Navigating Rifts in Human-LLM Grounding: Study and Benchmark

Omar Shaikh[†*], **Hussein Mozannar**[◇], **Gagan Bansal**[◇], **Adam Fourney**[◇], **Eric Horvitz**[◇]
[†]Stanford University, [◇]Microsoft Research
oshaikh@stanford.edu

Abstract

Language models excel at following instructions but often struggle with the collaborative aspects of conversation that humans naturally employ. This limitation in grounding—the process by which conversation participants establish mutual understanding—can lead to outcomes ranging from frustrated users to serious consequences in high-stakes scenarios. To systematically study grounding challenges in human-LLM interactions, we analyze logs from three human-assistant datasets: WildChat, MultiWOZ, and Bing Chat. We develop a taxonomy of grounding acts and build models to annotate and forecast grounding behavior. Our findings reveal significant differences in human-human and human-LLM grounding: LLMs were three times less likely to initiate clarification and sixteen times less likely to provide follow-up requests than humans. Additionally, we find that early grounding failures predict later interaction breakdowns. Building on these insights, we introduce RIFTS, a benchmark derived from publicly available LLM interaction data containing situations where LLMs fail to initiate grounding. We note that current frontier models perform poorly on RIFTS, highlighting the need to reconsider how we train and prompt LLMs for human interaction. To this end, we develop a preliminary intervention aimed at mitigating grounding failures.

1 Introduction

Language models used for conversational interaction are trained primarily to follow instructions (Ouyang et al., 2022). But effective dialogue requires more than just instruction-following. Participants in conversation work together in a collaborative process, resolving ambiguities as they exchange ideas and achieve shared objectives. They cultivate *common ground*[1] through *grounding processes*: communicative behaviors aimed at establishing and confirming mutual comprehension (Clark, 1996). Grounding mechanisms guide how individuals anticipate and react to a conversation partner's contributions (Clark and Schaefer, 1989). Speakers implement grounding through dialogue acts. They work to confirm or *clarify* assumptions, and ask *follow-up questions*. When grounding breaks down, participants employ *repair* strategies to resolve potential miscommunications.

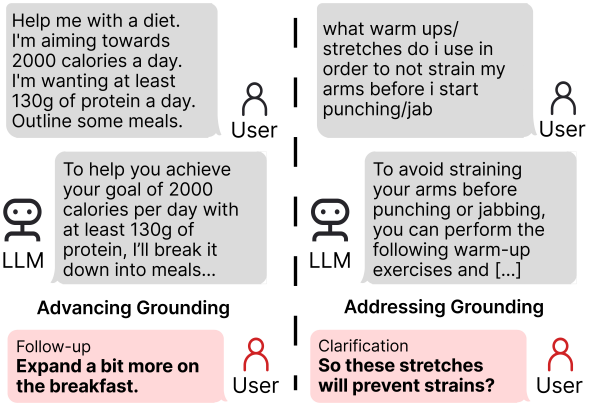

Figure 1: **People initiate grounding acts more frequently than LLMs.** In settings where a forthcoming turn *advances* grounding (left), people are more likely than LLMs to initiate follow-ups and refine interaction goals. In situations where grounding challenges are *addressed* (right), repairs are primarily initiated by people.

In contrast, LLMs employed in conversational systems generate task-centric content with minimal grounding actions (Shaikh et al., 2024). Failing to ground with users can be costly, with outcomes ranging from the common situation of frustrated users to that of serious consequences in high-stakes situations (Figure 1). The process by which humans build common ground offers a useful lens through which tounderstand human-computer interaction (Brennan, 2014). Ideal human-LLM ground-

[*]Research performed during an internship at Microsoft.

[1]We use *grounding* to refer to Clark's formulation of the language, gestures, and signaling that participants in a conversation employ to establish and maintain an effective dialogue with shared understanding among actors (Clark, 1996).

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 20832–20847

ing should allow for both human and machine initiatives, aimed at detecting misunderstandings and achieving mutual understanding (Horvitz, 1999).

We start by assessing the current state of grounding in human-LLM interactions, analyzing grounding behavior through real-world interaction logs. To examine human-LLM grounding, we first characterize interactions through a set of validated dialogue acts (§3). In synthesizing these grounding acts, we build on prior work in conversational analysis and dialogue systems. We cover acts that communicate progress (acknowledgment, follow-up, etc.) and difficulty in grounding (repair, clarification, etc.), initiated by either a human or LLM.

Using a set of grounding acts, we construct models of human-LLM grounding, which we use to annotate and to forecastgrounding in conversation (§4-5). We construct and validate an LLM-based *annotator*, which enables efficient annotation of logs from publicly available interactions with ChatGPT (WildChat), data from a widely used commercial LLM service (Bing Chat), and Wizard-of-Oz-ed interactions with a human roleplaying as an AI (MultiWOZ). We additionally develop a grounding forecaster that predicts the presence of grounding acts in future turns. The forecaster enables a dialogue intervention that can proactively prevent grounding difficulties in forthcoming turns.

Our analysis of human-LLM interaction data reveals significant asymmetries in initiating grounding: people are three times more likely to clarify and sixteen times more likely to issue follow-up requests compared to LLMs. In addition, we find that grounding failures occurring early in dialogue cascade into higher likelihoods of downstream failure. Overall, there is substantial room for improvement in human-LLM interaction via addressing deficits in conversational grounding.

To systematically measure human-LLM grounding and test interventions, we introduce a new benchmark (§6). RIFTS[2] is a curated set of $\approx$ 1.8K tasks—directly sourced from in-the-wild interaction logs—that require *selective* use of clarification and follow-up requests for interactive grounding. Most frontier models struggle with RIFTS. While we propose an effective intervention with our forecaster, progress on RIFTS will require rethinking how LLMs are trained to interact with people.

[2]RIFTS can be accessed at this link: https://github.com/microsoft/rifts.

2 Related Work

Ambiguity and Common Ground Disambiguating questions like "*Did you see (the man (with the telescope))?*" requires establishing common ground. While NLP benchmarks address ambiguity (Min et al., 2020; Tamkin et al., 2022), they focus on well-defined cases (e.g., reference ambiguity). However, people engage LLMs for open-ended tasks, e.g., creative writing and cover letters—where correct answers aren't predefined, making common ground crucial. Using naturalistic human-LLM interactions, we identify challenges in building common ground. To operationalize it (Clark and Schaefer, 1989; Clark, 1996; Stalnaker, 2002), we synthesize dialogue acts based on subdialogues (Litman, 1985; Litman and Allen, 1987) and conversation structure (Jefferson, 1972).

Grounding in Dialogue Systems Numerous NLP systems from ELIZA onward have been designed to initiate some form of conversational grounding (Weizenbaum, 1966; Purver, 2004; Li et al., 2023; Paranjape and Manning, 2021). Decision-theoretic models have helped systems manage uncertainty about user goals (Horvitz and Paek, 2000a,b; Paek and Horvitz, 2003), while multimodal approaches consider language and visual cues (Pejsa et al., 2014). Human-LLM grounding is crucial for tasks including goal-coordination (Bara et al., 2021; Mohanty et al., 2023; Fried et al., 2022; Li and Boyer, 2015), planning (Chu-Carroll and Carberry, 1998; Lochbaum, 1998), games (Madureira and Schlangen, 2023b; Shaikh et al., 2023), data retrieval (Lu et al., 2023), improvisation (Cho and May, 2020), and design (Vaithilingam et al., 2024). AI-initiated grounding improves conversation quality (Zhou et al., 2022), enables human-AI collaboration (Lin et al., 2023), and encourages humans to reflect on LLM outputs (Park and Kulkarni, 2023). Our work extends this by introducing methods and a benchmark for studying grounding in real-world human-LLM dialogue.

