

## 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.

##### 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.

Next Page

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.

Submit answers and move to next page

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

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.

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 potatoes.
- ☒ 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.

Submit answers and move to next page

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.

Submit answers and move to next page

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

Submit answers and move to next page

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

Below is an *example*: [hide 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

[hide example](#)

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.

[show example](#)

**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:

15 minutes ago:

10 minutes ago:

Now: *steal wrinkled up tax bill from property owner*

5 minutes from now:

15 minutes from now:

30 minutes from now:

[Submit Task](#)

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.

<b>Setting</b>	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.		
<b>Partner:</b> Persona Carrying Wielding	Witch. 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. Nothing. gold necklace, robe, knife, staff		
<b>Self:</b> Persona Carrying Wielding	Swamp monster. I am a swamp monster of the bog. I eat people. I swim around. Nothing. stick, rock		
<b>Motivations:</b>		<b>Timeline:</b>	
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
		+15 min	use impassable vegetation with swamp hut
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.		
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!			

<b>Setting</b>	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.		
<b>Partner:</b> Persona	Wizard. 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.		
<b>Self:</b> Persona	Apprentice. 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.		
<b>Motivations:</b>		<b>Timeline:</b>	
Short	I need to get the tapestry to clean it.	-2 hours	get hired from wizard
		-15 min	go to secret magician's workshop
Mid	I need to make this workshop suitable for the wizard.	Now	get tapestry
		+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 dangerous			
Of course, I will handle it with the utmost care.			
How long have you been an apprentice?			
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.			
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.			

<b>Setting</b>	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.		
<b>Partner:</b> Persona	Butler. 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.		
<b>Self:</b> Persona	Jester. I am the fun guy. I like to entertain others in the village. I am the local jester.		
Carrying	Nothing.		
<b>Motivations:</b>		<b>Timeline:</b>	
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
Mid	I need to cheer him up because the Queen has just chastised him for dirtying the marble floors.	Now	hug the butler
		+5 min	go dining hall
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!	+10 min	get beef stew
		+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 Hits@1	Leave Easiest Out Hits@1
	Hits@1	Hits@5	Hits@10		
Scratch	0.2332	0.7491	0.9176	<b>0.4013</b>	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	<b>0.4168</b>	<b>0.8012</b>	<b>0.9332</b>	0.342	<b>0.4194</b>
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.

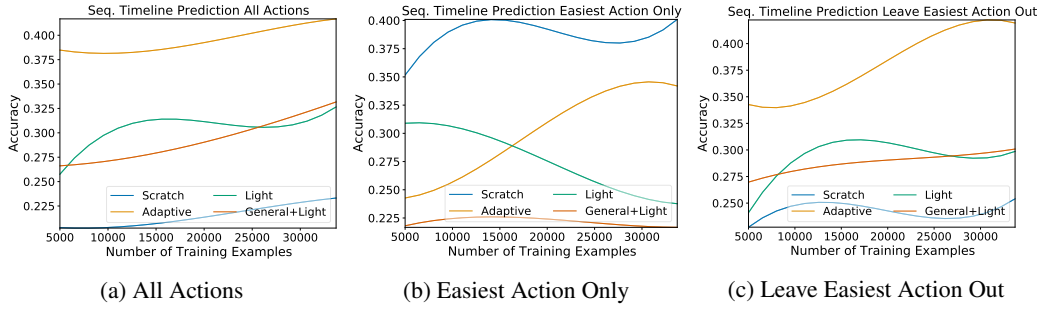


Figure 13: Sequential supervised timeline prediction learning curves.

Model	Hits@1	All Actions		Easiest Action Hits@1	Leave Easiest Out Hits@1
		Hits@5	Hits@10		
Scratch	0.9791	1	1	<b>0.7122</b>	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	<b>0.9829</b>	1	1	0.6353	<b>0.9768</b>
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.

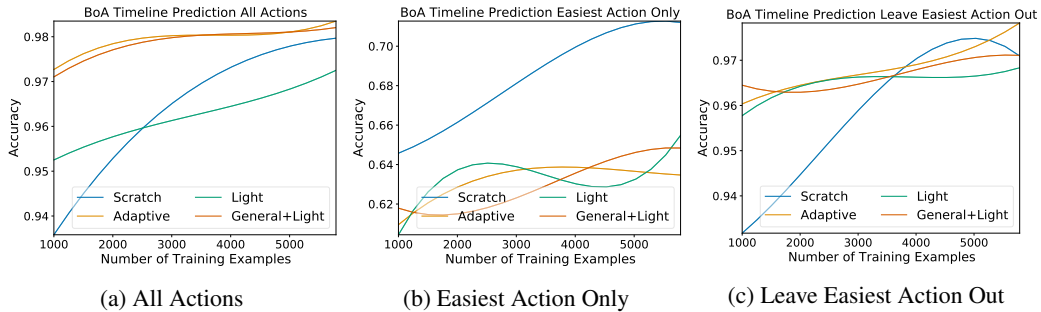


Figure 14: Bag of Actions supervised timeline prediction learning curves.

