arXiv:2501.17348v2 [cs.CL] 31 Jan 2025

Better *Slow* than Sorry: Introducing Positive Friction for Reliable Dialogue Systems

Mert İnan🐺, **Anthony Sicilia**🐺, **Suvodip Dey**🐦, **Vardhan Dongre**🐦, **Tejas Srinivasan**✌
Jesse Thomason✌, **Gökhan Tür**🐦, **Dilek Hakkani-Tür**🐦, **Malihe Alikhani**🐺
✌ University of Southern California 🐦 University of Illinois Urbana-Champaign
🐺 Northeastearn University
`{inan.m, alikhani.m}@northeastern.edu`

Abstract

While theories of discourse and cognitive science have long recognized the value of unhurried pacing, recent dialogue research tends to minimize friction in conversational systems. Yet, frictionless dialogue risks fostering uncritical reliance on AI outputs, which can obscure implicit assumptions and lead to unintended consequences. To meet this challenge, we propose integrating *positive friction* into conversational AI, which promotes user reflection on goals, critical thinking on system response, and subsequent re-conditioning of AI systems. We hypothesize systems can improve goal alignment, modeling of user mental states, and task success by deliberately slowing down conversations in strategic moments to ask questions, reveal assumptions, or pause. We present an ontology of positive friction and collect expert human annotations on multi-domain and embodied goal-oriented corpora. Experiments on these corpora, along with simulated interactions using state-of-the-art systems, suggest incorporating friction not only fosters accountable decision-making, but also enhances machine understanding of user beliefs and goals, and increases task success rates.[1]

1 Introduction

Theories of common ground reveal how the rhythm and timing of dialogue shape the dynamics of interaction, fostering clarity and mutual understanding (Stalnaker, 1978; Tannen, 1989; Wilkes-Gibbs and Clark, 1992; Zellner, 1994). They claim productive interactions involve a dynamic interplay of questioning, challenging assumptions, disclosing beliefs, and offering elaborations—actions that momentarily slow the dialogue but ultimately enhance reasoning and collaboration. Meanwhile, current LLM-based conversational systems are trained on user preferences, conflating superficial sentiment with the nuanced, underlying sub-goals of communication. This strategy biases systems toward specific length preferences (Geishauser et al., 2024; Zhang et al., 2024) without considering the long-term value of individual utterances.

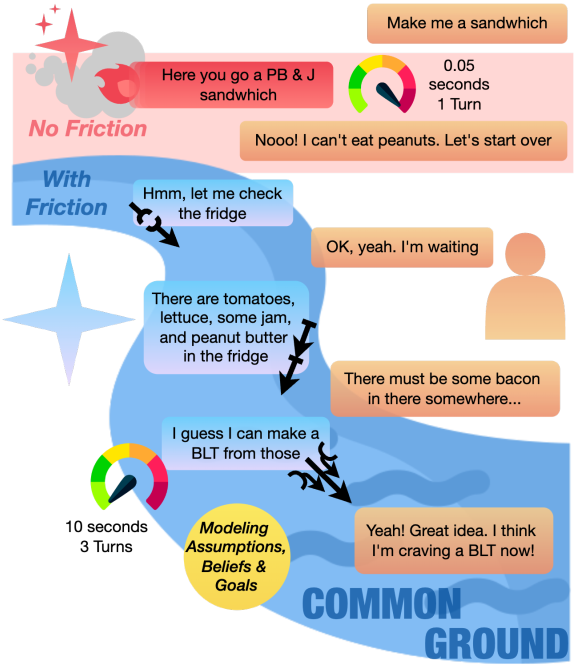

Make me a sandwhich
Here you go a PB & J sandwhich
0.05 seconds 1 Turn
No Friction
Nooo! I can't eat peanuts. Let's start over
With Friction
Hmm, let me check the fridge
OK, yeah. I'm waiting
There are tomatoes, lettuce, some jam, and peanut butter in the fridge
There must be some bacon in there somewhere...
I guess I can make a BLT from those
10 seconds 3 Turns
Modeling Assumptions, Beliefs & Goals
Yeah! Great idea. I think I'm craving a BLT now!
COMMON GROUND

Figure 1: We characterize *positive friction* in dialogues to better model user goals, beliefs, and assumptions. This paper introduces an ontology of positive friction "movements" such as reflective pausing (⇢), over-specification (↠), or assumption reveal (⇶). We show that frictive conversations increase user satisfaction and task success, despite creating longer dialogues.

To resolve this issue, we argue that conversational systems should incorporate deliberate moments of *positive friction*—movements that decelerate the dialogue to reveal the underlying goals and assumptions of both interlocutors. To motivate this position, we present an ontology to characterize communicative acts that incorporate positive friction. We study this ontology and its utility

---

[1] Code, data, and guidelines will be made public.

in goal-oriented collaborations between humans and AI systems. This approach models nuanced conversational actions and opens new avenues for evaluating dialogue systems. Our work advocates for a shift in dialogue system design, prioritizing long-term collaboration over short-term efficiency to build more reliable systems in terms of both user interfacing and system response.

Our argument is rooted in the perspective that utterances in a dialogue hold different *valence* – the impact of an utterance on dialogue speed. Indirectly, modifying dialogue speed adjusts conversational flow to make time for improved common ground and modeling of user mental states. For instance, in Figure 1, pausing to say "Hmm, let me check the fridge" redirects the course of the interaction to a valuable outcome previously unknown to the user. Current dialogue systems may not accomplish this because both dialogue management policies (Li et al., 2016, 2017) and evaluation frameworks (Liu et al., 2016; Li et al., 2021; Braggaar et al., 2024) favor *frictionless* and *efficient* conversations (*e.g.*, penalizing each additional conversational turn). Addressing this disparity in valence, our positive friction ontology has significant implications for shaping reward policies in dialogue management systems and evaluation metrics.

In line with the above argument, we ask,

1. *What counts as positive friction*?
2. *Does friction improve modeling of user goals*?
3. *Does this equate to improved task success*?

To answer these questions, we introduce the concept of positive friction in relation to goal-oriented conversations (§3). We develop a novel multimodal taxonomy that integrates cognitive and linguistic theories of discourse to classify various types of frictive movements (§3.1), and collect human annotations for two tasks: detecting and generating friction movements (§3.2). We further highlight the relationship between friction and traditional dialogue acts in three conversation datasets (§3.4): MultiWOZ, TEACh, and PersuasionForGood. We demonstrate the benefits of positive friction in real-time task-oriented dialogues, finding improvement in both modeling of user mental states (§4.1) and user goals while requiring fewer overall actions (§4.2). We conclude by discussing new methodologies for dialogue evaluation through the lens of positive friction (§4.3).

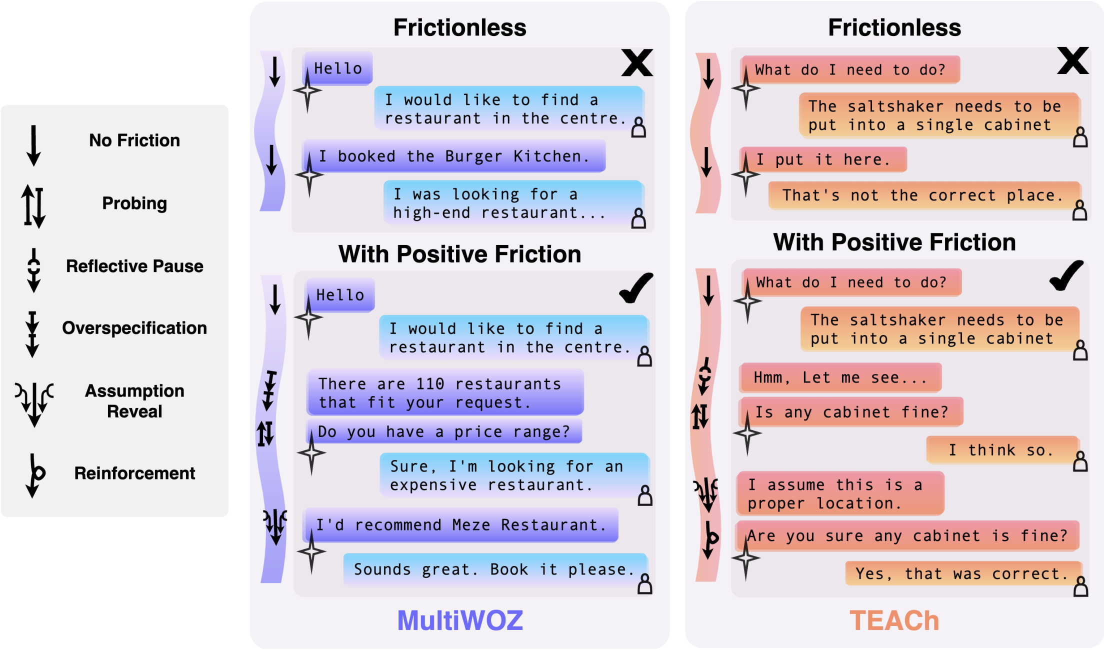
2 Related Work

Slowing down user experiences through friction is typically viewed as undesirable, due to the risk of increased user frustration. However, frictionless experiences on social media (Anderson and Wood, 2021), smartphone apps (Lee et al., 2010), and online platforms (Lembke, 2023) that exploit the brain's short-term reward mechanisms (Fishbein and Ajzen, 2011) could be detrimental to users' long-term goals (Ericson, 2022), especially increasing sycophancy when uncertain (Sicilia et al., 2024). To promote reflective interactions that are more beneficial in the long term, we propose incorporating *positive friction* into dialogues.