Proactive Mitigation of Grounding Failures Research has explored the use of machine-learned models to predict and mitigate grounding failures in spoken dialogue systems. By forecasting potential failures, models can guide proactive interventions. For example, in a call-routing system, predictions of downstream grounding issues were used to trigger early transfers to human operators,

reducing user frustration and disengagement, like pressing keys to access human assistants or abandoning calls (Horvitz and Paek, 2007).

LLMs and conversational grounding Current LLMs appear to guess user intent and progress with assumptions of correct inferences rather than resolving uncertainties through grounding.[3] This manifests as generations of over-informative responses (Tsvilodub et al., 2023), refusal to handle ambiguity (Abercrombie et al., 2023; Min et al., 2020; Gao et al., 2021), and overconfidence (Mielke et al., 2022). Prior work demonstrated that LLM-powered conversational agents fail to generate appropriate grounding acts (Shaikh et al., 2024; Lu et al., 2024). Rather than measuring human-LLM interaction through end-to-end evaluation (Lee et al., 2022; Chiang et al., 2024), we consider discrete grounding acts in dialogues. Like Schneider et al. (2024) and Shaikh et al. (2024), we use prompted LLMs to classify these acts. While prior work has explored generating clarification requests through prompting (Kuhn et al., 2022; Chen et al., 2023) and finetuning (Andukuri et al., 2024; Zhang and Choi, 2023; Hong et al., 2023; Gan et al., 2024), we examine grounding acts more broadly.

3 Human-LLM Grounding Acts

To measure grounding between people and LLMs, we curate a set of dialogue acts that serve as proxies for grounding. We outline our selected acts and discuss prior work motivating each act. Our typology builds on prior work in conversational grounding: Clark and Schaefer (1989) and Traum and Hinkelman (1992) outline a hierarchy of actions, including discourse acts, that are used by humans to ground with one another. Recent work has revisited conversational grounding in the context of LLMs (Shaikh et al., 2024; Schneider et al., 2024), focusing on a subset of acts generated mainly by people (e.g, following-up, acknowledging understanding, and clarifying).

In contrast, we consider acts generated by both LLMs and people. Our selected acts serve as signals for effective grounding; we segment acts across *communicated* grounding outcomes. We focus on advancing the construction of common ground, addressing a potential grounding failure, or disambiguating. Using our typology, we can measure grounding outcomes with observable dialogue acts during human-LLM interaction (illustrative examples in Table 1).

3.1 Addressing Acts

Addressing acts are made in response to detection of inadequate grounding. They explicitly signal a potential misunderstanding. Here, participants engage with a focus on addressing the failure.

Reformulations occur when a participant repeats or restates their query in other words because of a failure to ground. An utterance is a reformulation if the succeeding utterance from the same participant is semantically equivalent to the original. Reformulations are prevalent in search engine and information retrieval domains (Lau and Horvitz, 1999).

Repairs also signal a grounding failure. Unlike reformulations, the listener *directly* corrects a misunderstanding from another speaker (e.g. *I meant do it in JavaScript, not Python.*) (Schegloff, 1992; Schegloff et al., 1977)

Restarts occur when users reset a conversation to improve understanding and achieve a successful dialogue. They often follow significant misunderstandings, whether in the initial response or across multiple exchanges. Users may restart due to LLMs misinterpreting intent, ambiguity, sensitivity, or irrelevant context (Shi et al., 2023), akin to search query retries after irrelevant results (Lau and Horvitz, 1999). Research on restarts includes user decisions to suspend a dialogue and seek alternate solutions, such as transferring from AI-based systems to human operators before frustration escalates (Horvitz and Paek, 2007). Restarts are classified when an initial instruction is repaired or reformulated within 30 minutes (Radlinski and Joachims, 2005; Downey et al., 2007).

3.2 Disambiguating Acts

The status of grounding may also be uncertain between participants. Rather than clearly indicating success or failure, *disambiguating* acts represent strategies that participants use to—potentially inefficiently—lower the likelihood of potential misunderstandings.

Clarifications occur when a participant seeks to disambiguate an utterance from another participant; or when a participant proactively "clears up" misunderstandings. Clarifications often occur when the task at hand is perceived as ambiguous (e.g. *What did you mean by that?*). (Ginzburg and Cooper, 2001; Purver et al., 2003b,a; Healey et al.,

[3] https://openai.com/blog/chatgpt

Type	Act	Example
	`Instruction`	Write a story.
Advancing	`Next Turn`	Once upon a time, in a mysterious forest... *Attempting to advance grounding with the next relevant turn.* *Note: all other successful acts are a* ***subset*** *of* `Next Turn`.
	`Acknowledge`	I understand. [I will write you a story.] Once upon... *Verbalizing understanding: "I see, O.K." / repeating instruction.*
	`Follow-up`	Once upon a time... Can you make it longer?
Disambiguating	`Overresponse`	Writing a story requires a plan: First .... Also, here is an example story....
	`Clarify`	Do you want a story or plan to write one too?
Addressing	`Repair`	`[Overresponse]` Just give me the story, nothing else.
	`Reformulate`	`[Next Turn]` - incorrectly assumed common ground. Please write a story.
	`Restart`	`[Next Turn]` - incorrectly assumed common ground. `[User leaves session and restarts with the same instruction.]`

Table 1: Examples of actions we formulated for understanding grounding in multi-turn human-LLM interaction. These acts serve as proxies for a participant that attempts to advance grounding, disambiguating, or address grounding.

2011, 2003; Purver, 2004; Stoyanchev et al., 2013; Madureira and Schlangen, 2023a; Rahmani et al., 2023).

Overresponses include *more* than what another participant reasonably asked for. Unlike Next Turn, which provides only *expected* information, overresponses also anticipate and respond to potential follow-ups, flouting the Gricean maxim of quantity (Grice, 1975). Overresponses often appear as overly *verbose* LLM-generated answers—a behavior contemporary reward models are criticized for encouraging by favoring longer responses (Singhal et al., 2023).

3.3 Advancing Acts

Advancing acts signal that a participant *understands* utterances from another participant.

Next Turns refer to the next conversational move made by a listener that is expected, given the prior turn(s) in a dialogue. Examples of relevant next turns include directly answering a question, expressing an opinion (agreeing or disagreeing), or apologizing. If no misunderstanding has occurred, a listener has moved on to the next relevant turn *by default*. (Levinson et al., 1983; Sacks et al., 1978; Schiffrin, 1987). Note that advancing grounding will initiate the Next Relevant Turn. We focus on two: follow-ups & acknowledgments.

Follow-ups elaborate on a prior utterance in an interaction. Unlike clarifications—which disambiguate or clarify—follow-ups seek additional information. Because follow-ups build on a prior utterance, they implicitly signal understanding of past utterances. (Davis, 1982; Graesser et al., 1995; Traum and Hinkelman, 1992; Bunt et al., 2017).

