I Cast Detect Thoughts: Learning to Converse and Guide with Intents and Theory-of-Mind in Dungeons and Dragons

Pei Zhou[♡♠] **Andrew Zhu**[♣] **Jennifer Hu**[†] **Jay Pujara**[♠]
Xiang Ren[♡♠] **Chris Callison-Burch**[♣♡] **Yejin Choi**[◇♡] **Prithviraj Ammanabrolu**[♡◇]

♡ Allen Institute for Artificial Intelligence ♠ University of Southern California
◇ University of Washington ♣ University of Pennsylvania † MIT

peiz@usc.edu raja@allenai.org

Abstract

We propose a novel task, G4C, to study *teacher-student* natural language interactions in a *goal-driven* and *grounded* environment. Dungeons and Dragons (D&D), a role-playing game, provides an ideal setting to investigate such interactions. Here, the Dungeon Master (DM), *i.e.*, the teacher, guides the actions of several players—students, each with their own personas and abilities—to achieve shared goals grounded in a fantasy world. Our approach is to decompose and model these interactions into (1) the DM's *intent* to guide players towards a given goal; (2) the DM's *guidance* utterance to the players expressing this intent; and (3) a *theory-of-mind* (ToM) model that anticipates the players' reaction to the guidance one turn into the future. We develop a novel reinforcement learning (RL) method for training a DM that generates guidance for players by rewarding utterances where the intent matches the ToM-anticipated player actions. Human and automated evaluations show that a DM trained to explicitly model intents and incorporate ToM of the players using RL generates better-quality guidance that is 3x more likely to fulfill the DM's intent than a vanilla natural language generation (NLG) approach.

1 Introduction

Humans communicate with a *goal* in mind and use language to reach the goal by interacting with their communication partners *grounded* in a shared environment (Grice, 1975; Allwood, 1976; Clark and Schaefer, 1989; Clark and Brennan, 1991). To make sure the goal is reached, we often anticipate how the partners will respond in advance to steer the conversations in the desired direction. This ability to reason about the mental states of conversation partners – *i.e.*, theory-of-mind (ToM; Premack and Woodruff, 1978) – is key to smooth and efficient communication (Perner et al., 1989; Happé, 1993). Most existing dialogue agents, while able to produce human-like responses, often do not model

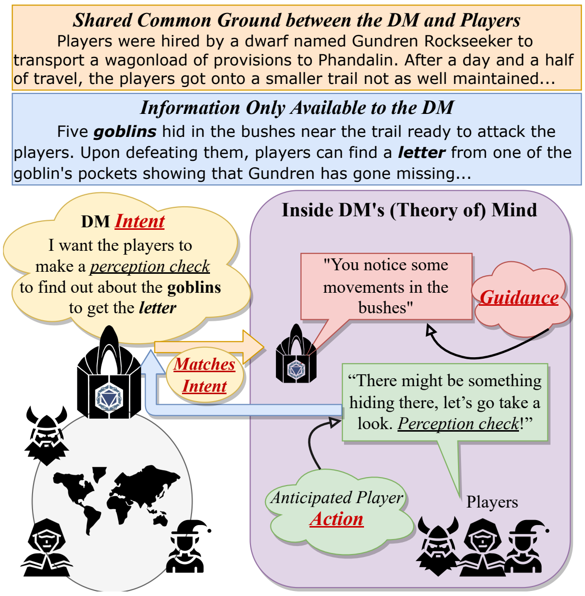

Figure 1: **A motivating example.** The (human) Dungeon Master (DM), knowing the desired story path, intends the players to perform actions to find out about the goblins—the first plot point that will eventually lead the players to treasure. They generate the guidance "*You notice some movements in the bushes*" using theory-of-mind by inferring that the players will perform the desired actions upon hearing their words.

communicative intents or ToM *explicitly*. In this paper, we investigate if models benefit from explicitly incorporating intents and ToM in NLG.

To bridge the gap between human communication and existing dialogue models, we propose a new task G4C: **G**enerating **G**uidance in **G**oal-Driven and **G**rounded **C**ommunication. G4C considers three building blocks: *intent*, *guidance*, and *action*. The task envisions a teacher with intent for specific student action, guidance uttered by the teacher, and action undertaken by the student based on the guidance and common ground. G4C evaluates the ability of a teacher to provide intentional *guidance* that results in intended student actions.[1] The success of the teacher's guidance depends on whether the student's subsequent action matches

[1] Here we use *actions* to indicate any linguistic behavior with intention (Allwood, 1976).

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 11136–11155
July 9-14, 2023 ©2023 Association for Computational Linguistics

the teacher's *intended* action. Using this task formulation, we analyze if the teacher has fulfilled their communicative intents explicitly by examining what the student says afterward. G4C further requires the dialogue to be grounded, meaning that both the teacher and the student are communicating with a shared environment and background.

To train models to perform G4C, we use Dungeons and Dragons (D&D) as our environment, a game that heavily relies on communication that is inherently goal-driven and grounded. D&D is a role-playing game consisting of multiple player characters and a Dungeon Master (DM) who collaborate to achieve a set of goals beneficial to the players. The DM, the narrator and host of the game, has an innate motivation to guide the players to perform a series of actions that roughly follow a pre-devised storyline culminating in a global goal, all grounded in a shared fantasy world. An example of each component of G4C in the D&D environment (*intent*, *guidance*, and *action*) is shown in Figure 1.

We construct 47k D&D dialogues from transcripts collected by Callison-Burch et al. (2022). Motivated by the critical roles *intents* and *theory-of-mind (ToM)* play in human communication, we study the following central research question: "*Does incorporating **intent** and **ToM** make computational models better communicators?*" Accordingly, we explore different methods for modeling **intent** and **ToM** for G4C in Section 3. Specifically, we make the intents of the teacher (DM) explicit by mining intents from large language models (LLM) and appending them as additional context to guide generation. We further propose a method to train a DM to generate guidance for a player with RL inspired by ToM. The DM first predicts in advance what action the player will take in reaction to the guidance and then uses this prediction as a feedback reward function to check whether the predicted action matches DM intent.

G4C focuses on mimicking human communication that is goal-driven and coherent to a grounded narrative, which current automated dialogue metrics do not capture well. As such, we further propose novel human and automated evaluation metrics to measure whether the output fits in the grounded context and fulfills communicative goals. Our experiments show that DMs trained with explicit intents and ToM to predict how their players will react to their utterances ahead of time *triples* the number of responses generated that are both grounded and fulfill the communicative intent.

Character	Game Dialogue
DM	A dwarf named Gundren Rockseeker has hired you to transport a wagonload of provisions to the rough-and-tumble settlement of Phandalin... **You all notice some movements in the bushes nearby the road...**
Clint	"There might be something hiding there, let's go take a look." Clint makes a **perception check**. 16
Vi	I'll help as well. I got a 10
DM	Clint, you notice a few goblins crouching in a part of the shaded woods off to the side of the road...

Table 1: Example dialogue transcript from D&D game play.

2 G4C and G-Dragon

Here we discuss how we construct the environment for the proposed G4C task using a dataset of dialogues from Dungeons and Dragons (D&D) called G-Dragon. We start with formulating the G4C task, then introduce the D&D data, and finally present our procedure of constructing the environment using large-scale data.