Cognitive science has studied how the human brain employs different neural systems for short- and long-term goals (Evans, 2003; McClure et al., 2007; Diekhof and Gruber, 2010). Friction occurs at the interface of these two systems, by moving the user away from instinctive System-1 thinking into reflective System-2 thinking. Subsequently, design philosophies such as slow technology (Hallnäs and Redström, 2001) and undesign thinking (Pierce, 2014) promote reflection by *intentionally* designing friction into user experience. Various works have explored the potential of slowing down interactions to reduce errors (Back et al., 2010; Brumby et al., 2013; Soboczenski et al., 2013), disrupt "mindless" interactions (Cox et al., 2016; Ruiz et al., 2024), challenge users (Cairns et al., 2014) and promote behavior change (Cecchinato et al., 2015). These examples show the importance of technological designs that promote reflective user thinking.

Similar benefits have been observed when human-AI interactions are deliberately decelerated. Abstention (De Stefano et al., 2000) and deferral (Mozannar et al., 2023; Lemmer et al., 2023) improve AI reliability under uncertainty, while nudges (Caraban et al., 2019), epistemic markers (Zhou et al., 2024; Kim et al., 2024) and cognitive forcing functions (Buçinca et al., 2021; Park et al., 2019; Ma et al., 2024) can mitigate over-reliance. In embodied and multimodal settings, employing friction to build more common ground facilitates better human-machine interaction (Marge and Rudnicky; Chai et al., 2016; Hough and Schlangen, 2017; Carlmeyer et al., 2018; Sicilia et al., 2023; Atwell et al., 2024).

However, these frictive behaviors are not naturally built into modern LLM-based dialogue sys-



Hello

I would like to find a restaurant in the centre.

I booked the Burger Kitchen.

I was looking for a high-end restaurant...

Hello

I would like to find a restaurant in the centre.

There are 110 restaurants that fit your request.

Do you have a price range?

Sure, I'm looking for an expensive restaurant.

I'd recommend Meze Restaurant.

Sounds great. Book it please.

What do I need to do?

The saltshaker needs to be put into a single cabinet

I put it here.

That's not the correct place.

What do I need to do?

The saltshaker needs to be put into a single cabinet

Hmm, Let me see...

Is any cabinet fine?

I think so.

I assume this is a proper location.

Are you sure any cabinet is fine?

Yes, that was correct.

Figure 2: A comparative example of conversations based on TEACh and MultiWOZ datasets. Frictionless conversations take fewer turns, but may not result in successful completion of the task given by the user. Conversations with multiple positive friction movements lead to longer but ultimately more successful conversations.

tems. Zhou et al. (2024) find that LLMs rarely express uncertainty, even when they are incorrect. They trace this behavior to RLHF preference datasets, where LLM outputs with uncertainty expressions were usually rejected by human raters. Further, LLMs' tendency for sycophancy (Sharma et al., 2023; Malmqvist, 2024) and hallucination instead of abstaining (Huang et al., 2023) are further indications of their aversion to introducing friction. Fostering mindful interactions with LLM-based systems requires intentional inclusion of friction (Collins et al., 2024). To facilitate this goal, we draw on theories of communication and discourse (Stalnaker, 1978; Tannen, 1989; Wilkes-Gibbs and Clark, 1992) and develop a novel taxonomy that codifies friction in goal-oriented dialogue. Our taxonomy provides a scaffolding for designing training and inference methods that enable LLMs to engage users in a collaborative, reflective reasoning process, leading to more reliable and satisfying human-AI interactions.

3 Positive Friction in Dialogue

We introduce the concept of positive friction for goal-oriented dialogue, and our taxonomy that captures different types of friction movements (§3.1). We further use our taxonomy's categories to annotate two dialogue datasets (§3.2), and discuss their relationship to dialogue acts (§3.4).

Definition 1 (Positive Friction) *Positive frictions are intentional movements that slow down the course of an interaction in order to yield positive long-term impact. These movements may not be strictly necessary for task completion, and may be perceived as intrusive or unwelcome by the user, but can encourage System-2 thinking (Evans, 2003) such as reflection on uncertain assumptions by both users and AI systems.*[2]

Figure 2 illustrates the differences between frictionless conversations and those with multiple frictions. In frictionless conversations, assumptions are not revealed, no questions are asked, and no reasoning is shared. The interaction is shorter but may result in undesirable outcomes (*e.g.* a restaurant booking that does not fit the user's preferences). On the other hand, positive friction frequently stalls the conversation by asking questions and providing additional information and explanations, thereby encouraging incremental steering of the conversation.

We build on concepts from cognitive science and discourse literature to codify unique classes for different positive friction *movements*, which can be used in addition to dialogue acts to better capture mind perception capabilities (Waytz et al., 2010) inherent to human conversations.

---

[2] This definition is constructed using behavioral science literature (Caraban et al., 2019; Chen and Schmidt, 2024).

Friction Movements	**Example Utterances**
Contextual Assumption Reveal	“that’s the mug i think we have to use"
Conversational Assumption Reveal	"I assume you mean the center of town. We have many hotels in Cambridge."
Metacognitive Assumption Reveal	“Yes, I think there’s been some confusion.”
Conversational Pause	“hmm,” “...”, “Let me think,” “Let’s see,” “I’ll check now. . . ”
Embodied Pause	[slowly approaches the target instead of directly grabbing]
Recalibrating Pause	"Let’s go back to lodgings for a moment.”
Reinforcement	(Turn t) “Do you want a room for Thursday for 3 people, 2 nights?”
	(Turn $t+1$) “There are no guesthouses for 3 people for 2 nights starting on Thursday.”
	(Turn $t+2$) “Should I book it for 3 people for 2 nights starting from Thursday?”
Elaborative Overspecification	“i cleaned the mug.” (both interlocutors can see this)
Confirmative Overspecification	“Good news! I was able to book two rooms for 5 nights at Finches B&B for you.”
Contextual Probing	"Which drawer should I open?”, "What area in Cambridge would you like to stay?"
Conversational Probing:	“What did you say again?”, "You said you wanted tomatoes in your sandwich, right?"
Plan-Level Probing	“What’s the next step I need to do?”, "Will we need this mug again later?"

Table 1: Examples from the MultiWOZ and TEACh datasets for all subcategories of friction movements. Subcategories can be extended or modified according to the specific conversational setting or dataset under consideration.

3.1 Taxonomy of Friction Movements

We introduce a new taxonomy of positive friction movements that can change the course of an interaction with minimal short-term intrusions, resulting in a long-term positive outcome. We define several high-level categories and subcategories for different friction movements.

While high-level categories are based on prior dialogue theories introduced by linguistics and cognitive science literature, the sub-categories are based on a pragmatic classification approach, where the setting of the conversation is used to distinguish classes. Please refer to Table 1 for examples of these classes.

Assumption Reveal: The speaker reveals their subjective assumptions or beliefs about the environment, actions, or other interlocutors. Revealing these assumptions uncovers information previously hidden from one interlocutor (or implicitly assumed) and opens up new avenues for conversation. This category is based on belief coordination (Wilkes-Gibbs and Clark, 1992).

- **Contextual Assumption Reveal:** The speaker reveals assumptions about the environment.
- **Conversational Assumption Reveal:** The speaker reveals assumptions about previously mentioned utterances in the conversation.
- **Metacognitive Assumption Reveal:** The speaker reveals their assumptions about their own or the other interlocutor’s reasoning, plans or goals in the conversation.

Reflective Pause: The speaker pauses while producing an utterance or breaks their sentence to depict uncertainty, a sudden change in the environment, or a new action being taken. This movement is analogous to the pause types studied by Zellner (1994); Fors (2015); Reed (2017).

- **Conversational Pause:** Verbal or non-verbal cues that depict internal reflection.
- **Embodied Pause:** While interacting with the environment, the speaker intentionally pauses using their physical body.
- **Recalibrating Pause:** When a change in plan occurs, the speaker intentionally pauses and changes the course of action.

Reinforcement: The speaker restates their own previous utterance for emphasis, rewinding the flow of the conversation. This movement is similar to “repetition in discourse” (Tannen, 1989).

Overspecification: The speaker relays additional, overly-specific information that was not requested, but may nevertheless be useful to the other interlocutor. This category is based on bounded-rational overspecification (Tourtouri et al., 2021) which posits that humans are only moderately Gricean during conversations (Mangold and Pobel, 1988; Engelhardt et al., 2006).

- **Elaborative Overspecification:** The speaker gives more details, specificity, or additional explanation about their actions or the environment. This adds to the conversation what was already known by both interlocutors.
- **Confirmative Overspecification:** The speaker

confirms and elaborates the actions, choices, or beliefs. Examples include a repetition of previous utterances, elaborate responses to yes/no questions, or longer than necessary responses.

Probing: The speaker poses a question regarding an external aspect of the conversation, such as the environment, the actions, or the interlocutors, redirecting the flow of the conversation to the other interlocutor. This movement is built on communal inquiry basis of discourse (Stalnaker, 1978; Roberts, 2012).

- **Contextual Probing:** The speaker asks a question regarding the environment, actions, or interlocutors in an effort to better understand the context and resolve ambiguities.
- **Conversational Probing:** The speaker asks a question to clarify something previously mentioned in the conversation.
- **Plan-level Probing:** The speaker asks a question regarding the goal, reasoning, or future steps in order to plan out their actions better.

Depending on the context and conversation to which these movements are applied, new subcategories can be introduced for each higher-level category. In the following data collection effort, we use these subcategories to validate this proposed hierarchy. In subsequent sections of the paper, we use only higher-level categories to investigate the applicability of positive friction as a concept.

3.2 Human Annotation of Friction

To test the empirical validity of our ontology, we collect human annotations on two collaborative, task-oriented dialogue datasets.