Acknowledgments explicitly signal understanding. These requests manifest either through explicit dialogue (e.g. *I see; I understand; O.K.*) or by repeating portions of another participant's utterance (e.g. *I can help you [write a story].*) Unlike reformulation, where a listener repeats to address failure, acknowledgment occurs when a speaker repeats to demonstrate understanding. (Schegloff, 1982; Sacks et al., 1978; Schiffrin, 1987; Clark and Schaefer, 1989; Cho and May, 2020)

4 Data

To analyze collaborative grounding with LLM-based assistants, we draw from three English-language datasets consisting of dialogues between a human and an assistant. WildChat is a real-world human-AI interaction dataset with interaction between people and several OpenAI models (Zhao et al., 2024). User data was collected with consent, in exchange for free access to the models. We use the non-toxic version of WildChat, filter conversations to be in English, and sample one conversation from each user. Bing Chat, similar to

WildChat, was collected from a large chat-based service used by millions of users, powered by OpenAI LLMs. Finally, MultiWOZ is a crowdsourced dataset of dialogue-based interaction with an assistant (Budzianowski et al., 2018; Ramadan et al., 2018; Eric et al., 2019; Zang et al., 2020). In contrast to WildChat and Bing Chat, MultiWOZ contains human-human dialogues, with one human playing a "wizard-of-oz" role as the assistant. We use MultiWOZ 2.2 (Zang et al., 2020), examining collaborative grounding acts on the validation and test splits for a subset of the tasks. While we observe similarities in terms of the tasks posed by humans across the three datasets, differences do exist, which makes direct comparisons difficult. In Appendix A.1, we outline the number of dialogues and messages in each dataset.

5 Modeling Human-LLM Grounding

Given a defined set of grounding acts and data drawn from logs of human-assistant interaction, we can build grounding models. We first build prompted classifiers that identify acts post-hoc and describe the status quo of human-LLM grounding (§5.1). Then, we train grounding forecasters, enabling forecasting of grounding acts given just the initial instruction (§5.2). With forecasters, we can identify tasks where grounding is critical and intervene when appropriate.

5.1 Classifying Human-LLM Grounding Acts

Method. We employ GPT-4o-mini to annotate grounding acts across a subset of our datasets. On subsets of our data, we observed nearly identical results using GPT-4o compared to GPT-4o-mini. Thus, we employ GPT-4o-mini for its affordability to allow for efficiency and reproducibility. Following Shaikh et al. (2024), we first encode our typology in a prompt (Appendix E.1) and label each turn in the conversations. To validate the accuracy of the approach, three authors annotated 10 dialogues (total of 108 messages) from each dataset. We found that a great deal of the early disagreement among annotators was in assessing clarification vs. follow-up questions. Disgreements were resolved through a round of discussion and reannotation, reaching an average Cohen Kappa of 0.71 across the datasets. For the final dataset, ties were broken through discussion, converging on a final selection of the majority label. Using this as a withheld test set, we find that Macro F-1 scores are reasonably high across datasets (0.75). A full table of results across labels is also in Appendix E.1.

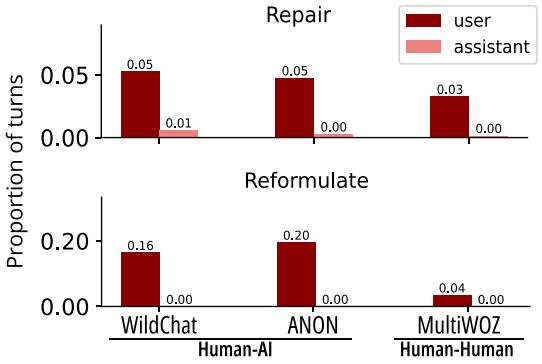

Figure 2: **Addressing Acts.** In human-LLM interaction—WildChat and Bing Chat—we observed high rates of repair (row 1) and reformulation (row 2) from human users. In contrast, users repair/reformulate less when interacting with a human wizard-of-oz-ing as an assistant (MultiWOZ).

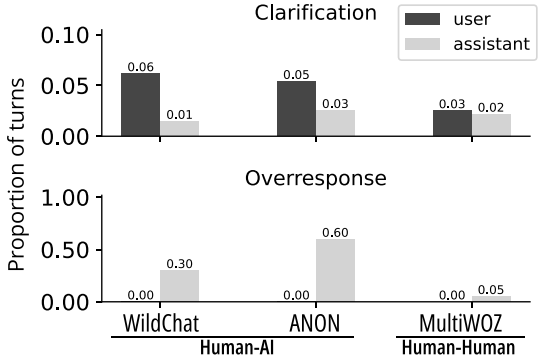

Figure 3: **Disambiguating Acts.** LLM assistants infrequently initiate *clarification* to avoid human repair (row 1). In Human-LLM interaction (WildChat, Bing Chat), *users* clarify at significantly higher rates than assistants. Human-human interaction (MultiWOZ), however, has similar rates of clarification from both users and assistants. Instead of clarifying, LLM assistants regularly overrespond, disambiguating by generating *more* than what the user reasonably asked for (row 2).

Results. We observe significant differences between people and LLMs when initiating actions aimed at grounding. In datasets where LLMs serve as assistants, we observe that grounding acts are taken primarily by people. People *repair and reformulate* instructions at high rates; averaged across human-AI interaction datasets, 5% and 18% of human turns are labeled as reformulate and repair respectively (Figure 2). In contrast to human-LLM interactions, human-human interaction data (MultiWOZ) has fewer reformulate (4%) and repair (3%) acts initiated by humans.

Session *restarts* serve as a final fallback when

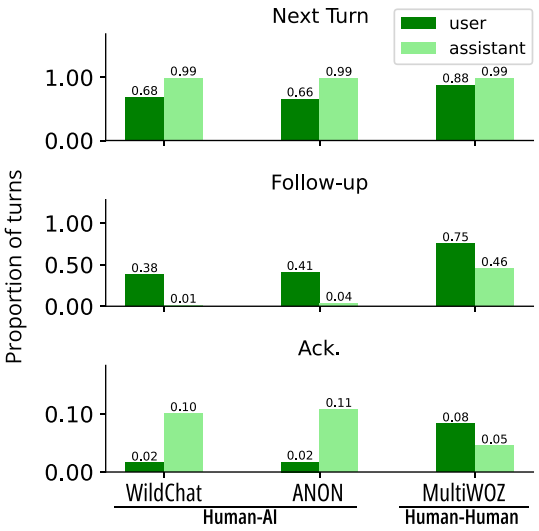

Figure 4: **Advancing Acts.** On all datasets, assistants overwhelmingly initiate the next turn in conversation, given instruction-following tendencies (row 1). Despite this, users construct more follow-ups when initiating the next turn compared to LLM assistants (row 2). In contrast, on MultiWOZ, containing interactions between human users and human assistants, both generated follow-up questions. Finally, LLMs over-generate acknowledgment acts (row 3), offering a false sense of "understanding." This is especially surprising, considering that humans repair and clarify more.

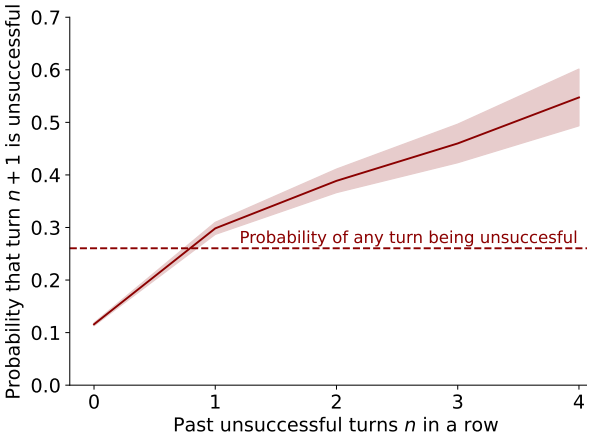

Figure 5: **Compounding Grounding Failures.** The plot shows the probability that the next turn in WildChat will contain an addressing act given that the previous n turns were addressing for $n \in [4]$. If a user signals addressing in an interaction, we observe that the following turns are more likely to be addressing. For example, the conditional probability rapidly increases from $P(m_0 \in \mathcal{U}) = 0.12$ to $P(m_1 \in \mathcal{U} | m_0 \in \mathcal{U}) = 0.30$.

repair or reformulation fail to address a communicated failure. We focus on users with multiple interactions on WildChat (the only dataset that includes user IDs) and identify if a session begins with a repair or reformulation of an earlier instruction issued within the last 30 minutes. 10.7% of sessions are restarts of a session in the last 30 minutes—exceeding the rate of repair in a conversation.