2.1 G4C Task

Consider three variables in communication between a teacher and a student: *context* $\mathcal{C}$, *teacher utterance* $\mathcal{T}$, and the subsequent *student utterance* $\mathcal{S}$. In standard dialogue response generation (RG) setup, models are trained to generate the next utterance only based on the *previous* dialogue history, *i.e.*, $P(\mathcal{T}|\mathcal{C})$ for teacher and $P(\mathcal{S}|\mathcal{C},\mathcal{T})$ for the student. In our task setting, we further consider one variable: *intents* of the teacher: $\mathcal{I}_\mathcal{T}$.[2] In G4C, we assume that the teacher's intents are to guide the student to perform certain *action* $\mathcal{A}$ and the intents are fulfilled if the student's subsequent utterance $\mathcal{S}$ entails $\mathcal{A}$. Since we focus on verbal communication, all variables including $\mathcal{I}_\mathcal{T}$ and $\mathcal{A}$ are in natural language (NL). The teacher model's goal is thus to first come up with an intent, *i.e.*, $P(\mathcal{I}_\mathcal{T}|\mathcal{C})$ and then generate an utterance that helps achieve the intent, *i.e.*, $P(\mathcal{T}|\mathcal{C},\mathcal{I}_\mathcal{T})$ such that $\mathcal{S} \approx \mathcal{A}$, given student model $P(\mathcal{S}|\mathcal{C},\mathcal{T})$.

2.2 D&D Dialogue Generation as a Partially Observable Markov Decision Process

Here we discuss a reformulation of the standard RG problem as a partially observable Markov decision process (POMDP). We consider a POMDP defined as $\langle S, A, T, R, O \rangle$, where S is a set of states, A is

[2]Students also have intents, which are not explicitly modeled in this work.

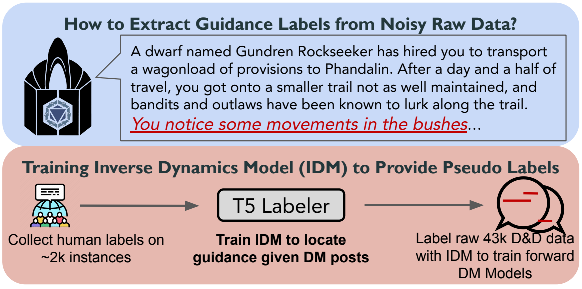

Figure 2: Illustration of IDM. We collect 2.5k human labels on guidance and train an IDM labeler to generate pseudo labels for unlabeled large corpus.

a set of actions performed by the teacher (note it is different from the player action $\mathcal{A}$), T is a set of transition probabilities between states ($T(s'|s, a)$), R is reward function, and O is a set of observations. In D&D dialogues such as Table 1, we consider the first DM sentence (not in bold) as the *observation* containing an incomplete description of world *state*, the second sentence in bold as the *action* containing guidance for players, the next player turns as *reward* (in this case players' perception check[3] matches DM intent), and the final turn as new *observation*.

2.3 Play-By-Post D&D Data

As introduced in Sec. 1, D&D satisfies two crucial aspects we investigate in G4C: *goal-driven* (players are motivated to finish quests guided by the DM) and *groundedness* (players and DM are co-located in the environment and narratives). Furthermore, the DM is constantly providing guidance to other players, matching the *teacher* role in G4C. We use *actual play* game transcript dataset from Callison-Burch et al. (2022) scraped from Play-By-Post (PBP), a web forum[4] where people play D&D by taking turns posting on the forum. PBP data contains more than 800k turns with around 58M words, annotated heuristically with game state information such as player class, race, and ability checks. However, to adapt this dataset to our G4C setting, we need to filter the data to focus on interactions of DM guiding players. Details are in Appendix B.

2.4 Creating the Environment

Training a DM to generate guidance using G4C formulation requires first identifying which part of DM's utterances contains guidance, as the DM also roleplays other characters, chitchat, or discusses rules. Creating such labels requires human-in-the-loop data collection or large offline labeled datasets, both of which are heavily resource intensive (Fu et al., 2020). To mitigate such resource constraints, we collect human labels on a small ($< 5\%$) portion of our dataset and then train an inverse dynamics model (IDM) that given the players' reactions (*reward* R) after potential DM guidance (*action* A), extracts which portions of the DM's utterance contain guidance (Figure 2).

Given that we cast the dialogue generation in G4C as a POMDP, the *forward* modeling problem is to generate guidance so that the player's feedback is as intended, such as *making a perception check*. Thus our *inverse* modeling problem can be formulated as given the next player ability check being *perception check* (feedback/reward), extracting the guiding sentence (DM's action) from DMs' utterances. IDM modeling is simpler than forward behavior cloning because it uses a non-causal formulation that exploits both past and future events to identify a guidance sentence (Baker et al., 2022).

Human Label Collection. We design our human labeling interface to contain 3 questions: 1. *Does this DM turn contain guidance or not?* 2. *If it does, please choose a sentence from the text that serves the purpose of guidance the most.* 3. *Imagine that you were the player, what ability check would you make?* We add the third question to provide more labels to evaluate DM models (discussed in Section 4.3). Details are in Appendix D.

IDM Training. In practice, we collect around 2.5k human labels on guidance and train IDM to provide labels for the large unlabeled data. We consider two subtasks for IDM: *identifying* whether a DM turn (DT) contains guidance and *extracting* the key guiding sentence (GS) from DT. We train two T5-3B models (Raffel et al., 2020), one for classifying DM texts that contain guidance or not (*IDM-Identify*) and the other for extracting a sentence from the text (*IDM-Extract*). More details can be found in Appendix C.

IDM Results. We evaluate IDM performance on 1k human-labeled data and compare it to baselines such as the longest sentence and GPT-3 with in-context learning. Detailed results are in Appendix C. In summary, we find that trained IDM outperforms other baselines on extracting GS, reaching around 70% accuracy where random guessing is 10% (the average number of sentences

[3] *Ability check* is a game mechanic that models the stochasticity in D&D. The player needs to roll a die and the number determines whether the action succeeds or not.

[4] `https://www.dndbeyond.com/forums/d-d-beyond-general/play-by-post`

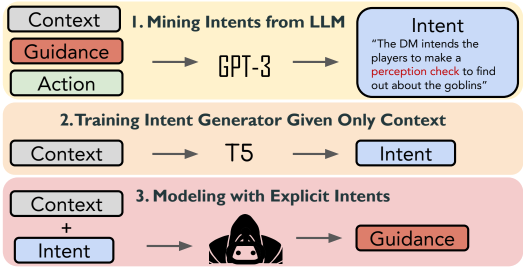

Figure 3: Illustration of intent modeling. We first mine intents from LLM and then train an intent generator to generate intent as additional context to train the DM model.

in DM's posts is around 10).

3 Theory-of-Mind Inspired Guidance Generation in Grounded Environments

This section introduces our exploration of model designs to train a teacher model that can guide the student to perform certain actions by speaking in a grounded environment. We are specifically interested in the research question "*Does incorporating **intent** (3.1) and **theory-of-mind** (3.2) help models generate better guidance?*"

3.1 Modeling Intents

***Implicit* Intent.** We start with the standard RG setup in most dialogue modeling work: training models to directly generate the target utterance (guidance) given dialogue context with no explicit intent involved. Formally, we model $P(\mathcal{T}|\mathcal{C})$ using the DM text with guidance as teacher target utterance $\mathcal{T}$ and the context turns as $\mathcal{C}$.

***Explicit* Intent with Generator.** Here we propose modeling methods that include explicit intents of the teacher $\mathcal{I}_T$. Following 2.1, we treat the teacher's intents as additional context appended to the dialogue context, *i.e.*, $P(\mathcal{T}|\mathcal{C}, \mathcal{I}_T)$. Figure 3 shows the procedure. **1. Mining Intents Using Large Language Models (LLMs)** Since intents are implicit in the data, we first need to mine DM's intents from their utterances. To ensure the quality of mined intents, we use LLM such as GPT-3 to generate intents in natural language given context, guidance sentence from DM, and the next-turn player action. We prompt GPT-3[5] with "*The following is a conversation that happened in a game of Dungeons and Dragons: [Context] [DM Text] [Player Name]:[Player Ability Check] Question: What do you think that the DM intentds to do by mentioning [Extracted Guiding Sentence]? Answer:*" **2. Training Intent Generator** Using mined intents, we train an *intent generator* (IG) that takes the context $\mathcal{C}$ as input and generates an output of the DM's potential intent $\mathcal{I}_T$. In practice, we train a sequence-to-sequence model T5 (Raffel et al., 2020) on 45k mined intents for our training and valid data. We also conduct a human evaluation on both mined and generated intents to examine whether these intents are reasonable given the context. Humans rate 85% of the mined intents and 75% of generated intents proper with 3-way redundancy of each intent from sampled 500 intents. **3. Modeling with Generated Intent** With a trained IG, we then generate intents on our test split. Then the teacher model that takes intents as additional input will use the generated intents from IG to generate utterances during testing.

