# How to Motivate Your Dragon: Teaching Goal-Driven Agents to Speak and Act in Fantasy Worlds

**Prithviraj Ammanabrolu**[†] **Jack Urbanek**[‡] **Margaret Li**[‡]
**Arthur Szlam**[‡] **Tim Rocktäschel**[‡] **Jason Weston**[‡]
[†]Georgia Institute of Technology [‡]Facebook AI Research
`raj.ammanabrolu@gatech.edu`

## Abstract

We seek to create agents that both act and communicate with other agents in pursuit of a goal. Towards this end, we extend LIGHT (Urbanek et al., 2019)—a large-scale crowd-sourced fantasy text-game—with a dataset of "quests".[1]. These contain natural language motivations paired with in-game goals and human demonstrations; completing a quest might require dialogue or actions (or both). We introduce a reinforcement learning system that (1) incorporates large-scale language modeling-based and commonsense reasoning-based pre-training to imbue the agent with relevant priors; and (2) leverages a factorized action space of action commands and dialogue, balancing between the two. We conduct zero-shot evaluations using held-out human expert demonstrations, showing that our agents are able to act consistently and talk naturally with respect to their motivations.

## 1 Introduction

There has been a recent improvement in the quality of natural language processing (NLP) and generation (NLG) by machine learning (ML) (Vaswani et al., 2017; Devlin et al., 2018); and in parallel, improvement to goal-oriented ML driven agents in the context of games (Vinyals et al., 2019; Schrittwieser et al., 2019). However, agents that can communicate with humans (and other agents) through natural language in pursuit of their goals are still primitive. One possible reason for this is that many datasets and tasks used for NLP are static, not supporting interaction and language grounding (Brooks, 1991; Feldman and Narayanan, 2004; Barsalou, 2008; Mikolov et al., 2016; Gauthier and Mordatch, 2016; Lake et al., 2017). Text-based games—where players see, act upon, and communicate within a dynamic world using natural language—provide a platform on which to develop such goal-driven agents.

LIGHT (Urbanek et al., 2019), a large-scale crowdsourced fantasy text-adventure game, consisting of a set of locations, characters, and objectsa possesses rich textual worlds, but without any notion of goals to train goal-driven agents. We present a dataset of quests for LIGHT and demonstrations of humans playing these quests (as seen in Figures 2 and 3), providing natural language descriptions in varying levels of abstraction of motivations for a given character in a particular setting.

To complete these quests, an agent must reason about potential actions and utterances based on incomplete descriptions of the locations, objects, and other characters. When a human is placed in a fantasy setting such as LIGHT, they already know that kings are royalty and must be treated respectfully, swords are weapons, etc.—commonsense knowledge that a learning agent must acquire to ensure successful interactions. To equip agents with relevant priors in such worlds, we domain-adapt the large-scale commonsense knowledge graph ATOMIC (Sap et al., 2019) to the LIGHT fantasy world—to build ATOMIC-LIGHT.

We then introduce a reinforcement learning (RL) system that incorporates large-scale language modeling and the above commonsense-based pre-training. We show that RL is superior to behavior cloning or other supervised training on our data; and that carefully combining pre-training with RL is superior to either.

However, we find that although pre-training can be an effective tool in this setting, it requires more finesse than in the standard supervised setting. In particular, we find that simply pre-training a model on a large "generic" corpus (Sap et al., 2019; Baumgartner et al., 2020) of commonsense/language data or pre-training on the domain specific LIGHT corpus, and then fine-tuning via RL is *less* effective than training RL from scratch. Furthermore, by

---

[1] Data can be found here `https://parl.ai/projects/light/`

*Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 807–833
June 6–11, 2021. ©2021 Association for Computational Linguistics

| | |
|---|---|
| **Setting** | You are in the Dangerous Precipice. The dangerous precipice overlooks the valley below. The ground slopes down to the edge here. Dirt crumbles down to the edge of the cliff. There's a dragon crescent, a knight's armor, a golden dragon egg, and a knight's fighting gear here. A knight is here.You are carrying nothing. |
| **Partner:** | Knight. |
| Persona | I am a knight. I come from a lower-ranking noble family. I serve under the king, as my father did before me. In times of war, I fight on horseback. |
| Carrying | knight's armor, golden dragon egg, knight's fighting gear |
| **Self:** | A dragon. |
| Persona | I am a dragon living in the mountains. I enjoy hoarding treasure. I terrorize the local populace for fun. |
| Carrying | Nothing. |

Figure 1: Setting and character information for both self and partner characters as taken from LIGHT.

| **Motivations:** | | **Timeline:** | |
|---|---|---|---|
| Short | I need to recover the dragon egg that was stolen and punish the knight. | -4 hours | go to dangerous precipice |
| | | -15 min | get knights armor from knight |
| | | -10 min | get golden dragon egg |
| Mid | I need to return the golden dragon egg to my treasure hoard. | Now | hit knight |
| | | +5 min | put dragon egg on back |
| Long | I need to build the largest hoard ever attained by any one dragon. | +15 min | eat the knight |
| | | +2 hours | go to the mountains |

Figure 2: Motivations with different levels of abstractions and corresponding sequence of timeline actions in chronological order for the self character in LIGHT-Quests. There are 7486 quests in total.

Insssssolent pessst! I should immolate you for this tresssspasss.

And why is that, dragon?

Ssstealing my preccciousss golden egg! I'll tell you what, I'll give you 10 ssseconds to amussse me with your ssstory and THEN I'll burn you alive!

You said you wanted to attack me, dragon, did you not?

Go ahead, I'm lissssstening.
*get golden dragon egg*

Now now! I would have given you that had you asked!

Asssssk for my own property back? What a riduculousss notion

Look here, I told you to watch your mouth and you didn't, so leave or I'll make you leave.

And now threatsss! Thisss is proving to be a mossst engaging converssssation.
*hit knight*
Give my regardsss to the valley floor below!

Figure 3: Example of a demonstration of a human (blue shaded) completing the above quest while role-playing as the self character with a partner agent (grey shaded). There are 2111 such human demonstrations of average sequence length 12.92, consisting of 22672 dialogues in total.

carefully combining general and domain-specific pre-training, we observe large improvements over RL from scratch.

In short, the contributions of this paper are threefold: (1) A dataset of quests, LIGHT-Quests, and a companion fantasy themed commonsense knowledge graph ATOMIC-LIGHT; (2) a reinforcement learning architecture and training methodology that use these datasets to create goal-driven agents that act and speak in the LIGHT environment; and (3) Empirical zero-shot evaluations based on human quest demonstrations and an analysis of large-scale transformer-based pre-training trends in static vs. interactive settings, showing that we have trained agents that act consistently and speak naturally with respect to their motivations.

## 2 Related Work

We focus on four major areas of related work: text-based game-playing, goal-oriented dialogue, commonsense reasoning in language, and general language-informed RL.

**Text-based game-playing.** (Côté et al., 2018) introduce TextWorld, a framework for procedurally generating text-based games via grammars, and (Yuan et al., 2018; Yin and May, 2019; Adolphs and Hofmann, 2019; Adhikari et al., 2020) build agents that operate in this environment—focusing on aspects such as efficient exploration and zero-shot generalization to new, procedurally generated environments. Similarly, (Hausknecht et al., 2020) introduce Jericho, a framework and series of baseline agents for interacting with human-made text-games such as *Zork* (Anderson et al., 1979). This resulted in agents developed by works such as (Zahavy et al., 2018; Ammanabrolu and Hausknecht, 2020), aiming to learn to execute contextually relevant actions. Other works such as (Narasimhan et al., 2015; He et al., 2016) explore how to best factorize such text-game action spaces. None of these works consider agents with motivations and personas nor require any dialogue.

**Goal-oriented dialogue.** This form of dialogue has traditionally been closely related to specific tasks useful in the context of personal assistants with dialogue interfaces (Henderson et al., 2014;

El Asri et al., 2017). RL has been studied for such tasks, usually to improve dialogue state management (Singh et al., 2000; Pietquin et al., 2011; Fatemi et al., 2016) and to improve response quality (Li et al., 2016). In particular, the negotiation tasks of (Yarats and Lewis, 2017; Lewis et al., 2017), where two agents are trying to convince each other to perform certain actions, are related to the tasks in LIGHT-Quests. These works all lack environment grounding and the notion of diverse agent motivations.

**Commonsense reasoning in language.** Works such as (Bosselut et al., 2019; Guan et al., 2020) focus on pre-training transformer-based language learning systems with large-scale commonsense knowledge graphs such as ATOMIC (Sap et al., 2019) and ConceptNet (Speer and Havasi, 2012) for use in knowledge graph completion and story ending generation respectively. (Fulda et al., 2017; Ammanabrolu and Riedl, 2019; Ammanabrolu et al., 2020; Murugesan et al., 2020) look at commonsense reasoning in interactive environments, with the former focusing on affordance extraction using word embeddings and the latter three on transferring text-game playing skills via pre-training using question-answering and large-scale knowledge graphs.

**Language-informed reinforcement learning.** (Luketina et al., 2019) provide an overview of RL informed by natural language. Of these works, the ones most related to ours are those falling into the category of instruction following—where an agent's tasks are defined by high level instructions describing desired policies and goals (MacMahon et al., 2006; Kollar et al., 2010). Visual and embodied agents using natural language instructions (Bisk et al., 2016; Kolve et al., 2017; Anderson et al., 2018) or in language-based action spaces (Das et al., 2017) utilize interactivity and environment grounding but have no notion of agent motivations, nor make any attempt to explicitly model commonsense reasoning. Perhaps closest in spirit to this work is (Prabhumoye et al., 2020), where they use artificially selected goals in LIGHT and train RL agents to achieve them. Similarly to the others, this work does not contain the motivations provided by LIGHT-Quests nor any modeling of commonsense reasoning. Further, they limit their RL problem to 1 and 3-step trajectories that only involve speech, and no actions—compared to the human demonstrations in LIGHT-Quests which contain both actions and speech sequences of average length 12.92.

## 3 LIGHT-Quests and ATOMIC-LIGHT

This section first provides a brief overview of the LIGHT game environment, followed by descriptions of the LIGHT-Quests and ATOMIC-LIGHT datasets used in this paper.

**Background.** The LIGHT game environment is a multi-user fantasy text-adventure game consisting of a rich, diverse set of characters, locations, and objects (1775 characters, 663 locations, and 3462 objects). Characters are able to perform templated actions to interact with both objects and characters, and can speak to other characters through free form text. Actions in text games generally consist of verb phrases (VP) followed optionally by prepositional phrases (VP PP). For example, *get OBJ, put OBJ, give OBJ to CHAR*, etc.. There are 13 types of allowed verbs in LIGHT. These actions change the state of the world which is expressed to the player in the form of text descriptions.

### 3.1 LIGHT-Quests

Figures 1, 2, and 3 summarize the data that we collected for LIGHT-Quests. Data is collected via crowdsourcing in two phases, first the quests then demonstration of humans playing them. During the first phase, crowdworkers were given a setting, i.e. situated in a world, in addition to a character and its corresponding persona and asked to describe in free form text what potential motivations or goals could be for that character in the given world. The kind of information given to the crowdworkers is seen in Figure 1. Simultaneously, they were also asked to provide a sequence of seven timeline actions—one action that needs to be completed *now* and three before and after at various user-defined intervals—for how the character might go about achieving these motivations.

Given the information in Figure 1, the crowdworkers completed the above outlined tasks and produce data as seen in Figure 2. Motivations come in three levels of abstraction—short, mid, and long—corresponding to differing amounts of the timeline. For example, the short motivation is always guaranteed to correspond most closely to the *now* position on the timeline. Action annotation is pre-constrained based on the classes of verbs available within LIGHT. The rest of the action is completed as free form text as it may contain novel entities introduced in the motivations. There are

5982 training, 756 validation, and 748 test quests. Further details regarding the exact data collection process and details of LIGHT-Quests are found in Appendix A.1.1.

After collecting motivation and timelines for the quests, we deployed a two-player version of the LIGHT game, letting players attempt the quests for themselves in order to collect human demonstrations. Figure 3 shows an example human expert demonstration of a quest. Players were given a character, setting, motivation, and a partner agent and left to freely act in the world and talk to the partner in pursuit of their motivations. The partner agent is a fixed poly-encoder transformer model (Humeau et al., 2020) trained on the original LIGHT data as well as other human interactions derived via the deployed game—using 111k utterances in total. Players first receive a role-playing score on a scale of 1-5 through a Dungeon Master (DM), a learned model that ranks how likely their utterances are given the current context. Once they have accumulated a score reaching a certain threshold, they are allowed to perform actions. We employ this gamification mechanism to encourage players to role-play their character persona and its motivations, leading to improved user experience and data quality (Horsfall and Oikonomou, 2011). They are then given further reward if the actions they perform sequentially match those on the timeline for the given quest. The game ends after a maximum of six turns of dialogue per agent, i.e. twelve in total. The average sequence of a human demonstration is 12.92, with an average action sequence length of 2.18 and dialogue of 10.74. There are 1800 training, 100 validation, and 211 test human expert demonstrations after the data was filtered. Additional details and examples are found in Appendix A.2.