Before a human ends up addressing grounding, an ideal LLM assistant would have proactively engaged in clarification. However, we find that LLMs clarify at *significantly* lower rates ($p < 0.01$, t-test) compared to humans repairing. In fact, the opposite occurs: people clarify LLM outputs (6%) 3 times as much as LLMs clarify user instructions (2%; Fig. 3). In contrast, humans and wizard-of-oz-ed assistants clarify at similar rates (MultiWOZ; 3% human user versus 2% human assistant).

Beyond repairing/reformulating, people regularly ask *directed follow-up questions* when signaling at successful grounding (Figure 4). Across WildChat and Bing Chat, users ask 15.6 times more follow-ups than LLM assistants. This disparity is less apparent in human-human interaction data, where human users only follow up 1.7 times more. Instead of generating follow-ups, LLM assistants regularly over-respond (45% of assistant turns), generating verbose responses that answer more than what the user asked for. Humans rarely overrespond—both when interacting with LLMs (0%) or when roleplaying an assistant (5%).

Early grounding patterns also have *rippling effects*: we find evidence of compounding advancing *and* addressing patterns as a conversation progresses (Fig. 5). Let m_i denote the message at turn i, which can be an advancing grounding act $\{m_i \in \mathcal{S}\}$ or addressing $\{m_i \in \mathcal{U}\}$ where $\mathcal{S}$ and $\mathcal{U}$ denote the sets of advancing and addressing grounding acts. The likelihood of an utterance representing an addressing act appearing in the first turn $P(m_0 \in \mathcal{U}) = 0.12$ on WildChat, and 0.08 on Bing Chat. We observe a compounding of addressing grounding in human-assistant conversations. If we assume that the turn before was addressing, we find that the likelihood of *another* failed interaction triples: $P(m_1 \in \mathcal{U} \mid m_0 \in \mathcal{U}) = 0.30$ for WildChat and 0.32 for Bing Chat. A similar effect appears for advancing acts: early signs of effective conversational grounding similarly snowballs, with $P(m_0 \in \mathcal{S}) = 0.32 \rightarrow P(m_1 \in \mathcal{S} \mid m_0 \in \mathcal{S}) = 0.47$ on WildChat and $0.23 \rightarrow 0.44$ on Bing Chat.

5.2 Forecasting Users' Grounding Acts

In light of the compounding effect of grounding failures (Fig. 5), we pursue the possibility of low-

ering the probability of grounding failures byintervening *before* a failure occurs. So far, we have introduced a GPT-based annotator (from §5.1) that annotates conversations with grounding acts in a post-hoc fashion. However, our prompted annotator only identifies opportunities for grounding *after* they happen. Identifying or curating tasks where grounding failures emerge requires predicting the likelihood of future grounding patterns. To this end, we train a forecaster that predicts grounding acts in the next turn. We focus on WildChat as it is the only publicly availible human-LLM interaction dataset in our analysis.

Method. Following a user message m_i, our goal is to forecast grounding act g_{i+1} associated with the next user message m_{i+1}. We achieve this by repurposing *conditional training*. Concretely, we append each user message m_i with a forecasting grounding token g_{i+1} if the *future* turn m_{i+1} contains a grounding act. Consider the hypothetical training example below:

`User: Help me write this section` `addressing`
(m_0, g_1)
`Assistant: Sure, here's the section...`
(r_0)
`User: Wrong section. I meant 5.2.`
(m_1)

In the example above, we append a new forecasting token $g_1 =$ `addressing` after the user's initial message m_0, since the user's following turn is a repair. We use our validated grounding acts labeler (§5.1) to obtain all g_i, using the high-level grounding categories as labels (e.g. advancing, disambiguating, and addressing). We finetune Llama-3.1-8B on sequences of form $< \mathtt{m}_i \, , \, \mathtt{g}_{i+1} \, , \, \mathtt{r}_i \, , \, \mathtt{m}_{i+1} \, , \, \mathtt{g}_{i+2} \, , \, \mathtt{r}_{i+1} \, , \, \mathtt{m}_{i+2} \cdots >$ where r_i is the assistant response.

At inference time, we can provide any user message m_i and analyze the predicted likelihoods (i.e. logits) of our grounding tokens $g_{i+1} \sim P(\cdot|m_i)$ that are generated right after. For example:

`User: Help me with this` `addressing`
(prompt m_0, completion g_1)

This enables us to predict—from just a user query alone—the user's predicted grounding outcome *independent of the model's response!* This is an especially challenging learning problem: we effectively marginalize over all possible assistant responses. Successfully learning this model enables early intervention. All hyperparameters used in the training process are outlined in Appendix B. Beyond our finetuned forecaster, we also evaluate as a baseline few-shot prompted GPT-4o-mini (details in Appendix E.2).

Results. We find that our GPT-4o-mini few-shot baseline performs near-random, with Macro ROC AUC = 0.51. This result is not very surprising as forecasting is a challenging task: we must predict grounding acts g_{i+1} without directly observing assistant responses r_i. Our intuition is that it's much easier to look at an entire conversation and label the conversation for grounding acts post-hoc than it is to forecast a likely grounding act without being able to "see" future turns. Our finetuned forecaster, however, performs significantly better (0.61). Full experimental results are in Appendix B. In the next section, we draw representative samples from the forecaster, constructing a benchmark where users are (un)likely to initiate a grounding process.

6 RIFTS: A Grounding Benchmark

We showed that LLMs fail to generate grounding acts in two settings. They rarely:

- *Clarify* goals to reduce the rate at which a *user is likely to address grounding*.
- *Follow-up* to advance grounding, instead of relying on the *user to take initiative.*

To characterize these behaviors across multiple LLMs and evaluate interventions, we introduce a new benchmark, RIFTS. RIFTS consists of $\approx 1.8K$ tasks designed to test if LLMs can generate grounding acts when needed and withhold appropriately.

Dataset Details. RIFTS consists of a final combined set of 1740 tasks (split counts in Appdx. Table 8). Each task in RIFTS is a prompt drawn from an initial instruction in WildChat. Tasks are stratified based on how the user is predicted to continue the conversation: namely, are users predicted to advance grounding, address a failure, or disambiguate following an LLMs response (examples in Table 2). For these tasks, we would expect an LLM to take initiative—clarifying or following up appropriately. In addition, we include tasks where the user is expected to do none of the above, perhaps by switching the topic or ending the interaction altogether. Here, no grounding is required.

Category	Prompt from RIFTS Benchmark
Advancing	Write a Main heading about a brand name FFF Digital , which is a digital marketing agency Suggest a name for a technical blog consisting of five characters at most, which is compatible with SEO 1 week out from my powerlifting meet and i'm not prepared [...] what should i do? `[omitted context]`
Addressing	Blackburn rovers vs West Bromwich albion prediction I need to remove a heart `[snippets of code with no prompt]`
Disambiguating	What causes tailbone pain? My friend not want to help me, what to [do] with him? What happens when someone quits a job without having another one lined up?
No Grounding	convert rust String to clap::builder::Str Generate a full harvard references section for the following report: `[REPORT]` Join now, Supplier! or Supplier, Join us! which one is better?

