Revealing User Familiarity Bias in Task-Oriented Dialogue via Interactive Evaluation

Takyoung Kim[1,†] **Jamin Shin**[1,2,‡] **Young-Ho Kim**[1] **Sanghwan Bae**[2] **Sungdong Kim**[1,2]

NAVER AI Lab[1] NAVER Cloud[2]

tk30@illinois.edu jayshin.nlp@gmail.com yghokim@younghokim.net

{sanghwan.bae, sungdong.kim}@navercorp.com

Abstract

Most task-oriented dialogue (TOD) benchmarks assume users that know exactly how to use the system by constraining the user behaviors within the system's capabilities via strict user goals, namely "user familiarity" bias. This data bias deepens when it combines with data-driven TOD systems, as it is impossible to fathom the effect of it with existing static evaluations. Hence, we conduct an interactive user study to unveil how vulnerable TOD systems are against realistic scenarios. In particular, we compare users with 1) detailed goal instructions that conform to the system boundaries (*closed-goal*) and 2) vague goal instructions that are often unsupported but realistic (*open-goal*). Our study reveals that conversations in open-goal settings lead to catastrophic failures of the system, in which 92% of the dialogues had significant issues. Moreover, we conduct a thorough analysis to identify distinctive features between the two settings through error annotation. From this, we discover a novel "pretending" behavior, in which the system pretends to handle the user requests even though they are beyond the system's capabilities. We discuss its characteristics and toxicity while showing recent large language models can also suffer from this behavior.

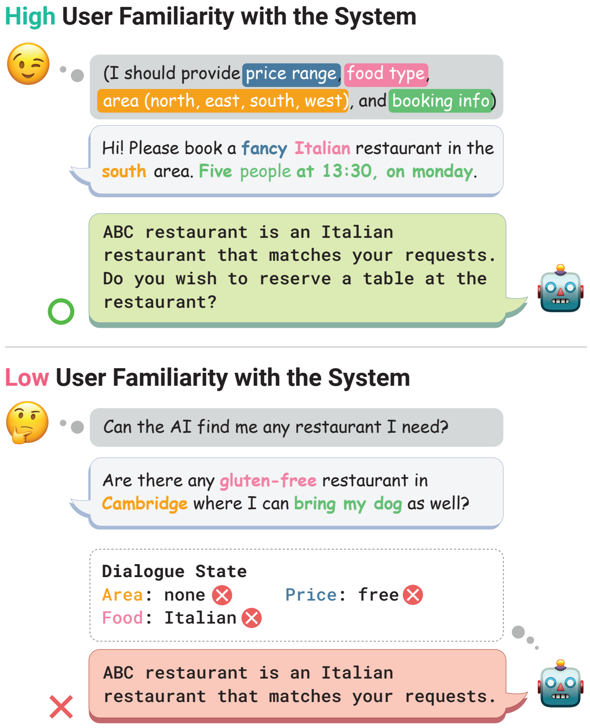

Figure 1: Contrastive dialogues according to user familiarity with the system. Users with high familiarity converse with the system within the predefined scenario since they already know the system well. However, users with low familiarity are more likely to talk about a broad range of topics beyond the system's capacities leading to the erroneous response of the system.

1 Introduction

Task-oriented dialogue (TOD) systems aim to accomplish specific user goals by comprehending their requests and making appropriate API calls or database (DB) searches (Young et al., 2013). TOD systems typically use a pipeline approach, connecting separate modules such as intent detection, dialogue state tracking, policy management, and natural language generation, often requiring complex rules or heuristics. End-to-end (E2E) TOD systems have been highlighted as a fully data-driven solution because of their concise implementation (Bordes et al., 2017; Wen et al., 2017). Recently, such TOD systems have significantly improved on top of pre-trained language models (Hosseini-Asl et al., 2020; Ham et al., 2020; Peng et al., 2021; He et al., 2022).

However, despite the numerous studies on TOD systems and the great successes of large language models, we argue that there is a huge gap between the TOD studies and deployable TOD systems. Among the many reasons hindering end-to-end systems from being widely adopted in the industry, the

[†]Work done as a research intern at NAVER AI Lab.
[‡]Work done while at NAVER AI Lab.

Proceedings of the 6th Workshop on NLP for Conversational AI (NLP4ConvAI 2024), pages 37–55
August 16, 2024

instability of such systems makes it harder to match the robustness of conventional pipeline systems.

We hypothesize that the major source of this instability lies in the naive assumption about the users during TOD data collection. We call this the *user familiarity* bias, as illustrated in Figure 1. For instance, during Wizard-of-Oz style data collection (Kelley, 1984), the user-role workers are provided with detailed instructions on the goal they need to achieve which conforms with the system capabilities (Budzianowski et al., 2018; Byrne et al., 2019). Hence, as the user behaviors are strictly constrained, this process simulates users who know exactly how to use the system (Larson et al., 2019). Other datasets based on user simulation, such as M2M and SGD (Shah et al., 2018; Rastogi et al., 2020), include the same user familiarity bias, as they simulate users based on predefined user goals and rules specifying how to converse. On the other hand, real users in the wild often have fairly creative or vague goals way beyond the system coverage, and this user-side bias prevents us from modeling such realistic conversations.

In this paper, we conduct an interactive user study on one of the most-used Wizard-of-Oz-styled TOD benchmarks, MultiWOZ (Budzianowski et al., 2018), to investigate the impact of *user familiarity*. The main objective of this study is to determine whether the familiarity of the user with the TOD system's capabilities influences the successful completion of a conversational task. To do this, we first divide the users into two groups: *closed-goal* and *open-goal*. The former user group is provided with detailed user goal instructions that are from the MultiWOZ, while the latter is given only a portion of the instructions along with some realistic goals that are not supported by the system, thereby simulating users who are not familiar with TOD systems. Based on each goal type they are assigned to, the users converse with a recent E2E TOD system, GALAXY (He et al., 2022), which is trained on diverse TOD datasets including MultiWOZ.

Our user study reveals that 92% of the dialogues in the open-goal setting has significant issues that often lead to failure in achieving the user goals. Moreover, we find that various inconveniences caused by the TOD system force users to quit the conversation regardless of the goal types. We thoroughly analyze the resulting conversations to identify the impact of user familiarity by annotating erroneous turns. In particular, we figure out six prevalent errors in both goal settings. As expected, open-goal dialogues contain more problematic turns, and the open-goal setup causes more irrelevant and unfaithful responses.

In addition, we identify unfaithful responses as "pretending" behaviors of the system that primarily arises in the open-goal setting. This is a phenomenon similar to hallucination, in which the dialogue system pretends to handle the user's requests even though they are beyond the system's boundaries, but more potentially harmful because it is **almost impossible** for users to verify the reliability of the information during conversation since the hallucinated pieces of information are usually *service-specific*. We believe this issue is relatively underexplored as we witness most previous works focused on the closed-goal setting, and our qualitative analysis of the open-goal dialogues demonstrates that such pretending behaviors are prevalent and crucial.

Finally, we conduct case studies to check whether recent large language models with strong zero-shot performance can mitigate each conversational error. We show that large language models are proficient to handle errors within given context, but preventing pretending problem highly depends on the system design, not only on language models' performance.

Our contributions are threefold: (1) interactive user study that breaks away from the closed-goal assumption; (2) examination of the characteristics of erroneous situations in both dialogue- and turn-levels; and (3) demonstration of the "pretending" problem of the TOD systems, especially as observed in an open-goal assumption, where the agent deceives users as if it handled their exceptional requests.