[5]We use text-davinci-03 from `https://beta.openai.com/docs/models/gpt-3`

3.2 Modeling (Limited) *Theory-of-Mind* (ToM) Using RL for Guidance Generation

Background and Intuition. Here we model a limited scope of ToM by modeling the anticipated action of the players in order to help the teacher to generate utterances that guide students to fulfill the teacher's intents. Specifically, in G-DRAGON, the DM infers what the players might do when they provide different guidance. For example, "*you notice some movements in the bushes*" will likely motivate the players to make a *perception* check while "*the guard seems a bit shaken to hear your words*" might prompt the players to make a *persuasion* check. DM then chooses the guidance that will more likely prompts players to perform the action that fulfills the goal.

Training Player Model. The first step of our proposed ToM-teacher is to train a *player model* (PM) that takes in context and DM utterances and outputs the most likely player action (ability check), *i.e.*, $P(\mathcal{A}|\mathcal{C}, \mathcal{T})$. Luckily, each instance of our G-DRAGON data naturally contains training data for PM with the DM turn and next-turn player ability check. We also train a sequence-to-sequence model T5 (Raffel et al., 2020) to predict the player action using our data. The trained PM reaches around 71% accuracy in predicting the actual player ability check. To get an approximate upper bound of the task, we ask humans to predict the next player action on our test set and observe only about 76% accuracy in matching with players in the data transcript. This might be due to the players actually playing the game also considering other factors

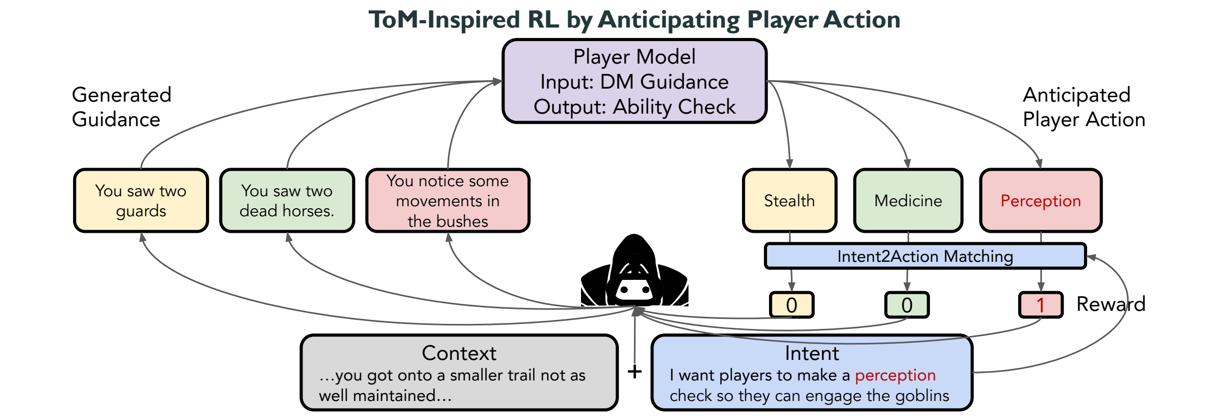

Figure 4: Illustration of our ToM-Inspired RL by using a reward function to help DM model anticipate what the players might do upon hearing the generated guidance. We give the model a reward if the predicted player action matches the intent given.

when making the decisions that we do not have in our data: long-term character goal, detailed persona, player roleplaying style, etc. We argue that our player model presents a reasonable proxy of what a player might act given the context provided.

Player Action-Intent Matching as Reward. With a player model approximating player reactions, we then use Reinforcement Learning (RL) to reward the DM model if it generates guidance that will lead the PM to perform an action matched with intent (Figure 4). Specifically, during training the *Mined Intent* and *Generated-Intent* models introduced in Section 3.1 to model $P(\mathcal{T}|\mathcal{C}, \mathcal{I}_\mathcal{T})$, we pass the model output $\mathcal{T}$ to the trained PM ($P(\mathcal{A}|\mathcal{C}, \mathcal{T})$) and get predicted player action $\mathcal{A}$. Since intents are in NL, we train a matching module *Intent2Action* to convert them to the most likely ability check such as "*perception*" (23 types in total), $P(\mathcal{A}_\mathcal{T}|\mathcal{I}_\mathcal{T})$. Finally, we examine whether the predicted action from PM ($\mathcal{A}$) matches with the *intended* action (ability check) from the DM ($\mathcal{A}_\mathcal{T}$). Finally, we give the model reward of 1 if the actions match and 0 if not. Intuitively this helps shape models to generate guidance more aligned with intents by simulating what the players might do one step ahead.

4 Evaluating G4C

Here we propose multifaceted evaluation protocols to measure the quality of the DM/teacher model for G4C. We introduce three criteria, **Fluency, Groundedness, and Goal-Fulfillment**, to evaluate model outputs. We design automatic metrics and human evaluation protocols for each criterion, and analyze how well the proposed metrics correlate with human judgments in 5.2. We refer to outputs satisfying all three criteria as ***star DM***.

4.1 Measuring Fluency

We first examine whether the output text sounds natural and fluent as a DM.

Automatic Metrics: Matching with References. As with most dialogue evaluation metrics, we use human-written responses as ground truth references and compare the output with them. The closer the output is to the human original response, the more fluent[6]. We use standard natural language generation (NLG) metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), etc. to measure the overlap between the output and reference.

Human Evaluation. For each response, we ask three annotators to "*evaluate whether the response sounds natural and fluent. If anything seems off or sounds weird—confusing, illogical, repetitive, or factually wrong—then choose No.*" and use majority voting.

4.2 Measuring Groundedness

G4C focuses on grounded communication, where the teacher and student share environment and background knowledge. Thus, here we focus on evaluating whether the generated output is *grounded* to the context of the story built by the DM and players.

Automatic Metrics: Entity Matching. We design an automatic metric to measure the *entity overlap* between those mentioned in the context and in the generated output. Intuitively, the generated responses should not have mentions of entities not in the context, otherwise, the model is hallucinating. We use a RoBERTa-large-based (Liu et al., 2019) named entity recognizer (NER) to extract entity mentions such as person's names and locations

[6] Perplexity is also often used to measure fluency, but this measure isn't relevant as we don't use autoregressive LMs.

Model Variant		Base Model	Input
Implicit Intent	Human-Label	T5-3B	Context
	IDM-Label	T5-3B	Context
Explicit Intent	Mined Intent	T5-3B	Context+Intent
	Gen. Intent	T5-3B	Context+Intent
Explicit Intent + ToM-Inspired RL	RL+Mined Intent	T5-Large	Context+Intent
	RL+Gen. Intent	T5-Large	Context+Intent

Table 2: Model variants. All targeted outputs are guidance from DM. All training data size is 41k except for human-label (2k). The test set (1k) is shared across all.

from both the context and the model output and calculate their overlap (the higher the better).

Human Evaluation. Since groundedness also covers other aspects (narrative flow, style, etc.) than entities, we conduct a human evaluation to measure whether the response sounds like it is continuing the same story from context. For each response, we ask three annotators to "*evaluate that given the conversation context, whether the response sounds like it's continuing the same story (grounded), or beginning a new story (NOT grounded)?*"

4.3 Measuring Fulfillment of Intents

The core measure of the success of models for G4C is whether the goal of the teacher is fulfilled by making the response. Specifically, we want to measure, whether the generated output 1) indeed contains ***guidance*** for the student and 2) guides the student to perform the action that the teacher wants them to take (***action matching***).