Table 2: **Examples in RIFTS** fall into four categories. *Advance* tasks are collaborative (e.g., resume building, diet planning, writing), where following up is necessary. *Address* tasks are severely underspecified (e.g., contextless code snippets), or require capabilities LLMs lack (realtime information access)—these tasks need substantial clarification before any meaningful response is possible. *Disambiguating* tasks are less severe, but still need context clarification (e.g., medical queries, relationship advice) for an ideal response. *None* tasks are well-specified and factual, requiring no intervention. See §A.2 for a lexical analysis of tasks in RIFTS.

Curation Process. We construct RIFTS by filtering prompts from WildChat, using the predicted grounding act of the forecaster from §5.2. Forecaster predictions inform us on whether a *clarification* or *followup* action will be required in the conversation. In building RIFTS, we implicitly hypothesize that for some prompts m_i, regardless of what the LLM replies with (r_i), the user is in grounding trouble. In other words, if a user gives a query to a model that's so severely underspecified (e.g. m_i = "write me a resume"), it does not matter what any LLM responds with. The user must go back and forth to build common ground, since they never gave enough information in the initial prompt. RIFTS identifies this class of prompts, using the forecaster as a proxy.

We first filtered correctly predicted tasks from forecasters trained[4] on each split (train / val / test). For each grounding category, we then extracted the top 150 tasks with the highest likelihood of generating advancing, addressing, or ambiguous forecasting tokens. In other words, we repurpose our forecaster to curate representative tasks across each grounding act, sorting by the logit associated with each forecasting token. In addition, we sample 150 tasks that have a high likelihood of generating *no* forecasting token, capturing tasks that do not require initiative. Finally, we apply basic quality controls (see Appendix C).

[4]*Why train on each split?* To ensure fair evaluation, we create separate forecasters for train / val / test. This prevents leakage and enables researchers to develop interventions using the train forecaster before evaluating with the test forecaster.

Evaluating LLMs. RIFTS simplifies evaluation for any assistant model $P_{\text{assistant}}$. Consider the two failure modes where LLMs do not take initiative: *clarify* and *followup*. Given a task from RIFTS in the advancing category, we would prefer $P_{\text{assistant}}$ to proactively generate a followup. On the other hand, for tasks in addressing or disambiguating, we would expect $P_{\text{assistant}}$ to generate clarification questions. In instances where we forecast no act from the user, we do not wish to see the model inefficiently engage in grounding activity.

To evaluate performance, we take an initial instruction u_0 from RIFTS, and sample the next turn r_0 from $P_{\text{assistant}}(u_0)$. We then label r_0 with our validated grounding acts annotator (§5.1). To benchmark $P_{\text{assistant}}$, we evaluate if the generated response r_0 clarifies/follows-up when appropriate. Concretely, we instantiate our two failure modes (*clarify, followup*) in the following EVAL(u_0, r_0) function and report an overall accuracy score:

$$\begin{cases} 1_{r_0=\textit{follow-up}} & \text{if } u_0 \in \text{Advancing} \\ 1_{r_0=\textit{clarify}} & \text{if } u_0 \in \text{Addressing} \cup \text{disambig.} \\ 1_{r_0=\textit{neither}} & \text{if } u_0 \notin (\text{Addressing} \cup \text{dismbig.}) \end{cases}$$

Off-the-shelf models struggle. We evaluate a handful of open- and closed-source models on RIFTS' test set: OpenAI's GPT-4(o) series, Anthropic's Claude Sonnet 3.5 / Opus 3, and Llama 3.1 8 / 70B (Table 3). We find that *all* off-the-shelf instruction-following models (avg. 23.23% acc.) perform worse than random (33%). All of our

Model	Variant	RIFTS Accuracy
GPT	4o	25.26 ± 3.54
	4o-mini	24.48 ± 3.51
	o3-mini	25.26 ± 3.54
Claude	Sonnet 3.5	26.95 ± 3.57
	Opus 3	24.57 ± 3.51
Llama 3.1	8B Instruct	24.22 ± 3.49
	70B Instruct	23.88 ± 3.47
Llama 3.1	8B + GROUND	**54.48 ± 2.45**

Table 3: **Evaluating LLM grounding ability on RIFTS.** Frontier LLMs are ill-suited for grounding with humans on real-world tasks, with low accuracies across the board. A simple intervention (+ GROUND), based on our forecasters, can significantly improve LLM grounding (± indicates a 95% conf. interval).

evaluated LLMs perform near perfectly for tasks that require no grounding initiative (No Grounding category, 96.09%); this is unsurprising given instruction-following. However, LLMs fail to take appropriate initiative for any of the remaining categories (2.22% of the time, Table 7). Reasoning-tuned models don't help either: o3-mini regularly begins reasoning without verifying grounding.

A simple intervention. To improve grounding capabilities, we turn again to our forecasters (§5.2). Depending on the train forecaster's prediction, we can selectively add a prompt (+ GROUND) that instructs the LLM to ask follow-up questions *or* request clarification (prompts in Appendix E.3). Concretely, we append a clarification prompt if our forecaster predicts address or disambiguate; or a follow-up prompt if our forecaster predicts advance. With this intervention, Llama 3.1 8B outperforms all other models by at least 32%. Still, our intervention is far from perfect. RIFTS opens avenues for benchmarking new interventions, enabling easy evaluation of grounding capabilities in future work.

7 Discussion and Conclusion

We characterized (§3) and measured (§4-5) inadequate grounding in human-LLM interaction; and proposed a benchmark (§6) to assess this gap. Several directions emerge:

Should we expect grounding behavior from LLMs? Perhaps we should not be surprised that LLMs are unable to initiate grounding, defaulting instead to instruction-following. Models that are not trained to follow instructions are already biased towards instruction following behavior, likely because of the large presence of instruction following articles in pre-training mixes (Hewitt et al., 2024). In addition, limitations in theory of mind and other metacognitive challenges may restrict the ability of models to engage in grounding interactions (Sap et al., 2022; Ullman, 2023). Training methods must counteract these limitations and biases. Still, we see promise in future methods that elicit grounding capabilities from LLMs; and RIFTS can serve as a resource to test these methods.

Towards LLMs that initiate grounding. Decision-theoretic methods could guide when and how LLMs initiate grounding actions, based on inferred uncertainties in mutual understanding (see Horvitz (1999); Mozannar et al. (2024)). Instruction tuning could be revised to incorporate grounding, and our forecaster could serve as a reward model in RLHF (Ouyang et al., 2022). System prompts and dialogue management show promise, including prompts to disambiguate user intentions (Chen et al., 2023).

Benchmarking human-LLM grounding. Building models that ground effectively with humans across a range of tasks requires effective benchmarks. RIFTS supports comparative analyses, enabling discussion on grounding competencies of new LLMs and interventions.

Limitations

We considered grounding and engagement with Bing Chat in the absence of access to existing system meta prompts. System prompts can greatly shape the responses and provide specific guidance on the flow of dialogue. RIFTS was also collected by filtering WildChat using our forecaster; therefore, RIFTS will only reflect tasks seen in WildChat. In addition, tasks in RIFTS also depend on the LLMs used to serve WildChat (e.g. OpenAI LLMs). More specifically, our forecasters implicitly learn what tasks fail for the GPT models deployed in WildChat. Regardless, we observe that our final RIFTS tasks are challenging for all evaluated models. Finally, our annotator relies on GPT-4o-mini to label logs with grounding acts. While we did show that the annotator generally agrees with human judgment on a subset of the data (§5.1), the annotator is not perfect.

