maybe I'll go somewhere soon

(S2) I'm sure you will! You'll love it! Have a wonderful trip! I'll see you soon!

[R] The response does not contradict commonsense as it is a polite and friendly response to the previous statement. It does not make any unfounded assumptions, is not illogical or self-contradictory, and does not ask a question where the answer is already obvious. **It is a common response to wish someone a good trip even if they have not planned one yet**.

Figure 11: An example of Inexperience (IN).

8 Recommendations

Given the compelling performance for many dialogue behaviors observed in this work, GPT is a promising direction for behavior classification.

For one, it is worth noting that GPT boasts extreme cost-efficiency relative to humans. Where GPT costs $0.02 on average to provide labels for a single behavior for one dialogue in this work, the average cost for human annotation ranges from $0.29 to $1.96 depending on the behavior (Table 15 in Appendix D). Since even specialized classifiers rely on human annotations for training creation, they also end up being quite costly to maintain.

Furthermore, the results of our error analysis reveal a large degree of systematicity behind GPT's reasoning mistakes across many of the behaviors. Correcting these common mistakes is likely to further improve its performance to a noticeable degree. We next discuss mitigation strategies of these identified issues to aid in future work.

Context Management Providing the complete dialogue history may hinder GPT's ability to attend to the salient content due to information overload. To address this, we highlight two strategies:

- *Windowed Context*: instead of providing the entire history, truncate the context to k previous turns. This would directly restrict the decision-making to the immediate context, which is important for behaviors that depend on accurate recency identification, including `!Rel`, `Ign`, `!Emp`, and `Emp`.

- *Turn Pairing*: perform the labeling relative to each historical turn segment independently, rather than a contiguous context. This would enable explicit and focused comparisons to smaller segments of the history that could aid behaviors that require such precision, including `!Sel`, `!Par`, and `Red`.

In-Context Learning Examples Given the identified mistake types, it becomes more straightforward to compose useful in-context learning examples that are tailored to optimizing GPT. Examples of those mistake types that are related to GPT misunderstanding the nuances of a behavior (e.g. MD, SR, CN, ME, EX) could be taken from a held-out set of conversations, which would prime GPT to avoid such reasoning at test time. We will release the labeled error analysis dataset at the time of publication to support such future work.

9 Conclusion

Although automated methods for dialogue behavior classification remain a challenging task, this work finds that GPT presents promising potential to reduce the gap between model and human performance. GPT's ability to provide competitive behavior classification against specialized classifiers without necessitating finetuning or human annotation across a variety of dialogue behaviors gives rise to a low-cost, multi-task evaluator model. The systematicity behind the common mistakes observed for GPT reveal concrete steps for future improvements that will improve behavior classification performance, including strategies for context management and better understanding of situational nuances. We look forward to future advancements in behavior classification that leverage GPT's unique capabilities.

References

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*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M Ponti, and Siva Reddy. 2022a. Faithdial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2022b. Evaluating Attribution in Dialogue Systems: The BEGIN Benchmark. *Transactions of the Association for Computational Linguistics*, 10:1066–1083.

Sarah E. Finch, James D. Finch, and Jinho D. Choi. 2023. Don't forget your abc's: Evaluating the state-of-the-art in chat-oriented dialogue systems. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Sarik Ghazarian, Nuan Wen, Aram Galstyan, and Nanyun Peng. 2022. DEAM: Dialogue coherence evaluation using AMR-based semantic manipulations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 771–785, Dublin, Ireland. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. Dialfact: A benchmark for fact-checking in dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3785–3801.

Ryuichiro Higashinaka, Luis F D'Haro, Bayan Abu Shawar, Rafael E Banchs, Kotaro Funakoshi, Michimasa Inaba, Yuiko Tsunomori, Tetsuro Takahashi, and Joao Sedoc. 2019. Overview of the dialogue breakdown detection challenge 4. In *Wochat Workshop at IWSDS 2019*.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q2:: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2022. Soda: Million-scale dialogue distillation with social commonsense contextualization.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2020. Will i sound like me? improving persona consistency in dialogues through pragmatic self-consciousness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 904–916.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2227–2240, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. 2023. Chatgpt: Jack of all trades, master of none. *arXiv preprint arXiv:2302.10724*.

Young-Jun Lee, Chae-Gyun Lim, and Ho-Jin Choi. 2022. Does gpt-3 generate empathetic dialogues? a novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 669–683.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021. Addressing inquiries about history: An efficient and practical framework for evaluating open-domain chatbot consistency. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1057–1067.

Qian Lin and Hwee Tou Ng. 2022. A semi-supervised learning approach with two teachers to improve breakdown identification in dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11011–11019.