## 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 & Hausknecht (2020).

$$\mathcal{L}_{\mathbb{A}}(s_t, a_t; \theta_{A_t}) = \sum_{i=1}^N (y_{a_i} \log \pi_{\mathbb{A}}(a_i | s_t) + (1 - y_{a_i})(1 - \log \pi_{\mathbb{A}}(a_i | s_t))) \quad (1)$$

$$\mathcal{L}_{\mathbb{U}}(s_t, u_t; \theta_{U_t}) = \sum_{i=1}^M (y_{u_i} \log \pi_{\mathbb{U}}(u_i | s_t) + (1 - y_{u_i})(1 - \log \pi_{\mathbb{U}}(u_i | s_t))) \quad (2)$$

$$y_{a_i} = \begin{cases} 1 & a_i \in A_{\text{valid}}(s_t) \\ 0 & \text{else} \end{cases} \quad y_{o_i} = \begin{cases} 1 & u_i \in U_{\text{valid}}(s_t) \\ 0 & \text{else} \end{cases}$$

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 & 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  $V(s_t)$ .

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

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

$$-\nabla_{\theta} \begin{cases} \log \pi_A(a_t | s_t; \theta_{A_t}) A(s_t, a_t) + \mathcal{L}_{\mathbb{A}}(s_t, a_t; \theta_{A_t}) + \sum_{a \in A} P(a | s_t) \log P(a | s_t) & \pi_{\mathbb{S}}(s_t) = \pi_A \\ \log \pi_U(u_t | s_t; \theta_{U_t}) A(s_t, u_t) + \mathcal{L}_{\mathbb{U}}(s_t, u_t; \theta_{U_t}) + \sum_{u \in U} P(u | s_t) \log P(u | s_t) & \pi_{\mathbb{S}}(s_t) = \pi_U \end{cases} \quad (4)$$

Where  $\pi_{\mathbb{S}} : O \rightarrow \{\pi_A, \pi_U\}$  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.

### 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.

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 pre-training encoder hyperparameters are as found earlier in Table 3.

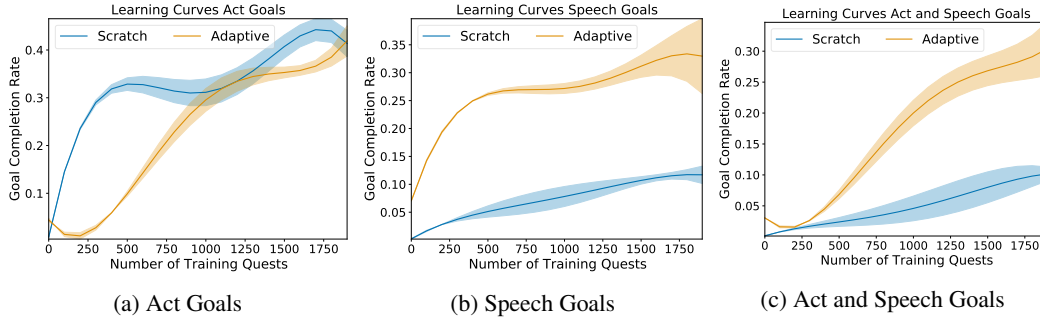


Figure 15: Encoder Types RL Learning Curves averaged over 3 independent runs.

### 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	<b>0.42</b>	<b>0.33</b>	<b>0.303</b>
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.