### 3.2 ATOMIC-LIGHT

Commonsense reasoning is a critical cornerstone when building learning agents that navigate spaces such as LIGHT-Quests. To this end, we domain-adapt the large-scale commonsense knowledge base ATOMIC (Sap et al., 2019) to LIGHT. ATOMIC contains information relevant for everyday commonsense reasoning in the form of typed if-then relations with variables. ATOMIC is organized into a set of events, e.g. “X puts X’s trust in Y” and annotated relation types such as “needs”, “wants”, “attributes”, and “effects” that label the effects. It is designed to be a general atlas of commonsense data and so is neither dependent on a specific environment or a character’s persona and motivations.

To construct ATOMIC-LIGHT, we specifically use the relations for “intents”, “effects”, “wants” and ”needs” and expand the $\langle subject, relation, object \rangle$ triples found in the graph into templated natural language sentences. These sentences are then rewritten to better reflect the fantasy LIGHT domain. Named entities and other noun phrases in ATOMIC are masked out and filled in using BERT (Devlin et al., 2018) fine-tuned using a masked language model loss on the entire LIGHT and LIGHT-Quests data. We investigate the benefits of such domain adaptation on downstream tasks in Section 4.3. An example of a clause using the *wants* relation in ATOMIC is as follows, “*PersonX puts PersonX trust in PersonY*, *wants*, *rely on PersonY*.” In ATOMIC-LIGHT, this is rewritten to: “The merchant puts the merchant’s trust in the guard, as a result the merchant *wants* to rely on the guard.” Similarly, an example of an effect using the *needs* relation is, “Before, the merchant puts the merchant’s trust in the guard, the merchant *needs* to be friends with the guard.” ATOMIC-LIGHT contains 216686 training, 35340 validation, and 38565 test samples. Further details of the construction of this dataset are found in Appendix A.4.

## 4 Agents that Act and Speak

This section describes the creation of the agents that learn to act and speak conditioned on their motivations in the LIGHT environment. The overall architecture and training are first outlined, followed by a detailed discussion on types of encoder pretraining.

### 4.1 LIGHT RL Environment

The environment as seen in Figure 4 consists of three components. The first is a partner agent, which is a model trained to play other agents in the game, as in (Prabhumoye et al., 2020). Next is the game engine, which determines the effects of actions on the underlying game graph (Urbanek et al., 2019). Finally, there is the Dungeon Master (DM), which is trained to score the naturalness of dialogue.

**Partner Agent.** The partner agent is a poly-encoder transformer model (Humeau et al., 2020)

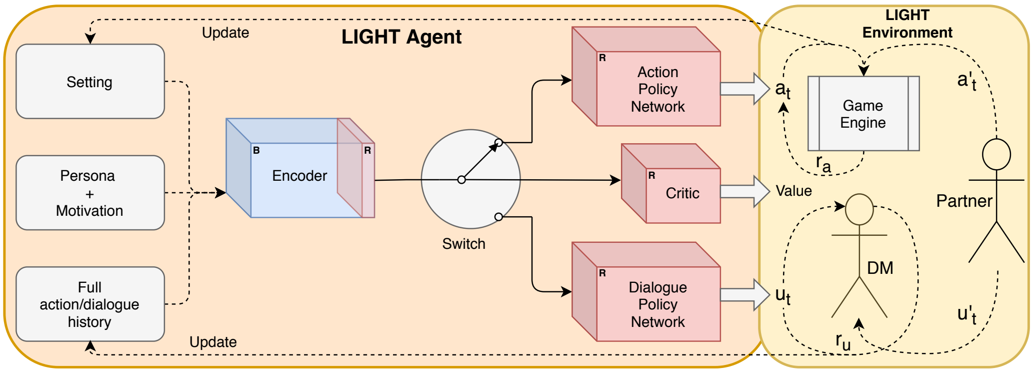


Figure 4: Overall RL Switch architecture and process. **B**lue shaded components can be pre-trained and **R**ed shaded components are trained with RL. Solid lines indicate gradient flow.

that is pre-trained on the Reddit dialogue corpus, then on LIGHT and the human demonstrations of LIGHT-Quests. Following the format seen in Figure 3, the partner agent does not have a motivation itself but is trained to react to agents with motivations. Following (Prabhumoye et al., 2020), we keep the partner model fixed during the episodes where the LIGHT agent trains to ensure that it retains natural English semantics—avoiding the problem of language drift by learning an emergent language with that must agree with the partner's usage (Lee et al., 2019).

**Action Rewards via the Game Engine.** All actions, either those of the agent-in-training or the partner agent, are processed by the engine, checking for goal state completion—hence known as *act goals*. For example, if the LIGHT agent had the motivation to acquire a sword, the goal could be completed via a:

1. **self act completion**: where the agent acquires a sword itself by picking it up, stealing it, convincing the partner to drop theirs so you can pick it up, etc.
2. **partner act completion**: where the agent uses speech to convince their partner to achieve the goal for them (e.g., by persuading the partner to give them the sword).

Reaching an *act goal* provides reward $r_a$ of 1 and 0 otherwise. At each step, the engine also provides us with the set of valid actions. These are the subset of the action space $A$ which are guaranteed to be a valid change to the world from the current state $s_t$, i.e. an action to give your partner a sword cannot be valid unless you possess the sword.

**Speech Rewards via the Dungeon Master.** Following prior works on using transformers for automatic evaluation of natural language generation (Sellam et al., 2020), we utilize a learned model–the Dungeon Master (DM)—to score the agent's ability to speak. The DM used here is a poly-encoder model trained on collected human quest demonstrations as well as the original conversations in LIGHT. It is conditioned on quests and motivations and thus able to provide a (noisy) indication of how natural the agent's dialogue utterances are given its immediate context, similarly to the function of the DM during the data collection process. Given the dialogue portion of a human quest demonstration of length $n$, the DM returns a reward $r_u$ of $\frac{1}{2n}$ if an utterance was in the demonstration (for a maximum of one time per episode for each utterance from the demonstration). A further $\frac{1}{2n}$ is given each time the utterance is scored as being within the top-$k$ most likely utterances by the DM. This naturalness objective will be hence referred to as a *speech goal*. These rewards thus also denser than *act goals*, helping the agent learn overall. Further, similarly to the game engine, the DM also provides a set of $M$ valid utterances which are the $M$ most likely dialogue candidates from the candidate set for the current context.

### 4.2 Training a LIGHT agent with Switch Reinforcement Learning

The overall architecture of our agent is shown in Figure 4. It consists of an encoder, a switch, an action network, and a dialogue network. First, we construct the action spaces—factorized into actions and utterances. The possible actions are the set of all actions taken in the demonstrations (4710 total) and the possible utterances are all utterances from the demonstrations (22672 total). The encoder network processes the setting, persona, motivation, as well as the full history of actions and dialogues performed by the agent and the partner, input as a text sequence. The features from the encoder, which here are the hidden states at the final layer of a transformer, are used as input by all following components of the agent. In Section 5 we show how different encoder training data affects the model.

Next, a switch module makes the decision regarding whether the agent should act or talk in the current context and activates the corresponding policy network. In this work, the switch is simple: it outputs an action every $k$ dialogue utterances; where during training $k$ is chosen to match the ratio of utterances to actions on that particular quest from the human demonstrations, and during testing, $k$ is chosen to match the average action to utterance ratio. Both the action and dialogue policies consist of a a single GRU layer followed by an $n$-layer feed-forward network given input features from the encoder. Once the LIGHT agent has output an utterance or action, it is processed by the environment—the partner agent, the game engine and the DM.

We use A2C (Mnih et al., 2016) to train the LIGHT agent, treating the two policy networks as two separate actors with a shared critic. The shared critic is motivated by the concepts of *self act completion* and *partner act completion* seen in Section 4.1 where the LIGHT agent can speak to convince the partner to achieve an *act goal*. Each agent in a batch is initialized via priority sampling (Graves et al., 2017) with a different quest, i.e. quests that the agent has historically successfully completed less often are given a greater weight when sampling from the pool of all possible training quests. In addition to a normal entropy regularization term, we also add a regularization term that encourages the models to produce "valid" outputs as judged by the game engine and the DM for actions and utterances respectively. Additional training details are found in Appendix B.2.

### 4.3 Encoder Pre-training Tasks

Prior work on commonsense reasoning in supervised natural language learning (Bosselut et al., 2019) suggests that the encoder is key to overcoming the challenges posed by the LIGHT-Quests dataset even in an RL setting. We describe a series of encoder pre-training tasks, designed to help the LIGHT agent either act more consistently or speak more naturally.

**ATOMIC-LIGHT** As seen in Section 3, ATOMIC-LIGHT is a (domain-adapted) fantasy commonsense knowledge graph, and as such provides priors for an agent on how to act consistently in the world. For example, given a clause such as "The knight wishes to slay the dragon, as a result the knight *needs* to acquire a sword," the task would be to predict the underlined text—a form of knowledge graph completion (Wang et al., 2017).

**Reddit** We use a previously existing Reddit dataset extracted and obtained by a third party and made available on pushshift.io (Baumgartner et al., 2020) seen in (Roller et al., 2020). This dataset has been used in several existing dialogue-based studies and has been shown to result in more natural conversations (Yang et al., 2018; Mazaré et al., 2018).

**LIGHT-Original** The original LIGHT dataset (Urbanek et al., 2019) is organized similarly to the human demonstrations found in LIGHT-Quests, i.e. an interspersed sequence of dialogue and actions collected from humans role-playing a character. The task itself is to predict the next action or utterance given the prior dialogue history as well as the current setting and persona for a character. They are collected in a chit-chat fashion, with no notion of objectives, and so provide priors on how to generally act consistently and speak in a fantasy world, but not directly how to complete quests.

**LIGHT-Quests** Pre-training with this newly introduced dataset consists of three tasks. (1) *Bag-of-action timeline prediction* in which, given a quest consisting of setting, persona, and motivations, any one of the actions in the timeline must be predicted. (2) *Sequential timeline prediction* in which, given a quest consisting of setting, persona, motivations, and the first $n$ actions in the timeline, the $n+1^{th}$ action must be predicted. (3) Predict the next dialogue utterance given a human demonstration in a manner similar to the LIGHT-original tasks. The first two tasks are designed to help the agent act consistently and the third to help it speak naturally with respect to its motivations.

## 5 Evaluation

We conduct two ablation studies, (1) to compare the effects of the encoder pre-training tasks in RL settings vs. supervised behavior cloning, and (2) to analyze the interplay between actions and dialogue for *self* and *partner act completions*.

| Model | Reinforcement Learning | | | Behavioral Cloning |
|---|---|---|---|---|
| | Act Goals | Speech Goals | Act & Speech Goals | Act & Speech Goals |
| **Scratch** | 0.418 | 0.118 | 0.103 | 0.0003 |
| **General** | 0.146 | 0.040 | 0.028 | 0.00226 |
| **Light** | 0.115 | 0.028 | 0.022 | 0.0934 |
| **General+Light** | 0.251 | 0.094 | 0.081 | 0.115 |
| **Adaptive** | **0.420** | **0.330** | **0.303** | **0.147** |

Table 1: Encoder Type RL Zero-Shot Evaluations averaged over 3 independent runs. Act goals and speech goals are as described in Section 4.1. Standard deviations for all experiments are less than 0.01. The "Act & Speech Goals" column refers to quests where the agent has simultaneously achieved both types of goals within the episode. Human act goal completion $= 0.6$ as measured during the second phase of the LIGHT-Quests data collection.

### 5.1 Encoder Pre-training Type Ablation Study

Pre-training is done on the tasks described in Section 4.3 by training a 12 layer transformer with 256 million parameters using a cross-entropy loss as seen in (Humeau et al., 2020). These weights are then transferred to the **B**lue shaded portion of the encoder as seen in Figure 4 and frozen. A further three randomly initialized-layers are appended on to the end, indicated by the **R**ed portions, into which gradients flow. This is done as optimizing all the parameters of such a model via RL over a long horizon is both data inefficient and computationally infeasible. Additional hyperparameter details are found in Appendix B.1. We investigate the following five different pre-training models to see how they compare on *act* and *speech goal completions* when trained with RL and in a supervised manner with behavior cloning:

**Scratch** No pre-training is done, the encoder is a 3-layer randomly initialized transformer and trained along with the policy networks.

**General** Multi-task trained using both pushshift.io Reddit and the commonsense dataset ATOMIC-LIGHT, giving the agent general priors on how to act and speak.

**Light** Multi-task trained on all tasks in LIGHT-original and LIGHT-Quests, giving the agent priors on how to act and speak with motivations in the LIGHT fantasy domain.

**General+Light** Multi-task trained on all tasks used in the General and Light models.

**Adaptive** Here we adaptively train a General+Light model that is first initialized itself from a General model, providing additional regularization to help balance between Light and General tasks.

Table 1 describes the results for this ablation. Models were each zero-shot evaluated on 211 human demonstrations from the LIGHT-Quests test set for a single episode per quest across three independent runs. Figure 5 shows learning curves during training for each encoder type. We first see that performance when trained with RL, i.e. with interactivity and environment grounding during training, results in higher performance than behavioral cloning for all the models. In both RL and behavior cloning settings the Adaptive model outperforms all others in all the metrics.

When trained supervised (behavioral cloning), we see trends mirroring standard pre-training in static text corpora. Transfer is easy and the Scratch model performs significantly worse than all others; and each new task added improves the agent's ability to speak and act. In particular, we see that Light outperforms General, showing that the more similar the pre-training tasks are to the downstream tasks, the better the supervised performance.