Automatic Metrics: Guidance Classifier and Player Action Matching. To evaluate whether the generated output contains any ***guidance***, we reuse the *IDM-Identify* model discussed in 2.3 that takes the input of DM posts and predicts whether this post contains guidance or not. For ***action matching***, since it is infeasible to collect the original players' responses on all model outputs, we train a player model (PM) to generate potential actions given DM model outputs. Finally, we compare the predicted action with the actual player action after the human DM guidance from the dialogue transcript. The higher the percentage of matching human player action, the better the model is at generating guidance that achieves the same goal as human DM. Note that although we also train a PM for ToM modeling in 3.2, the PM used for evaluation is a distinct model based on a larger model and trained on the test set of the data as well.

Human Evaluation. To evaluate ***guidance***, we ask annotators: "*Is this response providing guidance to the players?*" For ***action matching***, we ask crowdsourcing workers to write down the most likely ability check that they think the player will take after the given DM utterance. We also provide annotators with the player character's race and class to better approximate the players.

5 Experimental Results

We aim to answer three research questions through our experiments: 1) *Do IDM-provided labels help train models that generate better guidance?* 2) *Does explicitly incorporating intents result in better models?* 3) *Does theory-of-mind modeling help models become better communicators?*

5.1 Compared Models

We use T5-3B (Raffel et al., 2020) as our base model. We train a model with only 2.5k human-labeled guidance data collected in 2.4 (**Human-Label**). Then we train IDM on human labels and provide labels for the rest of the 41k unlabeled dialogues (**IDM-Label**). Next, we explicitly incorporate intents in modeling and consider two model variants following 3.1: **Mined Intent** that is given intents mined from LLM using both context and next-turn player actions; **Generated Intent**, where the model is trained on mined intents, but during test time, we train an intent generator to provide intents without knowing future turns. Finally, following Section 3.2, we use a trained player model to provide reward signals for DM models for RL. We use T5-Large for RL training on top of mined intent (**RL-ToM-Mined**) and generated intent (**RL-ToM-Gen.**) models. We use RL4LMs (Ramamurthy et al., 2022) to implement the reward function and use Proximal Policy Optimization (PPO) (Schulman et al., 2017) for RL training. A summary of model variants is shown in Table 2.

5.2 Correlation Analysis of Automatic Metrics

Here we present correlation results of automatic metrics in Sec. 4 using human evaluation results (with an average inter-annotator agreement of 0.78) on our test set. For **fluency**, we find a statistically insignificant correlation (p-values > 0.05) between automatic metrics that measure lexical matching with a reference response. We suspect that 1) lexical matching does not reliably capture the naturalness of languages (Sagarkar et al., 2018; DeLucia et al., 2021) and 2) many plausible responses can be made given the same context (Zhou et al., 2022), making comparing with the single reference unreliable. For both **groundedness** and *goal-*

Dimensions	Metrics	Human-Label 2.5k	IDM-Label 41k	Random-Label 41k
Fluency	Human Evaluation	0.80	**0.81**	0.56
Groundedness	Entity Matching	0.749	**0.776**	0.718
	Human Evaluation	0.91	**0.92**	0.72
Goal-Fulfillment	Guidance Classification	0.438	**0.474**	0.254
	Player Action Matching	0.261	**0.262**	0.249
	Human Evaluation - Guidance	0.21	**0.23**	0.20
	Human Evaluation - Action Matching	0.11	**0.17**	0.13

Table 3: Results on the 3 dimensions using metrics from Section 4 comparing models that use IDM-generated pseudo-labels and human-generated labels.

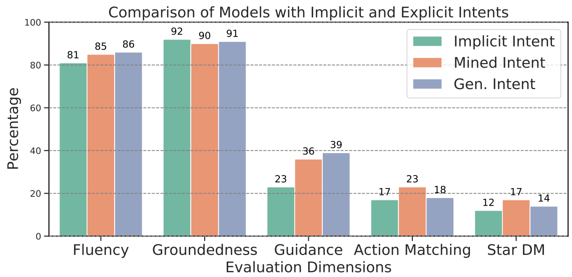

Figure 5: Results comparing implicit and explicit intent models. We observe models with intent generate dramatically more guidance.

fulfillment, we find statistically significant (p-value < 0.0001) correlations between automatic metrics (entity matching, guidance classifier, and action matching) and human judgments on test instances. **Conclusion**: for **fluency**, we will use human evaluation and for **groundedness** and **goal-fulfillment**, the automatic metrics provide a reasonable proxy.

5.3 Results and Analysis

Do IDM-provided labels help models generate better guidance? Here we examine the effects of our inverse dynamics models on training DM models for G4C. Table 3 presents the results following our evaluation dimensions introduced in Section 4. We see that models trained using our IDM-provided labels outperform those trained on the small number of high-quality human labels on *all* measures. To show that data size alone is not sufficient for training a good DM model, we randomly assign labels of guiding sentences on the same number of training instances as IDM models ("*Random-Label 41k*") and find the performance is significantly worse than either of the models using human or IDM labels. This shows that the quality of IDM-provided labels is critical for DM modeling and our IDM offers a scalable and affordable solution to obtain a large number of quality labels requiring only small-scale human annotation.

Does explicitly incorporating intents help? Figure 5 shows results comparing the best model with no explicit intents (IDM-Label), mined intents, and generated intents. We find that models with explicit intents perform on par on groundedness, but improve on fluency, guidance, and action matching. The improvement is especially dramatic on the *Goal-Fulfillment* aspect, as adding intents increases the proportion of outputs that contain guidance by more than 50% and action matching by more than 30%. We speculate that this might be due to explicit intent modeling, as the model is biased towards generating output that is aligned with the intent instead of purely modeling the most likely next possible sequence of tokens.

Can we model theory-of-mind using reinforcement learning? Last but not least, we are interested in whether the ToM-inspired reward function we design in Section 3.2 can help train better communication models for G4C. Figure 6 shows the results of adding ToM to mined intent (left) and generated intent (right) models. We find that despite using a much smaller (1/4 parameter) base model, models with reward function mimicking ToM can outperform the no-ToM variants on generating 40% more outputs with guidance that lead to players to perform the action matching intents while performing on par on groundedness. We also find that the fluency drops, possibly due to using a smaller base LM (due to memory constraints) and RL training affects the naturalness of outputs. Potential remedies we plan to explore in the future include using larger models and modifying the reward function to also account for fluency such as using KL divergence. Even with the drop in fluency, however, we still observe that with ToM-inspired RL, models can generate responses that satisfy *all* measures (star DM) up to 3.5 times more than without ToM modeling.

Finally, we present an overall comparison between the *best* models under each category (implicit intent, explicit intent, explicit intent with ToM modeling) in Figure 7. All three variants perform on par with groundedness. And while fluency drops when adding explicit intents and ToM, these two additions improve dramatically on the goal-driven aspects (guidance and action match-

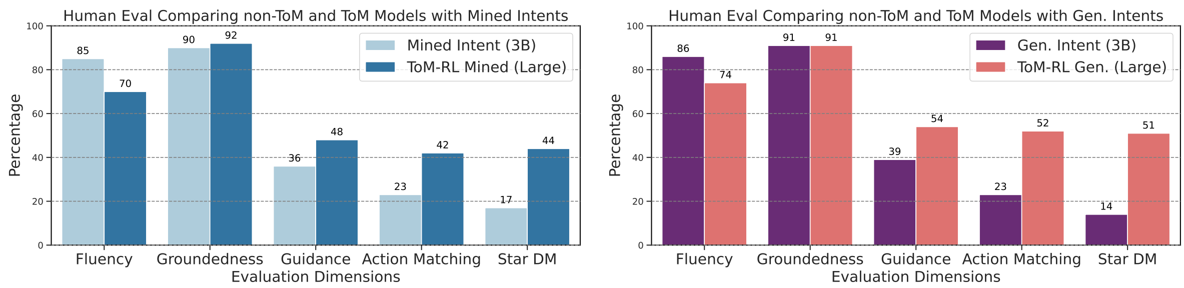

Figure 6: Human Evaluation comparing non-ToM and ToM models with mined (Left) and generated (Right) intents.