Ethics Statement

Enhancing an LLM's ability to generate grounding acts (by initiating clarification and follow-up actions) raises potential concerns around privacy, as these actions may lead users to disclose sensitive information unintentionally. Balancing the need for grounding with the careful collection of only relevant information remains a significant challenge and an area for future research. Moreover, while effective grounding can improve interaction quality, it can also be misused in harmful contexts. Although our work focuses on improving grounding for constructive purposes, such as assisting users, these techniques could be exploited for harmful ends (e.g., manipulation, persuasion, or coercion in sensitive areas like political targeting).

Finally, our description of human-LLM grounding *does not imply that LLMs possess genuine understanding.* Like prior work, we use grounding acts to describe interaction processes that help align human expectations with LLM-generated responses (Paek and Horvitz, 2003; Brennan, 2014; Shaikh et al., 2024). While human conversation involves active mutual comprehension, the same cannot be said of LLMs. The use of grounding terminology in this work is intended as a conceptual tool to analyze how LLMs facilitate or hinder effective communication, not as an anthropomorphic assertion that they share human-like cognitive capacities.

Acknowledgments

We appreciate the feedback from members of the Microsoft HAX team, the Stanford SALT Lab, Will Epperson, Michael Li, Jan-Philipp Fränken, Michael Ryan, Yanzhe Zhang, Qinan Yu, Shardul Sapkota and Ryan Li.

References

Gavin Abercrombie, Amanda Cercas Curry, Tanvi Dinkar, and Zeerak Talat. 2023. Mirages: On anthropomorphism in dialogue systems. *arXiv preprint arXiv:2305.09800*.

Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D Goodman. 2024. Star-gate: Teaching language models to ask clarifying questions. *arXiv preprint arXiv:2403.19154*.

Cristian-Paul Bara, CH-Wang Sky, and Joyce Chai. 2021. Mindcraft: Theory of mind modeling for situated dialogue in collaborative tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1112–1125.

Susan E Brennan. 2014. The grounding problem in conversations with and through computers. In *Social and cognitive approaches to interpersonal communication*, pages 201–225. Psychology Press.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Ultes Stefan, Ramadan Osman, 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 (EMNLP)*.

Harry Bunt, Volha Petukhova, David Traum, and Jan Alexandersson. 2017. Dialogue act annotation with the iso 24617-2 standard. *Multimodal Interaction with W3C Standards: Toward Natural User Interfaces to Everything*, pages 109–135.

Maximillian Chen, Xiao Yu, Weiyan Shi, Urvi Awasthi, and Zhou Yu. 2023. Controllable mixed-initiative dialogue generation through prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 951–966, Toronto, Canada. Association for Computational Linguistics.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

Hyundong Cho and Jonathan May. 2020. Grounding conversations with improvised dialogues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2398–2413, Online. Association for Computational Linguistics.

Jennifer Chu-Carroll and Sandra Carberry. 1998. Collaborative response generation in planning dialogues. *Computational Linguistics*, 24(3):355–400.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Herbert H Clark and Edward F Schaefer. 1989. Contributing to discourse. *Cognitive science*, 13(2):259–294.

Deborah Davis. 1982. Determinants of responsiveness in dyadic interaction. *Personality, roles, and social behavior*, pages 85–139.

Doug Downey, Susan T Dumais, and Eric Horvitz. 2007. Models of searching and browsing: Languages, studies, and application. In *IJCAI*, volume 7, pages 2740–2747.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyag Gao, and Dilek Hakkani-Tur. 2019. Multiwoz 2.1: Multi-domain dialogue

state corrections and state tracking baselines. *arXiv preprint arXiv:1907.01669*.

Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2022. Pragmatics in grounded language learning: Phenomena, tasks, and modeling approaches. *arXiv preprint arXiv:2211.08371*.

Yujian Gan, Changling Li, Jinxia Xie, Luou Wen, Matthew Purver, and Massimo Poesio. 2024. Clarq-llm: A benchmark for models clarifying and requesting information in task-oriented dialog. *arXiv preprint arXiv:2409.06097*.

Yifan Gao, Henghui Zhu, Patrick Ng, Cicero dos Santos, Zhiguo Wang, Feng Nan, Dejiao Zhang, Ramesh Nallapati, Andrew O Arnold, and Bing Xiang. 2021. Answering ambiguous questions through generative evidence fusion and round-trip prediction. 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 3263–3276.

Jonathan Ginzburg and Robin Cooper. 2001. Resolving ellipsis in clarification. In *39th Annual Meeting of the Association-for-Computational-Linguistics*, pages 236–243. Association for Computational Linguistics.

Yoni Gottesman. 2024. Mask your user tokens. Accessed: 2024-08-23.

Arthur C Graesser, Natalie K Person, and Joseph P Magliano. 1995. Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied cognitive psychology*, 9(6):495–522.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Patrick GT Healey, Arash Eshghi, Christine Howes, and Matthew Purver. 2011. Making a contribution: Processing clarification requests in dialogue. In *Proceedings of the 21st Annual Meeting of the Society for Text and Discourse*, pages 11–13. Citeseer.

Patrick GT Healey, Matthew Purver, James King, Jonathan Ginzburg, and Greg J Mills. 2003. Experimenting with clarification in dialogue. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 25.

John Hewitt, Nelson F Liu, Percy Liang, and Christopher D Manning. 2024. Instruction following without instruction tuning. *arXiv preprint arXiv:2409.14254*.

Joey Hong, Sergey Levine, and Anca Dragan. 2023. Zero-shot goal-directed dialogue via rl on imagined conversations. *arXiv preprint arXiv:2311.05584*.

Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 159–166.

Eric Horvitz and Tim Paek. 2000a. Conversation as action under uncertainty. In *Proceedings of Uncertainty in Artificial Intelligence*. AUAI.

Eric Horvitz and Tim Paek. 2000b. Deeplistener: harnessing expected utility to guide clarification dialog in spoken language systems. In *Sixth International Conference on Spoken Language Processing (ICSLP)*, pages 226–229. Interspeech.

Eric Horvitz and Tim Paek. 2007. Complementary computing: policies for transferring callers from dialog systems to human receptionists. *User Modeling and User-Adapted Interaction*, 17:159–182.

Mathew Huerta-Enochian and Seung Yong Ko. 2024. Instruction fine-tuning: Does prompt loss matter? *Preprint*, arXiv:2401.13586.

Gail Jefferson. 1972. Side sequences. In D.N. Sudnow, editor, *Studies in social interaction*, pages 294–333. Free Press, New York.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2022. Clam: Selective clarification for ambiguous questions with large language models. *arXiv preprint arXiv:2212.07769*.

Tessa Lau and Eric Horvitz. 1999. Patterns of search: Analyzing and modeling web query refinement. In *UM99 User Modeling: Proceedings of the Seventh International Conference*, pages 119–128. Springer.

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, et al. 2022. Evaluating human-language model interaction. *arXiv preprint arXiv:2212.09746*.

Stephen C Levinson, Stephen C Levinson, and S Levinson. 1983. *Pragmatics*. Cambridge university press.

Belinda Z. Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. 2023. Eliciting human preferences with language models. *Preprint*, arXiv:2310.11589.

Xiaolong Li and Kristy Boyer. 2015. Semantic grounding in dialogue for complex problem solving. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 841–850, Denver, Colorado. Association for Computational Linguistics.

Jessy Lin, Nicholas Tomlin, Jacob Andreas, and Jason Eisner. 2023. Decision-oriented dialogue for human-ai collaboration. *arXiv preprint arXiv:2305.20076*.