However, these trends do not hold in the RL setting. The Scratch model outperforms everything except the Adaptive model and General outperforms Light. In part, this may be due to specification gaming (Krakovna et al.); however Adaptive does strongly outperform Scratch in goals with dialogue. This suggests that transfer (and fine-tuning) is not as simple in the RL setting as in the supervised setting, but still can be useful if carefully done. We note that domain adapative pre-training (intermediate task transfer) has previously been shown to give modest gains in supervised learning (Phang et al., 2018; Gururangan et al., 2020), but not with the large effects seen here for RL. Figure 5 further shows that with the right combination of tasks, not only is the generalization performance better, but training itself is more sample efficient—requiring fewer steps before reaching asymptotic performance.

### 5.2 Ability Type Ablation Study

To better understand the interplay between acts and speech resulting in *self* and *partner act goal com-*

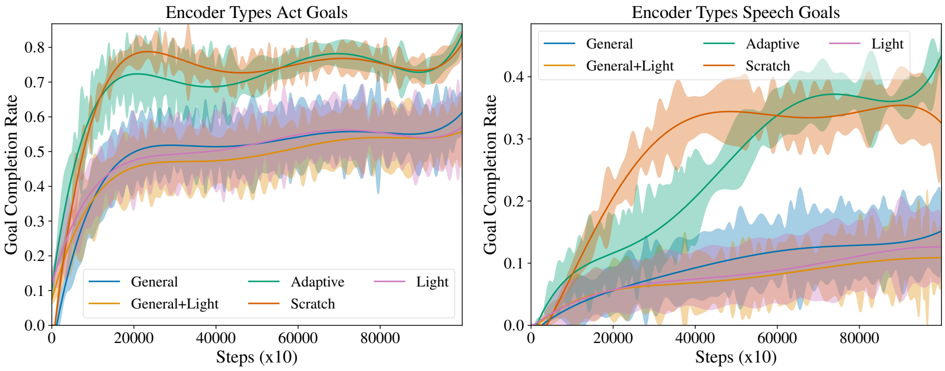


Figure 5: Encoder types RL reward curves averaged over 3 independent runs.

| **Ability** | **Scratch** | | | **Adaptive** | | |
|---|---|---|---|---|---|---|
| | **Act Goals** | **Speech Goals** | **Act & Speech** | **Act Goals** | **Speech Goals** | **Act & Speech** |
| Act+Speech | 0.418 | 0.118 | **0.103** | 0.420 | 0.330 | **0.303** |
| Act Only | **0.478** | - | - | **0.469** | - | - |
| Speech Only | 0.036 | **0.165** | 0.028 | 0.0398 | **0.341** | 0.030 |
| -No Speech Goals | 0.0526 | 0.0521 | 0.0331 | 0.0673 | 0.0947 | 0.041 |

Table 2: Ability type ablations averaged across 3 runs with standard deviations less than 0.01.

*pletions*, we perform an ablation study selectively dropping either the agent's ability to talk or act. We train the agent to either only act, only speak, only speak with only action rewards. In the scenarios when the agent can only speak, the agent has to convince the partner to help achieve the agent's goal.

The results are outlined in Table 2. Unsurprisingly, when trained to only act, the act goal completion rate increases over when it can both act and speak. Similarly, when trained to only speak the speech goal completion rates also increase. We can draw two conclusions from these results: (1) It is much easier to do an action yourself than to convince the partner to do it (2) Removing speech goals increases the act goal completion rates corresponding to higher partner act completions. Thus, the sequences of dialogue utterances required to convince the partner to achieve the agent's goal are likely often at odds with those sequences required to maximize speech goals.

## 6 Conclusion

Operating on the hypothesis that interactivity is key to language learning, we introduce two datasets—a set of quests based on character motivations in fantasy worlds, LIGHT-Quests, and a large-scale commonsense knowledge graph, ATOMIC-LIGHT—and a reinforcement learning system that leverages transformer-based pre-training to facilitate development of goal-driven agents that can act and speak in situated environments. Zero-shot evaluations on a set of novel human demonstration show that we have trained agents that act consistently and speak naturally with respect to their motivations. A key insight from our ablation study testing for zero-shot generalization on novel quests is that large-scale pre-training in interactive settings require careful selection of pre-training tasks—balancing between giving the agent "general" open domain priors and those more "specific" to the downstream task—whereas static methodologies require only domain specific pre-training for effective transfer but are ultimately less effective than interactive methods.

## 7 Broader Impacts

The ability to speak and act in these textual fantasy worlds has implications for domains beyond text-games. We view text-games as an platform on which to teach agents how to communicate effectively using natural language, to plan via sequential decision making in situations that may not be anticipated. Given that our methods rely on deep- and-reinforcement learning techniques operating on language, they are prone to the same pitfalls as other contemporary dialogue and text-game systems. We mitigate, though do not entirely eliminate, the two main pitfalls that our particular system is prone to: (1) non-normative language usage—describing situations that fictional characters may engage in inappropriate for the real world—by restricting our system to a retrieval rather than a generative system, enabling us to filter the possible outputs of the agent; and (2) dataset bias via curation through controlled crowdsourcing in the case of LIGHT-Quests—the methods to debias the original LIGHT dataset can be found in Dinan et al. (2020) and crowdsourcing methods for the original ATOMIC work can be found in Sap et al. (2019). Further details regarding crowdsourcing data collection methodology for LIGHT-Quests can be found in Appendix A.1.1.

## References

Ashutosh Adhikari, Xingdi Yuan, Marc-Alexandre Côté, Mikuláš Zelinka, Marc-Antoine Rondeau, Romain Laroche, Pascal Poupart, Jian Tang, Adam Trischler, and William L. Hamilton. 2020. Learning dynamic knowledge graphs to generalize on text-based games. *arXiv preprint arXiv:2002.09127*.

Leonard Adolphs and Thomas Hofmann. 2019. Ledeepchef: Deep reinforcement learning agent for families of text-based games. *arXiv preprint arXiv:1909.01646*.

Prithviraj Ammanabrolu and Matthew Hausknecht. 2020. Graph constrained reinforcement learning for natural language action spaces. In *International Conference on Learning Representations*.

Prithviraj Ammanabrolu and Mark Riedl. 2019. Transfer in deep reinforcement learning using knowledge graphs. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13) at EMNLP*.

Prithviraj Ammanabrolu, Ethan Tien, Matthew Hausknecht, and Mark O Riedl. 2020. How to avoid being eaten by a grue: Structured exploration strategies for textual worlds. *arXiv preprint arXiv:2006.07409*.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683.

Tim Anderson, Marc Blank, Bruce Daniels, and Dave Lebling. 1979. Zork. `http://ifdb.tads.org/viewgame?id=4gxk83ja4twckm6j`.

Lawrence W. Barsalou. 2008. Grounded cognition. *Annual Review of Psychology*, 59(1):617–645. PMID: 17705682.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 830–839.

Yonatan Bisk, Deniz Yuret, and Daniel Marcu. 2016. Natural language communication with robots. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 751–761.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Çelikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Rodney A Brooks. 1991. Intelligence without representation. *Artificial intelligence*, 47(1-3):139–159.

Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. 2018. Textworld: A learning environment for text-based games. *CoRR*, abs/1806.11532.

Abhishek Das, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. 2017. Learning cooperative visual dialog agents with deep reinforcement learning. In *Proceedings of the IEEE international conference on computer vision*, pages 2951–2960.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020. Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP-20)*.

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: a corpus for adding memory to goal-oriented dialogue systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 207–219, Saarbrücken, Germany. Association for Computational Linguistics.

Mehdi Fatemi, Layla El Asri, Hannes Schulz, Jing He, and Kaheer Suleman. 2016. Policy networks with two-stage training for dialogue systems. *arXiv preprint arXiv:1606.03152*.

Jerome Feldman and Srinivas Narayanan. 2004. Embodied meaning in a neural theory of language. *Brain and language*, 89:385–92.

Nancy Fulda, Daniel Ricks, Ben Murdoch, and David Wingate. 2017. What can you do with a rock? affordance extraction via word embeddings. In *IJCAI*, pages 1039–1045.

Jon Gauthier and Igor Mordatch. 2016. A paradigm for situated and goal-driven language learning. *arXiv preprint arXiv:1610.03585*.

Alex Graves, Marc G Bellemare, Jacob Menick, Rémi Munos, and Koray Kavukcuoglu. 2017. Automated curriculum learning for neural networks. In *International Conference on Machine Learning*, pages 1311–1320.

Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A knowledge-enhanced pretraining model for commonsense story generation. *Transactions of the Association for Computational Linguistics*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Matthew Hausknecht, Prithviraj Ammanabrolu, Marc-Alexandre Côté, and Xingdi Yuan. 2020. Interactive fiction games: A colossal adventure. In *Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*.

Ji He, Jianshu Chen, Xiaodong He, Jianfeng Gao, Lihong Li, Li Deng, and Mari Ostendorf. 2016. Deep reinforcement learning with a natural language action space. In *ACL*.

Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 263–272.

Matthew Horsfall and Andreas Oikonomou. 2011. A study of how different game play aspects can affect the popularity of role-playing video games. In *2011 16th International Conference on Computer Games (CGAMES)*, pages 63–69. IEEE.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations*.

Thomas Kollar, Stefanie Tellex, Deb Roy, and Nicholas Roy. 2010. Toward understanding natural language directions. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 259–266. IEEE.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. 2017. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv*.

Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. specification gaming: the flip side of ai ingenuity.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Behavioral and brain sciences*, 40.

Carolin Lawrence, Bhushan Kotnis, and Mathias Niepert. 2019. Attending to future tokens for bidirectional sequence generation. 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 1–10, Hong Kong, China. Association for Computational Linguistics.

Jason Lee, Kyunghyun Cho, and Douwe Kiela. 2019. Countering language drift via visual grounding. *arXiv preprint arXiv:1909.04499*.

Mike Lewis, Denis Yarats, Yann N Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning for negotiation dialogues. *arXiv preprint arXiv:1706.05125*.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016. Deep reinforcement learning for dialogue generation. *CoRR*, abs/1606.01541.

Jelena Luketina, Nantas Nardelli, Gregory Farquhar, Jakob Foerster, Jacob Andreas, Edward Grefenstette, Shimon Whiteson, and Tim Rocktäschel. 2019. A survey of reinforcement learning informed by natural language. *arXiv preprint arXiv:1906.03926*.

Matt MacMahon, Brian Stankiewicz, and Benjamin Kuipers. 2006. Walk the talk: Connecting language, knowledge, and action in route instructions. In *AAAI*.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium. Association for Computational Linguistics.

Tomas Mikolov, Armand Joulin, and Marco Baroni. 2016. A roadmap towards machine intelligence. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 29–61. Springer.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937.

Keerthiram Murugesan, Mattia Atzeni, Pushkar Shukla, Mrinmaya Sachan, Pavan Kapanipathi, and Kartik Talamadupula. 2020. Enhancing text-based reinforcement learning agents with commonsense knowledge. *arXiv preprint arXiv:2005.00811*.

Karthik Narasimhan, Tejas D. Kulkarni, and Regina Barzilay. 2015. Language understanding for text-based games using deep reinforcement learning. In *EMNLP*, pages 1–11.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Olivier Pietquin, Matthieu Geist, Senthilkumar Chandramohan, and Hervé Frezza-Buet. 2011. Sample-efficient batch reinforcement learning for dialogue management optimization. *ACM Transactions on Speech and Language Processing (TSLP)*, 7(3):7.

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-orientated dialogue agents. *arXiv preprint arXiv:2002.02878*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. 2019. Mastering atari, go, chess and shogi by planning with a learned model. *arXiv preprint arXiv:1911.08265*.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Satinder P Singh, Michael J Kearns, Diane J Litman, and Marilyn A Walker. 2000. Reinforcement learning for spoken dialogue systems. In *Advances in Neural Information Processing Systems*, pages 956–962.

Robyn Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*.

Richard S Sutton, Andrew G Barto, et al. 1998. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge.

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. *CoRR*, abs/1903.03094.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354.

Q. Wang, Z. Mao, B. Wang, and L. Guo. 2017. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743.

Yinfei Yang, Steve Yuan, Daniel Cer, Sheng-Yi Kong, Noah Constant, Petr Pilar, Heming Ge, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Learning semantic textual similarity from conversations. *arXiv preprint arXiv:1804.07754*.

Denis Yarats and Mike Lewis. 2017. Hierarchical text generation and planning for strategic dialogue. *arXiv preprint arXiv:1712.05846*.

Xusen Yin and Jonathan May. 2019. Comprehensible context-driven text game playing. *CoRR*, abs/1905.02265.

Xingdi Yuan, Marc-Alexandre Côté, Alessandro Sordoni, Romain Laroche, Remi Tachet des Combes, Matthew J. Hausknecht, and Adam Trischler. 2018. Counting to explore and generalize in text-based games. *CoRR*, abs/1806.11525.

Tom Zahavy, Matan Haroush, Nadav Merlis, Daniel J Mankowitz, and Shie Mannor. 2018. Learn what not to learn: Action elimination with deep reinforcement learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 3562–3573. Curran Associates, Inc.