Navonil Majumder, Deepanway Ghosal, Devamanyu Hazarika, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2022. Exemplars-guided empathetic response generation controlled by the elements of human communication. *IEEE Access*, 10:77176–77190.

Shikib Mehri and Maxine Eskenazi. 2020. Unsupervised evaluation of interactive dialog with dialogpt. In *21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 225.

Nathan Ng, Marzyeh Ghassemi, Narendran Thangarajan, Jiacheng Pan, and Qi Guo. 2020. Improving dialogue breakdown detection with semi-supervised learning. In *NeurIPS Workshop on Human in the Loop Dialogue Systems*.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2021. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1699–1713.

Yushan Qian, Bo Wang, Shangzhao Ma, Wu Bin, Shuo Zhang, Dongming Zhao, Kun Huang, and Yuexian Hou. 2023. Think twice: A human-like two-stage conversational agent for emotional response generation. *arXiv preprint arXiv:2301.04907*.

Michael V Reiss. 2023. Testing the reliability of chatgpt for text annotation and classification: A cautionary remark. *arXiv preprint arXiv:2304.11085*.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11229–11237.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276.

Weiyan Shi, Yu Li, Saurav Sahay, and Zhou Yu. 2021. Refine and imitate: Reducing repetition and inconsistency in persuasion dialogues via reinforcement learning and human demonstration. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3478–3492.

Kurt Shuster, Jack Urbanek, Arthur Szlam, and Jason Weston. 2022. Am i me or you? state-of-the-art dialogue models cannot maintain an identity. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2367–2387.

Haoyu Song, Yan Wang, Weinan Zhang, Xiaojiang Liu, and Ting Liu. 2020. Generate, delete and rewrite: A three-stage framework for improving persona consistency of dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5821–5831.

Ekaterina Svikhnushina and Pearl Pu. 2023. Approximating human evaluation of social chatbots with prompting. *arXiv preprint arXiv:2304.05253*.

Ekaterina Svikhnushina, Iuliana Voinea, Anuradha Welivita, and Pearl Pu. 2022. A taxonomy of empathetic questions in social dialogs. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2952–2973, Dublin, Ireland. Association for Computational Linguistics.

Longzheng Wang, Peng Zhang, Xiaoyu Lu, Lei Zhang, Chaoyang Yan, and Chuang Zhang. 2022. Qadialmoe: Question-answering dialogue based fact verification with mixture of experts. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3146–3159.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4886–4899, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

Yadong Xi, Jiashu Pu, and Xiaoxi Mao. 2021. Taming repetition in dialogue generation. *arXiv preprint arXiv:2112.08657*.

Jing Xu, Arthur Szlam, and Jason Weston. 2022a. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197.

Ruijian Xu, Chongyang Tao, Daxin Jiang, Xueliang Zhao, Dongyan Zhao, and Rui Yan. 2021. Learning an effective context-response matching model with self-supervised tasks for retrieval-based dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14158–14166.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022b. Long time no see! open-domain conversation with long-term persona memory. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2639–2650.

Haolan Zhan, Zhuang Li, Yufei Wang, Linhao Luo, Tao Feng, Xiaoxi Kang, Yuncheng Hua, Lizhen Qu, Lay-Ki Soon, Suraj Sharma, Ingrid Zukerman, Zhaleh Semnani-Azad, and Gholamreza Haffari. 2023. Socialdial: A benchmark for socially-aware dialogue systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Chen Zhang, Yiming Chen, Luis Fernando D'Haro, Yan Zhang, Thomas Friedrichs, Grandee Lee, and Haizhou Li. 2021. DynaEval: Unifying turn and dialogue level evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1:*

Long Papers), pages 5676–5689, Online. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213.

Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, and Minlie Huang. 2021. Comae: A multi-factor hierarchical framework for empathetic response generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 813–824.

Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. 2020. Towards persona-based empathetic conversational models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6556–6566.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*.

A Behavior Questions and Definitions

The Question (**Q**) and Definition (**D**) for each dialogue behavior label used for the final GPT prompts are shown in Tables 5 - 12, excluding `Red` which is shown in Table 2 in Section 5.

Q	Is this an empathetic response by Speaker 2:
D	A response is empathetic when Speaker 2 does ONE of the following: - clearly demonstrates an understanding of Speaker 1's emotions - reacts with the appropriate sentiment or emotion to Speaker 1's shared experience - understands or appropriately reacts to Speaker 1's experience or emotions - appropriately reassures, encourages, or supports Speaker 1

Table 5: `Emp`: behavior question and definition.