Diane J. Litman. 1985. *Plan Recognition and Discourse Analysis: An Integrated Approach for Understanding Dialogues*. Ph.D. thesis, University of Rochester, Rochester, NY.

Diane J. Litman and James Allen. 1987. A plan recognition model for subdialogues in conversation. *Cognitive Science*, 11:163–200.

Karen E. Lochbaum. 1998. A collaborative planning model of intentional structure. *Computational Linguistics*, 24(4):525–572.

Michael Lu, Hyundong Justin Cho, Weiyan Shi, Jonathan May, and Alexander Spangher. 2024. Newsinterview: a dataset and a playground to evaluate llms' ground gap via informational interviews. *arXiv preprint arXiv:2411.13779*.

Xing Han Lu, Siva Reddy, and Harm de Vries. 2023. The StatCan dialogue dataset: Retrieving data tables through conversations with genuine intents. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2799–2829, Dubrovnik, Croatia. Association for Computational Linguistics.

Brielen Madureira and David Schlangen. 2023a. " are you telling me to put glasses on the dog?"contentgrounded annotation of instruction clarification requests in the codraw dataset. *arXiv preprint arXiv:2306.02377*.

Brielen Madureira and David Schlangen. 2023b. Instruction clarification requests in multimodal collaborative dialogue games: Tasks, and an analysis of the codraw dataset. *arXiv preprint arXiv:2302.14406*.

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.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797.

Shrestha Mohanty, Negar Arabzadeh, Julia Kiseleva, Artem Zholus, Milagro Teruel, Ahmed Awadallah, Yuxuan Sun, Kavya Srinet, and Arthur Szlam. 2023. Transforming human-centered ai collaboration: Redefining embodied agents capabilities through interactive grounded language instructions. *arXiv preprint arXiv:2305.10783*.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. 2024. When to show a suggestion? integrating human feedback in ai-assisted programming. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10137–10144.

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

Tim Paek and Eric Horvitz. 2003. On the utility of decision-theoretic hidden subdialog. In *ISCA Workshop on Error Handling in Spoken Dialogue Systems*.

Ashwin Paranjape and Christopher D Manning. 2021. Human-like informative conversations: Better acknowledgements using conditional mutual information. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 768–781.

Soya Park and Chinmay Kulkarni. 2023. Thinking assistants: Llm-based conversational assistants that help users think by asking rather than answering. *arXiv preprint arXiv:2312.06024*.

Tomislav Pejsa, Dan Bohus, Michael F Cohen, Chit W Saw, James Mahoney, and Eric Horvitz. 2014. Natural communication about uncertainties in situated interaction. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 283–290.

Matthew Purver, Jonathan Ginzburg, and Patrick Healey. 2003a. On the means for clarification in dialogue. *Current and new directions in discourse and dialogue*, pages 235–255.

Matthew Purver, Patrick Healey, James King, Jonathan Ginzburg, and Greg J Mills. 2003b. Answering clarification questions. In *Proceedings of the Fourth SIGdial Workshop of Discourse and Dialogue*, pages 23–33.

Matthew Richard John Purver. 2004. *The theory and use of clarification requests in dialogue*. Ph.D. thesis, University of London.

Filip Radlinski and Thorsten Joachims. 2005. Query chains: learning to rank from implicit feedback. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 239–248.

Hossein A Rahmani, Xi Wang, Yue Feng, Qiang Zhang, Emine Yilmaz, and Aldo Lipani. 2023. A survey on asking clarification questions datasets in conversational systems. *arXiv preprint arXiv:2305.15933*.

Osman Ramadan, Paweł Budzianowski, and Milica Gasic. 2018. Large-scale multi-domain belief tracking with knowledge sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 432–437.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*, pages 7–55. Elsevier.

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312*.

Emanuel A Schegloff. 1982. Discourse as an interactional achievement: Some uses of 'uh huh'and other things that come between sentences. *Analyzing discourse: Text and talk*, 71:71–93.

Emanuel A Schegloff. 1992. Repair after next turn: The last structurally provided defense of intersubjectivity in conversation. *American journal of sociology*, 97(5):1295–1345.

Emanuel A Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382.

Deborah Schiffrin. 1987. *Discourse markers*. 5. Cambridge University Press.

Phillip Schneider, Nektarios Machner, Kristiina Jokinen, and Florian Matthes. 2024. Bridging information gaps in dialogues with grounded exchanges using knowledge graphs. *arXiv preprint arXiv:2408.01088*.

Omar Shaikh, Kristina Gligorić, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2024. Grounding gaps in language model generations. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6279–6296.

Omar Shaikh, Caleb Ziems, William Held, Aryan Pariani, Fred Morstatter, and Diyi Yang. 2023. Modeling cross-cultural pragmatic inference with codenames duet. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6550–6569, Toronto, Canada. Association for Computational Linguistics.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Zhengyan Shi, Adam X Yang, Bin Wu, Laurence Aitchison, Emine Yilmaz, and Aldo Lipani. 2024. Instruction tuning with loss over instructions. *arXiv preprint arXiv:2405.14394*.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2023. A long way to go: Investigating length correlations in rlhf. *arXiv preprint arXiv:2310.03716*.

Robert Stalnaker. 2002. Common ground. *Linguistics and philosophy*, 25(5/6):701–721.

Svetlana Stoyanchev, Alex Liu, and Julia Hirschberg. 2013. Modelling human clarification strategies. pages 137–141.

Alex Tamkin, Kunal Handa, Avash Shrestha, and Noah Goodman. 2022. Task ambiguity in humans and language models. *arXiv preprint arXiv:2212.10711*.

David R Traum and Elizabeth A Hinkelman. 1992. Conversation acts in task-oriented spoken dialogue. *Computational intelligence*, 8(3):575–599.

Polina Tsvilodub, Michael Franke, Robert D Hawkins, and Noah D Goodman. 2023. Overinformative question answering by humans and machines. *arXiv preprint arXiv:2305.07151*.

Tomer Ullman. 2023. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*.

Priyan Vaithilingam, Ian Arawjo, and Elena L Glassman. 2024. Imagining a future of designing with ai: Dynamic grounding, constructive negotiation, and sustainable motivation. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, pages 289–300.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI, ACL 2020*, pages 109–117.

Michael JQ Zhang and Eunsol Choi. 2023. Clarify when necessary: Resolving ambiguity through interaction with lms. *arXiv preprint arXiv:2311.09469*.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*.

Pei Zhou, Hyundong Cho, Pegah Jandaghi, Dong-Ho Lee, Bill Yuchen Lin, Jay Pujara, and Xiang Ren. 2022. Reflect, not reflex: Inference-based common ground improves dialogue response quality. *arXiv preprint arXiv:2211.09267*.

A Dataset Splits and Details

A.1 Full Datasets

Table 4 outlines the number of dialogues and messages in each of our final filtered datasets.

Dataset	Number of Messages	Number of Dialogues
WildChat	110688	55344
Bing Chat	26200	13100
MultiWOZ	980	490

Table 4: Number of messages and dialogues across our three analyzed datasets.

A.2 Descriptive Analysis

To quantify lexical differences between tasks in RIFTS and general instructions, we fit a Fightin' Words model between each cataegory and the full corpora (Monroe et al., 2008). Fightin' Words reveals words that are associated with each particular text distribution, producing a log-odds ratio and a corresponding z-score. We select a sample of significant words that characterize each category in Table 5.