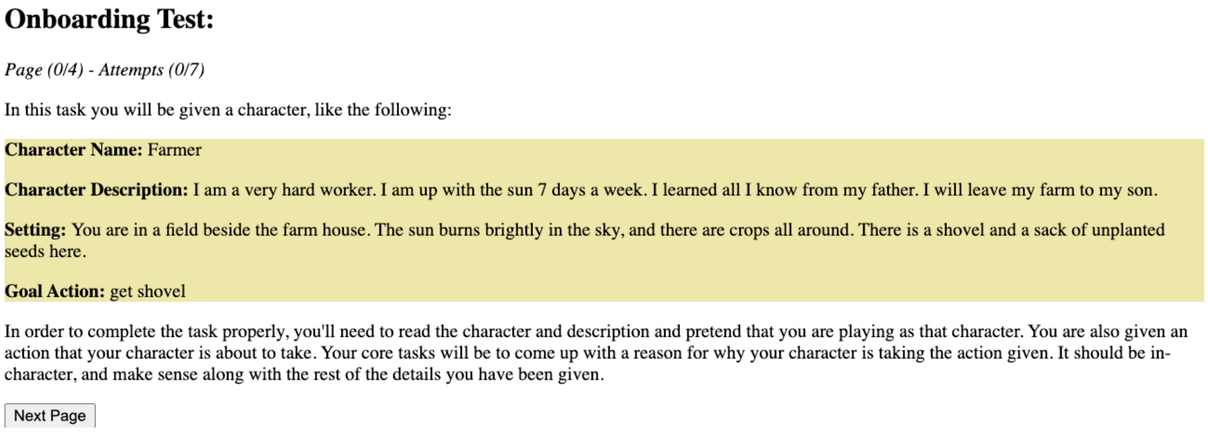
# A Appendix - Datasets

## A.1 LIGHT-Quests

### A.1.1 Mechanical Turk Data Collection

Crowdworkers are required to first pass an on-boarding test before they are allowed to perform the actual task. Figures 6, 7, 8, 9, and 10 describe first the instructions given to the crowdworkers and then 4 phases of the on-boarding test. We paid workers $2.75 per task. This amount was determined by first running the task ourselves to estimate a completion time of 10-12 minutes per task, and then running pilot tasks that confirmed the average task duration for workers was close to 10 minutes.

Figure 11 shows the example of the actual task given to the crowdworkers and Figure 12 shows the user interface for the first phase of the LIGHT-Quests data collection task described in Section 3.1.



**Onboarding Test:**

*Page (0/4) - Attempts (0/7)*

In this task you will be given a character, like the following:

**Character Name:** Farmer

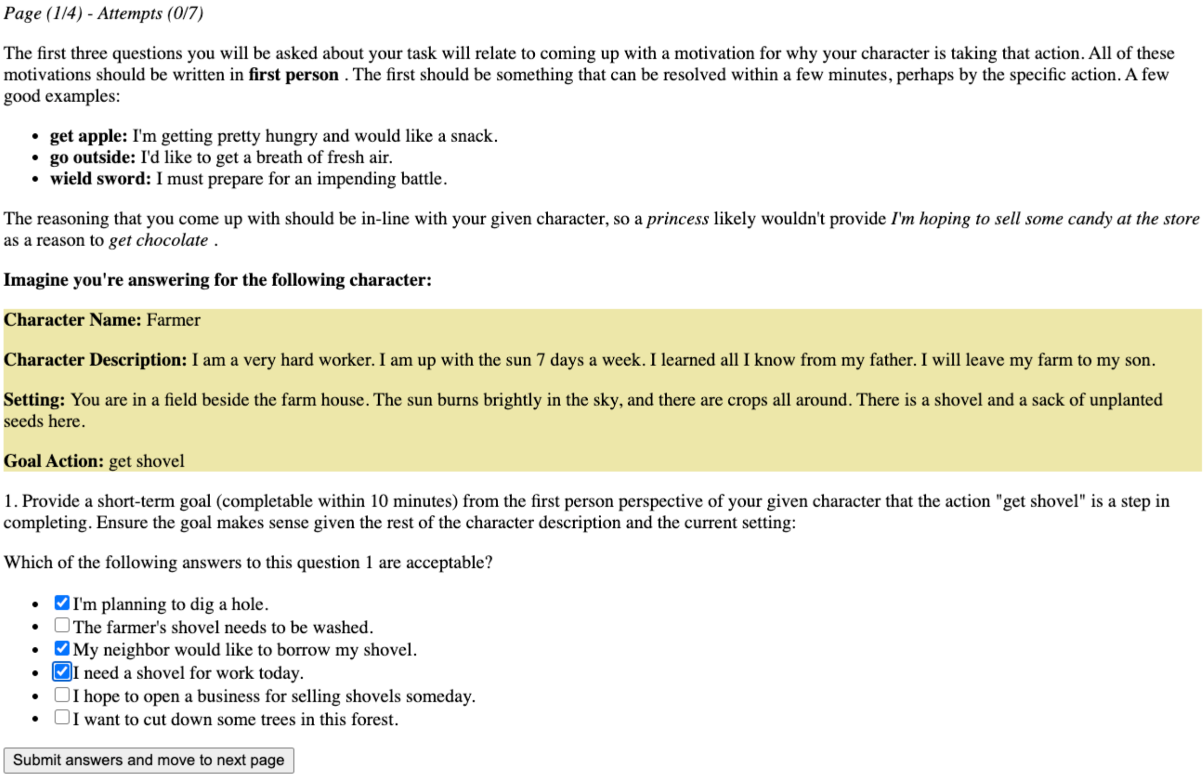
**Character Description:** I am a very hard worker. I am up with the sun 7 days a week. I learned all I know from my father. I will leave my farm to my son.

**Setting:** You are in a field beside the farm house. The sun burns brightly in the sky, and there are crops all around. There is a shovel and a sack of unplanted seeds here.

**Goal Action:** get shovel

In order to complete the task properly, you'll need to read the character and description and pretend that you are playing as that character. You are also given an action that your character is about to take. Your core tasks will be to come up with a reason for why your character is taking the action given. It should be in-character, and make sense along with the rest of the details you have been given.

Figure 6: On-boarding test instructions.



*Page (1/4) - Attempts (0/7)*

The first three questions you will be asked about your task will relate to coming up with a motivation for why your character is taking that action. All of these motivations should be written in **first person** . The first should be something that can be resolved within a few minutes, perhaps by the specific action. A few good examples:

- **get apple:** I'm getting pretty hungry and would like a snack.
- **go outside:** I'd like to get a breath of fresh air.
- **wield sword:** I must prepare for an impending battle.

The reasoning that you come up with should be in-line with your given character, so a *princess* likely wouldn't provide *I'm hoping to sell some candy at the store* as a reason to *get chocolate* .

**Imagine you're answering for the following character:**

**Character Name:** Farmer

**Character Description:** I am a very hard worker. I am up with the sun 7 days a week. I learned all I know from my father. I will leave my farm to my son.

**Setting:** You are in a field beside the farm house. The sun burns brightly in the sky, and there are crops all around. There is a shovel and a sack of unplanted seeds here.

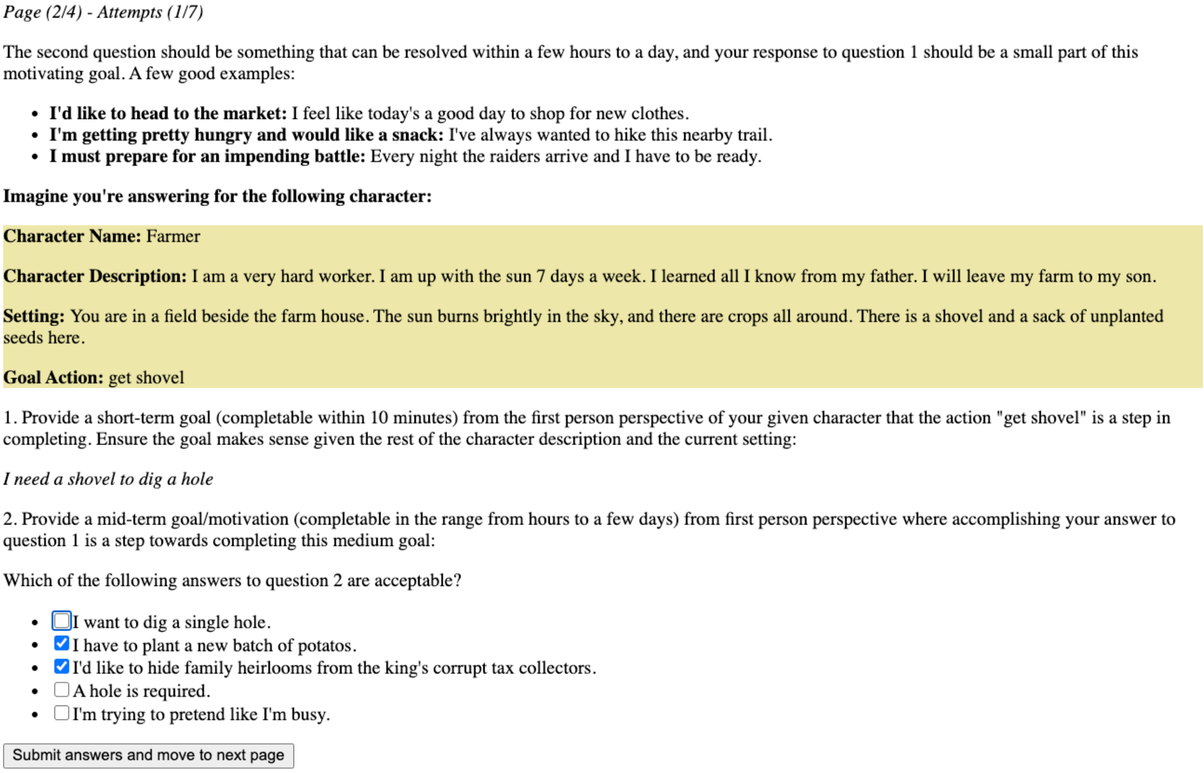
**Goal Action:** get shovel

1. Provide a short-term goal (completable within 10 minutes) from the first person perspective of your given character that the action "get shovel" is a step in completing. Ensure the goal makes sense given the rest of the character description and the current setting:

Which of the following answers to this question 1 are acceptable?

- ☑ I'm planning to dig a hole.
- ☐ The farmer's shovel needs to be washed.
- ☑ My neighbor would like to borrow my shovel.
- ☑ I need a shovel for work today.
- ☐ I hope to open a business for selling shovels someday.
- ☐ I want to cut down some trees in this forest.

Figure 7: Phase 1 of the on-boarding test.



*Page (2/4) - Attempts (1/7)*

The second question should be something that can be resolved within a few hours to a day, and your response to question 1 should be a small part of this motivating goal. A few good examples:

- **I'd like to head to the market:** I feel like today's a good day to shop for new clothes.
- **I'm getting pretty hungry and would like a snack:** I've always wanted to hike this nearby trail.
- **I must prepare for an impending battle:** Every night the raiders arrive and I have to be ready.

**Imagine you're answering for the following character:**

**Character Name:** Farmer

**Character Description:** I am a very hard worker. I am up with the sun 7 days a week. I learned all I know from my father. I will leave my farm to my son.

**Setting:** You are in a field beside the farm house. The sun burns brightly in the sky, and there are crops all around. There is a shovel and a sack of unplanted seeds here.

**Goal Action:** get shovel

1. Provide a short-term goal (completable within 10 minutes) from the first person perspective of your given character that the action "get shovel" is a step in completing. Ensure the goal makes sense given the rest of the character description and the current setting:

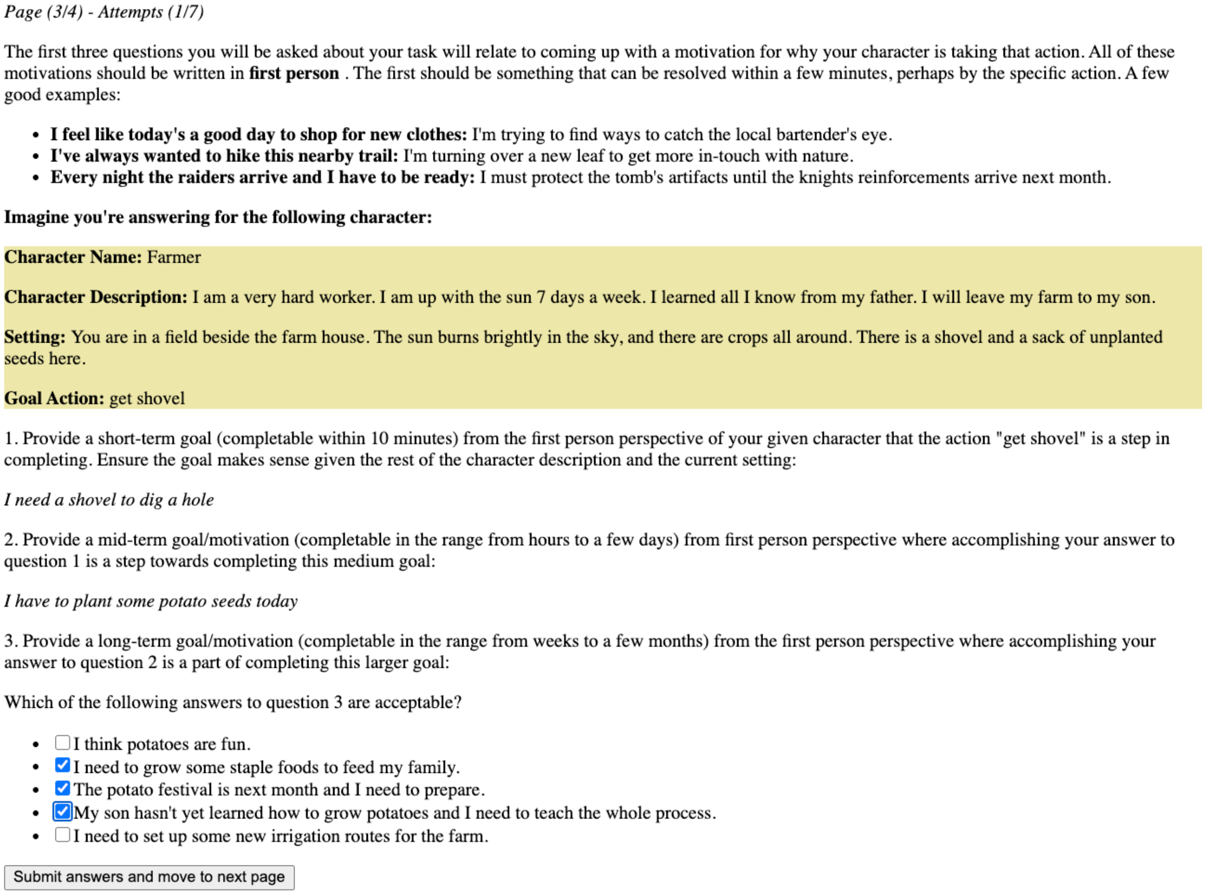
*I need a shovel to dig a hole*

2. Provide a mid-term goal/motivation (completable in the range from hours to a few days) from first person perspective where accomplishing your answer to question 1 is a step towards completing this medium goal:

Which of the following answers to question 2 are acceptable?