2 Related Works

2.1 TOD Benchmarks

MultiWOZ is one of the largest TOD benchmarks containing about 10,000 multi-domain dialogues (Budzianowski et al., 2018), following the Wizard-of-Oz protocol (Kelley, 1984). In this setup, human workers participate in a conversation as either a user or system role. To guide the user-side workers to behave within the desired scenarios, the authors provide *goal instructions* to the user-side worker detailing what to inform and request to the system-side worker. Meanwhile, the Taskmaster-1 dataset is more severe in that each conversation of the dataset is collected by one human worker playing both user and system roles, namely the self-play method (Byrne et al., 2019). Thus, these datasets naturally contain significant *user familiarity* bias. Similarly, other datasets constructed by an automatic user simulator also contain the same bias since the simulation is based on predefined goals and rules *bound* to the system's coverage (Shah et al., 2018; Rastogi et al., 2020).

2.2 Benchmark Reality in TOD

Recently, there have been studies concerning the reality of the benchmark dataset in the area of TOD. Kim et al. (2020) incorporate an external unstructured knowledge (*i.e.*, FAQ) to complement the language model trained on limited scenarios. Even though the dataset includes knowledge-seeking turn detection to handle out-of-scope requests, it still assumes high user familiarity with the system in that users require information specified in the external knowledge. Qian et al. (2022); Kim et al. (2022); Yang et al. (2022) point out the limited coverage of dialogue expression by modifying the utterances of the user and system. Furthermore, Sun et al. (2021); Li et al. (2022); Young et al. (2022) improve the model's natural conversation skills in terms of engagingness by combining with open-domain dialogue. However, we believe the combination cannot be a solution for dealing with users who have open-ended goals. On the other hand, Qin et al. (2021) argue inconsistent responses can be a more critical problem in TOD, and propose an accompanying new dataset to mitigate it.

2.3 TOD in Deployment

Potential issues related to interaction or deployment were discussed among communities. For example, Larsson (2017) mainly discussed technical and architectural difficulties in deploying dialogue systems. More similarly, Leuski and Artstein (2017) presented challenges where dialogue systems do not properly handle users' sub-dialogues of different topics or domains (*i.e.*, lack of affordance). However, these discussions were mainly tested on proprietary products such as Siri and Alexa.

2.4 Evaluation of TOD System

Many recent works evaluate performance using quantitative metrics for predefined slots and responses. Specifically, Budzianowski et al. (2018) define two task-specific metrics, Inform and Success rate, which measure how often the system has provided the appropriate entities and answered all the requested attributes. In addition, BLEU (Papineni et al., 2002) is used to measure the fluency of the generated responses. However, Nekvinda and Dušek (2021) report inconsistencies in data preprocessing for these metrics in an attempt to make standardized evaluation scripts. Furthermore, Cheng et al. (2022) build a user simulator capable of dynamic evaluation to solve the static evaluation problem for TOD systems. However, the evaluation is still limited to the closed-goal setup.

Apart from the automatic quantitative evaluation, there are consistent works of user evaluation in spoken dialogue research (Walker et al., 1998; Ai et al., 2007; Gašić et al., 2008). Our work is more closely inspired by user studies in the Human-Computer Interaction (HCI) area that investigated live interactions between chatbots and users. In particular, Yeh et al. (2022) investigate the impacts of various guidance types and timing on performance metrics for TOD systems. Li et al. (2020) analyze conversation logs between users and TOD chatbots and claimed that identifying conversational "non-progress" is crucial for improving chatbot usability.

3 Interactive User Study

In this section, we explain the experimental setups of our interactive user studies on the current state-of-the-art data-driven TOD model. Our focus lies on creating realistic scenarios, breaking away from evaluation solely based on TOD benchmarks. In particular, we are curious about the influence of user familiarity on the TOD system. We describe the details of the study in the following sections.

3.1 User Goal

Most TOD systems assume the users have specific goals in a given domain, *e.g.*, restaurant or hotel reservations. Typically, such goals can be represented by sentences in natural language to control user-side human participants when collecting dialogue data (Budzianowski et al., 2018; Byrne et al., 2019). The following is one of the user goal instructions provided in Budzianowski et al. (2018).

> *You are looking for a* ***place to stay****. The hotel should be in the* ***north*** *and should* ***include free parking****. The hotel should be in the type of* ***guesthouse****. Once you find the hotel, you want to book it for* ***3 people*** *and* ***2 nights*** *starting from* ***wednesday****. ... Make sure you get the* ***reference number****.*

However, all user goals in most TOD benchmarks are based on a naive assumption that the users have sufficient knowledge about the dialogue system in advance. Thus, conversations based on such goals are always within expected scenarios from the TOD system's point of view. On the other hand, we argue that most real users are not familiar with TOD systems, and such users are prone to making exceptional requests beyond the system's capacity. To investigate the impact of user familiarity, we set up two user groups that have different types of goals considering their familiarity with TOD systems, which we refer to as closed-goal and open-goal, respectively.

Closed Goal Closed goals contain predefined scenarios which TOD systems can accomplish easily. In other words, it does not include any exceptional requests or actions from the perspective of the tested TOD system. As we mentioned, most dialogues in existing TOD datasets constructed based on such predefined user goals fall within the capacity that the system can correspond to. We use these user goals from the *restaurant* and *hotel* domains of MultiWOZ (Budzianowski et al., 2018) as our set of closed goals.

Open Goal Contrary to the closed-goal setting, open-goal settings are used to simulate realistic situations for users who have little idea about the TOD systems except for the domain. Real-world users may have a wider range of purposes than the predefined situations because the system capacity cannot include every possible scenario within its boundaries. Thus, we include exceptional requests which are not covered by the original dataset. Specifically, we create an open-goal by inserting the exceptional requests into a subset of closed-goal. By doing so, we are able to cover essential user requirements covered by the system (*e.g.*, time to visit), while also simulating real-life requests that are unsupported. In our experiments, we limit the number of exceptional requests in a single open-goal to a maximum of two.

To construct the set of exceptional requests, we use InstructGPT (Brown et al., 2020; Ouyang et al., 2022) `text-davinci-003` by OpenAI API. First, we input high-level task information as a prompt and let InstructGPT come up with the remaining requests to complete the task instruction. Table 1 is an example input prompt and output of the generated exceptional request. Then, we manually validate the generated outputs to filter out noisy outputs and categorize commonly observed requests for the target domains into an ontology, which is shown in Appendix Listing 1.

Input Prompt
`Imagine that you are planning to travel UK. The following goal is user specification to find information from the bot. Freely fill in the remaining specification.` `(Goal) You are looking for a` **`place to stay`**`.`
Example Output
`Find a hotel that is` **`nearby Cambridge city, close to public transportation, good customer reviews`** `from past guests, include` **`daily meals`** `in the cost,` **`WiFi`** `included, and` **`reasonable cost`**`.`

Table 1: An example input and output of exceptional requests generation based on InstructGPT. We guide InstructGPT to perform text completion from the given input prompt, and manually select exceptional requests not handled in MultiWOZ. After collecting generated outputs, we construct an open-goal ontology demonstrated in the Appendix Listing 1.

3.2 Participants

We recruit 20 participants working at a tech company who meet our inclusion criteria[1] : (1) having some experience with AI chatbots and (2) feeling comfortable carrying on written conversations in English. In terms of the number of participants, our sample size is congruent with the guidelines and standards for the sample size for HCI studies (Hwang and Salvendy, 2010; Caine, 2016). Following suggestions by Hwang and Salvendy (2010), we aim to recruit around 10 participants per group. We randomly assign participants one of the two conditions: Open-goal ($N = 10$; referred to as `O1`–`O10`; 7 females) or Closed-goal ($N = 10$; referred to as `C1`–`C10`; 5 females). Each group of participants is provided with the corresponding type of goal instructions.

3.3 Procedure

We implement a chat interface on Gradio (Abid et al., 2019) web platform illustrated in Appendix Figure 4. For the system agent, we use one of the most performant[2] E2E TOD model (He et al., 2022) trained on diverse benchmarks including MultiWOZ 2.1 (Eric et al., 2020). Please note that

[1]Refer to Appendix A for exhaustive demographics.