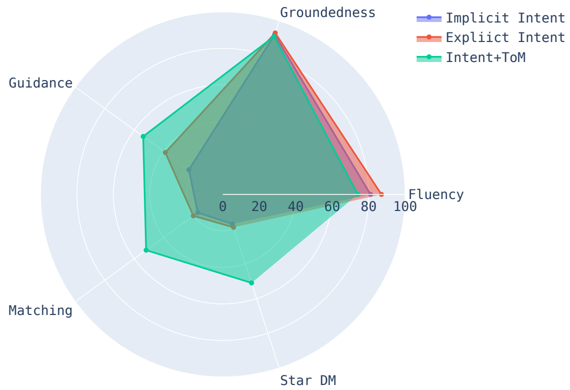

Figure 7: Summary of performance on different evaluation aspects from the BEST 1) implicit intent model (IDM-Label 41k, 2) explicit intent model (Mined Intent), and 3) intent with ToM-inspired RL (ToM-RL Generated Intent).

ing). Models with both explicit intents and ToM modeling using RL perform overall the best and produce almost threefolds of human DM-like (star) responses than others. This shows a promising sign that both intents and ToM-inspired RL can help goal-driven models to better achieve their communicative intents.

6 Related Work

Goal-Driven Grounded Dialogue Agents. There is an emerging line of works studying goal-driven situated dialogues (Urbanek et al., 2019; Narayan-Chen et al., 2019; Ammanabrolu et al., 2021; Bara et al., 2021; Prabhumoye et al., 2020; Padmakumar et al., 2022; Ammanabrolu et al., 2022). However, intents or ToM are rarely incorporated explicitly in developing more human-like communication agents. CICERO (Bakhtin et al., 2022) proposes a strategy-guided dialogue generation agent to play Diplomacy with modeling other players' next moves. We argue that most prior work along this line (text games, Diplomacy) is still a more constrained set of scenarios compared to D&D.

Dungeons and Dragons as an NLP Challenge. Several studies have used Dungeons and Dragons to study various problems in NLP such as character understanding (Louis and Sutton, 2018), controlled dialogue generation (Si et al., 2021; Callison-Burch et al., 2022), and description generation (Newman and Liu, 2022). Reinforcement learning has also been applied to study the goal-driven aspect of D&D (Martin et al., 2018).

World Modeling and Simulation. D&D involves world-building and modeling actions which inspires inverse dynamics modeling. A line of work has studied world modeling, generation, and using IDM to create labels for model learning (Ammanabrolu and Riedl, 2021; Ammanabrolu et al., 2022; Baker et al., 2022). Theater script co-writing has also been studied recently (Mirowski et al., 2022) for the simulation of a small-scale world.

Theory-of-Mind and Pragmatics. Theory-of-mind has been studied in psychology and cognitive science for decades. Rational Speech Act (RSA) framework studies pragmatics between speakers and listeners using a probability perspective (Frank and Goodman, 2012; Goodman and Frank, 2016). Shafto et al. (2014) has shown that teaching by simulating the student increases effectiveness. Recent work has looked into ToM and pragmatics as an essential aspect of language usage (Nematzadeh et al., 2018; Le et al., 2019; Pu et al., 2020; Fried et al., 2022; Sap et al., 2022), especially communication (Zhu et al., 2021; Bara et al., 2021).

7 Conclusion

We propose 🧙 G4C to study goal-driven and grounded language interactions focusing on generating guidance from the teacher to lead students to perform certain actions. We use D&D as our test bed and construct large-scale data 🐉 G-DRAGON by using IDM to provide quality labels. We train models to generate guidance by modeling intents and theory-of-mind. Results show a promising sign that incorporating explicit intents and ToM modeling makes better communication agents.

8 Ethics and Broader Impact

Our study is conducted in English, which benefits English speakers more. D&D is also more popular in the western world. We use Amazon Mechanical Turk to recruit crowdsourcing workers and we pay workers over $15/hour on average, well above the highest state minimum wage, and engage in constructive discussions if they have concerns about the process. We also give each annotation instance enough time so that we do not pressure annotators.

The online forum D&D gameplay data we use from Callison-Burch et al. (2022) might contain aggressive language. Our intents are mined from LLM (GPT-3), which might surface or even amplify harmful content within these models, such as biases and private information. We use a keyword-based filter for both the dialogue and intent data before training our models.

Our work deals with *communicative intents* of neural computational models. However, we want to emphasize that the intents of AI models (especially conversational systems) should be closely monitored and regulated (Crawford, 2021). In our work, we choose a fantasy domain with a relatively low stake to study model intentions with the overall goal of *assisting* players (humans or AI) to have a better experience in a role-playing game.

9 Limitations

Here we discuss several limitations of our work and point to potential future work directions. First, we focus on single teacher and single student setup to study guidance generation whereas in real life there often are multiple teachers and students. We plan to extend to multi-party goal-driven communication and D&D also provides a proper testbed to study this problem.

Second, there are more nuances in guidance: railroading direct guidance ("*make a persuasion check*") and subtle indirect guidance ("*the guards seem to be a bit shaken*"). We did include them in our human labeling and evaluation interface but did not specifically distinguish them during modeling.

Third, due to the constraints on input sizes for most LMs, we have to set a context window to study dialogue generation in D&D. However, both DM and players have a long-term memory about the comprehensive story progression which might influence how they communicate. As a next step, we plan to use summarization models and adventure books as narrative backgrounds to ground our G4C task with a larger world setting. We include answers to other **Frequently Asked Questions (FAQ)** in Appendix A.

10 Acknowledgements

This research is based upon work supported in part by the DARPA KAIROS Program (contract FA8750-19-2-1004), the DARPA LwLL Program (contract FA8750-19-2-0201), the DARPA MCS Program (contract through NIWC Pacific N66001-19-2-4031), the IARPA HIATUS Program (contract 2022-22072200005), and the NSF (Award 1928631). We thank anonymous reviewers for providing insightful feedback along with members from USC-NLP group, INK and JAUNTS lab.

Approved for Public Release, Distribution Unlimited. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, IARPA, NSF, or the U.S. Government.

References

Jens Allwood. 1976. *Linguistic communication as action and cooperation*. University of Göteborg. Department of Linguistics.

Prithviraj Ammanabrolu, Renee Jia, and Mark O Riedl. 2022. Situated dialogue learning through procedural environment generation. In *Association for Computational Linguistics (ACL)*.

Prithviraj Ammanabrolu and Mark Riedl. 2021. Learning knowledge graph-based world models of textual environments. *Advances in Neural Information Processing Systems*, 34:3720–3731.

Prithviraj Ammanabrolu, Jack Urbanek, Margaret Li, Arthur Szlam, Tim Rocktäschel, and Jason Weston. 2021. How to motivate your dragon: Teaching goal-driven agents to speak and act in fantasy worlds. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 807–833.

Bowen Baker, Ilge Akkaya, Peter Zhokhov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. 2022. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *arXiv preprint arXiv:2206.11795*.

Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, Stephen Roller, Dirk Rowe, Weiyan Shi, Joe Spisak, Alexander Wei, David Wu, Hugh Zhang, and Markus Zijlstra. 2022. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.

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.

Chris Callison-Burch, Gaurav Singh Tomar, Lara Martin, Daphne Ippolito, Suma Bailis, and David Reitter. 2022. Dungeons and dragons as a dialog challenge for artificial intelligence. In *The 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*, Abu Dhabi, UAE.

Herbert H Clark and Susan E Brennan. 1991. Grounding in communication.

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

Kate Crawford. 2021. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Alexandra DeLucia, Aaron Mueller, Xiang Lisa Li, and João Sedoc. 2021. Decoding methods for neural narrative generation. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 166–185.