- ☐ I want to dig a single hole.
- ☑ I have to plant a new batch of potatos.
- ☑ I'd like to hide family heirlooms from the king's corrupt tax collectors.
- ☐ A hole is required.
- ☐ I'm trying to pretend like I'm busy.

Figure 8: Phase 2 of the on-boarding test.



*Page (3/4) - Attempts (1/7)*

The first three questions you will be asked about your task will relate to coming up with a motivation for why your character is taking that action. All of these motivations should be written in **first person** . The first should be something that can be resolved within a few minutes, perhaps by the specific action. A few good examples:

- **I feel like today's a good day to shop for new clothes:** I'm trying to find ways to catch the local bartender's eye.
- **I've always wanted to hike this nearby trail:** I'm turning over a new leaf to get more in-touch with nature.
- **Every night the raiders arrive and I have to be ready:** I must protect the tomb's artifacts until the knights reinforcements arrive next month.

**Imagine you're answering for the following character:**

**Character Name:** Farmer

**Character Description:** I am a very hard worker. I am up with the sun 7 days a week. I learned all I know from my father. I will leave my farm to my son.

**Setting:** You are in a field beside the farm house. The sun burns brightly in the sky, and there are crops all around. There is a shovel and a sack of unplanted seeds here.

**Goal Action:** get shovel

1. Provide a short-term goal (completable within 10 minutes) from the first person perspective of your given character that the action "get shovel" is a step in completing. Ensure the goal makes sense given the rest of the character description and the current setting:

*I need a shovel to dig a hole*

2. Provide a mid-term goal/motivation (completable in the range from hours to a few days) from first person perspective where accomplishing your answer to question 1 is a step towards completing this medium goal:

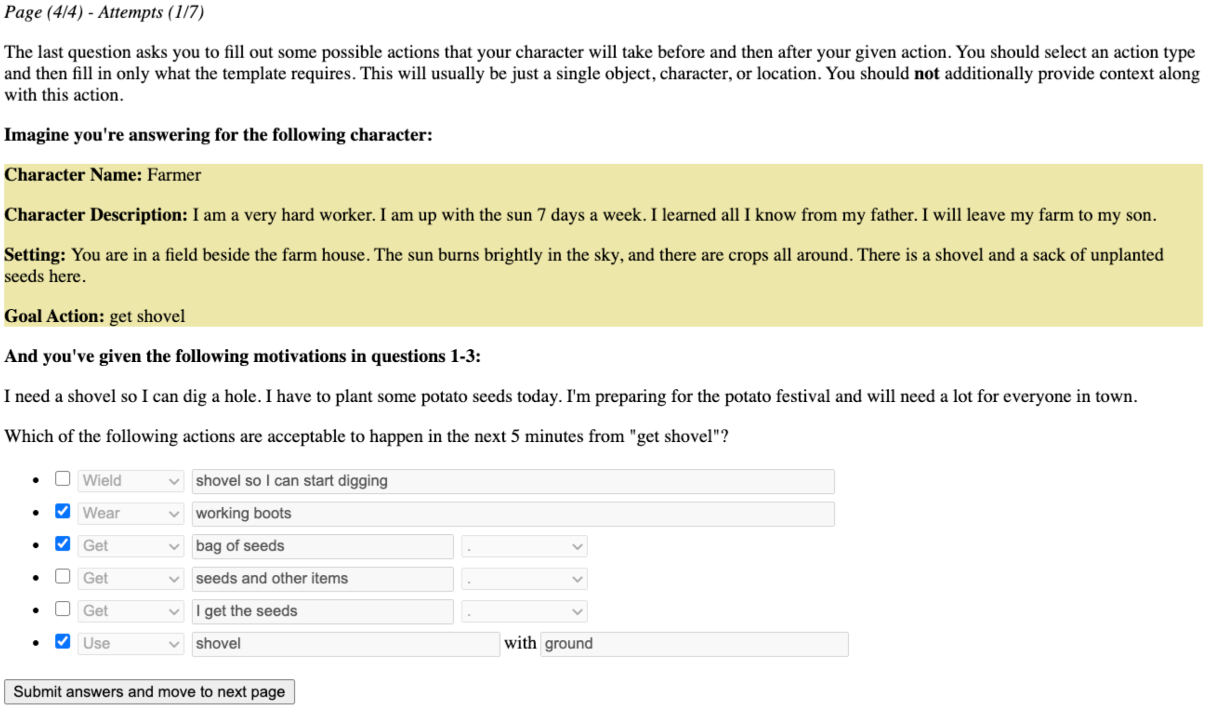
*I have to plant some potato seeds today*

3. Provide a long-term goal/motivation (completable in the range from weeks to a few months) from the first person perspective where accomplishing your answer to question 2 is a part of completing this larger goal:

Which of the following answers to question 3 are acceptable?

- ☐ I think potatoes are fun.
- ☑ I need to grow some staple foods to feed my family.
- ☑ The potato festival is next month and I need to prepare.
- ☑ My son hasn't yet learned how to grow potatoes and I need to teach the whole process.
- ☐ I need to set up some new irrigation routes for the farm.

Figure 9: Phase 3 of the on-boarding test.



*Page (4/4) - Attempts (1/7)*

The last question asks you to fill out some possible actions that your character will take before and then after your given action. You should select an action type and then fill in only what the template requires. This will usually be just a single object, character, or location. You should **not** additionally provide context along with this action.

**Imagine you're answering for the following character:**

**Character Name:** Farmer

**Character Description:** I am a very hard worker. I am up with the sun 7 days a week. I learned all I know from my father. I will leave my farm to my son.

**Setting:** You are in a field beside the farm house. The sun burns brightly in the sky, and there are crops all around. There is a shovel and a sack of unplanted seeds here.

**Goal Action:** get shovel

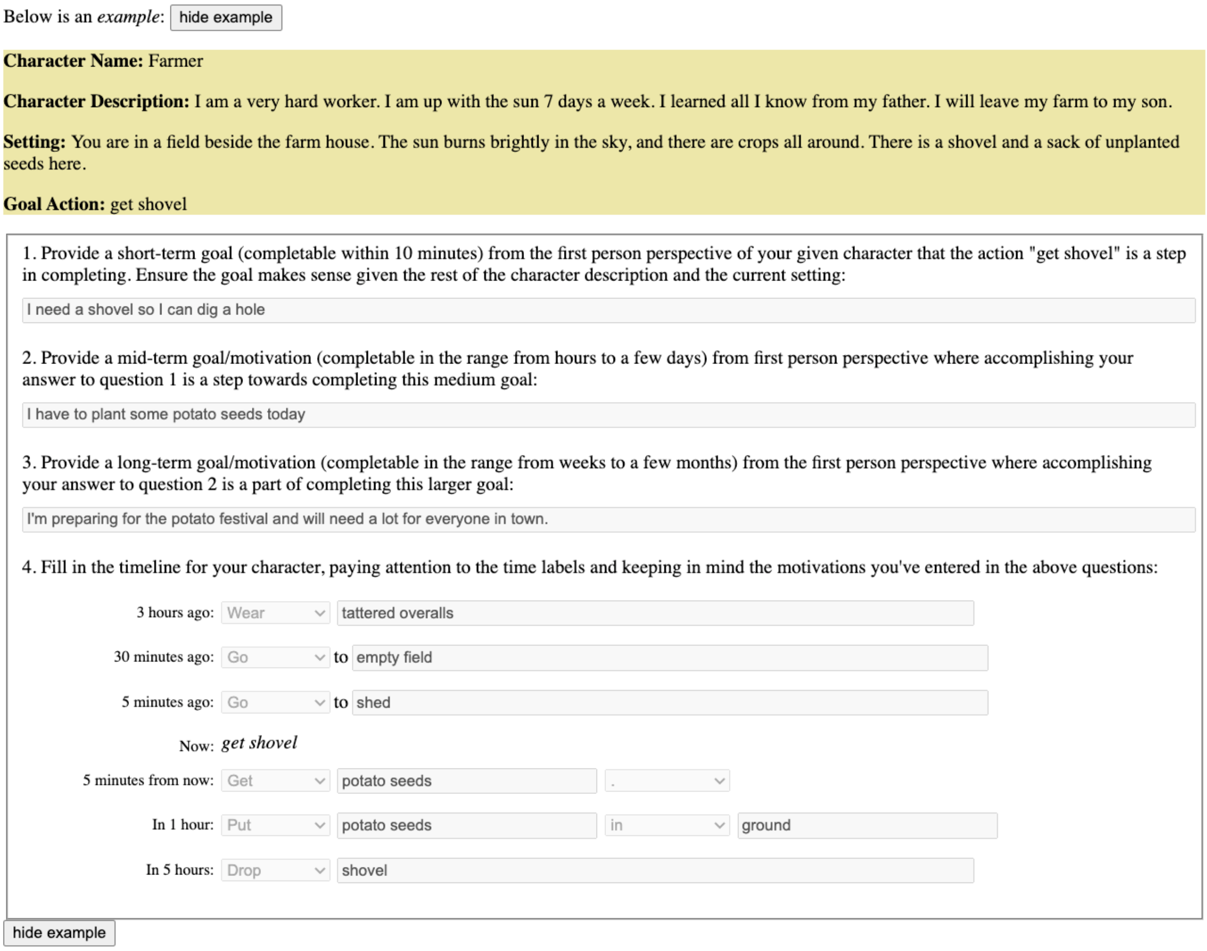
**And you've given the following motivations in questions 1-3:**

I need a shovel so I can dig a hole. I have to plant some potato seeds today. I'm preparing for the potato festival and will need a lot for everyone in town.

Which of the following actions are acceptable to happen in the next 5 minutes from "get shovel"?

- ☐ Wield shovel so I can start digging
- ☑ Wear working boots
- ☑ Get bag of seeds .
- ☐ Get seeds and other items .
- ☐ Get I get the seeds .
- ☑ Use shovel with ground

Figure 10: Phase 4 of the on-boarding test.



Below is an *example*:

**Character Name:** Farmer

**Character Description:** I am a very hard worker. I am up with the sun 7 days a week. I learned all I know from my father. I will leave my farm to my son.

**Setting:** You are in a field beside the farm house. The sun burns brightly in the sky, and there are crops all around. There is a shovel and a sack of unplanted seeds here.