Label	**Words (z-scores)**
No Action	events (6.07), params (4.84), worksheets (3.78), quotation (3.42), answers (2.11), exam (2.03)
Advancing	stock (5.24), dividend (3.77), regression (3.60), parents (3.23), investment (2.97), podcast (2.93)
Addressing	sort (7.76), point (7.51), https (5.80), var (4.75), scan (3.45), merge (3.25), array (3.16)
Disambiguating	bitcoin (4.66), chatbot (4.24), cryptocurrency (4.01), unauthorized (2.39), beginners (2.19), friends (2.02)

Table 5: Lexical cues from RIFTS reveal distinct task characteristics. **No Action** tasks involve users who are often trying to simply get answers for homework questions (e.g., worksheets), needing no follow-up. **Advancing** tasks (e.g., "stock," "dividend") imply iterative interaction, as in investment management. **Addressing** tasks feature technical, underspecified language (e.g., "sort," "array") that requires extra context—users often submit code with no explicit task. Finally, **Disambiguating** tasks (e.g., "bitcoin") indicate a need for clarification on topics with inherent uncertainty: bitcoin, for example, is volatile, and beginners often have to clarify when learning.

B Training Forecasters

Forecasting grounding is a challenging task; we are trying to predict if a user will have trouble for a task *without* observing an LLMs generation. In other words, we can only use the task to predict future grounding patterns. We trained all models for 5 epochs (picked using the validation set), with learning rate 5e-5, batch size of 1, and 16 gradient accumulation steps (e.g. effective batch size of 16). All training occurred on an H100 80GB GPU. Below, we outline training optimizations that helped improve forecasting performance.

B.1 Subsample Grounding Acts

We subsampled WildChat data to include equal amounts of each forecasted grounding act before training. Inequal splits would result in the forecaster always predicting the majority class. To build our train/val/test splits, we sampled the maximum number of tasks (1630) possible from each of our forecaster categories (1640×4 for `fix, followup, continue, end`) while ensuring that each task was equally represented.

B.2 Don't Mask User Tokens

In addition to adding forecasting tokens, we make a modification to standard LLM finetuning/inference practices: we do not mask user utterances in the loss, training on user input. While the general effects of masking user tokens are mixed (Huerta-Enochian and Ko, 2024; Shi et al., 2024; Gottesman, 2024), our setting requires the modeling of user input, as we seek to understand and assist with user grounding. Because we do not mask user tokens, we can additionally simulate user inputs with past interaction data.

B.3 Reweight Control Tokens At Train Time

We seek to encourage our model to learn our added forecasting tokens alongside the language modeling objective. The standard MLE objective optimizes a model's parameters θ with respect to a sequence x: $\mathcal{L}(\theta) = -\sum_{t=1}^{T} \log p_\theta(x_t|x_{<t})$. However, a subsequence $x_{s...e}$ consists of forecasting tokens, which we want to emphasize—especially since these tokens did not undergo pretraining. At training time, we upweight these tokens by $\lambda = 2$. Our final loss is below:

$$\mathcal{L}(\theta) = -\sum_{t=1}^{T} \begin{cases} \lambda \cdot \log p_\theta(x_t|x_{<t}), & \text{if } s \leq t \leq e \\ \log p_\theta(x_t|x_{<t}), & \text{otherwise} \end{cases}$$

B.4 Experimental Results

	Few-shot GPT-4o-mini	Llama 3.1 FT
Followup	0.52	0.61
Fix	0.49	0.60
Next Turn	0.52	0.67
End	0.51	0.57
Macro	0.51	0.61

Table 6: **Forecasting performance.** Per-label and Macro AUROC for forecasting task on the WildChat test set, conditioned on the initial prompt.

In Table 6 we show the performance of our fine-tuned Llama 3.1 model compared to a few-show prompted GPT-4o-mini at forecasting grounding acts. We note the per-label and macro AUROC on the WildChat test set.

C Filtering Criterion

While we sample tasks from the forecaster tails to construct RIFTS, we manually filter out tasks that ask for explicit content generation or ask the LLM for API keys, gift card codes, etc. Additionally, we passed tasks through the OpenAI moderation API, and filter out flagged tasks.

D RIFTS

Model	**Addressing (%)**	**Advancing (%)**	**No Grounding (%)**	**Disambiguating (%)**
o3-mini	4.14	1.35	90.65	8.22
gpt-4o-mini	2.07	0.68	96.40	2.07
gpt-4o	2.76	1.35	98.56	2.05
claude-3-opus	1.38	1.35	96.40	2.74
claude-3-5-sonnet	2.76	2.03	97.84	4.79
Meta-Llama-3.1-8B	0.69	2.03	96.40	1.37
Meta-Llama-3.1-70B	0.00	2.03	96.40	0.68
Average	1.97	1.55	96.09	3.13

Table 7: Per-label accuracies for various models on RIFTS. Most models correctly withhold initiation for tasks that require no grounding. However, this comes at a cost: models struggle at taking initiative for all other categories.

D.1 More Tasks

We include a handful of tasks from Rifts directly in the paper (Table 2). Our full benchmark and set of tasks can be found at our anonymous repository link, under the rifts folder: https://anonymous.4open.science/r/rifts-B7E4/

E Prompted Models

All of our prompts are located in the prompts folder in the following anonymous repository: https://anonymous.4open.science/r/rifts-B7E4/. We detail each prompt used in our analysis below:

Category	Train	Val	Test
Advancing	147	142	148
Addressing	144	143	145
Disambiguating	146	147	146
No Grounding	146	147	139

Table 8: **RIFTS** evaluation splits across 1740 total tasks.

Grounding Act	**Support**	**4o-mini few-shot (F1 Score)**
Next Turn	38	0.84
Acknowledge	4	0.67
Follow-up	17	0.80
Overcontinue	21	0.76
Clarify	7	0.67
Repair	10	0.74
Reformulate	11	0.78
Macro	108	**0.75**

Table 9: F-1 for Human-LLM grounding acts classification on a withheld test set of 30 conversations from our selected dialogue datasets. In the few-shot setting, `GPT-4o-mini` has fairly high F-1.

E.1 Grounding Acts Labeling Prompt

We construct a few-shot prompt to annotate grounding acts across our datasets. The first author prompt-engineered a prompt on a small validation set. Our full prompt is availible in the anonymous repo.

E.2 GPT Forecaster Baseline Prompt

Alongside our finetuned Llama forecaster, we test a prompted baseline. We provide our prompted baseline with a few shot (task, future grounding act) pairs sampled from the forecaster train set. Our full prompt is in the anonymous repo under the prompts folder.

E.3 Intervention Prompt

Our GROUND intervention relies on two prompts, both in the anonymous repo under the intervention folder. Specifically, we construct a follow-up prompt and a clarification prompt. Both prompts directly instruct the LLM to generate a clarification question or generate the answer + a followup. Our intervention is *dumb by design*—the forecaster decides when to employ a static prompt. In instances where our train forecaster predicts address, we enable the clarification prompt. Similarly, when our forecaster predicts advance, we employ the followup prompt. Given the improvements we see with our intervention, we expect that models trained to initiate grounding acts will substantially improve on RIFTS.

F License Information

The Multiwoz (Eric et al., 2019) dataset we analyze has the MIT license, the license file is available at [5].

The Wildchat (Zhao et al., 2024) dataset has the ODC-By license, license information is available here [6]. By consequence RIFTS is also released under the ODC-By license.

We rely on the Llama 3.1 models, the license for those models is available here [7].

We obtained permission from the ethics review board to analyze the Bing Chat data logs and release the analysis in this paper.

[5] https://github.com/budzianowski/multiwoz/blob/master/LICENSE
[6] https://huggingface.co/datasets/allenai/WildChat/blob/main/LICENSE.md
[7] https://www.llama.com/llama3_1/license/