Q	If this were the next response in the dialogue, would Speaker 1 feel like their feelings are not being understood by Speaker 2:
D	A response displays a lack of empathy when: - it indicates a misunderstanding of how Speaker 1 feels based on what Speaker 1 just said - the tone, emotion, or sentiment of the response is clearly inappropriate for what Speaker 1 just said - the response has an inappropriate lack of emotion to what Speaker 1 just said Do NOT consider its empathy relative to previous topics in the conversation if the dialogue has moved on from them. Instead, only consider the most recent dialogue context when evaluating the empathy of a response.

Table 6: `!Emp`: behavior question and definition.

Q	If this were the next response in the dialogue, would it contradict commonsense:
D	To identify contradictions of commonsense, judge whether a vast majority of people would agree that the response doesn't make sense because the response: - contradicts common knowledge - makes unfounded assumptions - is highly illogical or self-contradictory - asks a question where the answer is already obvious Do NOT mark responses that don't make sense because they: - are off-topic or irrelevant as responses - don't have any clear meaning (e.g. overly vague or ill-formed responses)

Table 7: `!Com`: behavior question and definition.

Q	If this were the next response in the dialogue, does it completely ignore the immediate last turn from Speaker 1:
D	Responses that are completely off-topic, fail to address the asked question, or are otherwise completely inappropriate in the context are considered to be ignoring the other speaker.

Table 8: `Ign`: behavior question and definition.

Q	If this were the next response in the dialogue, is it a self-contradiction by Speaker 2:
D	Self contradictions occur when Speaker 2 says something that is a contradiction of what they have said previously or it is extremely implausible based on the information they have already shared. Self contradictions may also occur within a single turn if Speaker 2 shares two contradictory things. If Speaker 2 shares world knowledge that is factually incorrect this is NOT enough on its own to warrant a self contradiction. If Speaker 2 contradicts something the other speaker Speaker 1 has said, this is NOT a self-contradiction.

Table 9: `!Sel`: behavior question and definition.

Q	Does this response include an incorrect fact:
D	Incorrect facts occur when the response includes information that is either: - false - unproven - highly controversial - highly implausible - clearly misleading If an organization, person, place, etc. is mentioned as a part of public knowledge, but it does not exist or it is inaccurately represented, then this is an incorrect fact. Do NOT consider a turn as an incorrect fact if the turn could be interpreted as expressing: - preference or value judgements - estimates or predictions - personal information about the speaker or their partner - information about things in either speaker's life that are not publicly relevant

Table 10: `!Fac`: behavior question and definition.

Q	Is Speaker 2 saying something about Speaker 1 that is contradicting what Speaker 1 has already shared:
D	Partner contradictions occur when Speaker 2: - shares an assumption about Speaker 1 that is impossible to know based on what has already been said - shares an inference about Speaker 1 that is implausible based on what has already been said - contradicts something Speaker 1 shared about themselves - asks a repetitive question about Speaker 1 when the answer is already known based on what has already been said If Speaker 2 says something that makes it seem like they have forgotten or misremembered what their partner Speaker 1 has said earlier in the dialogue, this is a partner contradiction. If Speaker 2 shares a difference of opinion or situation in their own life as compared to Speaker 1, this is NOT a partner contradiction.

Table 11: `!Par`: behavior question and definition.

Q	If this were the next response in the dialogue, is it completely irrelevant to what was just said:
D	If a response fails to continue the current discussion or jumps to a new and off-topic discussion, it is considered to be irrelevant. Responses that are irrelevant feel abrupt and interrupt the discussion, usually because they present questions or ideas that are unrelated to the previous turn. Short reactions to or acknowledgements of the previous turn are NOT irrelevant.

Table 12: `!Rel`: behavior question and definition.

B Full Prompt Example

Table 13 shows an example of the full GPT prompt utilized in this work. This full version preserves the 'DIALOGUE' header, turn numbers, whitespace newlines, and full speaker names, which were removed from the prompt in Table 2 in Section 5 due to spacing constraints.

```
----------
DIALOGUE
----------
1. Speaker 1: Hi!
2. Speaker 2: Hi! I am happy to be talking to you. Do you have any
   pets at home?
3. Speaker 1: I don't, my mom is allergic to most pets. I used to have
   a pet fish when I was much younger.
4. Speaker 2: Do you want a pet at some point in the future? No
   judgment here, I'm just curious.
5. Speaker 1: I would love to have pets in the future! Once I have
   my own place, I'd like to get a dog or two.
----------

Is this response repeating something that has already been said:

Speaker 2: Would you want to get a cat or a dog?

A response is repetitive if:
- it repeats something from earlier in the dialogue
- it includes asking a question whose answer has been
  already shared

If any part of the response is repetitive, then it should be
labeled as repetitive.
Note that sometimes repetition is useful, such as for emphasis,
acknowledgement, clarification, or elaboration, and in these
cases it should NOT be labeled as repetitive.

Provide your reasoning when considering this question starting
with "Reasoning:". Then, finish by writing your final
decision as one of: "Decision: [YES]" or "Decision: [NO]".
Do NOT fill in your decision with any terms other than YES or NO.
```

Table 13: An example of an unmodified GPT prompt.

C Full Results

Table 14 extends Table 3 from §6.2 to include the precision and recall scores for the automated models. Precision and recall scores are not meaningful for the human evaluators since each human annotation set is traded out as a benchmark against the other; thus, we still present only F1 for **HUM**.

D GPT Cost

We compare the average cost of labeling a single dialogue from ABC-Eval for each behavior using GPT and human judges. Table 15 contains the calculated costs.

GPT The GPT cost for a single dialogue is calculated from the OpenAI API pricing[6] on the sum total number of tokens used for obtaining labels for each bot response for a particular behavior. These costs are then averaged over all dialogues used in this work to obtain the average cost per dialogue. Because there is not much difference in prompt length for the different behavior prompts, the average cost per behavior is essentially identical.

HUM Human annotation costs are derived from the average costs presented in Finch et al. (2023). Since the behavior labels were grouped into annotation tasks for the human judges, we divide each task cost by the number of behaviors contained within that task. The cost for a single label is then the resulting quotient for its respective task.

	Model	P/R/F1+	P/R/F1-	Acc.	$\#_+$
`!Fac`	FC	12.3 / 22.4 / 15.9	93.3 / 87.1 / 90.1	82.2	223
	GPT	37.7 / 44.9 / 41.0	95.5 / 94.0 / 94.7	**90.3**††	146
	HUM	67.8	97.4	**95.2****	122
`Red`	GPT	30.7 / 35.5 / 32.9	94.3 / 93.2 / 93.8	88.6	148
	HUM	58.7	96.4	**93.5****	129
`!Com`	GPT	43.8 / 29.1 / 34.9	83.3 / 90.5 / 86.7	78.0	219
	HUM	55.6	88.6	**81.9***	333
`!Rel`	S2T2	24.0 / 33.5 / 27.9	86.3 / 79.8 / 82.9	**72.4**†	365
	GPT	30.1 / 62.5 / 40.6	91.0 / 72.3 / 80.6	70.8	543
	HUM	54.3	91.3	**85.4****	261
`!Par`	GPT	27.2 / 14.2 / 18.6	91.6 / 96.1 / 93.8	88.5	79
	HUM	48.8	94.8	**90.5****	151
`!Sel`	DEC	22.8 / 49.1 / 31.1	96.3 / 89.2 / 92.6	**86.6**††	215
	GPT	14.6 / 35.9 / 20.7	95.3 / 86.1 / 90.5	83.0	250
	HUM	44.3	96.3	**93.1****	101
`!Emp`	EPI	12.0 / 15.1 / 13.4	85.4 / 81.8 / 83.5	72.3	291
	GPT	21.1 / 36.2 / 26.6	88.1 / 77.7 / 82.6	71.8	396
	HUM	51.5	92.0	**86.3****	231
`Ign`	S2T2	18.5 / 39.5 / 25.2	91.9 / 79.7 / 85.3	**75.5**††	365
	GPT	15.5 / 63.4 / 24.9	93.3 / 59.8 / 72.9	60.2	696
	HUM	61.6	95.5	**92.0****	170
`Emp`	EPI	39.6 / 86.0 / 54.2	70.3 / 20.1 / 31.3	45.0	1343
	GPT	50.7 / 11.9 / 19.3	63.4 / 92.9 / 75.4	**62.3**††	146
	HUM	69.7	81.6	**77.1****	618

Table 14: Precision, recall, F1 and accuracy achieved by each model, where **HUM** stands for human judges. **#+**: num. positive labels predicted. $\dagger | \dagger\dagger$ denote significance between *automated* models on one or both annotation sets. $\star | \star\star$ denote significance against best automated model on one or both annotation sets, respectively.

	GPT	HUM
`!Fac`	0.02	1.96
`Red`	0.02	0.29
`!Com`	0.02	0.92
`!Rel`	0.02	0.47
`!Par`	0.02	0.29
`!Sel`	0.02	0.29
`!Emp`	0.02	0.58
`Ign`	0.02	0.47
`Emp`	0.02	0.58

Table 15: Cost ($) per dialogue for each behavior using GPT or humans (**HUM**).

[6] https://openai.com/pricing