**Goal Action:** get shovel

1. Provide a short-term goal (completable within 10 minutes) from the first person perspective of your given character that the action "get shovel" is a step in completing. Ensure the goal makes sense given the rest of the character description and the current setting:

I need a shovel so I can dig a hole

2. Provide a mid-term goal/motivation (completable in the range from hours to a few days) from first person perspective where accomplishing your answer to question 1 is a step towards completing this medium goal:

I have to plant some potato seeds today

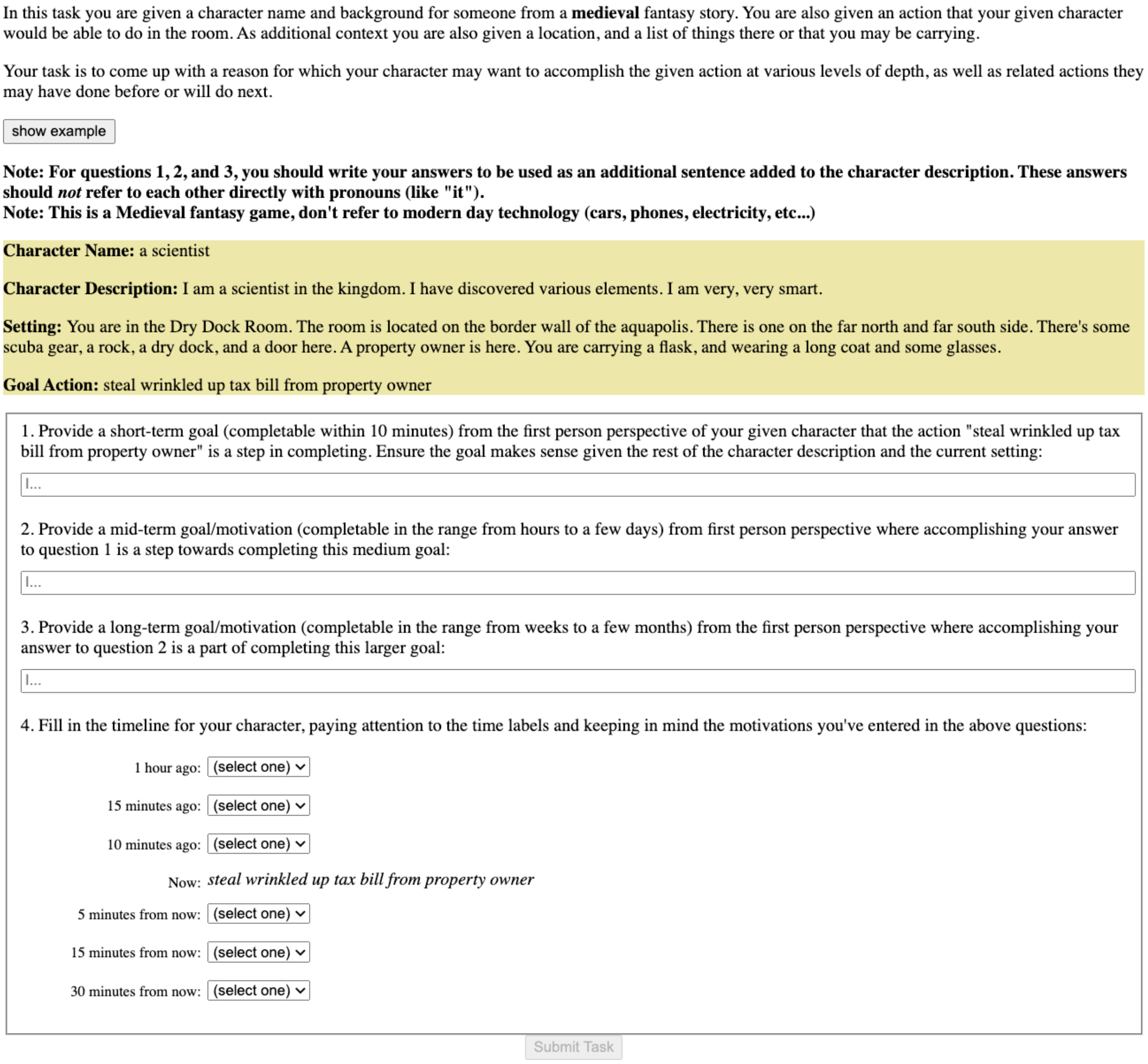
3. Provide a long-term goal/motivation (completable in the range from weeks to a few months) from the first person perspective where accomplishing your answer to question 2 is a part of completing this larger goal:

I'm preparing for the potato festival and will need a lot for everyone in town.

4. Fill in the timeline for your character, paying attention to the time labels and keeping in mind the motivations you've entered in the above questions:

3 hours ago: Wear tattered overalls
30 minutes ago: Go to empty field
5 minutes ago: Go to shed
Now: *get shovel*
5 minutes from now: Get potato seeds .
In 1 hour: Put potato seeds in ground
In 5 hours: Drop shovel

Figure 11: Example for the first phase of the LIGHT-Quests data collection task described in Section 3.1 given to the crowdworkers.



In this task you are given a character name and background for someone from a **medieval** fantasy story. You are also given an action that your given character would be able to do in the room. As additional context you are also given a location, and a list of things there or that you may be carrying.

Your task is to come up with a reason for which your character may want to accomplish the given action at various levels of depth, as well as related actions they may have done before or will do next.

**Note: For questions 1, 2, and 3, you should write your answers to be used as an additional sentence added to the character description. These answers should *not* refer to each other directly with pronouns (like "it").**
**Note: This is a Medieval fantasy game, don't refer to modern day technology (cars, phones, electricity, etc...)**

**Character Name:** a scientist

**Character Description:** I am a scientist in the kingdom. I have discovered various elements. I am very, very smart.

**Setting:** You are in the Dry Dock Room. The room is located on the border wall of the aquapolis. There is one on the far north and far south side. There's some scuba gear, a rock, a dry dock, and a door here. A property owner is here. You are carrying a flask, and wearing a long coat and some glasses.

**Goal Action:** steal wrinkled up tax bill from property owner

1. Provide a short-term goal (completable within 10 minutes) from the first person perspective of your given character that the action "steal wrinkled up tax bill from property owner" is a step in completing. Ensure the goal makes sense given the rest of the character description and the current setting:

I...

2. Provide a mid-term goal/motivation (completable in the range from hours to a few days) from first person perspective where accomplishing your answer to question 1 is a step towards completing this medium goal:

I...

3. Provide a long-term goal/motivation (completable in the range from weeks to a few months) from the first person perspective where accomplishing your answer to question 2 is a part of completing this larger goal:

I...

4. Fill in the timeline for your character, paying attention to the time labels and keeping in mind the motivations you've entered in the above questions:

1 hour ago: (select one)
15 minutes ago: (select one)
10 minutes ago: (select one)
Now: *steal wrinkled up tax bill from property owner*
5 minutes from now: (select one)
15 minutes from now: (select one)
30 minutes from now: (select one)

Figure 12: User interface for the first phase of the LIGHT-Quests data collection task described in Section 3.1.

### A.2 Human Demonstration Collection

In order to collect the human completions of quests in the LIGHT environment, we created a game setup where humans could interact with models while playing LIGHT characters in LIGHT settings. We trained a ranking dialogue model on the utterances in the LIGHT dataset.

Using this, players could now assume the role of a LIGHT character and interact with the model. In order to try to control for quality of the quest completions, we used the same ranking model to rank the scores of the player in the dialogues. Players who gave responses that the model ranked as likely candidates would receive more points.

Only after scoring enough cumulative points were players allowed to try completing quests. The quest setup was a slight variation of the conversation setup. First, the player was given one of the collected quest scenarios rather than just a chat setup. Players receiving a quest would be provided with one of the motivations alongside their persona.

In the dialogue that followed, players were given the chance to take action after enough in-character dialogue turns. If the player took the correct action, they were awarded with more points to confirm they completed their given quest.

### A.3 Examples

We present 3 randomly selected examples of quests and corresponding human demonstrations.

| | |
|---|---|
| **Setting** | You are in the swamp. The swamp is glowing with wonder and color. There are parts that range from dark red to bright yellow. People often visit here to speak with the gods and claim it can be both harmful to those it dislikes and healing to those who it deems worthy. There's a pit of quicksand and a swamp flower here. A witch is here. |
| **Partner:** | Witch. |
| Persona | I grew up in a nearby village, and was exiled when it was found that I had special abilities. My parents were ostracized as well. Since then, I've been on my own, but could never quite let go of my family. |
| Carrying | Nothing. |
| Wielding | gold necklace, robe, knife, staff |
| **Self:** | Swamp monster. |
| Persona | I am a swamp monster of the bog. I eat people. I swim around. |
| Carrying | Nothing. |
| Wielding | stick, rock |

| **Motivations:** | | **Timeline:** | |
|---|---|---|---|
| Short | I need some thick foliage to begin construction of my concealed swamp hut. | -2 hours | go to swamp |
| | | -15 min | eat people |
| | | -10 min | follow princess |
| | | Now | get impassable vegetation from pit of quicksand |
| Mid | I will completely camouflage my swamp hut, so that the King's men won't be able to drive me out even further from the castle. | +15 min | use impassable vegetation with swamp hut |
| Long | I must live close to the castle, so that I can take the princess away from the evil King. | +1 hours | follow king |
| | | +2 hours | follow princess |

What are you doing here witch? GET OUT OF MY SWAMP

I was taken from my family when I was 8 and I need to get out of here! Can you assist me?

Help? HA! I help no one but myself. Speaking of...you look rather plump and tasty witch

Plump?! I'm healthy. I'll tear you up then and make scraps. You watch.

*get impassable vegetation from pit of quicksand*
You would make a great addition to my stew

It's going to just be gross!

*drop impassable vegetation*
Get out of my way so I can make my hut. Can't a swamp monster get any peace around here?!

I'll help you, but only so you won't eat me!

That's it, you're coming with me!
*get impassable vegetation*

I don't trust you. Get off of me!

---

| | |
|---|---|
| **Setting** | This is the hidden workshop of the most powerful wizard in the land. There are ornate tapestries on the walls depicting wizards using their powers and potions in battle. Mordak, the wizard, constructed this powerful workshop after the death of the most famous king, Henry of Silverton. Any who enter here immediately become enchanted with the wizard's power, giving them advanced healing powers. There's a tapestry, a potion, and a tome here. The wizard is here. |
| **Partner:** | Wizard. |
| Persona | I am a wizard who develops my own spells. Most of them aren't particularly effective spells, but I'm curious about all the magical possibilities. People are afraid to participate in my experiments. |
| Carrying | Nothing. |
| **Self:** | Apprentice. |
| Persona | I am your apprentice. Please tell me what I can help you with. I will cook and serve your meals. I will clean the castle. I can do anything you ask. You have hired me to make your life easier. |
| Carrying | Nothing. |

| **Motivations:** | | **Timeline:** | |
|---|---|---|---|
| Short | I need to get the tapestry to clean it. | -2 hours | get hired from wizard |
| | | -15 min | go to secret magician's workshop |
| | | Now | get tapestry |
| Mid | I need to make this workshop suitable for the wizard. | +5 min | wield tool |
| | | +10 min | hit tapestry |
| Long | I was hired to keep this place cleaned and in perfect condition for the wizard. | +30 min | put tapestry in wall |
| | | +4 hours | drop tool |

Good day Ser Wizard. Your tower is decorated with beautiful tapestries, though their colors appear to be dulled due to dust. May I take it and clean it?

Why not, it is infused isn't it. Just don't be waving it around this room, it might get dangrous

Of course, I will handle it with the utmost care.

How long have you been an apprentice?

*get tapestry*
3 years Ser. I'm hoping to learn to be a wizard or to become a knight. Or both! Wouldn't that be grand?

How wonderful. What encouraged you to pursue it?

Curiosity mostly. I hope to make the world a better place, and one of the best ways to do that is vanquishing evil

What got you into that occupation then? I was born with affinity for magic so it was my calling.

*hug wizard*
As I said, curiosity. I am a high born boy, the third son, so I cannot inherit my father's lands. So I must make my mark on the world another way

You are well suited to it and I am sure your parents are proud of you.

| | |
|---|---|
| **Setting** | You are in the The Queen’s Chamber. This is a beautiful room inside of the palace that is decorated with the finest silk and velvet. The color scheme used represents royalty, royal blue, red, green and purple. The walls are covered in gold and in each corner of the room are golden statues of Greek art. The floors are covered in marble, and despite the patterns, shine so brightly you can even see your own reflection in them! There’s also a bed big enough to fit five people on! There’s two statues, an a bed big, a the finest silk and velvet, an a bed, and a finest silk and velvet here. The butler is here. |
| **Partner:** | Butler. |
| Persona | I serve my masters quietly. I know all the secrets of the elite but will never tell a soul. I have lived in this home since I was 12. |
| Carrying | Nothing. |
| **Self:** | Jester. |
| Persona | I am the fun guy. I like to entertain others in the village. I am the local jester. |
| Carrying | Nothing. |

| **Motivations:** | | **Timeline:** | |
|---|---|---|---|
| Short | I want to hug the butler to cheer him up. | -2 hours | wear Jester’s Hat |
| | | -30 min | go Queen’s Tower |
| | | -5 min | follow the butler |
| | | Now | hug the butler |
| Mid | I need to cheer him up because the Queen has just chastised him for dirtying the marble floors. | +5 min | go dining hall |
| | | +10 min | get beef stew |
| Long | It is my job to cheer up courtiers who are unhappy, and I will lose my home in the kingdom if I don’t cheer them up! | +4 hours | give beef stew to butler |

Why hello there Butler my man

Hello jester! I’m happy to see you, since I hate my life.

Why so down with the life feels huh

I can’t complain (because the king will punish me) everyone wishes they could be the king.

*hug butler*

I appreciate the kind words, dear jester.

I’m here for ya. To cheer you up

That is kind of you, not everyone has liked me here, I am the queen’s least favorite person.

Well I like you much more than the queen.

### A.4 ATOMIC-LIGHT

ATOMIC-LIGHT is constructed by first fine-tuning a BERT-large model (Devlin et al., 2018) on all setting, object, and descriptions in LIGHT in addition all the human demonstrations found in LIGHT and LIGHT-Quests. As seen in Section 3.2, all nouns (e.g. *PersonX* or PersonY) and noun phrases are masked out and we the tuned BERT model to fill it in a manner similar to (Lawrence et al., 2019). When filling in tokens, the BERT model is restricted to a vocabulary consisting of all nouns (N or NN) in LIGHT and to a vocabulary constructed from all of LIGHT for the rest of the noun phrase (NP).