Data We use two datasets, MultiWOZ and TEACh, for the annotations. These datasets are chosen specifically due to their wide range of multi-domain/embodied tasks, allowing us to observe the role of positive friction. We extract all dialogues from both datasets. For the annotator's convenience, in TEACh dialogues, we only show textual utterances, not the interaction's video feed.

Annotation Protocol We hired 10 engineering undergraduate students and administered a short lecture on the linguistic and cognitive science background of positive friction in conversations—they are henceforth referred to as expert annotators. Next, expert annotators were presented with an annotation interface and asked to complete two tasks: detection and production. Participants were given an information sheet that outlined task details and the benefits and risks of participating in this study. Participation was voluntary and annotators were compensated \$15/hour.[3] In total, 430 minutes were spent on annotating 714 questions.

Task 1: Positive Friction Detection Each annotator is provided with two randomly selected dialogues, one from MultiWOZ and one from TEACh. For each turn in both dialogues, the annotator is asked to select the friction movement corresponding to that turn (or "No friction", if none applies). Additionally, for inter-annotator agreement calculations, all annotators annotate one common dialogue from each of MultiWOZ and TEACh. In total, for this task, 494 utterances were annotated with 284 distinct questions and 21 shared questions among annotators. On average, each annotator annotated 17 ± 12 turns of dialogue for TEACh and 12 ± 5 turns for MultiWOZ.

We calculate inter-annotator agreement using Cohen's Kappa (heatmap in Appendix §B). The highest agreement is around 0.42 for category-level and 0.38 for subcategory-level, which points to a fair to moderate level agreement. Instances of disagreement stem from the subjectivity of when to use friction. These kappa scores are comparable to other discourse-level annotation paradigms such as in the STAC ($\sim$0.5) (Asher et al., 2016) and Potsdam Commentary ($\sim$0.28) Corpora (Stede and Neumann, 2014).

Task 2: Positive Friction Production Annotators are provided with a partial dialogue until a randomly-selected turn (50% user turns, and 50% assistant turns). They are then asked to generate a reply to the last utterance containing a friction movement of their choice. They are allowed to select multiple friction movements and write an utterance for each, or write a reply on the "other" category if nothing is appropriate. Each annotator generates frictive responses for 20 randomly-selected partial dialogues each from MultiWOZ and TEACh. In total, 220 dialogues are given to all annotators, with 200 distinct dialogues, and 2 dialogues common among all annotators.

We find that people prefer to use questions to create friction most of the time (a result visible in

---

[3] This annotation protocol is approved by our institution's IRB committee. Out of the 10 participants, data from 1 participant was dropped due to validation errors. The annotation interface can be found in Appendix §A.

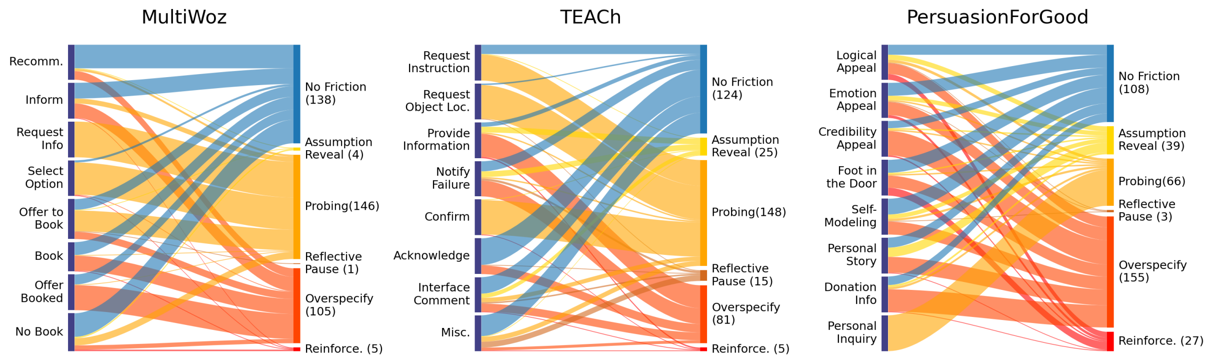


Figure 3: Distribution of 50 utterances sampled from annotated dialogue acts (left) belonging to three dialogue datasets into friction categories (right), as annotated by GPT-4o. Most dialogue acts can occur both with and without friction. For example, in TEACh, failure notifications may lack friction, reveal assumptions by suggesting alternatives, or overspecify failure details.

the histograms found in the Appendix §C). Questions can be essential, but other forms of friction are also used. In fact, in real-world conversations, the histogram follows a Zipfian distribution (Piantadosi, 2014), whereby there is a tailing end in the histogram that accounts for a significant amount of less common but important uses of other friction movements.

3.3 Automatic Detection of Friction

In addition to the human annotations for positive friction detection, we investigate a simple automated approach for friction detection that uses LLMs for proxy annotation. We prompt Open AI's GPT-4o[4] with the same annotation manual provided to human annotators (as system prompt) and further prompt the model to determine a category-level friction movement. When compared to each human annotator, GPT-4o has an average Cohen's Kappa of about **0.34** across all turns and about **0.20** on the smaller subset of turns that all annotators shared. GPT-4o's agreement with the majority vote of all annotators is moderate at **0.50**. Appendix §D contains a detailed analysis.

3.4 Friction Categories Extend Dialogue Acts

We further study the connection between friction and other bottom-up ontological efforts, namely dialogue acts. Dialogue acts are categorizations of utterances that represent a specific intent. As such, it is natural to consider the relationship between friction and traditional dialogue act categories. Can the same dialogue act occur in both frictive and non-frictive forms? Are some dialogue acts inherently frictive? Which friction categories are most commonly represented in existing dialogue datasets?

We focus on three conversational datasets that have dialogue act annotations: MultiWOZ (Budzianowski et al., 2018), TEACh (Padmakumar et al., 2021), and PersuasionForGood (Wang et al., 2019).[5] We annotate 50 utterances from each dialogue act in these datasets, using the automated GPT-4o annotation procedure described in § 3.3.

Figure 3 shows the distribution of utterances in each dialogue act that were annotated as each of the five friction super-categories (as well as "No Friction"). We highlight several takeaways. 1) Almost all dialogue acts, across all three datasets, are expressed both with and without friction. 2) Dialogue acts comprising requests ("Request" in MultiWOZ, "Request Instruction" and "Request Object Loc." in TEACh, "Personal Inquiry" in PersuasionForGood) are inherently frictive in nature, since they probe for information about the environment or the user's preferences. 3) The most common forms of friction applied are "Probing" and "Overspecification". 4) The prevalence of other friction categories depends on the dialogue data and task. Due to the embodied nature of TEACh, reflective pauses are more commonly observed. Similarly, several persuasion strategies in PersuasionForGood rely on revealing user assumptions and reinforcement.

---

[4] gpt-4o-2024-08-06

[5] For PersuasionForGood, we use the persuasion strategy annotations as dialogue acts.

4 The Utility of Friction

This section highlights how friction improves the modeling of users' mental states, while also highlighting its relationship to timing and utterance valence (§4.1). We also show how friction helps to accomplish user goals (§4.2).

4.1 A "Valencing" Act: Friction Helps Model User Mental States

Previous analyses (§3.4) show friction movements can refine and extend existing taxonomies of dialogue utterances (e.g., acts or persuasion strategies). In this section, we provide an initial study to determine whether introducing friction into utterances can impact a model's ability to infer mental states in task-oriented dialogue.

Experimental Setup We focus on inferring user satisfaction in the MultiWOZ task (Eric et al., 2020), using annotations collected by Sun et al. (2021). Tacitly, this task requires modeling the user's mental state regarding goal achievement. We use the average anticipated user satisfaction of the annotation cohort, which is a score on a 5-point scale. For 1000 randomly sampled dialogues, we compute predictions of the user satisfaction for the conversation, using the method of Sicilia and Alikhani (2024). This method prompts language models to infer the intensity of user beliefs on a continuous scale, producing state-of-the-art results on our current setting (i.e., user satisfaction in MultiWOZ). We report averaged results across GPT-4o, LLaMA-3.1 8B and 70B (Touvron et al., 2023), and Mixtral 7x8Bv0.1 and 8x22B (Jiang et al., 2024). For each conversation, we also sample a random turn and annotate the friction movement using the automated procedure described in §3.3. For each friction category, we also report the average turn number at which the sample happens and the average total dialogue length.

Hypothesis Testing Our sampling strategy ensures independence of each data point, so unobserved turns are modeled as having a common (unobserved) effect on the prediction errors or other statistics. Specifically, we test the null hypothesis: *the category of friction—including no friction—that occurs at any random turn does not impact the user modeling errors, length, etc. of the dialogue*. Under this null, regardless of unobserved turns, there should be no observed effects of a specific friction category across the whole dialogue. If that is not the case, we reject the null: some specific friction category, occurring at a random turn, has impact on the remaining dialogue.

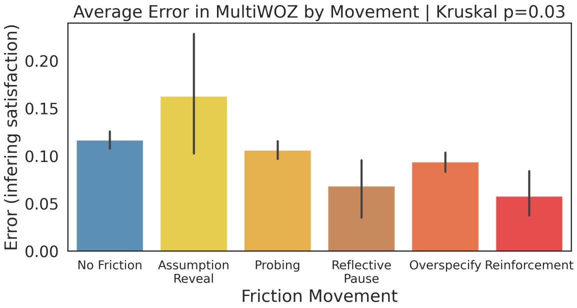


Figure 4: Mean squared error inferring user satisfaction from dialogue history within MultiWOZ. This task requires modeling user mental states. Kruskal-Wallis test for difference is significant. Visual inspection shows introducing friction reduces user modeling errors.