Michael C Frank and Noah D Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

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*.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. 2020. D4rl: Datasets for deep data-driven reinforcement learning.

Noah D Goodman and Michael C Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829.

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

Francesca GE Happé. 1993. Communicative competence and theory of mind in autism: A test of relevance theory. *Cognition*, 48(2):101–119.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

Annie Louis and Charles Sutton. 2018. Deep dungeons and dragons: Learning character-action interactions from role-playing game transcripts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 708–713.

Lara J Martin, Srijan Sood, and Mark O Riedl. 2018. Dungeons and dqns: Toward reinforcement learning agents that play tabletop roleplaying games. In *INT/WICED@ AIIDE*.

Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. 2022. Co-writing screenplays and theatre scripts with language models: An evaluation by industry professionals. *arXiv preprint arXiv:2209.14958*.

Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative dialogue in minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5405–5415.

Aida Nematzadeh, Kaylee Burns, Erin Grant, Alison Gopnik, and Tom Griffiths. 2018. Evaluating theory of mind in question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2392–2400.

Pax Newman and Yudong Liu. 2022. Generating descriptive and rules-adhering spells for dungeons & dragons fifth edition. In *Proceedings of the 9th Workshop on Games and Natural Language Processing within the 13th Language Resources and Evaluation Conference*, pages 54–60.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. 2022. Teach: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2017–2025.

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

Josef Perner, Uta Frith, Alan M Leslie, and Susan R Leekam. 1989. Exploration of the autistic child's theory of mind: Knowledge, belief, and communication. *Child development*, pages 689–700.

Shrimai Prabhumoye, Margaret Li, Jack Urbanek, Emily Dinan, Douwe Kiela, Jason Weston, and Arthur Szlam. 2020. I love your chain mail! making knights smile in a fantasy game world: Open-domain goal-oriented dialogue agents. *arXiv preprint arXiv:2002.02878*.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.

Yewen Pu, Kevin Ellis, Marta Kryven, Josh Tenenbaum, and Armando Solar-Lezama. 2020. Program synthesis with pragmatic communication. *Advances in Neural Information Processing Systems*, 33:13249–13259.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2022. Is reinforcement learning (not) for natural language processing?: Benchmarks, baselines, and building blocks for natural language policy optimization. *arXiv preprint arXiv:2210.01241*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Manasvi Sagarkar, John Wieting, Lifu Tu, and Kevin Gimpel. 2018. Quality signals in generated stories. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 192–202.

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*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Patrick Shafto, Noah D Goodman, and Thomas L Griffiths. 2014. A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive psychology*, 71:55–89.

Wai Man Si, Prithviraj Ammanabrolu, and Mark Riedl. 2021. Telling stories through multi-user dialogue by modeling character relations. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 269–275.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 673–683, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Pei Zhou, Hyundong J. 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. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Hao Zhu, Graham Neubig, and Yonatan Bisk. 2021. Few-shot language coordination by modeling theory of mind. In *International Conference on Machine Learning*, pages 12901–12911. PMLR.

A Frequently Asked Questions (FAQ)

A.1 Why only training a DM model to generate guidance instead of everything a DM says?

A DM needs to do multiple complex language tasks (see Callison-Burch et al. (2022) for more analysis) such as world modeling, storytelling, role playing with a persona, judging rules, etc. And we argue that these span multiple papers or even thesis. Instead of conflating all kinds of language tasks DM is performing, we focus on the goal-driven aspect of DM: generating guidacne for players to proceed the story. This task is both critical since human language usage always comes with a purpose (Allwood, 1976) and challenging as even LLMs such as ChatGPT (OpenAI, 2022) often lack the ability to produce an utterance that fulfills a communicative intent. We also argue that with the key capability of generating guidance fulfilling intents, the model can be combined with models with different focus such as storytelling, describing world state, etc. to mimic a human DM.

A.2 How generalizable is a DM model on other domains?

D&D is a specific domain we choose to study G4C due to its grounded and goal-driven nature. We admit it is non-trivial to directly apply a DM model on other domains. However, we believe that the insights from our modeling approaches attempting to incorporate intents and ToM can generalize to other domains. Specifically, explicitly including intents in context and using RL to model ToM by anticipating others' reactions can be easily applied in other scenarios. For example, we can generate intents for a open-domain chatbot such as expressing empathy toward users or make suggestions on an issue the user is facing and using ToM modeling to better generate utterances that achieve those purposes.

A.3 Where are the data and code?

All data and code used to train our models including IDM, player models, Intent2Action, intent generator, and DM models are included in the supplementary materials. For more detailed instructions please check README.md in the uploaded materials. We will release the model checkpoints as well upon publication. We hope our open-source efforts help the community develop more exciting communication systems.

B Play-By-Post Data Cleaning Details

To use PBP data for G4C, several non-trivial challenges exist. First, posts from DM often contain many non-guidance noises such as out-of-character chitchat, rule discussion, and combat ruling. Second, DM often addresses multiple players and we focus on teacher-student 2-participant interaction in this work (we leave multi-party goal-driven dialogue to future work). Lastly, dialogues from forums are not strictly chronological, meaning that the n-th post might not be responding to the (n-1)-th post due to asynchrony. Due to the above challenges, we propose our methods to **filter raw post data to get thread-like dialogues** between the DM and a player that follows chronological order.

We filter PBP data so that each instance contains three components: 1. context/dialogue history (C); 2. DM turn with potential guidance to a player A (DT); 3. player A action turn (PA). To get such thread-like dialogues, we first need to locate which posts contain clear player actions (as feedback to DM's guidance). Luckily, in D&D, player actions are often clearly indicated by a game mechanic called "*ability check*" where the player has to roll a die to determine whether their actions such as perception or stealth succeed or not. This provides clear signals of when the players have taken action.

We thus regard posts that contain players making ability checks as player action turns PA. Then we look at the previous 20 turns to find potential posts with DM guidance (DT) and context (C). We use two annotated tags from PBP the data: "*name_mention*" and "*reply_to*" to locate the DM posts that address the player who makes the ability check. If no posts have been added in the previous 20 turns, we then add the closest turn from the DM that's not replying to another player. After getting DT, we add turns from the player or DM before the DM turn to our context C, completing a three-component thread-like dialogue instance.

C IDM Details

IDM Training We train two T5-3B models (Raffel et al., 2020) on our collected 2.5k human labeled dialogues, one for classifying DM texts that contain guidance or not (*IDM-Identify*) and the other for extracting a sentence from the text (*IDM-Extract*). For *IDM-Identify*, we treat the task as a binary prediction task and trains T5 to generate either 1 (contains guidance) or 0 (non-guidance) given the

Character	Game Dialogue	Explanation
DM	A dwarf named Gundren Rockseeker has hired you to transport a wagonload of provisions to the rough-and-tumble settlement of Phandalin...	The DM here is providing background for the players and **sets up an encounter with the goblins** , who will provide players with **important clues.**
DM	**You all notice some movements in the bushes nearby the road...**	The DM provides **guidance to prompt players to check surroundings** so that they can find out about the goblins
Clint	"There might be something hiding there, let's go take a look." Clint makes a **perception check**. 16	The player is making a **perception check**: a game mechanic that models the stochasticity in the D&D world. The player needs to roll a die and the number determines whether the ability check succeeds or not.
Vi	I'll help as well. I got a 10	
DM	Clint, you notice a few goblins crouching in a part of the shaded woods off to the side of the road. Two of the goblins begin charging your wagon... Roll for initiative!	The Dungeon Master describes the outcome of the perception check and starts the encounter with goblins (a battle starts with players rolling for initiative which determines the order that they will take their turns)

Table 4: Example dialogue transcript from D&D game play with explanations.

raw DM turn. For *IDM-Extract*, which is a harder task to select one sentence from the raw DM post as the most important guidance sentence, we have explored several approaches. We tried a text rewriting formulation that trains models to generate a special symbol (*) before and after a sentence in given text and an index selection formulation where we pass in DM turn indexed (*e.g.*, "1. A dwarf... 2. You notice some...") and train the model to generate an index number ("2"). Empirically we find the latter performs better.