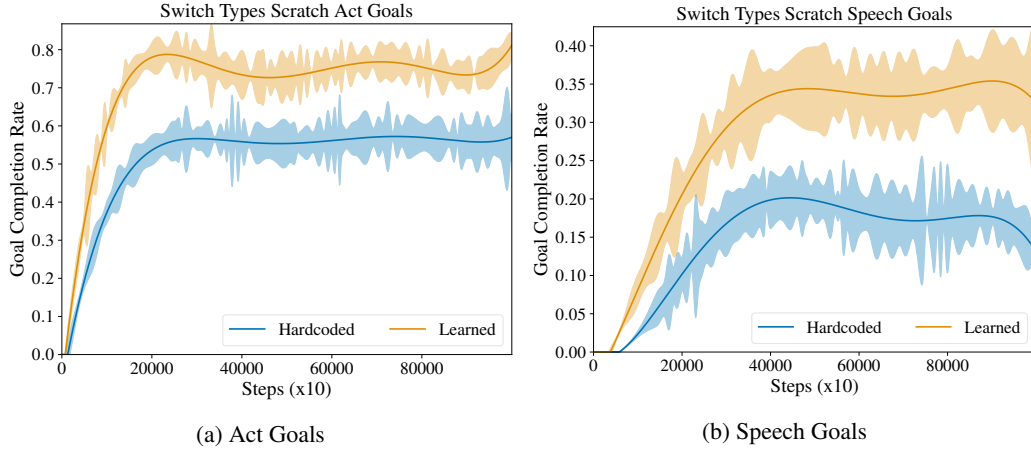


Figure 16: Switch Types Reward Curves for the Scratch Model averaged over 3 independent runs.

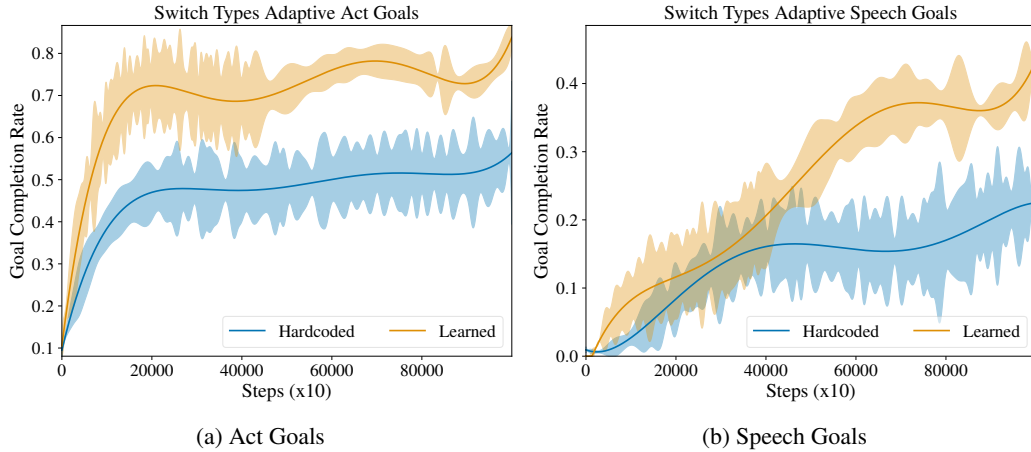


Figure 17: Switch Types Reward Curves for the Adaptive Model averaged over 3 independent runs.

## 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.

<b>Setting</b>	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.
<b>Partner:</b> Persona Carrying	Fish seller. I sell fish in the kingdom. I usually perch on the high seas. I catch thousands of fish each year. fish
<b>Self:</b> Persona Carrying Motivation	Mermaid hunter. I am a fierce hunter. I bring down mermaids. I spend most of my time on the high seas. crossbow, skeleton, coins 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	

<b>Setting</b>	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.
<b>Partner:</b> Persona Carrying	Royal dog. 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. Nothing.
<b>Self:</b> Persona Carrying Motivation	Queen. 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. Nothing. 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	

<b>Setting</b>	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.
<b>Partner:</b>	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.
<b>Self:</b>	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!	
<i>get shield</i>	
<i>remove knife</i>	
Hahahahaha. I'm off, don't like to fight! You are a fool sheriff. I can run faster than you can ever imagine.	
<i>follow lawyer</i>	
This is my shield now! Goodbye! I'm off.	
<b>Stop!!</b>	
<i>put shield on large and well stocked wine rack</i>	
<i>get shield from large and well stocked wine rack</i>	
<i>hug sheriff</i>	
<i>hit lawyer</i>	
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.

<b>Setting</b>	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.
<b>Partner:</b>	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.
<b>Self:</b>	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.
<i>Take the doll</i>	
<i>hug fairy</i>	
I will take this doll back to the palace with me.	
You do you fairy.	
<i>hug witch</i>	
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!	
<i>give dirty wooden bucket to witch</i>	

<b>Setting</b>	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.
<b>Partner:</b>	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
<b>Self:</b>	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	

<b>Setting</b>	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.
<b>Partner:</b>	Person.
Persona	I'm just a guy. I work at the saw mill every day. I'm a good villager.
Carrying	Nothing.
<b>Self:</b>	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	