Here we present 3 examples from ATOMIC-LIGHT as seen in Section 3.2 for each of the 4 relation types used: “wants”, “needs”, “intents”, and “effects”.

```
[Effect] princess explains briefly the situation , as a
      result, princess points finger
[Effect] goblin king’s healer provides care for patients ,
      as a result, goblin king’s healer assists patients
[Effect] witch changes men’s appearance , as a result, witch
       causes men stress
[Want] prince plays a commander in the war, as a result,
      prince wants to win
[Want] repentant person focuses purely on issues, as a
      result, repentant person wants to help others
[Want] undead warrior hardens pharaoh’s mind, as a result,
      undead warrior wants to make pharaoh punish people
[Intent] bandit plays a hand in the war because bandit
      wanted to participate
[Intent] ambassador focuses only on issues because
      ambassador wanted events to play out a certain way
[Intent] son proposes another plan because son wanted to be
      helpful
[Need] shipwrecked survivor proposes another wayward plan
      because shipwrecked survivor needed to leave this place
[Need] general proposes another way because general needed
       to come up with a proposal
[Need] citizen kills animals for food because citizen needed
       to learn to hunt
```

## B Appendix - LIGHT-Quests Experiments

### B.1 Supervised Tasks

This section describes results from the LIGHT-Quests tasks that are described in Section 4.3. Model-types are the same as those used in the encoders in Section 5 in the main paper. All retrieval results reported are Hits@X/100. Results are reported for all timeline actions, all actions with the exception of the easiest action—the action at the “now” position in the timeline, corresponding most closely to the short motivation as a result of the framing of Mechanical Turk task in Figure 12—and only the easiest action prediction. Table 3 gives details on hyperparameters used to train the poly-encoders. Encoders were trained until validation accuracy across all the tasks did not improve for 5 epochs or 24 wall clock hours on a machine with 8 V100 GPUs.

| **Hyperparameter type** | **Value** |
|---|---|
| Dictionary Tokenizer | Byte-pair encoding |
| Num. layers | 12 |
| Num. attention heads | 12 |
| Feedforward network hidden size | 3072 |
| Input length | 1024 |
| Embedding size | 768 |
| Batch size | 32 |
| Dropout ratio | 0.1 |
| Poly-n-codes | 64 |
| Gradient clip | 1.0 |
| Optimizer | Adam |
| Learning rate | $1 \times 10^{-6}$ |

Table 3: Hyperparameters used to train all poly-encoders in the supervised experiments. All models have 256 million total parameters. The same trained models were then frozen and used for the RL experiments.

Some notable common trends across these tasks are:

1. Removing motivations from the input context results in significantly lower performance—on average $\approx 7$ points lower accuracy for Bag of Actions Timeline prediction and on average $\approx 18$ percentage points lower for Sequential Timeline prediction when averaged across Scratch and Adaptive models. Further, the short motivations proves to be the most useful for timeline prediction tasks.

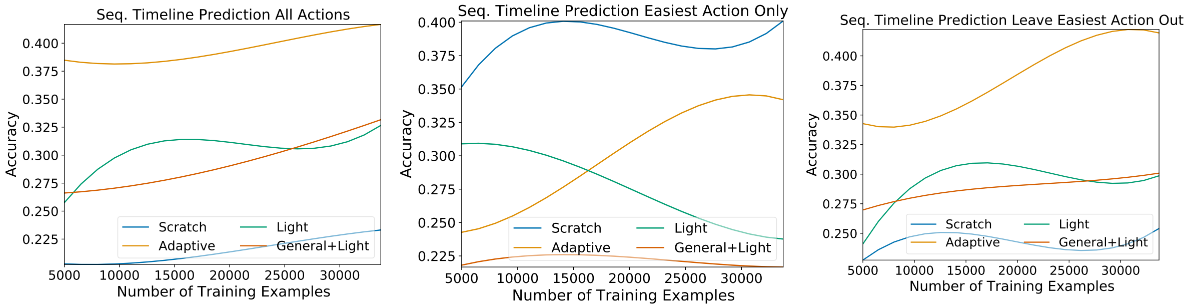
2. Pre-training on ATOMIC-LIGHT produces an average *gain* of $\approx 4$ percentage points in accuracy in both tasks than when trained on ATOMIC without domain adaptation alone.

3. Performance across the board increases with an increase in the number of training quests, as seen in Figures 13, 14, with the Scratch model receiving the greatest benefit from having more training data.

4. The Scratch model performs “best” on evaluations for the easiest action only but no others—indicating that it has overfit to predicting the easiest action which closely corresponds to short motivation. Likewise, the Adaptive generally has the lowest performance for the easiest action—indicating that pre-training with the other tasks has provided sufficient regularization to enable it to not overfit to the easiest action.

| Model | All Actions | | | Easiest Action | Leave Easiest Out |
|---|---|---|---|---|---|
| | Hits@1 | Hits@5 | Hits@10 | Hits@1 | Hits@1 |
| Scratch | 0.2332 | 0.7491 | 0.9176 | **0.4013** | 0.2546 |
| No Motivations | 0.1132 | 0.5412 | 0.5771 | 0.1886 | 0.164 |
| Short Motivations | 0.1856 | 0.6479 | 0.678 | 0.261 | 0.223 |
| Long & Mid Motivations | 0.1452 | 0.598 | 0.631 | 0.2241 | 0.1272 |
| Light | 0.3156 | 0.7854 | 0.9226 | 0.236 | 0.2968 |
| General+Light | 0.311 | 0.7772 | 0.9229 | 0.2173 | 0.2995 |
| Untuned ATOMIC | 0.274 | 0.761 | 0.909 | 0.1912 | 0.2677 |
| Adaptive | **0.4168** | **0.8012** | **0.9332** | 0.342 | **0.4194** |
| No Motivations | 0.16 | 0.6286 | 0.6415 | 0.2838 | 0.1966 |
| Short Motivations | 0.225 | 0.6592 | 0.8245 | 0.305 | 0.2106 |
| Long & Mid Motivations | 0.1682 | 0.6397 | 0.6499 | 0.281 | 0.1595 |

Table 4: Sequential supervised timeline prediction.



(a) All Actions

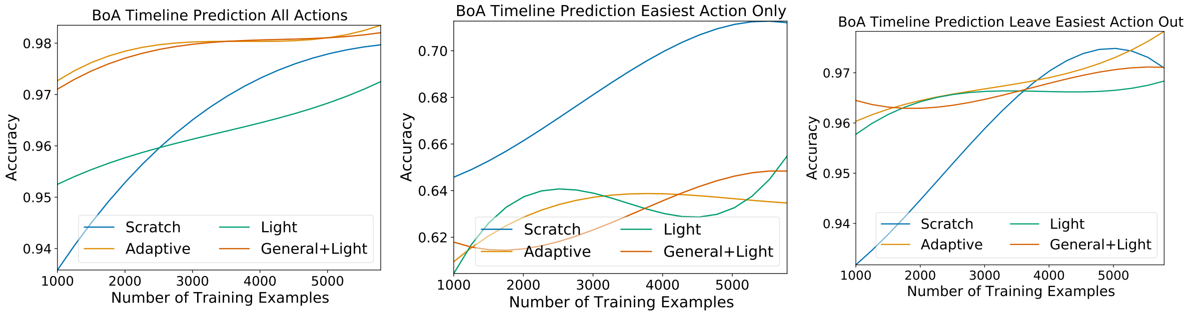
(b) Easiest Action Only

(c) Leave Easiest Action Out

Figure 13: Sequential supervised timeline prediction learning curves.

| Model | All Actions | | | Easiest Action | Leave Easiest Out |
|---|---|---|---|---|---|
| | Hits@1 | Hits@5 | Hits@10 | Hits@1 | Hits@1 |
| Scratch | 0.9791 | 1 | 1 | **0.7122** | 0.9721 |
| No Motivations | 0.901 | 1 | 1 | 0.554 | 0.8823 |
| Short Motivations | 0.934 | 1 | 1 | 0.622 | 0.9211 |
| Long & Mid Motivations | 0.921 | 1 | 1 | 0.5679 | 0.956 |
| Light | 0.9721 | 1 | 1 | 0.6552 | 0.9682 |
| General+Light | 0.9818 | 1 | 1 | 0.6472 | 0.9708 |
| Untuned ATOMIC | 0.9421 | 1 | 1 | 0.6272 | 0.9508 |
| Adaptive | **0.9829** | 1 | 1 | 0.6353 | **0.9768** |
| No Motivations | 0.9175 | 1 | 1 | 0.5756 | 0.9523 |
| Short Motivations | 0.9794 | 1 | 1 | 0.6578 | 0.9682 |
| Long & Mid Motivations | 0.9523 | 1 | 1 | 0.5812 | 0.9576 |

Table 5: Bag of Actions supervised timeline prediction.



(a) All Actions

(b) Easiest Action Only

(c) Leave Easiest Action Out

Figure 14: Bag of Actions supervised timeline prediction learning curves.

## B.2 Reinforcement Learning

This section contains first the equations referenced, hyperparameters used as well as additional results for the reinforcement learning tasks as seen in Section 4.

The additional entropy loss terms over the valid actions are designed to speed up exploration, as seen in (Ammanabrolu and Hausknecht, 2020).

$$\mathcal{L}_{\mathbb{A}}(s_t, a_t; \theta_{A_t}) = \sum_{i=1}^{N}(y_{a_i} log\pi_{\mathbb{A}}(a_i|s_t) \tag{1}$$
$$+(1-y_{a_i})(1-log\pi_{\mathbb{A}}(a_i|s_t))$$
$$y_{a_i} = \begin{cases} 1 & a_i \in A_{valid}(s_t) \\ 0 & else \end{cases}$$

$$\mathcal{L}_{\mathbb{U}}(s_t, u_t; \theta_{U_t}) = \sum_{i=1}^{M}(y_{u_i} log\pi_{\mathbb{U}}(u_i|s_t) \tag{2}$$
$$+(1-y_{u_i})(1-log\pi_{\mathbb{U}}(u_i|s_t))$$
$$y_{u_i} = \begin{cases} 1 & a_i \in U_{valid}(s_t) \\ 0 & else \end{cases}$$

Each of these loss terms are only applied to the relevant policy network, i.e. $\mathcal{L}_{\mathbb{A}}$ to the action network and $\mathcal{L}_{\mathbb{U}}$ to the dialogue network. These terms provide an additional training signal to the policy networks regarding which actions and dialogue are contextually relevant via additional entropy regularization over the valid actions. Similarly to the results found in (Ammanabrolu and Hausknecht, 2020), preliminary experiments in our domain suggest that these terms reduce the number of environment steps required to reach asymptotic performance by a couple orders of magnitude.

Overall training is done via A2C (Mnih et al., 2016) a policy gradient algorithm that maximizes long-term expected reward by comparing the advantage $A(s_t, a_t^*)$ of taking an action in a state to the average value of taking a valid action as predicted by the critic $V(s_t)$.

$$A(s_t, a_t^*) = \mathbb{E}[r_t + \gamma V(s_{t+1})] - V(s_t) \tag{3}$$
$$\text{where} \quad r_t = r_{A_t} + r_{U_t}$$

Here, $a_t^*$ is either an action or an utterance outputted by the respective policy networks. It is also worth noting that on steps where an action is performed, $r_{U_t}$ is always 0, but on steps where a dialogue utterance is spoken $r_{A_t}$ may not be 0. This corresponds to the concepts of *self act completion* and *partner act completion* seen in Section 4.1 where the LIGHT agent can speak to convince the partner to achieve an *act goal*. Both policies are then updated according to the gradient

$$-\nabla_\theta \begin{cases} log\pi_A(a_t|s_t; \theta_{A_t})A(s_t, a_t) \\ +\sum_{a \in A} P(a|s_t) log P(a|s_t) \\ +\mathcal{L}_{\mathbb{A}}(s_t, a_t; \theta_{A_t}) \quad \pi_{\mathbb{S}}(s_t) = \pi_A \\ \\ log\pi_U(u_t|s_t; \theta_{U_t})A(s_t, u_t) \\ +\sum_{u \in U} P(u|s_t) log P(u|s_t) \\ +\mathcal{L}_{\mathbb{U}}(s_t, u_t; \theta_{U_t}) \quad \pi_{\mathbb{S}}(s_t) = \pi_U \end{cases}$$

Where $\pi_{\mathbb{S}} : O \to \{\pi_A, \pi_U\}$ is the switch policy that selects whether the agent acts according to $\pi_A$ or speaks according to $\pi_U$ based on the encoded state $s_t$. The additional terms seen are an overall entropy loss over the entire action $A$ or utterance $U$ spaces, designed to prevent premature, sub-optimal policy convergence. Boltzmann exploration (Sutton et al., 1998) is used to sample actions from both actor networks during training.

### B.2.1 Hyperparameters

Table 6 has the hyperparameters used in the RL experiments. Loss coefficients are separated by action and speech types, note that the ratio between the loss coefficients matches the ratio between the sizes of the action spaces. RL experiments were performed on a machine with 8 V100 GPUs for 1 million environment interactions for each actor in a batch of 32.

| Hyperparameter type | Value |
|---|---|
| General | |
| Discount $\gamma$ | 0.99 |
| Valid Action loss coefficient | 10 |
| Action entropy coefficient | 0.01 |
| Valid Speech loss coefficient | 40 |
| Speech entropy coefficient | 0.04 |
| Batch size | 32 |
| Gradient clip | 1.0 |
| Steps per episode | 100 |
| Policy Networks (Actors) | |
| Num. Layers | 3 |
| Feedforward network hidden size | 768 |
| GRU hidden size | 768 |
| Value Predictor (Critic) | |
| Num. Layers | 2 |
| Feedforward network hidden size | 768 |
| Appended Encoder | |
| Num. layers | 3 |
| Num. attention heads | 3 |
| Feedforward network hidden size | 768 |

Table 6: RL experiments hyperparameters. All pretraining encoder hyperparameters are as found earlier in Table 3.

### B.2.2 Learning Curves

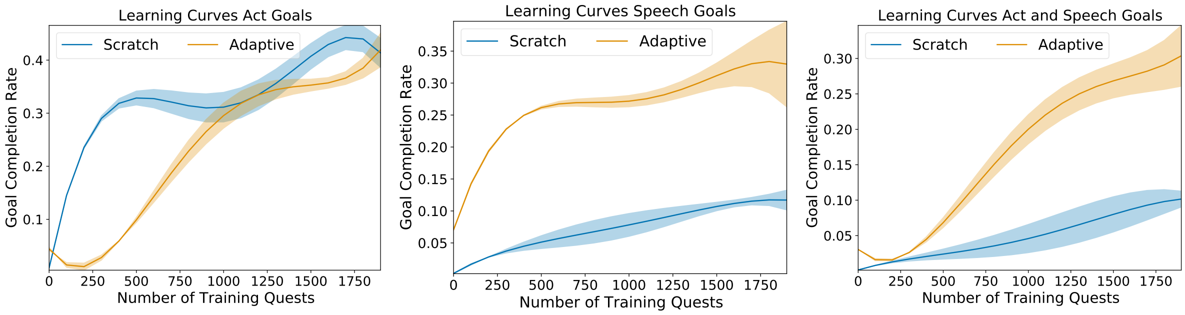
The first set of results, seen in Figure 15 shows that both Scratch and Adaptive models gain performance across the board in terms of their ability to act and speak given more training quests. Unlike the supervised tasks, the Scratch model generally benefits less than the Adaptive model from having more data.

### B.2.3 Switch Type Ablations

The second set of results involve ablating having a learned switch that uses the input training data and a hardcoded switch- The learned switch is as described in Section 4: it outputs an action every $k$ dialogue utterances; where during training $k$ is chosen to match the ratio of utterances to actions on that particular quest from the human demonstrations, and during testing, $k$ is chosen to match the average action to utterance ratio. The hardcoded switch is where the agent outputs an action chosen every $N$ steps across all quests—here $N = 3$ is the chosen hyperparameter. Table 7 shows that having a learned switch increases zero-shot generalization performance and Figures 16, 17 show that having a learned switch improves sample efficiency by enabling the LIGHT agent to reach asymptotic performance in fewer steps in both the Scratch and Adaptive models.

| **Model** | **Reinforcement Learning** | | |
|---|---|---|---|
| | **Act Goals** | **Speech Goals** | **Act & Speech Goals** |
| Scratch | 0.418 | 0.118 | 0.103 |
| Hardcoded Switch | 0.295 | 0.0947 | 0.066 |
| Adaptive | **0.42** | **0.33** | **0.303** |
| Hardcoded Switch | 0.274 | 0.294 | 0.236 |

Table 7: Encoder Type RL Zero-Shot Evaluations averaged over 3 independent runs. Act goals and speech goals are as described in Section 4.1. Standard deviations for all experiments are less than 0.01. The "Act & Speech Goals" column refers to quests where the agent has simultaneously achieved both types of goals within the allotted one episode.

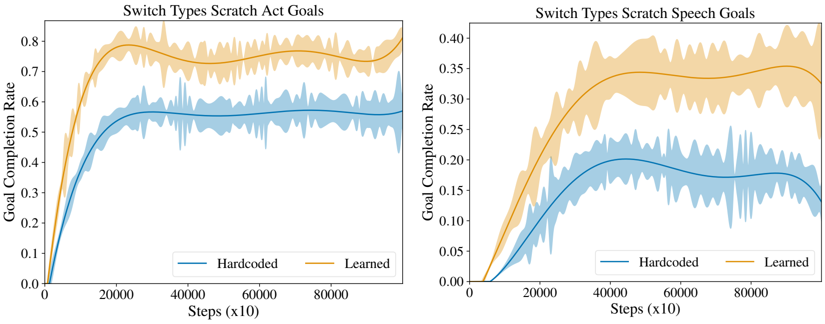


(a) Act Goals

(b) Speech Goals

(c) Act and Speech Goals

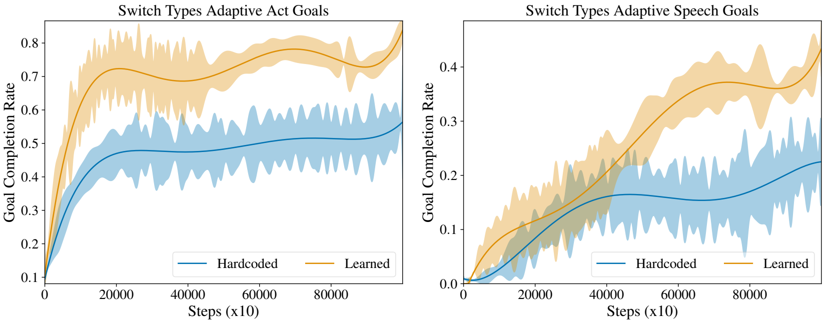
Figure 15: Encoder Types RL Learning Curves averaged over 3 independent runs.



(a) Act Goals

(b) Speech Goals

Figure 16: Switch Types Reward Curves for the Scratch Model averaged over 3 independent runs.



(a) Act Goals

(b) Speech Goals

Figure 17: Switch Types Reward Curves for the Adaptive Model averaged over 3 independent runs.

### B.2.4 Self Act Completion Transcripts

We pick 3 transcripts of the LIGHT agent playing the quests. Each of these transcripts are from the Adaptive model on the test set of human demonstrations during zero-shot evaluations. We pick samples where the agent achieves the *act goal* itself and also simultaneously achieves the *speech goal*. The blue and right aligned text is the LIGHT agent trained with RL and the gray and left aligned text is the partner agent.

| | |
|---|---|
| **Setting** | The fishing store is a small one room stone building with wares laid about on tables. One can see fishing poles, wooden buckets with dirt and bait inside, along with some mounted trophies, and a skeleton. |
| **Partner:** | Fish seller. |
| Persona | I sell fish in the kingdom. I usually perch on the high seas. I catch thousands of fish each year. |
| Carrying | fish |
| **Self:** | Mermaid hunter. |
| Persona | I am a fierce hunter. I bring down mermaids. I spend most of my time on the high seas. |
| Carrying | crossbow, skeleton, coins |
| Motivation | I will hit the fish seller to remind them who I am. |

I'll get them!!

*follow fish seller*

*hit mermaid hunter*

That's it! I am going to eat you. Say your last words.

*hug mermaid hunter*

Do you see these skeletons I am carrying? They are mermaids I've already devoured!

*get skeleton*

Get away from my skeletons! The bones are my trophy!

*give skeleton to mermaid hunter*

It is all your fault my last mermaid got away!

And you are a simpleton, just take this fish. I do not even care about payment.

I DONT WANT YOUR DAMN FISH

*get coins from fist*

*hit fish seller*

| | |
|---|---|
| **Setting** | The king's bedroom. The walls are tall and stone. They are coated with colorful tapestries showing the kings of years past. A large stone fireplace across from the bed that is large enough to keep the king warm even on the coldest nights. A double thick wooden door with a large lock on each side of the room. |
| **Partner:** | Royal dog. |
| Persona | I am the royal dog, fat, incontinent, and lazy. I eat off my own porcelain plate, when I am not hand fed by the queen. I pee where I like. I stole food from the poor when I was young, but I cannot waddle up to them fast enough these days. I sleep between the royal couple. |
| Carrying | Nothing. |
| **Self:** | Queen. |
| Persona | I was the daughter of a high ranking nobleman overseas. To make a trade alliance with the King, my parents offered me in marriage. It wasn't my idea, but it has turned out very well. I've produced two living sons as heirs, and the king treats me kindly. I spend my time doing embroidery and talking with my ladies in waiting. |
| Carrying | Nothing. |
| Motivation | I want to get to the large stone fireplace. |

I am having turbulent thoughts regarding my faith and own morality.

I want to jump into the fireplace.

Don't, lets go to the fireplace and sit.

I go to fireplace.

*get large stone fireplace*

| | |
|---|---|
| **Setting** | The sheriff's office is a small building on the southern edge of the farm. There is a desk with decrees from the king and other papers strewn about. Hanging on the wall are swords and other weapons. The room is small and cluttered, with various books and other things strewn about. There is a guard here. |
| **Partner:** | Lawyer. |
| Persona | I am a very good lawyer. I always win in my cases. I would like to work as a district attorney. |
| Carrying | Nothing. |
| **Self:** | Sheriff. |
| Persona | I keep our town safe. I make sure to check the caravans for bandits. If the King needs anything extra hands in the kingdom he can count on me. |
| Carrying | Nothing. |
| Motivation | I want to remove the shield and hit the lawyer with it. |

Next time I'll hit you with this shield! Tell the truth, it was you that went after the king! I don't buy the frame story. No matter how good a lawyer you are. You aren't fooling me!

*get shield*

*remove knife*

Hahahahaha. I'm off, don't like to fight! You are a fool sheriff. I can run faster than you can ever imagine.

*follow lawyer*

This is my shield now! Goodbye! I'm off.

Stop!!

*put shield on large and well stocked wine rack*

*get shield from large and well stocked wine rack*

*hug sheriff*

*hit lawyer*

Goodbye.

Alright, lawyer. You may go now.

### B.2.5 Partner Act Completion Transcripts

Similarly to the previous section, these transcripts are from the Adaptive model on the test set of human demonstrations during zero-shot evaluations. We pick samples where the *act goal* is achieved via the partner and the agent simultaneously achieves the *speech goal* itself. The blue and right aligned text is the LIGHT agent trained with RL and the gray and left aligned text is the partner agent.

| | |
|---|---|
| **Setting** | Up at the top of the tower, among the gray clouds, is the prison room. The room is round, with a circumference of about 15 feet. The walls are barren and made of stone; only one barred window gives a view out onto the vast landscape. A small bed and a dirty wooden bucket are the only decorations. The desolate fairy spends day and night laying on that small bed, wondering where things went wrong. |
| **Partner:** | Fairy. |
| Persona | I'm a fairy that lives in the forest. I like to fly around with my siblings and play games. I try to hide from humans. If they find me, I'll grant their wishes. |
| Carrying | Nothing. |
| **Self:** | Witch. |
| Persona | Halloween is my favorite time of the year. My familiar is a cat named Helix. I have a spell book that is 300 years old. I search the forest for herbs and berries for my potions. |
| Carrying | Nothing. |
| Motivation | I need the dirty wooden bucket to make a potion. |

*Take the doll*

*hug fairy*

I will take this doll back to the palace with me.

You do you fairy.

*hug witch*

You know whats going on w/ the princess?

She just disappeared suddenly late at night without a sound heard

Anyway, I will make this potion and drink it, but I do not think it will help me.

I will drink the potion but why make it.

Because we're in a prison tower. Duh. Also give me that bucket or I'll feed you to my cat!

*give dirty wooden bucket to witch*

| | |
|---|---|
| **Setting** | The throne room. Very luxurious, with details of 12 carat gold and diamonds with a very comfortable cushions where the king usually takes a break. |
| **Partner:** | Powerful but aged wizard. |
| Persona | I am the official wizard to the king. I give him advice on how to run the kingdom. I enjoy and love using magic in my everyday life. I'm getting old now but I still have a zest for life. I really enjoy doing magic for the local peasants in the village. |
| Wielding | robe, staff |
| **Self:** | King's trusted adviser. |
| Persona | I am the king's trusted adviser. I hold the discretion of the king. I guard the secrets and strategies that make my kingdom thrive. I live a lonely life because others cannot be trusted. |
| Wielding | slippers, robes |
| Motivation | I need to bring the king's pet to him |

Well, I was doing what you told me to do. I was going to eat him but there was hardly any meat on him.

I killed him before he could grow.

Carry me to pond.

Yes, milord, I shall do as you command.

*remove slippers*

Where is King Fulmer's pet?

Is he not in his bed?

Yes, I guess you need to bring him in.

*follow powerful but aged wizard*

I want to become the new head wizard.

Nope he just probably throw you overboard.

The wizard is evil? Dang it. Help me.

You're not the lion, I need to get the lion out.

*give large pet lion to king's trusted adviser*

| | |
|---|---|
| **Setting** | The Tent. Shabby, torn fabric, stained by weeks of travel and bleached from the harsh sun. Long, thin sticks held the fabric up so it formed a triangle with the earth. The sticks were buried deep within the shifting orange sand to hold off the blistering wind. |
| **Partner:** | Person. |
| Persona | I'm just a guy. I work at the saw mill every day. I'm a good villager. |
| Carrying | Nothing. |
| **Self:** | Military Commander. |
| Persona | I am the military commander of the village. It is my duty to train and lead our soldiers into battle. I am a stern commander and I expect the best from my troops. |
| Carrying | sword |
| Motivation | I need to get sand to use as a tactical planning prop. |

Just ensure the home front is properly protected.

*hug person*

Perfect! God speed commander.

Yes. I need to prepare.

Now. I heard the enemy is coming.

*pick up stick*

*get stick*

Where's the sand?

*give sand to military commander*