[2]Based on the official MultiWOZ leaderboard: `https://github.com/budzianowski/multiwoz`.

this TOD model also contains the user familiarity bias in TOD benchmark datasets, and our experiment can be generalized to every model trained with datasets constructed with similar manner (refer to Section 2.1). We give users structured goals instead of the sentence format in order to prevent copying biases brought on by sentences. After completing each conversation, participants are instructed to specify whether they finish the conversation until the end (whether the goal is achieved) and whether they have encountered any significant inconveniences.

If participants mention that they cannot properly complete the conversation or they experience any inconvenience, we prompt a follow-up checkbox field to ask the categories of inconveniences: (1) **Repetitions** for repeatedly responding with the same text, (2) **Unrelated** for irrelevant responses to what users request, (3) **Not-aligned** for responses contradicting with previous context, and (4) **Awkward** for grammatically wrong or unfluent responses. We also add a free-form answer field where participants can describe the situations that do not fall within the above four categories. For each study session, we invite one or two participants to a Zoom video call, where a moderator briefs the study and instructs participants to complete **five** conversations with the TOD model. During the group session, participants are not allowed to disclose anything related to their conversations with a chatbot. The moderator supports participants only when they encounter technical issues. When the chatbot provides wrong responses, participants are guided to repeat their original intent up to two times, as we expect the TOD model to recover from its mistakes. Moreover, participants can continue the conversation with their own arbitrary goals if the chatbot cannot provide services related to the given goals because it is possible for the chatbot to fail to search entities satisfying all requests from users (even in closed-goal settings).

4 Analysis

4.1 Dataset and Descriptive Statistics

We collect 49 open-goal and 50 closed-goal dialogues from 20 participants; due to technical issues, one open-goal participant missed a conversation. The open-goal dialogues consisted of an average 10.53 turns (an adjacent pair of the user and system messages; SD [3] $= 4.33$), whereas the closed-goal dialogues had 8.92 turns on average ($SD = 3.62$).

[3]Standard deviation.

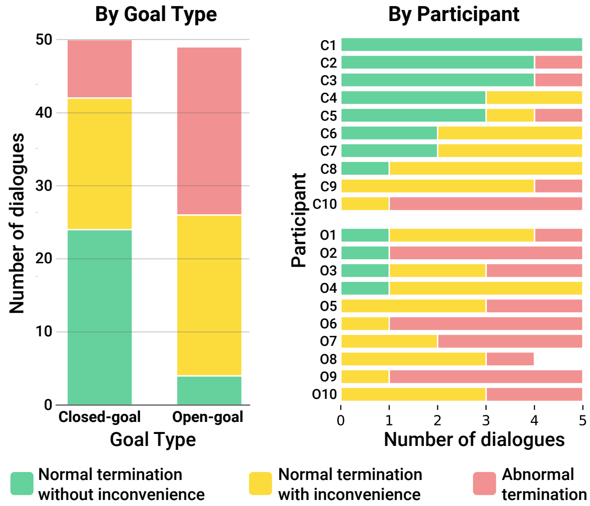

Figure 2: Distributions of the dialogue termination by goal type (left) and participant (right). The green bar refers to situations that users finish the conversation with satisfaction, and the yellow bar refers to situations that users finished the conversation but experience some errors or inconveniences. Lastly, the red bar expresses the proportion of users' strong dissatisfaction by forcedly stopping the conversation. `C1-C10` and `O1-O10` denote users with a closed-goal and open-goal, respectively.

4.2 Dialogue Stability

Figure 2 shows the proportion of forced termination during our experiment. We find that only 8% (4 out of 49) of the total open-goal dialogues have finished without any inconveniences, while almost half of the closed-goal dialogues (24 out of 50) show normal termination without any inconveniences. Meanwhile, it is important to note that more than half of the dialogues in both goal types had problematic situations for participants. Statistical tests (and Figure 2) reveal that open-goal settings result in significantly more erroneous dialogues. We describe the analysis method below, but, in short, we find interactive conversations in the wild to have a clear difference from static benchmark evaluations for both goal settings and especially for the open-goal setting.

Termination Pattern To assess the difference in termination patterns between the two goal types, we use *mixed-effect models*. These multi-level linear regression models can model the effect of the independent variables (*i.e.*, fixed effect) while controlling the random variance among subjects (*i.e.*, random effect) where multiple data points came from the same subject (Pinheiro and Bates, 2000). Treating each dialogue as a data point, we fit a

Error Type	Decription	Proportion	
		Closed	Open
Irrelevant	Irrelevant responses from the given dialogue context.	14.6%	**23.4%**
Self-Contradiction	Contradictory responses with previous bot's responses.	4.2%	5.8%
Repetition	Unnecessary repeated responses with the same semantics.	4.8%	6.6%
Poor Fluency	Awkward or grammatically broken responses	4.8%	3.7%
Pretending	Hallucinated responses on unverifiable requests	2.4%	**23.6%**
Miscellaneous	All other less frequent errors	2.9%	2.7%

Table 2: Definitions of each error type and corresponding proportion by two goal-types. As demonstrated in the bold text, while other error types occur with the similar proportion, the irrelevancy and pretending problems occur significantly often in the open-goal circumstance (8.8%p and 21.2%p more often, respectively).

mixed-effect model to the termination type mapped to a numeric scale (0: normal termination, 1: normal termination with inconvenience, and 2: abnormal termination) in increasing order of severity. We put participants as a random effect and the goal type as a fixed effect to see whether the average severity levels of each group are different.

The maximum-likelihood test reveals that there is a significant random effect of participants ($p < .0001$, $t(17.98) = 5.06$), and a significant fixed effect of goal type ($p = .002$, $t(18.14) = 3.71$). The estimated mean of the severity scale is 0.68 for closed-goal ($SE = 0.13$) and 1.39 for open-goal ($SE = 0.14$) with 95% confidence. This indicates that the severity levels of termination of the two groups are significantly different, and the open-goal dialogues tend to fall in either normal termination with inconvenience or abnormal termination.

Learning Effect We also assess whether there is a learning effect on carrying on a successful conversation as participants go through more conversations. We use a mixed-effect model similar to the above analysis but put the order of dialogue per participant (i.e., 0 to 4) as a fixed effect, instead of the goal type. The maximum likelihood test reveals that the order does not have a significant impact on the model ($p = .85$). This implies that participants do not seem to "learn" the system's conversation and request capabilities within 5 times of trials.

4.3 Erroneous Turn Analysis

In this section, we thoroughly analyze erroneous dialogue turns that occurred in both goal types by categorizing dominant error types.

We qualitatively annotate the erroneous turns. First, all authors of this paper separately annotate a subset (10 dialogues; ~10%) of the user study data with candidates for the error category. Then they iterate several discussions to generate a codebook of the errors. After the codebook is defined and agreed upon, the main author revisits the entire dataset to annotate erroneous turns. As a result, we identify six types of errors that commonly appeared in the conversations (Table 2). We further describe each error category and corresponding example in the following paragraphs. We denote as a user and as a system in the examples.

Irrelevant If the system response is irrelevant to the dialogue context, especially to the previous user utterance, we annotate such responses as `Irrelevant`. It is the most frequent error type in the user study. Moreover, the error proportion from open-goal dialogues is significantly higher, 8.8% points, than that from closed-goal dialogues. The following quote is an illustrative example by `O1`:

: *Please let me know the **cancellation policies** of Alexander.*

: `Yes, I will work on getting this booked for you.`

Self-Contradiction `Self-Contradiction` represents cases that the system response is contradictory with the previous system's responses. This error is severe because it makes the users confused and also causes doubt over the systems in terms of proceeding given task, as illustrated in the below quote by `C9`. We find that open-goal dialogues contain 1.6% points more self-contradiction than closed-goal dialogues.

: *how about **3 stars**?*

: **`Bridge guest house is available`**`, would you like to try that?`

`...`

: *I want 3 stars guesthouse which has a parking lot.*

: **I don’t have anything 3 stars.** would you like to try a different area?

...

: ***Thank you. Can you repeat my reservation?***

Repetition Repetition refers to cases that the system repeats unnecessary responses across multiple turns. This error is also one of the main obstacles to achieving task success as it often annoys the users to quit the conversation. The following quote shows a severe repetition error that happened during our user study (by O5):

: *Yes. Can you make a reservation for 11:45?*

: ··· Can you tell me the day for the reservation, **the number of people** in your party, and **the number of people**?

: *Tomorrow. 2 people.*

: I have your table booked for 2 people on **Tuesday** at 11:45, ··· **What day** would you like to dine?

: *You already said Tuesday! Thank you for your help.*

Poor Fluency When the system’s response is awkward from a given dialogue context or itself, we judge the response as a Poor Fluency. Also, grammatical error belongs to this type. Closed-goal dialogues include about 1% more Poor Fluency errors than open-goal dialogues. The following example represents this type of error (It is from a conversation with C8).

: *Hmm, that’s not that I want. Can you cancel the current reservation and make one for an African restaurant?*

: **No, it does not.** I can try a new restaurant if you would like.

Pretending Pretending represents system behaviors pretending to handle user’s requests beyond the system’s scope or answer user’s non-verifiable questions. For instance, in the below example by O4, the system provides false information to the users because the presence of a pet-friendliness is not verifiable information for the system. Pretending is the most prominent error type of the open-goal setup; meanwhile, it occurs about 10 times more than in the closed-goal setting.

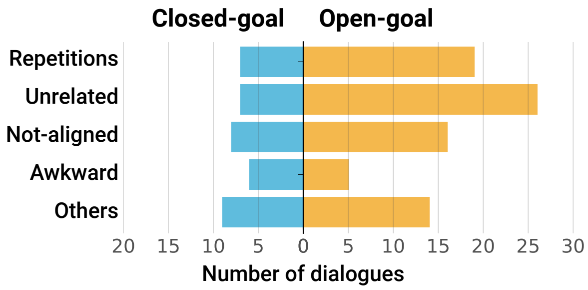

Figure 3: Inconvenience categories collected from participants by two goal types. It can be seen as an implicit or explicit reason to terminate a conversation from the perspective of the user.

This indicates that this Pretending error might have been overlooked in previous works that focused on static closed-goal evaluations. We further describe and discuss more details of this pretending error type in Section 4.5.

: *Is Wagamama **pet-friendly**?*

: **Yes it is.** Would you like to book it for you?

: *Yes. Today at 7pm please.*

Miscellaneous The Miscellaneous category, which is less frequent, usually includes arbitrary system behaviors like suddenly making a booking request and ignoring the user’s intent. The following example by C10 describes such a situation.

: *Then, is the Indian Cuisine you mentioned earlier available for reservation on Wednesday at 15:45?*

: Yes it is. **I have booked it for you** and your reference number is 64o8zpsl.

: ***Suddenly?** The person I’m going with just now asks if we can change it to the same time as Thursday?*

: Ok. I will try to book that for you.

4.4 Participant Feedback

We collect negative feedback from participants at a dialogue-level if there are any inconveniences during the conversation. Figure 3 shows the inconvenience categorization results. As expected, open-goal dialogues show a significantly higher proportion in most inconvenience categories than closed-goal dialogues because the model must have difficulty in handling novel requests in the open-goals. In addition to the inconvenience categories we specify, we also collect feedback from users in a text form (Others category in Figure 3), and the full table after removing duplicated feedback will be made public in the Appendix. Interestingly, we find that most inconveniences align with our error categories of Section 4.3 well except for the Pretending error type. In other words, the pretending behaviors are *not* identified as inconveniences from the users' perspective.

4.5 Why does Pretending Matter?

In this subsection, we take a deeper look at the "pretending" behaviors of the TOD systems. The pretending behavior is similar to the hallucination problem (Ji et al., 2022) prevalent in generative models in that it provides false information to the users. However, one distinct feature of the pretending behavior is that it is **hard to be recognized as an error at the time of occurrence** by only its surface form because most TOD systems rely on variable service-specific knowledge that users cannot easily access while using the service. It also differs from the knowledge base inconsistency (KBI) in Qin et al. (2021). While the KBI only regards the wrong responses based on "verifiable" knowledge, Pretending indicates responses over "non-verifiable" knowledge beyond the system's scope.

It is a severe problem for both agents and users since it interrupts accurate decision-making to achieve users' goals. For example, other error types such as Irrelevant, Self-Contradiction, and Repitition can be easily recognized as superficial problems by the users. In those cases, the users can avoid unwanted conversation flow by complaining and terminating the conversations. However, when the TOD system naturally responds to users' exceptional requests and does not take corresponding action behind, users have no way to perceive the fact that the ongoing conversation is wrong from their initial requests. For this reason, the Pretending is not exposed in *any* user evaluation shown in Figure 2, Figure 3, or Appendix Table 6. In other words, even users who normally terminated dialogue without any inconvenience (*i.e.*, green bar in Figure 2) can suffer from the pretended dialogues.

4.6 Can Large Language Models Solve TOD Problems?

Although previous studies demonstrate the imperfectness of large language models in TOD systems (Jakobovits et al., 2022; Bang et al., 2023), we conduct case studies to verify large language models' capacities to resolve aforementioned conversational errors. Since it is impossible to equalize every experimental setup between fine-tuned language models and proprietary large language models[4], we proxy intermediate modules of TOD systems with instruction prompts by modifying those of Chung et al. (2023). On top of that, we pinpoint erroneous turns in Section 4.3 and compare generated responses of each model. We describe detailed setup such as prompts and action definition in Section C.1.

Based on the result of our case studies, as illustrated in Table 5, we observe that large language models mitigate most conversational inconvenience, presumably due to their strong zero/few-shot capabilities on unseen domains. However, regarding Pretending, responses of both cases contain significant flaws. Specifically, according to Table 5, both conventional TOD model and large language model do not recognize the fact that pet-friendliness does not belong to the service range and provide untrustworthy responses, which can lead to physical harm (*e.g.*, wrong reservation) in real services. One of the expected causes is language models' overconfidence of unseen scenarios, but we also find that predefined actions given to TOD models are confined to deal with diverse situations in a flexible manner[5]. Controlling overconfidence in language models (Miao et al., 2021; Mielke et al., 2022) can partially resolve conversational errors, but defining available actions mostly belongs to the range of system design, especially in service-specific scenarios. We further discuss the future direction in Section D.

5 Conclusion

In this work, we demonstrate user familiarity bias in current TOD benchmarks, which the recent TOD research community has overlooked. To effectively unveil the bias, we contrast two user groups with different user goals via an interactive user study. Against the closed-goal within the constrained scenarios, we introduce a control user group by assigning unconstrained scenarios to the participants, namely open-goal. Users in the two groups converse with the academically-discussed TOD chatbot following the given closed or open-goals. Our

[4] gpt-4-turbo in our case studies.

[5] Notably, there are roughly-defined actions in conventional TOD scenarios, such as inform, request, recommend, etc.

study reveals the TOD system exposed to the user familiarity bias significantly fails to converse with the users with open-goals. We identify prevalent error types by analyzing the resulting conversations. Furthermore, we highlight the pretending behaviors of the TOD system with its characteristics and toxicity, which are not easily solved by simply utilizing large language models.

6 Limitations

Regarding the participants of our user study, all of them are internal employees of a giant tech company, the majority of whom are highly educated (60% hold a master's or doctoral degree). However, they show various experiences with chatbots, not correlated with educational degree. Some of them are bilingual, while others are not native English speakers. Furthermore, since we assume a traveling situation, the conversational scenario was not challenging, even for non-native speakers.

7 Ethics Statement

In our user study, we collected demographic information such as name, age, gender, the highest level of education, occupation, native language, and experiences with the AI chatbot, after informing them that it would be used only for research purposes and acquiring their consent. We clearly introduced the purpose of our study and the usage of collected information before experiments, and all participants consented to our instructions.

Throughout the interaction with the chatbot, we instructed participants to play the role of potential users only, without disclosing any personally identifiable information about themselves. Collected dialogues were de-identified by giving anonymized user IDs. Throughout the annotating process, the authors examined all the gathered conversations, and no offensive content was found. Participants took part in the chat for roughly 30 minutes and were compensated with a 5,000 KRW (equivalent to 3.7 USD) gift card, which was somewhat higher than the Korean minimum wage at that time.

References

Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*.

Hua Ai, Antoine Raux, Dan Bohus, Maxine Eskenazi, and Diane Litman. 2007. Comparing spoken dialog corpora collected with recruited subjects versus real users. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 124–131, Antwerp, Belgium. Association for Computational Linguistics.

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA. Association for Computing Machinery.

Sanghwan Bae, Donghyun Kwak, Soyoung Kang, Min Young Lee, Sungdong Kim, Yuin Jeong, Hyeri Kim, Sang-Woo Lee, Woomyoung Park, and Nako Sung. 2022. Keep me updated! memory management in long-term conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3769–3787, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. In *International Conference on Learning Representations*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4516–4525, Hong Kong, China. Association for Computational Linguistics.

Kelly Caine. 2016. Local standards for sample size at chi. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 981–992, New York, NY, USA. Association for Computing Machinery.

Qinyuan Cheng, Linyang Li, Guofeng Quan, Feng Gao, Xiaofeng Mou, and Xipeng Qiu. 2022. Is MultiWOZ a solved task? an interactive TOD evaluation framework with user simulator. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1248–1259, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hyunsoo Cho, Choonghyun Park, Jaewook Kang, Kang Min Yoo, Taeuk Kim, and Sang-goo Lee. 2022. Enhancing out-of-distribution detection in natural language understanding via implicit layer ensemble. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 783–798, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Willy Chung, Samuel Cahyawijaya, Bryan Wilie, Holy Lovenia, and Pascale Fung. 2023. InstructTODS: Large language models for end-to-end task-oriented dialogue systems. In *Proceedings of the Second Workshop on Natural Language Interfaces*, pages 1–21, Bali, Indonesia. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Milica Gašić, Simon Keizer, Francois Mairesse, Jost Schatzmann, Blaise Thomson, Kai Yu, and Steve Young. 2008. Training and evaluation of the HIS POMDP dialogue system in noise. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 112–119, Columbus, Ohio. Association for Computational Linguistics.

Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. 2020. End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 583–592, Online. Association for Computational Linguistics.

Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, Jian Sun, and Yongbin Li. 2022. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10749–10757.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. In *Advances in Neural Information Processing Systems*, volume 33, pages 20179–20191. Curran Associates, Inc.

Wonil Hwang and Gavriel Salvendy. 2010. Number of people required for usability evaluation: The 10±2 rule. *Commun. ACM*, 53(5):130–133.

Alice Shoshana Jakobovits, Francesco Piccinno, and Yasemin Altun. 2022. What did you say? task-oriented dialog datasets are not conversational!?

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *ACM Comput. Surv.* Just Accepted.

J. F. Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Trans. Inf. Syst.*, 2(1):26–41.

Seokhwan Kim, Mihail Eric, Karthik Gopalakrishnan, Behnam Hedayatnia, Yang Liu, and Dilek Hakkani-Tur. 2020. Beyond domain APIs: Task-oriented conversational modeling with unstructured knowledge access. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 278–289, 1st virtual meeting. Association for Computational Linguistics.

Takyoung Kim, Yukyung Lee, Hoonsang Yoon, Pilsung Kang, Junseong Bang, and Misuk Kim. 2022. Oh my mistake!: Toward realistic dialogue state tracking including turnback utterances. In *Proceedings of the Towards Semi-Supervised and Reinforced Task-Oriented Dialog Systems (SereTOD)*, pages 1–12, Abu Dhabi, Beijing (Hybrid). Association for Computational Linguistics.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A.

Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316, Hong Kong, China. Association for Computational Linguistics.

Staffan Larsson. 2017. User-initiated sub-dialogues in state-of-the-art dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 17–22, Saarbrücken, Germany. Association for Computational Linguistics.

Anton Leuski and Ron Artstein. 2017. Lessons in dialogue system deployment. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 352–355, Saarbrücken, Germany. Association for Computational Linguistics.

Chi-Hsun Li, Su-Fang Yeh, Tang-Jie Chang, Meng-Hsuan Tsai, Ken Chen, and Yung-Ju Chang. 2020. A conversation analysis of non-progress and coping strategies with a banking task-oriented chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA. Association for Computing Machinery.

Miaoran Li, Baolin Peng, Michel Galley, Jianfeng Gao, and Zhu Zhang. 2022. Enhancing task bot engagement with synthesized open-domain dialog.

Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the ai: Informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–15, New York, NY, USA. Association for Computing Machinery.

Mengqi Miao, Fandong Meng, Yijin Liu, Xiao-Hua Zhou, and Jie Zhou. 2021. Prevent the language model from being overconfident in neural machine translation. 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 3456–3468, Online. Association for Computational Linguistics.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

Tomáš Nekvinda and Ondřej Dušek. 2021. Shades of BLEU, flavours of success: The case of MultiWOZ. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 34–46, Online. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2021. Soloist: Building task bots at scale with transfer learning and machine teaching. *Transactions of the Association for Computational Linguistics*, 9:807–824.

José Pinheiro and Douglas Bates. 2000. *Mixed-Effects Models in S and S-PLUS*, 1 edition. Statistics and Computing. Springer-Verlag, New York.

Kun Qian, Satwik Kottur, Ahmad Beirami, Shahin Shayandeh, Paul Crook, Alborz Geramifard, Zhou Yu, and Chinnadhurai Sankar. 2022. Database search results disambiguation for task-oriented dialog systems. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1158–1173, Seattle, United States. Association for Computational Linguistics.

Libo Qin, Tianbao Xie, Shijue Huang, Qiguang Chen, Xiao Xu, and Wanxiang Che. 2021. Don't be contradicted with anything! CI-ToD: Towards benchmarking consistency for task-oriented dialogue system. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2367, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–13, New York, NY, USA. Association for Computing Machinery.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8689–8696.

Pararth Shah, Dilek Hakkani-Tür, Bing Liu, and Gokhan Tür. 2018. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 41–51, New Orleans - Louisiana. Association for Computational Linguistics.

Ashish Shrivastava, Kaustubh Dhole, Abhinav Bhatt, and Sharvani Raghunath. 2021. Saying No is An Art: Contextualized Fallback Responses for Unanswerable Dialogue Queries. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 87–92, Online. Association for Computational Linguistics.

Kai Sun, Seungwhan Moon, Paul Crook, Stephen Roller, Becka Silvert, Bing Liu, Zhiguang Wang, Honglei Liu, Eunjoon Cho, and Claire Cardie. 2021. Adding chit-chat to enhance task-oriented dialogues. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1570–1583, Online. Association for Computational Linguistics.

David Thulke, Nico Daheim, Christian Dugast, and Hermann Ney. 2021. Efficient retrieval augmented generation from unstructured knowledge for task-oriented dialog. *arXiv preprint arXiv:2102.04643*.

Marilyn A. Walker, Jeanne Fromer, Giuseppe Di Fabbrizio, Craig Mestel, and Don Hindle. 1998. What can i say? evaluating a spoken language interface to email. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '98, page 582–589, USA. ACM Press/Addison-Wesley Publishing Co.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

Puhai Yang, Heyan Huang, Wei Wei, and Xian-Ling Mao. 2022. Toward real-life dialogue state tracking involving negative feedback utterances. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 2222–2232, New York, NY, USA. Association for Computing Machinery.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Su-Fang Yeh, Meng-Hsin Wu, Tze-Yu Chen, Yen-Chun Lin, XiJing Chang, You-Hsuan Chiang, and Yung-Ju Chang. 2022. How to guide task-oriented chatbot users, and when: A mixed-methods study of combinations of chatbot guidance types and timings. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D. Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Tom Young, Frank Xing, Vlad Pandelea, Jinjie Ni, and Erik Cambria. 2022. Fusing task-oriented and open-domain dialogues in conversational agents. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11622–11629.

Jianguo Zhang, Kazuma Hashimoto, Yao Wan, Zhiwei Liu, Ye Liu, Caiming Xiong, and Philip Yu. 2022a. Are pre-trained transformers robust in intent classification? a missing ingredient in evaluation of out-of-scope intent detection. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 12–20, Dublin, Ireland. Association for Computational Linguistics.

Yuwei Zhang, Haode Zhang, Li-Ming Zhan, Xiao-Ming Wu, and Albert Lam. 2022b. New intent discovery with pre-training and contrastive learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 256–269, Dublin, Ireland. Association for Computational Linguistics.

A Participant Information

Participants are aged between 24 to 35 ($Median = 29.5$), and 12 are female. Three participants report that they are native English speakers. Eight participants have used AI chatbots for less than one month. Participants consist of 7 infrastructure engineers, 6 software engineers, 3 AI research scientists, 2 self-employed, a UX designer, and a data scientist. Detailed information can be found in Table 3.

User ID	Age Range	Gender	Education	Occupation	Chatbot Proficiency
C1	26-30	Male	Bachelor's	Infrastructure Engineer	Less than 1 month
C2	26-30	Female	Master's	Software Engineer	Less than 1 month
C3	26-30	Male	Master's	Software Engineer	Less than 1 month
C4	26-30	Male	Master's	Data scientist	1 year to 3 year
C5	26-30	Female	Bachelor's	Software Engineer	1 year to 3 year
C6	26-30	Female	Bachelor's	Software Engineer	Less than 1 month
C7	31-35	Male	Master's	UX Designer	1 year to 3 year
C8	31-35	Female	Master's	Research Scientist	1 year to 3 year
C9	31-35	Male	Ph.D./M.D.	Infrastructure Engineer	Less than 6 months
C10	31-35	Female	Bachelor's	Self-employed	1 year to 3 year
O1	26-30	Male	Ph.D./M.D.	Infrastructure Engineer	Less than 1 month
O2	21-25	Female	Bachelor's	Software Engineer	Less than 1 month
O3	21-25	Female	Bachelor's	Infrastructure Engineer	Less than 1 year
O4	26-30	Female	Bachelor's	Self-employed	1 year to 3 year
O5	26-30	Female	Master's	Infrastructure Engineer	Less than 1 month
O6	21-25	Female	Bachelor's	Infrastructure Engineer	Less than 1 month
O7	21-25	Female	Master's	Infrastructure Engineer	Less than 6 months
O8	31-35	Male	Ph.D./M.D.	Research Scientist	More than 3 years
O9	31-35	Male	Master's	Software Engineer	1 year to 3 year
O10	31-35	Female	Ph.D./M.D.	Research Scientist	More than 3 years

Table 3: Participant information of our user study. We anonymize the name of each participant by assigning user ID and categorizing the range of their age. Users whose ID starts with `C` conduct closed-goal conversation, whereas those whose ID starts with `O` conduct open-goal conversation.

B Model Implementation Details

For the TOD system in our experiments, we use the public implementation of GALAXY[6] (He et al., 2022). The model specification follows He et al. (2022); initialized with UniLM (Dong et al., 2019), which has a transformer-based architecture with 109M parameter size. We fine-tune this model on MultiWOZ 2.1[7] (Eric et al., 2020). We follow the default hyper-parameter settings provided by the authors. Training is completed within a few hours using 1 NVIDIA A100. PyTorch[8] library is used for model training, and NLTK[9] and spaCy[10] are for text processing. We implemented a chat interface on Gradio[11] (Abid et al., 2019) web platform. At inference time, greedy search is used for output prediction.

[6] `https://github.com/siat-nlp/GALAXY`. Apache license 2.0.
[7] `https://github.com/budzianowski/multiwoz`. MIT license.
[8] `https://pytorch.org/`
[9] `https://www.nltk.org/`
[10] `https://spacy.io/`
[11] `https://gradio.app/`

C Case Studies on Large Language Model

C.1 Setup

With a slight modification of Chung et al. (2023), we build prompts for action decision and response generation, demonstrated in Table 4, in order to simulate TOD systems utilizing large language models. We adopt predefined actions in GALAXY model as available actions. For end-to-end simulation, large language models generate responses by referring to conversation context and selected action (which is also generated).

Prompt for Action Decision

```
In a task-oriented dialogue setting, generate an appropriate system ACT to the USER query in the
conversation provided in CONTEXT. A single system ACT should be selected within the list of
available actions. Make sure that selected action should not contradict with previous conversation.
Only generate the selected action.

You can follow this example:

Conversation domain: restaurant
Available actions: ["inform", "request", "nooffer", "recommend", "select", "offerbook"]
CONTEXT:
USER: I need a place to eat foods.
SYSTEM: Any preference in the type of food?
USER: Preferably Italian, but Chinese is also fine.
ACT: recommend

Conversation domain: {{domain}}
Available actions: {{action}}
CONTEXT:
{{context}}
ACT:
```

Prompt for Response Generation

```
In a task-oriented dialogue setting, generate a natural and helpful SYSTEM response to the USER
query in the conversation provided in CONTEXT. You should follow the information provided in ACT to
generate this answer. Do not mention that you are referring to a dataframe and don't overload the
user with too many choices.

You can follow this example:

USER: I need a place to eat foods.
SYSTEM: Any preference in the type of food?
USER: Preferably Italian, but Chinese is also fine.
ACT: [restaurant] recommend
SYSTEM: How about in Bouly's Italy?


CONTEXT:
{{context}}
ACT: {{action}}
SYSTEM:
```

Table 4: Prompts used for large language model simulation.

C.2 Response Comparison

We compare responses of GALAXY model with those of proprietary large language model on sampled dialogues, specifically mentioned in Section 4.3. Table 5 demonstrates each example with partial dialogue turns.

<table>
<tr><th>Error Type</th><th>Response (GALAXY)</th><th>Response (<code>gpt-4-turbo</code>)</th></tr>
<tr><td rowspan="2">Irrelevant</td><td colspan="2">No. tell me the other option.
The alexander is located at 56 saint barnabas road and the phone number is 01223525725.
Please let me know the cancellation policies of Alexander.</td></tr>
<tr><td>Yes, I will work on getting this booked for you.</td><td>I apologize, but I don't have information about cancellation policies for the Alexander Bed and Breakfast. Would you like me to help you with anything else?</td></tr>
<tr><td rowspan="2">Self-Contradiction</td><td colspan="2">How about 3 stars?
Bridge guest house is available, would you like to try that ?
Okay. I have a car. Is there a parking lot?
I am sorry , there isn't anything that meets your criteria .
I want 3 stars guesthouse in the south which has a parking lot.</td></tr>
<tr><td>I don't have anything 3 stars. would you like to try a different area?</td><td>I'm sorry, but currently, there are no 3-star guesthouses in the south area that offer parking facilities. Is there anything else I might assist you with or perhaps another area you would consider?</td></tr>
<tr><td rowspan="2">Repetition</td><td colspan="2">Yes. Can you make a reservation for 11:45?
I would be happy to. can you tell me the day for the reservation, the number of people in your party, and the number of people.
Tomorrow. 2 people.</td></tr>
<tr><td>I have your table booked for 2 people on Tuesday at 11:45 , your reference number is zbznwq27. What day would you like to dine?</td><td>Great! Your reservation at Curry Prince for 2 people tomorrow at 11:45 AM is all set. If you need to make any changes or have more requests, feel free to let me know. Enjoy your meal!</td></tr>
<tr><td rowspan="2">Poor Fluency</td><td colspan="2">Does the place offer halal or African food?
It serves modern european food.
Hmm, that's not that I want. Can you cancel the current reservation and make one for an African restaurant?</td></tr>
<tr><td>No, it does not. I can try a new restaurant if you would like.</td><td>I've cancelled your reservation at the Cambridge Chop House. I can book a table for you at Bedouin, which serves African cuisine. Would you like me to proceed with the booking for 6 people this Monday at 17:15?</td></tr>
<tr><td rowspan="2">Pretending</td><td colspan="2">It's just going to be me. Make sure its pet-friendly though.
It is expensive. What time would you like to dine?
Is Wagamama pet-friendly?</td></tr>
<tr><td>Yes it is. Would you like to book it for you?</td><td>Wagamama does not allow pets inside. Would you like to find another restaurant or something else?</td></tr>
</table>

Table 5: Comparison of GALAXY and `gpt-4-turbo` (LLM) on sampled dialogues in our user studies. Upper part of conversation on each error type illustrates partial context. For large language model prompt, full dialogues become an input as a context. **Irrelevant:** While GALAXY does not handle cancellation policies, LLM correctly understands the requirement and provides relevant response. **Self-Contradiction:** By providing explicit context (*i.e.*, no 3-star guesthouses with parking), response of LLM does not contradict with previous conversation. **Repetition:** LLM mostly does not unnecessarily ask the same question. **Poor Fluency:** LLM shows fluent conversational skills by understanding the current topic of the conversation. **Pretending:** Both GALAXY and LLM have no capability to verify whether Wagamama is enrolled as a pet-friendly restaurant. Therefore, both responses are not true.

D Discussion

As the importance of real user study or interactive evaluation gets bigger, we discuss the future direction in two folds: enhancing transparency to let users trust the chatbot, and managing fallback situations to detect users' exceptional requests.

D.1 Enhancing Transparency

Our participants often struggle to confirm their booking options as in Appendix Table 6. As a way to improve the reliability of TOD systems, we suggest enhancing *transparency* of the system, which has been actively discussed in the HCI community (Amershi et al., 2019). Transparency is a mechanism exposing hidden (*i.e.*, non-obvious) information to users who have difficulty in getting the information directly (Rader et al., 2018). As our findings show that the lack of *user familiarity* provokes various inconveniences including the pretending problem, TOD systems in natural language processing field should also be designed to display intermediate by-products during the conversation in order to provide explainable rationales for their decisions (Amershi et al., 2019; Liao et al., 2020).

In the era of billion-scale large language models, the necessity of transparency is still valid. Although emerging works on grounded LLMs (*e.g.*, Yao et al. (2023), ChatGPT with plugins (OpenAI, 2023) try to enhance trustworthiness using executable sources, they are still exposed to familiarity bias problem as long as they keep black-boxed service pipeline.

D.2 Managing Fallback Situation

Users with low familiarity with the system inevitably make exceptional requests. As we can find in user comments in Appendix Table 6, a large number of users in an open-goal setup go through irrelevant and pretending responses from the bot. We emphasize the need to recognize exceptional requests and manage fallback situations towards robust TOD systems.

Out-of-Scope Detection In the field of intent classification, previous literature has studied detecting out-of-scope intents to prevent generating erroneous responses from the original intent (Larson et al., 2019; Zhang et al., 2022a,b; Cho et al., 2022). Moreover, Shrivastava et al. (2021) try to generate contextualized fallback responses to users' exceptional requests. However, more datasets for fallback detection are required especially for multi-turn and multi-domain TOD scenarios beyond the single-turn detection scenarios.

Handling Request as Unstructured Form Kim et al. (2020) combine unstructured knowledge, FAQ pairs, with structured knowledge, DBs. The work includes (unstructured) knowledge-seeking turn detection to handle domain-specific requests with FAQs beyond the scope of structured knowledge. However, the work still assumes high user familiarity, *i.e.*, it always contains relevant knowledge for a given request. We believe retrieval-augmented detection leveraging the FAQ pairs can be a promising approach to strengthen the approach towards a low user familiarity setup effectively (Thulke et al., 2021).

On the other hand, typical dialogue state tracking to access structured knowledge is not robust in terms of handling exceptional requests since it works based on *predefined* slots. Bae et al. (2022) adopts a text-formed dialogue state by summarising the dialogue context for effective memory management in multi-session open-domain dialogue. We believe that dialogue management based on unstructured information can have advantages not only in avoiding exceptional requests but also in leveraging advanced language understanding abilities of recent language models at a scale, as its generalizable text format.

E Subjective User Feedback

<table>
<tr><th>Goal Type</th><th>Error Type</th><th>Feedback</th></tr>
<tr><td rowspan="14">Closed</td><td rowspan="3">Relevancy</td><td>I told about the reservation conditions, but the chatbot answered irrelevantly.</td></tr>
<tr><td>There was an answer that seemed to have forgotten the context of the past,
but generally the conversation ended without any problems.</td></tr>
<tr><td>I was asked how many people would visit, so I said I was alone. But the chatbot said
it didn’t have a room, and it couldn’t continue conversation after that.</td></tr>
<tr><td>Awkward</td><td>The chatbot said “Yes, I can” when I asked the parking availability in the hotel.</td></tr>
<tr><td>Repetition</td><td>The chatbot repetitively asked “What area would you like to stay in?”.</td></tr>
<tr><td>Contradiction</td><td>The chatbot said there is no place to park, but it reversed its saying.
It also told there is a 3-star hotel, then reversed.</td></tr>
<tr><td>Redundant</td><td>I told the model that I can look up the address by myself, but it gave me the address.</td></tr>
<tr><td rowspan="7">Booking</td><td>The chatbot just ended conversation by just recommending, not booking.</td></tr>
<tr><td>I asked the chatbot to recommend, but it arbitrarily booked it.</td></tr>
<tr><td>It booked without any options I prefer.</td></tr>
<tr><td>It did not confirm my requests.</td></tr>
<tr><td>I wanted to confirm that my reservation is at 9:30 but chatbot did not say.</td></tr>
<tr><td>I wished to reserve Varsity, but the chatbot booked Bloomsbury and did not fix.</td></tr>
<tr><td>I requested to confirm my reservation because I did not trust, but it could not.</td></tr>
<tr><td rowspan="28">Open</td><td rowspan="16">Relevancy</td><td>The bot couldn’t understand my additional requests.</td></tr>
<tr><td>The bot couldn’t understand and answer my question about additional information.</td></tr>
<tr><td>After being asked whether the Asian restaurant serves Italian wines,
it keeps answering that the Asian restaurant serves Italian food.
Following correction questions did not work.</td></tr>
<tr><td>The model does not understand the question correctly.</td></tr>
<tr><td>It does not get back with the list of menus from the pizza hut city centre.</td></tr>
<tr><td>The model keeps saying about night clubs information instead of accommodation.</td></tr>
<tr><td>The chatbot doesn’t understand additional requests on gluten-free and pet-friendliness.</td></tr>
<tr><td>The chatbot understood “Slightly more expensive [than cheap]” expression as “expensive,”
which is wrong.</td></tr>
<tr><td>The chatbot asked whether I wish for a different cuisine,
when I never stated any in the firstplace.</td></tr>
<tr><td>I asked for hotel amenities, but the chatbot thought I was asking for the address.</td></tr>
<tr><td>I asked whether a certain restaurant serves gluten-free,
but the chatbot didn’t directly address the request.</td></tr>
<tr><td>It would not answer my question.</td></tr>
<tr><td>It suddenly says “Your booking was successful , the reference number is i23gx1yf”.</td></tr>
<tr><td>I don’t feel like the model remembers the conversation context.</td></tr>
<tr><td>It often made weird responses.</td></tr>
<tr><td>I stopped conversation because it never answer what I asked.</td></tr>
<tr><td>Awkward</td><td>Sometimes the bot would repeat the same options twice in the same sentence.</td></tr>
<tr><td rowspan="3">Repetition</td><td>It made a reservation for tuesday, but still asked me what day I’d like to dine.</td></tr>
<tr><td>Although I answered, it would ask me the same thing again.</td></tr>
<tr><td>Regardless of my answer it just repeats the same thing.</td></tr>
<tr><td>Contradiction</td><td>“Since there are several hotel-s in the centre of town we have only 2 guest house.”
didn’t make sense.</td></tr>
<tr><td rowspan="2">Redundant</td><td>Right after booking a guest house,
the model asked about hotel booking which is unnecessary.</td></tr>
<tr><td>The chatbot said relevant, but unnecessary questions.</td></tr>
<tr><td rowspan="4">Booking</td><td>I am not sure the chatbot truly understand my booking requests.</td></tr>
<tr><td>The chatbot unnecessarily tried to push me into booking the places/restaurants
when my goal is to simply get information.</td></tr>
<tr><td>I am not sure if the model really booked successfully.</td></tr>
<tr><td>The chatbot seems to be obsessed with the purpose of booking something.</td></tr>
</table>

Table 6: A dialogue-level user feedback based on goal types. Similar feedbacks are categorized with error types.

F Ontology Used for Open-Goal Dialogues


```
{
    "hotel": {
        "Requests": [
            "Outdoor seating",
            "Rooms with exceptional views",
            "24-hour front desk",
            "Breakfast options",
            "Check-in/out policies",
            "Cancellation policies",
            "Cost around $150 per nights"
        ],
        "Facilities": [
            "Facilities: Gym",
            "Facilities: Spa",
            "Facilities: Swimming pool",
            "Facilities: Outdoor terrace",
            "Facilities: Non-smoking room"
        ],
        "Amenities": [
            "Amenities: Mini-fridge",
            "Amenities: Tea and coffee facilities",
            "Amenities: Private bath"
        ]
    },
    "restaurant": {
        "Requests": [
            "Cost under $50 per person",
            "Open until late",
            "Italian wine serving",
            "List of menus",
            "Comfortable seating with enough space",
            "Delivery or take-out service"
        ],
        "FoodOptions": [
            "Food option: Gluten-free",
            "Food option: Vegetarian"
        ]
    },
    "common": {
        "Requests": [
            "Wheelchair access",
            "Weather forecast",
            "Website link",
            "Good customer review",
            "Located in safe environment",
            "Accept payment via cards",
            "Easy access to public transportation",
            "Near the tourist attraction"
        ],
        "Friendly": [
            "Pet-friendly", "Kid-friendly", "Family-friendly"
        ]
    }
}
```


Listing 1: A postprocessed ontology for exceptional requests generated from GPT-3. We collect raw task descriptions and categorize common scenarios which are not seen during the model training.

G Demo Interface for User Study

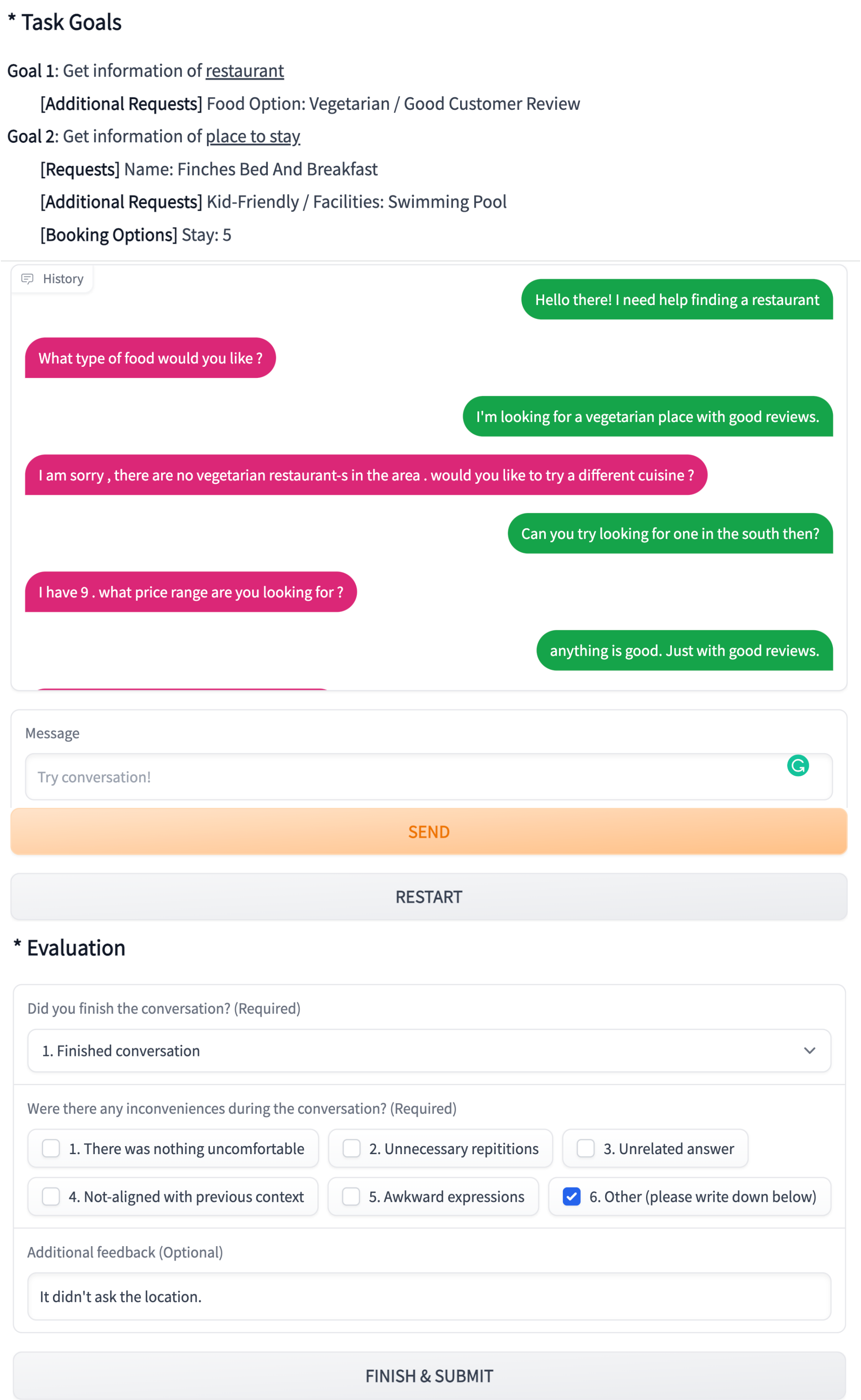

Figure 4: A Gradio demo example of an interactive user study. Unlike sentence-based user goal guidance of conventional Wizard-of-Oz setup, we provide user goals in a structured form.