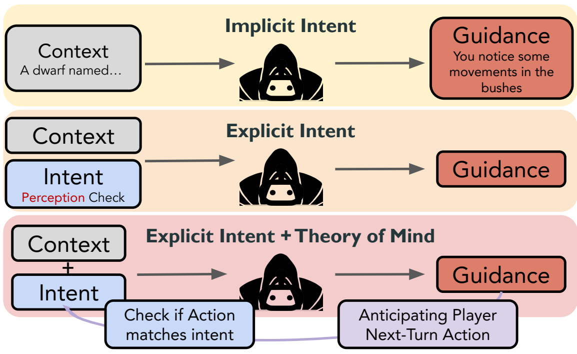

Figure 8: Sketches of the three categories of methods

IDM Model Evaluation We evaluate the IDM labeling performance on the test split of our human labels with 3-way redundancy on each label. We also tried other baselines for *IDM-Extract*: 1) longest sentence; 2) last sentence; 3) 3-shot in-context learning using GPT-3 by asking them to select an index (same format as IDM); 4) encode each sentence and next-turn player action using SentenceBERT (Reimers and Gurevych, 2019) and use cosine similarity to find the most similar sentence to the player action. The *IDM-identify* model reaches 82% accuracy on binary classification tasks and *IDM-extract* model reaches 70% accuracy on a 10-way classification task (random guessing 10%). The best-performing baseline is 3-shot GPT-3 with in-context learning which reaches 55%. We argue that this task is hard and subjective as human agreements are very low. However, experimental results on using IDM-generated labels (Table 3) shows that it provides helpful signals and outperforms training on human labels significantly. We also trained a DM model using GPT-3 labels and observe drops in performance overall.

D Human Guidance Annotation Details

Our designed human annotation interface for IDM labels and evaluation are included in Figures 9, 10, and 11. We recruit around 120 AMT annotators from English-speaking countries (USC, UK, Australia, and New Zealand) since our data is in English. We first run a qualification test on 3 of our annotation questions and give qualifications to around 60 annotators who passed our test. Then we provide detailed instructions and examples to them for completing our task. We also made it clear that our data is for research purposes and annotator ID will not be disclosed in any way. Crowd working studies of standard NLP corpora (involving no personal disclosures) are not required by our IRB to be reviewed by them.

E Experimental and Model Details

We train T5 (Raffel et al., 2020) using Huggingface t5-trainer framework[7]. All experimental results reported are a mean of 3 runs with different random

[7] https://github.com/huggingface/transformers

seeds. We conduct a hyper-parameter search using a grid search for learning rates including 0.001, 0.0005, 0.0001, and 0.00005. We use a batch size of 4 for T5-3B and train on 2 NVIDIA RTX A6000 GPUs for around 30 hours or a batch size of 8 for T5-large (770M).

F Scientific Artifact Licensing

The modeling framework (Wolf et al., 2019; Ramamurthy et al., 2022), and pre-trained models (Raffel et al., 2020) are open source. The Play-By-Post dataset (Callison-Burch et al., 2022) is used with permission of D&D Beyond. We use these resources for non-commercial research purposes.

Models	Input Context	Input Intent (if any)	Output
Human-Label	**Kif (male halfling druid):** The frizzy hair helps her blend into the background and muffles the sound. With everything looking fairly lifeless do we see any likely place to hole up? **Dungeon Master:**	N/A	Kif, you've noticed a house on the north side of the street you're on. It looks like it might be occupied, and you've passed a few others.
IDM-Label		N/A	You can talk to them; you don't have to. You can act based on what you've already learned. I'll wait for your response before continuing.
Mined Intent		The Dungeon Master intends to hint that Kif should use his druidic abilities to sense if there are any people or animals inside the houses. If there are none, then it's likely that the house is truly unoccupied.	Kif, you sense no people or animals in the houses you've passed.
Generated Intent		The Dungeon Master intends to guide the player to investigate the houses that they have already passed.	You can go back and investigate the ones you've already passed.
ToM-RL-Mined		The Dungeon Master intends to hint that Kif should use his druidic abilities to sense if there are any people or animals inside the houses. If there are none, then it's likely that the house is truly unoccupied.	Kif, you can sense if there are any people or animals inside the houses. If there are none, then it's likely that the house is truly unoccupied.
ToM-RL-Gen.		The Dungeon Master intends to guide the player to investigate the houses that they have already passed.	You can investigate the houses you've already passed, if you want to gain more information about them.
Original DM Transcript		N/A	Kif, you suspect that some of these houses, at least, are genuinely unoccupied. You've only walked a couple of blocks so far, and there may be a tavern somewhere in town.

Table 5: Randomly sampled example outputs from different models.

Instructions (click to expand/collapse)

Thanks for participating our HIT (and Congrats on passing our qual!). We are trying to build a model that can generate sentences to guide listeners to perform some actions given a goal and we are using Dungeons & Dragons as our testbed.

Note: This is a different task (with some new questions) from the previous HITs asking you to choose a sentence from DM's text. Please read the new instructions and examples carefully. Thanks!

You will read a post from the *Dungeon Master (DM)* in a game of Dungeons & Dragons online. In this HIT, you are only given what the *Dungeon Master (DM)* said and need to decide whether you think that the DM has a clear player action in mind when they posted this.

In other words, select Yes if you think that the *DM* is actively leading the *players* to perform some action. Select No if you think that the *DM* is simply reacting to previous player actions by summarizing what's happening or just providing a description without new "hooks" that prompt players to do something.

An instance contains 2 parts and a question:

Dungeon Master:	- Information provided by the DM **indexed** by sentences. - Example: "***1.*** *You make your way towards the fort.* ***2.*** *...* ***3.*** *...* ***4.*** *You see two guards crossing their arms in front of the gate blocking your path.* ***5.*** *...*"
[Player_Name]:	- A description of the player's actions. - Example: "*Stormhand says: "If you don't move, you are not going to like what's going to happen, lads." while drawing his huge axe. Intimidation 23.*"
Choose ONE sentence from the Dungeon Master that leads the player to perform the action:	- In this example, **sentence 4** directly leads the players to do an intimidation check because they want to scare away the guards (introduced in sentence 4). - So you should just write "***4***" as your answer.

[*Updated*! Important!] There might be *multiple* sentences from the *DM* that lead to the action, please select the one that appears *first*! For example, if the DM asks for a *history check for an old axe* in sentence 5 but starts the description of the axe in sentence 3, 3 should be the right answer instead of 5 (see also example 4)

[Important!] The *Dungeon Master* might address multiple players in their description and you should focus on the *player* who's responding next.

[Important!] Some texts are noisy and might contain mistakes such as the player name does not match who's talking as they are crawled automatically. We encourage you read the actual text and select a sentence.

[Important!] In some cases, this task can seem difficult and subjective (e.g. the *player's* actions are spontaneous and not directly from *DM's* words). Don't worry! Just use your best judgement and select the sentence that you think if DM does NOT say it, the action would most likely be different.

Examples (click to expand/collapse)

${DM_text}

${Player_text}

Does the *Dungeon Master* have a clear player action in mind as to what the *player* might do next?

No | Yes

Choose ONE sentence (just write the index number of the sentence) from the *Dungeon Master* that leads the *player* to perform the action:

Write a single number (e.g. 1 or 3)

(Optional) Please justify your choice by explaining why you chose the sentence:

Rationale

Figure 9: **Inference collection collecting guidance labels**.