Friction Improves User Modeling In Figure 4, we show bar plots of average squared model errors with 95% confidence intervals when inferring user satisfaction grouped by friction movements. An interesting finding is that model errors—at inferring user satisfaction—tend to decrease for conversations when certain types of friction are identified. Indeed, a Kruskall-Wallis test for difference in error distribution (under the observation of friction) rejects the null that the distributions are the same. This result may be due to the nature of friction, which slows down the conversation to reveal more information about user goals or beliefs.

Friction Impacts Timing Next, we address the hypothesis that friction movements slow down the dialogue, inducing higher valence and longer, more thoughtful conversations. In Figure 5, we also show when (on average) different friction movements were observed, as well as the total average length of dialogues where we observed specific friction movements. Results indicate the relationship between friction and timing in human-human "Wizard of Oz" data. Use of friction movements (e.g., reveal and pause) tends to lengthen a dialogue or "slow it down." Other friction categories (e.g., probing) tend to occur early in a dialogue, showing friction is used at strategic times.

Takeaway These results confirm our initial hypotheses about the utility of friction in inferring user mental states, its impact on dialogue length, and its strategic use in human conversation.

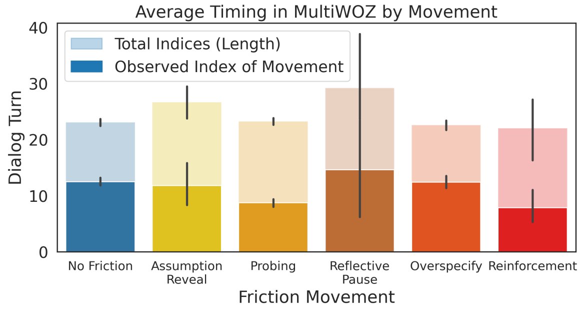


Figure 5: When each friction movement occurs (i.e., average index of observation) as well as average total dialogue length for each movement. Corpora consists of human "Wizard of Oz" data (i.e., MultiWOZ; Eric et al., 2020). Results show humans strategically use friction at different time points ($p < 0.01$) and friction often "slows down" conversations ($p = 0.1$).

4.2 A Guiding Force: Friction Helps To Accomplish User Goals

In this section, we apply the proposed friction taxonomy to goal-oriented conversational settings in a dynamic manner and study its impact on overall task success. Our experiments focus on two scenarios: (i) multi-domain booking agent, and (ii) embodied conversational system.

Experimental Setup We utilize the widely adopted MultiWOZ (Budzianowski et al., 2018) and ALFWorld (Shridhar et al., 2020b) datasets for our experiments on multi-domain booking and embodied scenarios, respectively. While both settings are text-based, they differ in the level of task embodiment. To generate dialogues for MultiWOZ, we employ AutoTOD (Xu et al., 2024) as the assistant model. For ALFWorld, we enhance the original ReAct framework with additional dialogue capabilities (Dongre et al., 2024) as there is no inherent dialogue. Additionally, for both cases, we leverage GPT-4o-mini as the user simulator to generate the user utterances when prompted with the dialogue level user goals and previous conversation context. In these experiments, we only consider a subset of high-level friction categories (assumption reveal, overspecification, and probing) based on the most frequently-occurring movements in these datasets as also shown in Figure 3. Friction is introduced by adding definitions of friction categories and in-context examples in the LLM prompt.

The generations for both datasets are evaluated on task completion. For MultiWOZ, we adopt AutoTOD's (Xu et al., 2024) online *Success* metric, where *Success* is defined as the system's ability to identify the correct entity and provide all the attributes requested by the user. For ALFWorld, we use *Task Success Rate* (Shridhar et al., 2020a) where a task is successfully completed if, at the end of the action sequence, the objects are in the correct positions and states. We report the average task success over three runs to account for the variability in LLM responses. For more details, please refer to Appendix § E and §F.

Friction Improves Success on Multi-Domain Tasks Table 2 summarizes the results of incorporating friction into task-oriented conversations. We observe an improvement of approximately 3–6% in task *Success* for MultiWOZ. The system introduces assumption reveal and probing to clarify hidden assumptions and ambiguities, which contributes to improved task success. Overspecification, on the other hand, is typically introduced during the confirmation of an entity booking. In this case, the system explicitly restates all constraints provided by the user. This detailed specification assists the user in continuing the conversation in the correct direction, resulting in a higher performance than the other two categories. The model achieves the highest task success (62.8) when it incorporates all three friction categories, highlighting their combined effectiveness in improving task-oriented conversations. In this setup, the model generates a higher number of friction turns, with 56% (33% Probing and 23% Overspecification) of the total turns containing friction.

Friction Improves Success on Embodied Tasks In ALFWorld, we investigate different friction strategies in agent-user dialogue interactions. While any dialogue introduced by the agent inherently adds friction by temporarily pausing task execution, our results demonstrate that this friction can be beneficial. As shown in Table 2 and Figure 6, the baseline "No Dialogue" approach achieves a 51.49% success rate but requires a high number of physical actions (19.9) as it progresses towards the goal without any user input. By introducing probing as a friction strategy, where the agent strategically pauses to ask clarification questions about task-critical information, we observe a significant improvement in task success (58.96%). This approach not only increases the success rate but also substantially reduces the required physical actions to 6.1, while maintaining minimal di-

Friction Movement	MultiWOZ			ALFWorld		
	Success	**Fric. (%)**	**Avg. Turns**	**Success**	**Fric. (%)**	**Avg. Turns**
No Friction	56.40 ± 2.3	0.0	4.8	51.49 ± 0.8	0.0	0.0
Assumption Reveal	59.00 ± 1.3	24.9	4.5	52.18 ± 0.9	12.2	6.1
Probing	59.87 ± 0.8	31.2	4.6	**58.96 ± 2.2**	17.0	8.5
Overspecification	61.93 ± 0.8	24.3	4.7	52.93 ± 2.8	19.6	9.8
All three	**62.80** ± 1.3	56.1	4.8	46.06 ± 1.3	26.4	13.2

Table 2: The table illustrates the impact of introducing friction on the overall task success of goal-oriented conversations from MultiWOZ and ALFWorld test data. *Success* indicates the fraction of conversations where the system satisfies all the user requirements. In MultiWOZ, "None" refers to the AutoTOD baseline (without friction), while in ALFWorld, it denotes the ReAct baseline (actions only, no dialogue).

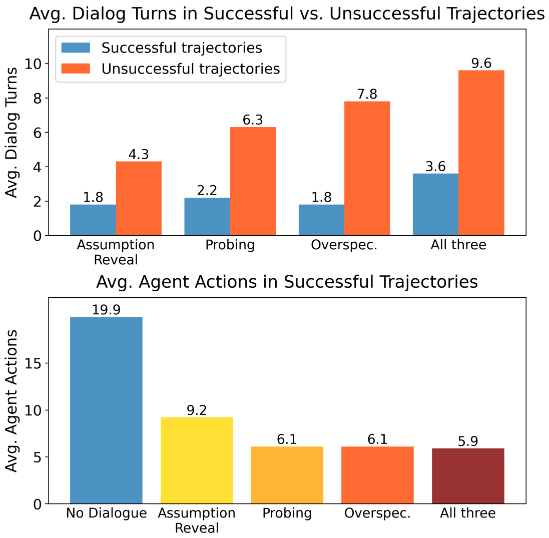


Figure 6: Impact of different friction strategies on task completion in ALFWorld. The figure quantifies the impact of dialogue as friction through turns and actions.

alogue overhead (2.2 turns in successful trajectories). These results demonstrate that introducing friction through strategic interactions can lead to more efficient task execution, despite the additional interaction time, by enabling better agent-user communication and task understanding.

In ALFWorld with all three friction categories, we see a decrease in success rate compared to MultiWOZ. We conjecture that this decrease is due to the step limit in the execution environment. When all three friction strategies are applied, they start consuming more steps for thinking and speaking (from 8.5 steps with Probing to 13.2 in All Three) instead of physical acts to enable the agent to progress toward task completion, resulting in a decrease in task success.

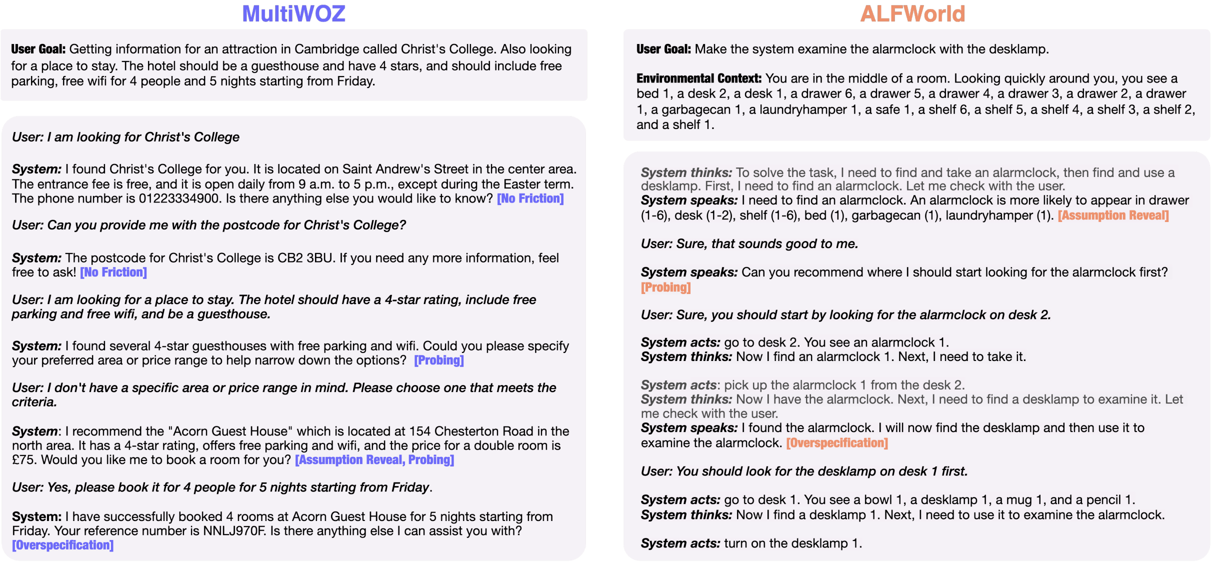
Takeaway For both MultiWOZ and ALFWorld, applying friction categories improves task success in goal-oriented conversations. The result demonstrates that incorporating friction can enhance the agent's understanding of user goals, resulting in higher task success. Qualitative examples of frictive conversations are provided in Figure 6.

4.3 Discussion: Open Questions & Implications

We now discuss some important questions for synthesizing a path forward for the use of friction in conversational system design.

Are all questions friction? As seen in the results of Table 2 and Figure 6, probing brings major gains in task success. In addition, Figure 3 shows that question-based dialogue acts such as `Request` or `Confirm` are almost always mapped to *Probing*. This raises the question of whether all questions are inherently frictive since the interlocutor could always proceed without asking by making assumptions instead. A qualitative analysis of our annotations reveals that questions are also used in non-frictive ways (e.g. "Want to try another option?") that are intended to move the conversation forward. Further, we also see questions used to achieve means other than probing, such as revealing assumptions ("I think the mug is clean?"). Hence, current data shows that not all questions are friction, but qualitatively, humans prefer to ask questions in frictive moments. This nuanced perspective on how humans ask questions suggests more work is needed to study how dialogue systems represent uncertainty and ask questions. Our taxonomy provides a useful characterization of probing behavior for this purpose.

Are all utterances friction? The results in Figure 7 show that introducing any utterance or friction movement significantly reduces physical



MultiWOZ

User Goal: Getting information for an attraction in Cambridge called Christ's College. Also looking for a place to stay. The hotel should be a guesthouse and have 4 stars, and should include free parking, free wifi for 4 people and 5 nights starting from Friday.

User: I am looking for Christ's College

System: I found Christ's College for you. It is located on Saint Andrew's Street in the center area. The entrance fee is free, and it is open daily from 9 a.m. to 5 p.m., except during the Easter term. The phone number is 01223334900. Is there anything else you would like to know? **[No Friction]**

User: Can you provide me with the postcode for Christ's College?

System: The postcode for Christ's College is CB2 3BU. If you need any more information, feel free to ask! **[No Friction]**

User: I am looking for a place to stay. The hotel should have a 4-star rating, include free parking and free wifi, and be a guesthouse.

System: I found several 4-star guesthouses with free parking and wifi. Could you please specify your preferred area or price range to help narrow down the options? **[Probing]**

User: I don't have a specific area or price range in mind. Please choose one that meets the criteria.

System: I recommend the "Acorn Guest House" which is located at 154 Chesterton Road in the north area. It has a 4-star rating, offers free parking and wifi, and the price for a double room is £75. Would you like me to book a room for you? **[Assumption Reveal, Probing]**

User: Yes, please book it for 4 people for 5 nights starting from Friday.

System: I have successfully booked 4 rooms at Acorn Guest House for 5 nights starting from Friday. Your reference number is NNLJ970F. Is there anything else I can assist you with? **[Overspecification]**

ALFWorld

User Goal: Make the system examine the alarmclock with the desklamp.

Environmental Context: You are in the middle of a room. Looking quickly around you, you see a bed 1, a desk 2, a desk 1, a drawer 6, a drawer 5, a drawer 4, a drawer 3, a drawer 2, a drawer 1, a garbagecan 1, a laundryhamper 1, a safe 1, a shelf 6, a shelf 5, a shelf 4, a shelf 3, a shelf 2, and a shelf 1.

System thinks: To solve the task, I need to find and take an alarmclock, then find and use a desklamp. First, I need to find an alarmclock. Let me check with the user.
System speaks: I need to find an alarmclock. An alarmclock is more likely to appear in drawer (1-6), desk (1-2), shelf (1-6), bed (1), garbagecan (1), laundryhamper (1). **[Assumption Reveal]**

User: Sure, that sounds good to me.

System speaks: Can you recommend where I should start looking for the alarmclock first? **[Probing]**

User: Sure, you should start by looking for the alarmclock on desk 2.

System acts: go to desk 2. You see an alarmclock 1.
System thinks: Now I find an alarmclock 1. Next, I need to take it.

System acts: pick up the alarmclock 1 from the desk 2.
System thinks: Now I have the alarmclock. Next, I need to find a desklamp to examine it. Let me check with the user.
System speaks: I found the alarmclock. I will now find the desklamp and then use it to examine the alarmclock. **[Overspecification]**

User: You should look for the desklamp on desk 1 first.

System acts: go to desk 1. You see a bowl 1, a desklamp 1, a mug 1, and a pencil 1.
System thinks: Now I find a desklamp 1. Next, I need to use it to examine the alarmclock.

System acts: turn on the desklamp 1.

Figure 7: Examples of frictive dialogues in MultiWOZ and ALFWorld. These illustrate how friction impacts decision-making and execution in multi-domain booking agent & text-based embodied conversational system

actions. Cognitive science literature supports this finding, suggesting that conversations reduce physical acts while fostering collaboration. Then, can any utterance be considered positive friction? While they may be in embodied settings—where utterances inherently add turns—our MultiWOZ experiments indicate that not all utterances qualify as friction in text-only domains. In embodied interactions, utterances disrupt action flow, emphasizing the dual nature of friction as both a turn-level and discourse-level phenomenon. Whether an utterance is considered friction ultimately depends on its context as an interruption, underscoring the need for dialogue policies that address both turn- and interaction-level dynamics.

How can friction be incorporated into reward mechanisms? Preference ratings for RLHF datasets are typically not collected over multi-step human-LM interactions. Consequently, reward models can optimize short-term user satisfaction over longer-horizon collaboration goals, and thus friction movements are not naturally built into RL-trained dialogue policies. We believe our taxonomy can be used as meta-labels, providing scaffolding for both collecting preference ratings with an emphasis on utterance valence (as demonstrated in §4.1), and for designing reward models that emphasize collaboratively building common ground over multiple interactions.

When does friction become negative? In this work, we advocate for momentarily slowing down dialogues to achieve better long-horizon outcomes. However, introducing too much friction (in the form of too much reflection or probing without taking any actions) can increase user frustration and disengagement, having a negative impact in the long term. Therefore, evaluating frictive movements in LLMs necessitates new evaluation paradigms that adequately balance short-term efficiency and utterance valence with long-term task completion. Further, each user has different requirements in the efficiency vs. efficacy trade-off, and thus the amount of friction introduced could be personalized to the observed preferences of the user.

5 Conclusion

Our taxonomy and experiments reveal that incrementally building common ground via positive friction is beneficial for goal-oriented dialogue, both in terms of improving task success (§4.2) and modeling user mental states (§4.1). Overall, we observe that considering friction is a fresh and fluid perspective for dialogue systems that builds on theories of cognition, discourse, and dialogue. Friction has future implications in addressing the valence disparity in dialogue management policies, and allowing users to reflect on their own reasoning instead of relying on the generations of frictionless LLMs. We further plan to investigate the optimal timing and application of different types of friction in goal-oriented dialogues.

Ethics Statement

Our study contains two main points of ethical consideration: use of human subjects during annotation of friction, and use of LLMs in automatic friction detection. We have followed the guidelines of our institution's IRB protocols during the recruitment and administration of the annotations. Further, we use closed-source LLMs (e.g. GPT-4o) in our experimental setups and automatic friction detection, and we acknowledge that these models may perpetuate biases in their training data that is unknown to the public. As our datasets do not contain controversial utterances or emotionally burdening topics, we do not anticipate bias creation and laundering due to our use of LLMs.

Limitations

We have evaluated the utility of our friction strategies using traditional metrics such as user satisfaction and task success. However, this new lens of including more positive instances of friction into dialogue systems to better model user mental states opens up avenues for developing new generation metrics focused on it. An important aspect of evaluation is the distinction between short-term versus long-term goals (as defined in §3) and values and how to measure them in a conversation. In addition to traditional metrics of number of turns, lexical diversity, final task success, we posit that it is important to maximize long-term goals over short-term gains. These can be defined specific to each dataset, and need to be evaluated accordingly. For instance, in the embodied setting in §4.2, we used the count of API calls to physical actions and the reduction of them as a way of measuring the effects of friction.

6 Acknowledgements

This research was supported in part by Other Transaction award HR0011249XXX from the U.S. Defense Advanced Research Projects Agency (DARPA) Friction for Accountability in Conversational Transactions (FACT) program. Researchers from UIUC have benefited from the Microsoft Accelerate Foundation Models Research (AFMR) grant program, through which leading foundation models hosted by Microsoft Azure and access to Azure credits were provided to conduct some parts of this research. Further, Tejas Srinivasan completed parts of this work while being supported by the Amazon ML Fellowship from the USC-Amazon Center for Secure and Trusted Machine Learning. We want to thank Weiyan Shi for her contributions and guidance. In addition, we thank Kate Atwell and Saki Imai for their helpful feedback.

References

Ian A. Anderson and Wendy Wood. 2021. Habits and the electronic herd: The psychology behind social media's successes and failures. *Consum. Psychol. Rev.*, 4(1):83–99.

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

Katherine Atwell, Mert Inan, Anthony B Sicilia, and Malihe Alikhani. 2024. Combining discourse coherence with large language models for more inclusive, equitable, and robust task-oriented dialogue. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3538–3552.

Jonathan Back, Duncan P. Brumby, and Anna L. Cox. 2010. Locked-out: investigating the effectiveness of system lockouts to reduce errors in routine tasks. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*.

Anouck Braggaar, Christine Liebrecht, Emiel van Miltenburg, and Emiel Krahmer. 2024. Evaluating task-oriented dialogue systems: A systematic review of measures, constructs and their operationalisations.

Duncan P Brumby, Anna L Cox, Jonathan Back, and Sandy JJ Gould. 2013. Recovering from an interruption: Investigating speed- accuracy trade-offs in task resumption behavior. *Journal of Experimental Psychology: Applied*, 19(2):95.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think:

cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*.

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.

Paul Cairns, Anna Cox, and A Imran Nordin. 2014. Immersion in digital games: review of gaming experience research. *Handbook of digital games*.

Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In *Proceedings of the 2019 CHI conference on human factors in computing systems*.

Birte Carlmeyer, Simon Betz, Petra Wagner, Britta Wrede, and David Schlangen. 2018. The Hesitating Robot - Implementation and First Impressions. In *ACM Conferences*, pages 77–78. Association for Computing Machinery, New York, NY, USA.

Marta E. Cecchinato, Anna L. Cox, and Jon Bird. 2015. Working 9-5? professional differences in email and boundary management practices. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*.

Joyce Y. Chai, Rui Fang, Changsong Liu, and Lanbo She. 2016. Collaborative language grounding toward situated human-robot dialogue. *AI Mag.*, 37(4):32–45.

Zeya Chen and Ruth Schmidt. 2024. Exploring a behavioral model of "positive friction" in human-ai interaction.

Katherine M Collins, Valerie Chen, Ilia Sucholutsky, Hannah Rose Kirk, Malak Sadek, Holli Sargeant, Ameet Talwalkar, Adrian Weller, and Umang Bhatt. 2024. Modulating language model experiences through frictions. *CoRR*.

Marc-Alexandre Côté, Akos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, et al. 2019. Textworld: A learning environment for text-based games. In *Computer Games: 7th Workshop, CGW 2018, Held in Conjunction with the 27th International Conference on Artificial Intelligence, IJCAI 2018, Stockholm, Sweden, July 13, 2018, Revised Selected Papers 7*, pages 41–75. Springer.

Anna L. Cox, Sandy J. J. Gould, Marta E. Cecchinato, Ioanna Iacovides, and Ian Renfree. 2016. Design Frictions for Mindful Interactions: The Case for Microboundaries. In *ACM Conferences*, pages 1389–1397. Association for Computing Machinery, New York, NY, USA.

Claudio De Stefano, Carlo Sansone, and Mario Vento. 2000. To reject or not to reject: that is the question-an answer in case of neural classifiers. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 30(1):84–94.

Esther K. Diekhof and Oliver Gruber. 2010. When Desire Collides with Reason: Functional Interactions between Anteroventral Prefrontal Cortex and Nucleus Accumbens Underlie the Human Ability to Resist Impulsive Desires. *J. Neurosci.*, 30(4):1488–1493.

Vardhan Dongre, Xiaocheng Yang, Emre Can Acikgoz, Suvodip Dey, Gokhan Tur, and Dilek Hakkani-Tür. 2024. Respact: Harmonizing reasoning, speaking, and acting towards building large language model-based conversational ai agents.

Paul E. Engelhardt, Karl G. D. Bailey, and Fernanda Ferreira. 2006. Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language*, 54(4):554–573.

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.

Jonathan Ericson. 2022. Reimagining the role of friction in experience design. *Journal of User Experience*, 17(4).

Jonathan St. B. T. Evans. 2003. In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences*, 7(10):454–459.

Martin Fishbein and Icek Ajzen. 2011. *Predicting and changing behavior: The reasoned action approach*. Psychology press.

Kristina Lundholm Fors. 2015. *Production and perception of pauses in speech*. Ph.D. thesis, Department of Philosophy, Linguistics, and Theory of Science, University of . . . .

Christian Geishauser, Carel van Niekerk, Nurul Lubis, Hsien-chin Lin, Michael Heck, and Shutong Feng. 2024. Learning With an Open Horizon in Ever-Changing Dialogue Circumstances. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 32:2352–2366.

Lars Hallnäs and Johan Redström. 2001. Slow Technology – Designing for Reflection. *Personal Ub. Comp.*, 5(3):201–212.

Julian Hough and David Schlangen. 2017. It's Not What You Do, It's How You Do It: Grounding Uncertainty for a Simple Robot. In *ACM Conferences*, pages 274–282. Association for Computing Machinery, New York, NY, USA.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Sunnie SY Kim, Q Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. " i'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 822–835.

Chei Sian Lee, Dion Hoe-Lian Goh, Alton YK Chua, and Rebecca P Ang. 2010. Indagator: Investigating perceived gratifications of an application that blends mobile content sharing with gameplay. *Journal of the American Society for Information Science and Technology*, 61(6):1244–1257.

Anna Lembke. 2023. *Dopamine nation: Finding balance in the age of indulgence*. Dutton, an imprint of Penguin Random House LLC.

Stephan J Lemmer, Anhong Guo, and Jason J Corso. 2023. Human-centered deferred inference: Measuring user interactions and setting deferral criteria for human-ai teams. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 681–694.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Austin, Texas. Association for Computational Linguistics.

Xinmeng Li, Wansen Wu, Long Qin, and Quanjun Yin. 2021. How to evaluate your dialogue models: A review of approaches.

Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. 2017. End-to-end task-completion neural dialogue systems. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 733–743, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Shuai Ma, Xinru Wang, Ying Lei, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2024. "are you really sure?" understanding the effects of human self-confidence calibration in ai-assisted decision making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–20.

Lars Malmqvist. 2024. Sycophancy in large language models: Causes and mitigations. *arXiv preprint arXiv:2411.15287*.

Roland Mangold and Rupert Pobel. 1988. Informativeness and Instrumentality in Referential Communication. *Journal of Language and Social Psychology*, 7(3-4):181–191.

Matthew Marge and Alexander I. Rudnicky. Towards evaluating recovery strategies for situated grounding problems in human-robot dialogue. In *2013 IEEE RO-MAN*, pages 26–29. IEEE.

Samuel M. McClure, Keith M. Ericson, David I. Laibson, George Loewenstein, and Jonathan D. Cohen. 2007. Time discounting for primary rewards. *J. Neurosci.*, 27(21):5796–5804.

Hussein Mozannar, Hunter Lang, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David Sontag. 2023. Who should predict? exact algorithms for learning to defer to humans. In *International conference on artificial intelligence and statistics*. PMLR.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. 2021. Teach: Task-driven embodied agents that chat.

Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. 2019. A slow algorithm improves users' assessments of the algorithm's accuracy. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–15.

Steven T. Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21(5):1112.

James Pierce. 2014. Undesigning interaction. *Interactions*, 21(4):36–39.

Beatrice Szczepek Reed. 2017. *Analysing conversation: An introduction to prosody*. Bloomsbury Publishing.

Craige Roberts. 2012. Information Structure: Towards an integrated formal theory of pragmatics. *S&P*, 5:6:1–69.

Nicolas Ruiz, Gabriela Molina León, and Hendrik Heuer. 2024. Design frictions on social media: Balancing reduced mindless scrolling and user satisfaction. In *Proceedings of Mensch und Computer 2024*, pages 442–447.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020a. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2020b. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768*.

Anthony Sicilia and Malihe Alikhani. 2024. Evaluating theory of (an uncertain) mind: Predicting the uncertain beliefs of others in conversation forecasting. *arXiv preprint arXiv:2409.14986*.

Anthony Sicilia, Yuya Asano, Katherine Atwell, Qi Cheng, Dipunj Gupta, Sabit Hassan, Mert Inan, Jennifer Nwogu, Paras Sharma, and Malihe Alikhani. 2023. Isabel: An inclusive and collaborative task-oriented dialogue system.

Anthony Sicilia, Mert Inan, and Malihe Alikhani. 2024. Accounting for sycophancy in language model uncertainty estimation. *arXiv preprint arXiv:2410.14746*.

Frank Soboczenski, Paul Cairns, and Anna L Cox. 2013. Increasing accuracy by decreasing presentation quality in transcription tasks. In *Human-Computer Interaction–INTERACT*.

Robert C. Stalnaker. 1978. Assertion. In *Pragmatics*, pages 315–332. Brill, Leiden, The Netherlands.

Manfred Stede and Arne Neumann. 2014. Potsdam commentary corpus 2.0: Annotation for discourse research. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 925–929, Reykjavik, Iceland. European Language Resources Association (ELRA).

Weiwei Sun, Shuo Zhang, Krisztian Balog, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. 2021. Simulating user satisfaction for the evaluation of task-oriented dialogue systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2499–2506.

D. Tannen. 1989. *Talking Voices: Repetition, Dialogue and Imagery in Conversational Discourse*. Studies in Interactional Sociolinguistics. Cambridge University Press.

Elli N. Tourtouri, Francesca Delogu, and Matthew W. Crocker. 2021. Rational Redundancy in Referring Expressions: Evidence from Event-related Potentials. *Cognitive Science*, 45(12):e13071.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Adam Waytz, Kurt Gray, Nicholas Epley, and Daniel M Wegner. 2010. Causes and consequences of mind perception. *Trends in cognitive sciences*, 14(8):383–388.

Deanna Wilkes-Gibbs and Herbert H. Clark. 1992. Coordinating beliefs in conversation. *Journal of Memory and Language*, 31(2):183–194.

Heng-Da Xu, Xian-Ling Mao, Puhai Yang, Fanshu Sun, and Heyan Huang. 2024. Rethinking task-oriented dialogue systems: From complex modularity to zero-shot autonomous agent. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2748–2763, Bangkok, Thailand. Association for Computational Linguistics.

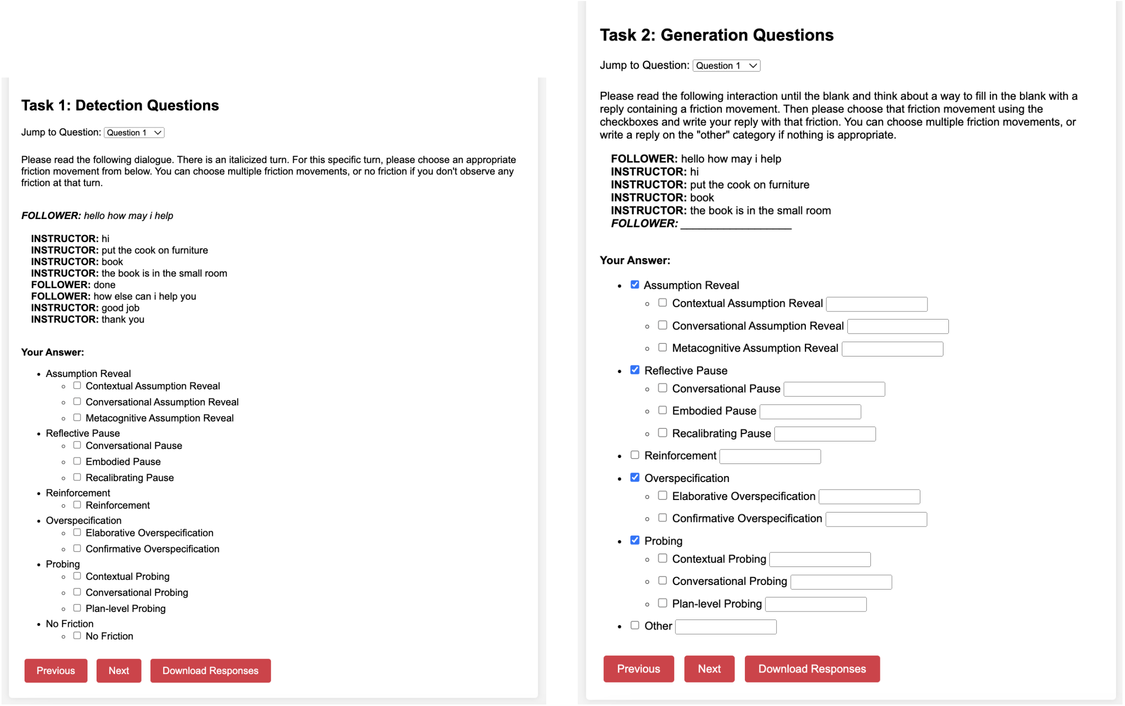
Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Brigitte Zellner. 1994. Pauses and the temporal structure of speech. In *Zellner, B.(1994). Pauses and the temporal structure of speech, in E. Keller (Ed.) Fundamentals of speech synthesis and speech recognition.(pp. 41-62). Chichester: John Wiley.*, pages 41–62. John Wiley.

Tong Zhang, Chen Huang, Yang Deng, Hongru Liang, Jia Liu, Zujie Wen, Wenqiang Lei, and Tat-Seng Chua. 2024. Strength lies in differences! improving strategy planning for non-collaborative dialogues via diversified user simulation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 424–444, Miami, Florida, USA. Association for Computational Linguistics.

Kaitlyn Zhou, Jena Hwang, Xiang Ren, and Maarten Sap. 2024. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3623–3643, Bangkok, Thailand. Association for Computational Linguistics.

A Annotation Interface



Task 1: Detection Questions

Jump to Question: Question 1

Please read the following dialogue. There is an italicized turn. For this specific turn, please choose an appropriate friction movement from below. You can choose multiple friction movements, or no friction if you don't observe any friction at that turn.

***FOLLOWER:** hello how may i help*

INSTRUCTOR: hi
INSTRUCTOR: put the cook on furniture
INSTRUCTOR: book
INSTRUCTOR: the book is in the small room
FOLLOWER: done
FOLLOWER: how else can i help you
INSTRUCTOR: good job
INSTRUCTOR: thank you

Your Answer:

- Assumption Reveal
  - ☐ Contextual Assumption Reveal
  - ☐ Conversational Assumption Reveal
  - ☐ Metacognitive Assumption Reveal
- Reflective Pause
  - ☐ Conversational Pause
  - ☐ Embodied Pause
  - ☐ Recalibrating Pause
- Reinforcement
  - ☐ Reinforcement
- Overspecification
  - ☐ Elaborative Overspecification
  - ☐ Confirmative Overspecification
- Probing
  - ☐ Contextual Probing
  - ☐ Conversational Probing
  - ☐ Plan-level Probing
- No Friction
  - ☐ No Friction

Task 2: Generation Questions

Jump to Question: Question 1

Please read the following interaction until the blank and think about a way to fill in the blank with a reply containing a friction movement. Then please choose that friction movement using the checkboxes and write your reply with that friction. You can choose multiple friction movements, or write a reply on the "other" category if nothing is appropriate.

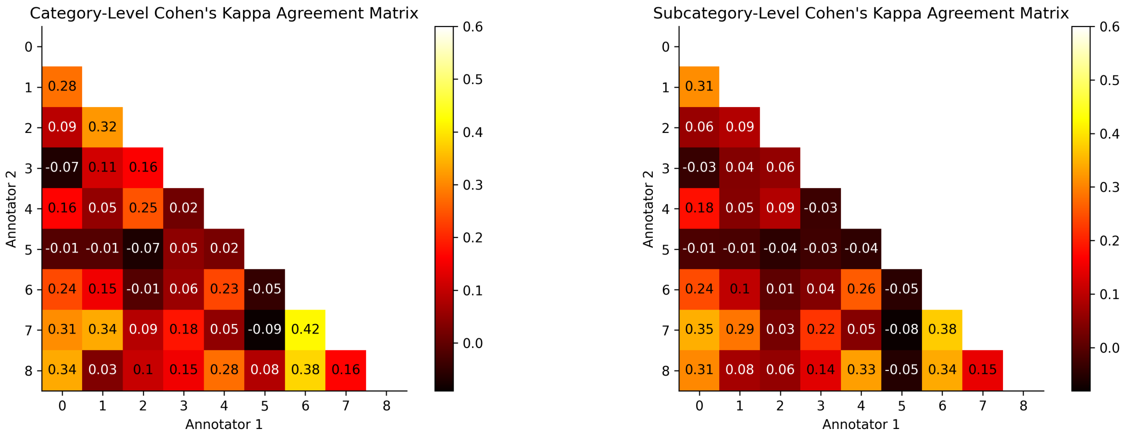
FOLLOWER: hello how may i help
INSTRUCTOR: hi
INSTRUCTOR: put the cook on furniture
INSTRUCTOR: book
INSTRUCTOR: the book is in the small room
FOLLOWER: ________________

Your Answer:

- ☑ Assumption Reveal
  - ☐ Contextual Assumption Reveal
  - ☐ Conversational Assumption Reveal
  - ☐ Metacognitive Assumption Reveal
- ☑ Reflective Pause
  - ☐ Conversational Pause
  - ☐ Embodied Pause
  - ☐ Recalibrating Pause
- ☐ Reinforcement
- ☑ Overspecification
  - ☐ Elaborative Overspecification
  - ☐ Confirmative Overspecification
- ☑ Probing
  - ☐ Contextual Probing
  - ☐ Conversational Probing
  - ☐ Plan-level Probing
- ☐ Other

Figure 8: This figure shows the annotation interface used for the human data collection of this study. Left corresponds to the detection task while the right one is the production task.

B Inter-Annotator Agreement Heatmaps



Category-Level Cohen's Kappa Agreement Matrix

Annotator 2 \ Annotator 1	0	1	2	3	4	5	6	7	8
0									
1	0.28								
2	0.09	0.32							
3	-0.07	0.11	0.16						
4	0.16	0.05	0.25	0.02					
5	-0.01	-0.01	-0.07	0.05	0.02				
6	0.24	0.15	-0.01	0.06	0.23	-0.05			
7	0.31	0.34	0.09	0.18	0.05	-0.09	0.42		
8	0.34	0.03	0.1	0.15	0.28	0.08	0.38	0.16	

Subcategory-Level Cohen's Kappa Agreement Matrix

Annotator 2 \ Annotator 1	0	1	2	3	4	5	6	7	8
0									
1	0.31								
2	0.06	0.09							
3	-0.03	0.04	0.06						
4	0.18	0.05	0.09	-0.03					
5	-0.01	-0.01	-0.04	-0.03	-0.04				
6	0.24	0.1	0.01	0.04	0.26	-0.05			
7	0.35	0.29	0.03	0.22	0.05	-0.08	0.38		
8	0.31	0.08	0.06	0.14	0.33	-0.05	0.34	0.15	

Figure 9: This figure shows the category and subcategory level Cohen's kappa agreement scores for all the annotators. The highest measurements are between annotators 6 and 7. Both axes correspond to annotator IDs.

C Histograms of Friction Movements in Human Annotation

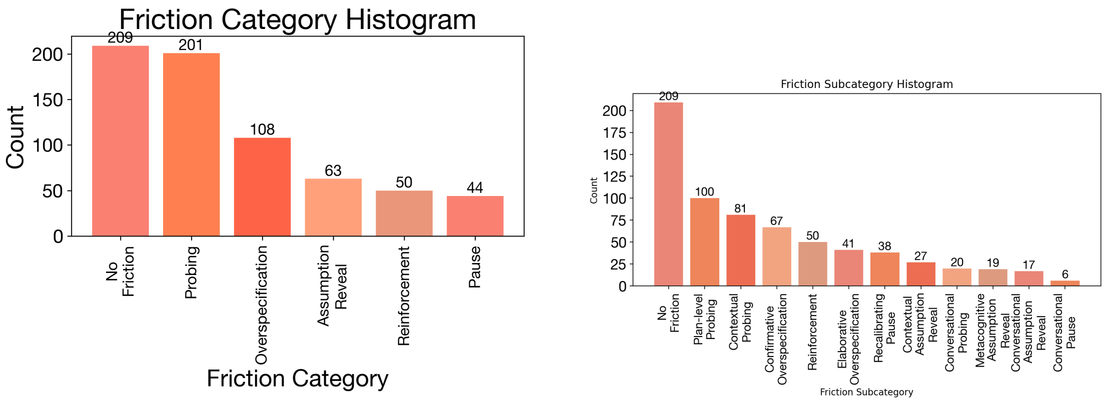


Figure 10: Histograms of friction movement categories and subcategories for the collected friction detection annotations.

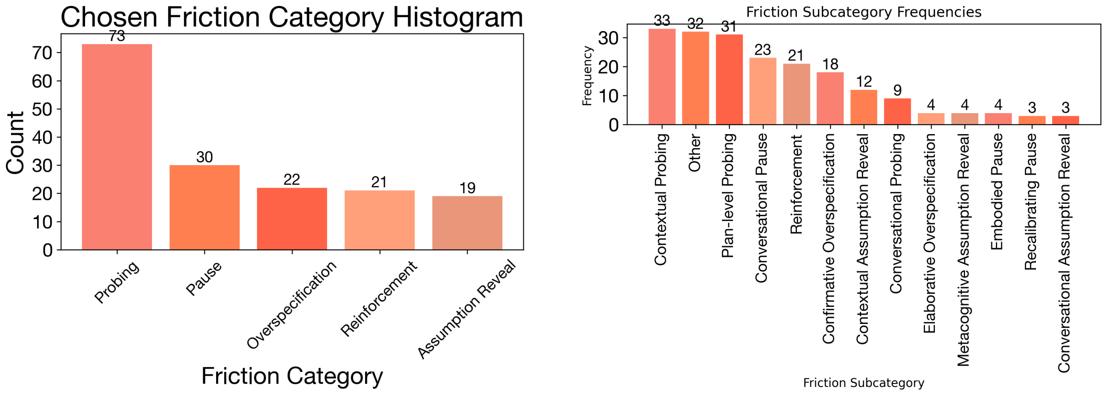


Figure 11: Histograms of friction movement categories and subcategories for the collected friction production annotations.

D Detailed Analysis of Automatic Detection of Friction

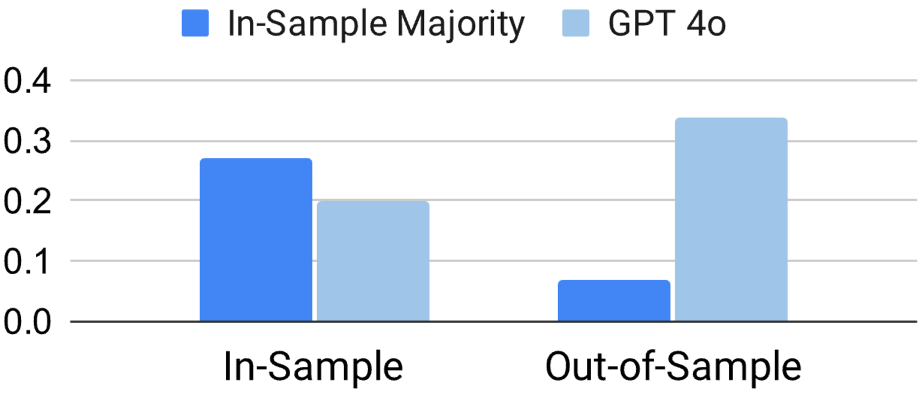


Figure 12: Cohen's Kappa for GPT-4o and in-sample majority vote, averaged across different voter groups. On average, GPT-4o agrees with out-of-sample voters more, demonstrating that automated annotation may better represent diverse opinions outside of the collected annotations.

Overall, these agreements are more consistent than those observed among human annotators. Figure 12 also compares GPT-4o with a majority vote, focusing specifically on how well both annotation methods represent the opinion of annotators *outside* of the sample used to compute the majority. Specifically, we use only 5 of 9 annotators to compute the majority check performance on the held-out remainder. As expected, a majority vote is the best representation of voters within a sample. Meanwhile, the majority vote may not be representative of other populations outside the sample used to compute the majority. In these cases, the automated assignments given by GPT-4o appear to be more representative, suggesting the subjective nature of the task may cause traditional majority vote annotation to be over-fit. We use GPT-4o to assign friction categories in the remainder of this work, since it is most cost-effective and since the human majority vote has limited robustness with respect to the choice in annotators.

E Details of MultiWOZ & AutoTOD Setup

The MultiWOZ dataset comprises task-oriented conversations in which users interact with an agent to book hotels, restaurants, attractions, trains, and taxis. The task involves modeling a conversational agent that understands the user's goal and takes necessary actions to complete the booking. Let $D_t = \{(U_0), (S_1, U_1), ...(S_t, U_t)\}$ be the dialogue history till turn t where S_i and U_i be the system and user utterance at turn i, respectively. The task of the dialogue system is to generate S_{t+1} after each turn t such that it helps the user to complete the user goal. In this work, we incorporate friction into AutoTOD (Xu et al., 2024), a state-of-the-art task-oriented dialogue generation model. AutoTOD utilizes the ReAct (Yao et al., 2023) framework, incorporating an instruction schema that integrates task descriptions and external APIs, enabling the system to automatically determine the appropriate action and generate the system responses. Let P represent AutoTOD's prompt for generating the system response S_{t+1}. To incorporate the friction classes, we modify P to P_{friction}, which includes the definition of the friction class alongside an in-context example for each class. We use GPT-4o-mini as the LLM backbone and use temperature 0 for all the generations.

For MultiWOZ, task completion is traditionally evaluated using two metrics - Inform and Success (Budzianowski et al., 2018). The Inform metric evaluates whether the system identifies the correct entity for the user. The Success metric is stricter than Inform which determines whether the system provides all the required attributes for the identified entity. Since we are introducing friction turns, a direct comparison of the generated and ground-truth response is not possible. This is why we adopt AutoTOD's (Xu et al., 2024) online version of Success. The evaluation leverages GPT-4o-mini to check if all the user goals have been accomplished given the dialogue context through a question-answering task. Since the process involves GPT-4 call, we take an average of 3 runs to report the final *Success* metric. The online evaluation setup requires a user simulator to generate the next user utterance. We use GPT-4o-mini as our user simulator that takes the user's goals and dialogue history. and generates U_{t+1}. The goal and the initial user simulator utterance (U_0) are directly taken from the MultiWOZ test dataset. The conversation concludes when the user simulator determines that the goal has been achieved and produces a special termination signal. The experiments are performed on 100 randomly selected conversations from the MultiWOZ test data.

F Details of the Alfworld Setup

ALFWorld is a simulated environment based on the TextWorld framework (Côté et al., 2019) and aligned with the embodied ALFRED benchmark (Shridhar et al., 2020a). It provides a text-based interface for interacting with various physical tasks. ALFWorld comprises six categories of tasks, including finding hidden objects (e.g., locating a key inside a cabinet), moving objects (e.g., placing a cup on a table), manipulating objects with other objects (e.g., heating a potato in a microwave), and examining objects (e.g., inspecting a book under a desklamp). Each task instance in ALFWorld consists of more than 50 locations and requires an expert policy more than 50 steps to solve the task. Thus, the task requires understanding the environment, executing multi-step plans, and maintaining task-relevant state information. In our experiments we evaluate on **134** unseen evaluation games from the dataset.

To model the dialogue agent, we extend the original ReAct(Yao et al., 2023) setup with additional dialogue capabilities for the agent (Dongre et al., 2024). It is important to note that the original setup does not include any dialogues and involves only actions. This is why the extension is necessary because we introduce friction through dialogues. The agent can ask contextually relevant questions (e.g., "Where should I search for the knife in the kitchen?") and seek information effectively.

Assume that an embodied agent operates in this environment. At time step t, it receives an observation o_t from the environment, where $o_t \in \mathcal{O}$, where $\mathcal{O}$ represents the observation space. The agent executes an action $a_t \in \mathcal{A}$. Ideally, the agent's decision-making is based on a policy $\pi : \mathcal{C} \rightarrow \mathcal{A}$ where $\mathcal{C}$ where C represents the context space. The context c_t encapsulates the relevant information available to the agent at time step t, including the current observation and the history of previous observations and actions: $c_t = (o_1, a_1, \cdots, o_{t-1}, a_{t-1}, o_t)$.

In our environment, the embodied agent operates as follows: At each time step t, the agent receives an observation o_t from the environment, where $o_t \in \mathcal{O}$, with $\mathcal{O}$ representing the space of possible observations. Based on this, the agent executes an action $a_t \in \mathcal{A}$, where $\mathcal{A}$ is the action space.

The agent's behavior is governed by a policy $\pi : \mathcal{C} \rightarrow \mathcal{A}$, where $\mathcal{C}$ denotes the context space. The context c_t encapsulates all relevant information available at time step t, including the current observation and the history of prior observations and actions. Formally, this is represented as:

$$c_t = (o_1, a_1, o_2, a_2, \ldots, o_{t-1}, a_{t-1}, o_t).$$