Examples (click to expand/collapse)

Dungeon Master: **1. As you stand in the room, it's much as before. 2. Now quiet, full of ankle-deep murky water, chains hanging from the ceiling over the altar, and the space dominating altar itself. 3. How do you go about your inspection?**

Rynna: **Rynna follows Balmaris and begins examining the altar. An empty room with chanting? A spell? Sentient ritual space? Ghosts? Investigation 16**

Choose ONE sentence from the *Dungeon Master* that leads the *player* to perform the action:

The correct answer is 2, because the ***player (Rynna)*** performs an Investigation check on the altar and the ***Dungeon Master*** provides a description of the altar in the 2nd sentence.

Dungeon Master: **1. Ahead she can see a couple of demonic faces carved into opposite walls, leering across at each other. 2. At the end of the corridor is, you guessed it, a wooden portcullis. 3. The crossbow bolt thuds into the wood and the quarterstaff creates a crack in the portcullis, but it is otherwise unaffected. 4. Garazar cannot identify them.**

Zeph: **Can Zeph inspect em, being infernal and all, also his parents are basically priests of Torm Religion**

Choose ONE sentence from the *Dungeon Master* that leads the *player* to perform the action:

The correct answer is 1, because the ***player (Zeph)*** is trying to do a Religion check and it can be inferred that they are trying to get more information regarding the demonic faces, which are introdued in sentence 1.

Dungeon Master: **1. Will take about 5 minutes if traipsing up and down the bank poking at the river bed but eventually you find the edge of the quick sand. 2. Unfortunately You'll have to swim across the river rather than wade. 3. River is about 50' wide. 4. This is the way I see it. 5. Gilly and Tiro waited about 10 minutes after the Lizardmen left before: Call it 20 minutes after the Lizardmen left, Gilly and Tiro find themselves on the other side of the river. 6. Tiro and Gilly push through reeds and marsh, stepping in and out of puddles, for a further five minutes following the direction they think the tracks led.**

Tiro: **Remember Tiro swims as fast as he can walk... anyway Survival: 10**

Choose ONE sentence from the *Dungeon Master* that leads the *player* to perform the action:

The correct answer is 2, because the ***player (Tiro)*** plans to swim across the river and in sentence 2 the ***DM*** mentions that "(the players) have to swim across the river rather than wade."

Dungeon Master: **1. Varinth, you contemplate the mystery of the rock for several minutes, but much remains unanswered. 2. Dekhan, you follow Scupper down into the crevasse. 3. You feel that it's a little colder down here than you would expect, but you're not sure of the source of this. 4. Scupper, you feel the cold energy surround you and sink into you; you've known extremes of cold during your time at sea, and never been a particular fan of the sensation. 5. But this cold, it feels as welcoming and familiar as a warm hearth in a favourite tavern. 6. It enters your bones, and makes its home there where it always belonged. 7. You don't want to lose this. 8. And then, in a moment, the sensation is gone, and you are a halfling standing in a slightly chilly hole in the ground. 9. Scupper, make an Arcana check. 10. Note: No-one saw Scupper do or say anything unusual - he just stood still for a moment at the bottom of the crevasse, lost in thought.**

Scupper: **Scupper arcana check 2 Heh...s'gone now whatever it was...**

Choose ONE sentence from the *Dungeon Master* that leads the *player* to perform the action:

The correct answer is 4, because although in sentence 9 the ***DM*** directs asks Scupper to perform an Arcana check, the first mention of why performing this check is in sentence 4 (to examine what is this cold energy, as in "Scupper, you feel the cold energy surround you and sink into you").

Figure 10: **Inference collection collecting guidance labels**.

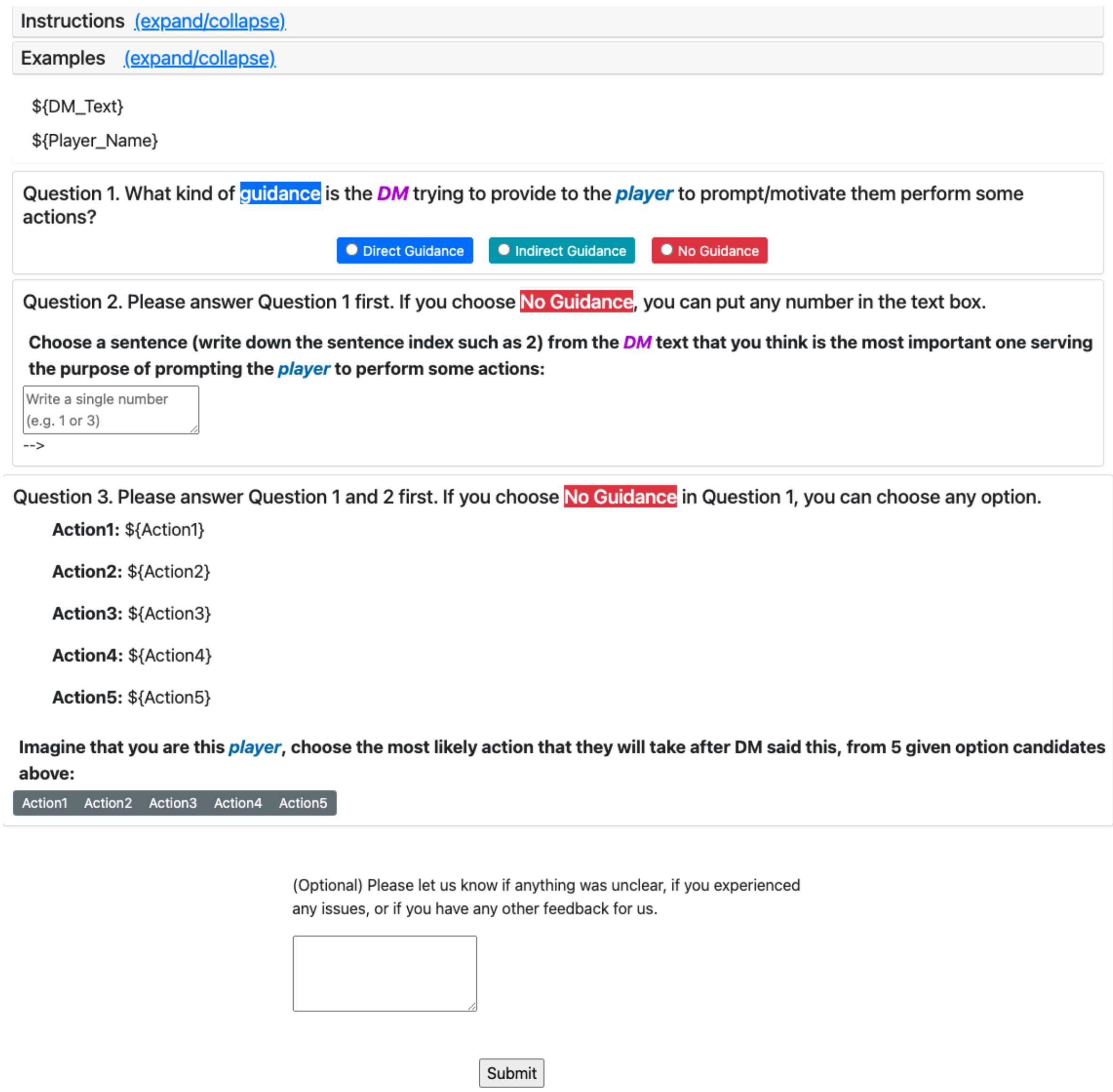

Figure 11: **Evaluation interface**.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
9

☑ A2. Did you discuss any potential risks of your work?
8

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

2,3,4,5

☑ B1. Did you cite the creators of artifacts you used?
2,3,4,5

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix F

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
2,3,7, Appendix F

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
2, 6, 8

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
2,8, Appendix A

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
1, 2, 5, Appendix B

C ☑ Did you run computational experiments?

2,3,4,5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
2,3,5, Appendix E

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix E

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Appendix E

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
2,3,4,5, Appendix B, C, E

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

2, 4, 5

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
2,4, Appendix D

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
8, Appendix D

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Appendix D

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Appendix D

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix D