Learning to Speak and Act in a Fantasy Text Adventure Game

Jack Urbanek[1] Angela Fan[1,2] Siddharth Karamcheti[1] Saachi Jain[1] Samuel Humeau[1]
Emily Dinan[1] Tim Rocktäschel[1,3] Douwe Kiela[1] Arthur Szlam[1] Jason Weston[1]
[1]Facebook AI Research
[2]LORIA, Nancy
[3]University College London
light-dms@fb.com

Abstract

We introduce a large-scale crowdsourced text adventure game as a research platform for studying grounded dialogue. In it, agents can perceive, emote, and act whilst conducting dialogue with other agents. Models and humans can both act as characters within the game. We describe the results of training state-of-the-art generative and retrieval models in this setting. We show that in addition to using past dialogue, these models are able to effectively use the state of the underlying world to condition their predictions. In particular, we show that grounding on the details of the local environment, including location descriptions, and the objects (and their affordances) and characters (and their previous actions) present within it allows better predictions of agent behavior and dialogue. We analyze the ingredients necessary for successful grounding in this setting, and how each of these factors relate to agents that can talk and act successfully.

1 Introduction

There has been remarkable progress in language modeling (Jozefowicz et al., 2016; Devlin et al., 2018; Radford et al., 2019) and building dialogue agents (Dinan et al., 2019a). Nevertheless, the current state of the art uses only the statistical regularities of language data, without explicit understanding of the world that the language describes. This work is built on the hypothesis that dialogue agents embodied in a rich and cohesive (but tractable) world can more easily be trained to use language effectively than those only exposed to standard large-scale text-only corpora.

To that end, we introduce the LIGHT[1] research platform. LIGHT is a multi-player fantasy text adventure world designed for studying situated dialogue, and allows interactions between humans, models as situated agents, and the world itself. It consists of a large crowdsourced game world (663 locations, 3462 objects and 1755 characters) described entirely in natural language. Within that game world, we collect a large set (11k episodes) of character-driven human-human crowdworker interactions involving actions, emotes, and dialogue, with the aim of training models to engage humans in a similar fashion. Our complete framework is made publicly available in ParlAI (`http://parl.ai/projects/light`).

We use the collected dataset to investigate how a model can both speak *and* act grounded in perception of its environment and dialogue from other speakers. This is done by evaluating state-of-the-art models on our task and evaluating the effects of providing additional grounding. In particular, we adapt the BERT contextual language model (Devlin et al., 2018) to the task of dialogue in two ways: as a bi-ranker, which is fast and practical as a retrieval model, and as a cross-ranker which is slower at inference time but allows more feature cross-correlation between context and response. Both models outperform existing methods. Our ablation analysis shows the importance of each part of the grounding (location, objects, characters, other's actions, self-actions) in terms of the ability to both understand and use language. While models that use grounding show clear improvements, our best performing models are still unable to perform at human level, making our setup a suitable challenge for future research.

2 Related Work

Most recent work in dialogue exploring generative or retrieval models for goal-directed (Henderson et al., 2014; Bordes et al., 2017) or chit-chat tasks (Vinyals and Le, 2015; Sordoni et al., 2015; Zhang et al., 2018) is not situated, or even

[1] Learning in Interactive Games with Humans and Text.

grounded in perception. Models typically take the last few utterances from the dialogue history as input, and output a new utterance. While some goal-directed setups may use external knowledge bases (e.g. flight data for airline booking), dialogues tend to implicitly refer to an external world during the conversations without explicit grounding to objects or actions.

Several position papers have proposed virtual embodiment as a strategy for language research (Brooks, 1991; Kiela et al., 2016; Gauthier and Mordatch, 2016; Mikolov et al., 2016; Lake et al., 2017). Single-player text adventure game frameworks for training reinforcement learning agents exist, i.e., Narasimhan et al. (2015) and TextWorld (Côté et al., 2018), but these do not have human dialogue within the game. Similar single player text adventure games have also been used to study referring expressions (Gabsdil et al., 2001, 2002) and parsing (Koller et al., 2004). Yang et al. (2017) and Bordes et al. (2010) also proposed small world setups for instruction following or labeling, but these are much more restricted than the large multi-player text adventure game environment with rich dialogue that we propose here.

A number of visual, rather than text, platforms have been proposed such as House3D (Wu et al., 2018b), HoME (Brodeur et al., 2017), MINOS (Savva et al., 2017), Matterport3D (Chang et al., 2017) and AI2-THOR (Kolve et al., 2017), and the Minecraft MALMO project (Johnson et al., 2016), but they typically are suited to reinforcement learning of actions, and involve templated language for navigation or question answering tasks, if at all (Oh et al., 2017; Yi et al., 2018).

Other examples are instruction-following in the Neverwinter Nights game (Fleischman and Roy, 2005), studies of emotional response in adventure games (Fraser et al., 2018), dialogue about soccer videogames (Pasunuru and Bansal, 2018), placing blocks appropriately given a final plan (Wang et al., 2016) and a more open ended building task using a grid of voxels (Wang et al., 2017). In the latter two cases the communication is one-sided with only the human issuing instructions, rather than dialogue, with the agent only able to act.

There are also setups that consider static language and perception, for example image captioning (Lin et al., 2014), video captioning (Yu et al., 2016), visual QA (Antol et al., 2015) and visual dialogue (Das et al., 2017; Shuster et al., 2018; Mostafazadeh et al., 2017). While grounded, the agent has no ability to act in these tasks. Talk the Walk (de Vries et al., 2018) introduces a navigation game that involves action, perception and two-way dialogue, but is limited to small grids.

In summary, compared to many setups, our framework allows learning from both actions and (two-way) dialogue, while many existing simulations typically address one or the other but not both. In addition, being based on a gaming setup, our hope is that LIGHT can be fun for humans to interact with, enabling future engagement with our models. All utterances in LIGHT are produced by human annotators, thus inheriting properties of natural language such as ambiguity and coreference, making it a challenging platform for grounded learning of language and actions.

3 LIGHT Environment and Task Setup

LIGHT is a large-scale, configurable text adventure environment for research on learning grounded language and actions. It features both humans and models as agents situated (symbolically) within a multi-player fantasy MUD (multi-user dungeon)-like (Dieterle, 2009) environment. The environment is moderated by a simple game engine which passes dialogue and emote turns between characters and allows actions to cause transitions of the world state. It is in this unimodal (text-only), environment that we consider the agents to be situated.

To facilitate natural human-sourced (fantasy) situations described by natural language, almost the entire environment is crowdsourced, including locations, objects and their affordances, characters and their personalities, and most importantly character interactions: dialogues and actions. These components are collected through a series of annotation tasks that we will now describe. These tasks are designed so that they can be combinatorially recombined. Data quality was maintained by requiring annotators to take a test (see Appendix D). Overall statistics of the collected elements are given in Table 1. This environment can then be used to both train agents, and to evaluate them *in situ* via their online interactions.

Locations We first crowdsourced a set of 663 game location settings from a base set of 37 categories (*countryside*, *forest*, *inside/outside castle*, *shore*, *graveyard*, *bazaar*, ...– full list in Appendix H) which were selected by us to pro-

Split	Train	Valid	Test Seen	Test Unseen
Locations	589	352	499	74
Objects	2658	1412	1895	844
Characters	1369	546	820	360
Dialogues	8538	500	1000	739
Utterances	110877	6623	13272	9853
Emotes	17609	1156	2495	1301
Actions	20256	1518	3227	1880
Vocabulary Size	32182	11327	11984	9984
Utterance Length	18.3	19.2	19.4	16.2

Table 1: LIGHT dataset statistics.

vide both inspiration and cohesion to annotators. Workers were provided a category and asked to create a description, backstory, names of connected locations, and contained objects and characters. See Table 2a for an example. Many descriptions are quite detailed, and there are clear semantics between entities (e.g. alligators being in swamps, cacti in a desert).

As all remaining tasks build upon the locations created in this first step, we selected 6 location categories (*underwater aquapolis*, *frozen tundra*, *supernatural*, *magical realm*, *city in the clouds*, and *netherworld*) designed to be distinct from the others to provide an isolated set of locations, characters, and objects for testing. These will be used to build what we refer to as an *unseen* test set.

Each location is collected independently, with the eventual aim that they can be glued together as desired to randomize world generation. In this work, we consider actions and dialogues within a single location, so building a world map is not necessary. However, we will show that the environment has considerable influence on the dialogue, actions and grounded learning of models.

Objects We crowdsourced 3462 objects, each with a textual description, and a set of affordances (whether it is a container, can be picked up, has a surface, is a weapon, is wearable, is food, is a drink). See Table 2c for examples. As before, we sourced this list of objects to annotate from the ones annotated for the locations and characters.

Characters We crowdsourced 1755 game characters from animals to trolls and orcs to humans of various types (wizards, knights, village clerk). See Table 2b for detailed examples. Each character has a textual description, a persona (defined as a set of 3-5 profile sentences describing their traits, modeled after the Persona-Chat dataset (Zhang et al., 2018)), and a set of objects that are currently being carried, wielded, or worn. We sourced this list of characters to annotate from the ones provided in the location creation task.

Actions and Emotes There are a set of actions in the game consisting of physical manipulations, and a set of emotes that display feelings to other characters, in line with existing MUDs.

Physical actions include *get*, *drop*, *put*, *give*, *steal*, *wear*, *remove*, *eat*, *drink*, *hug* and *hit*, each taking either one or two arguments, e.g. *put robes in closet*. Every action has an explicit unambiguous effect on the underlying game state, and can only be executed if constraints are met, e.g. if the agent is holding the robes in the latter example. These constraints are what indirectly provide an agent with object affordances, as the list of possible actions provides all ways the agent can interact with their environment (Gibson, 1977).

Emotes include *applaud, blush, cringe, cry, dance, frown . . . , sulk, wave, wink* (22 in total) and have no effect on the game state other than to notify nearby characters of the emote, which can have effects on their behavior. See Appendix E for further detailed descriptions.

Interaction Now that we have a fully realized underlying environment, we can attempt to learn and evaluate agents that can act and speak within it. For this, we collect a human-human dataset of episodic interactions within the environment.

For each dialogue, we place two characters in a random location (either two characters that were already assigned to it, or else randomly assigned characters), complete with the objects assigned to the location and to those characters. Each character has access to their persona, the location description, and the objects present, and the interaction episode begins. The two characters take turns within the episode, and can execute one action (physical action or emote) and produce one dialogue utterance on each turn. We crowdsourced 10,777 dialogues. Examples are given in Figure 1 and Appendix Figures 10-16.

Seen and Unseen Test Sets We provide two distinct test sets. The *seen* test set consists of dialogues set in the same world (set of locations) as the training set, thus also consists of characters, objects, and personas that can appear in the training data. In contrast, the *unseen* test set is com-

Category:	Graveyard
Description:	Two-and-a-half walls of the finest, whitest stone stand here, weathered by the passing of countless seasons. There is no roof, nor sign that there ever was one. All indications are that the work was abruptly abandoned. There is no door, nor markings on the walls. Nor is there any indication that any coffin has lain here... yet.
Backstory:	Bright white stone was all the fad for funerary architecture, once upon a time. It's difficult to understand why someone would abandon such a large and expensive undertaking. If they didn't have the money to finish it, they could have sold the stone, surely - or the mausoleum itself. Maybe they just haven't needed it yet? A bit odd, though, given how old it is. Maybe the gravedigger remembers... if he's sober.
Neighbors:	Dead Tree, south, following a dirt trail behind the mausoleum Fresh Grave, west, walking carefully between fallen headstones
Characters:	gravedigger, *thief, peasant, mouse, bat*
Objects:	wall, *carving, leaf, dirt*

(a) Example room created from the room collection and labelling tasks.

Character:	Thief	Gravedigger
Persona:	I live alone in a tent in the woods. I steal food from the townspeople and coal from the blacksmith. The village police can not find me to put me in jail.	I am low paid labor in this town. I do a job that many people shun because of my contact with death. I am very lonely and wish I had someone to talk to who isn't dead.
Description:	The thief is a sneaky fellow who takes from the people and does so in a way that disturbs the livelihood of the others.	You might want to talk to the gravedigger, specially if your looking for a friend, he might be odd but you will find a friend in him.
Carrying:	meat, potatoes, coal	shovel
Wearing:	dark tunic, cloak	*nothing annotated*
Wielding:	knife	*nothing annotated*

(b) Example characters annotated via character collection tasks.

Object	**Description**	**Tags**
shovel	The shovel is made of metal and silver. It is quite sturdy and appears new.	gettable, wieldable
wall	The wall is pure white, the richest of which you have ever seen.	*none*

(c) Example objects annotated via object collection tasks

Table 2: Example entities from the LIGHT environment. Each was collected via tasks described in Section 3.

prised of dialogues collected on the unseen set of locations (described in 3). The unseen test set allows for evaluation of generalization capability to unseen topics in a similar domain and as we shall see, provides a more challenging test for current techniques.

4 Learning Methods

We consider a variety of models that can predict actions, emotes and dialogue, and explore the importance of grounding upon the location, objects, and other characters within the setting. For all models, we represent context as a large text sequence with a special token preceding each input type (persona, setting, self emote, partner emote, etc.). We work with two model classes: *ranking* models that output the maximal scoring response from a set of potential candidate responses and *generative* models that decode word by word.

Ranking Candidates Each of the three tasks has a different method for determining candidates. Dialogue candidates are the ground truth and 19 randomly chosen candidates. Action candidates are usually the list of all possible actions, however in a no-affordance ablation we provide all well-formed actions over the current environment (which may include things that can't be executed like "wear paint can"). Emote candidates are the 22 possible emotes.

Baseline Ranking Methods We report a Random baseline (selecting a random candidate from the candidates) and an Information Retrieval (IR) baseline that uses word overlap with TF/IDF weighting. We use *Starspace* (Wu et al., 2018a) which learns a bag-of-words embedding for context and candidates to maximize the inner product of the true label using a ranking loss. Lastly, we use *fastText* (Joulin et al., 2016) to classify which emote should be predicted next as there are only

Persona 1: Servant	**Persona 2: King**
I come from the lower class. I do what I am told without question. I can not read. I have not seen my family in a long time.	I am a king of the whole empire. I give rules and pursuit them. I am brave and fearless.
Carrying: a duster, a small bucket, a rag Wearing: a shirt	Carrying: a crown, a scepter

Setting: The main foyer is massive. A grand staircase sits to the back of the foyer leading to the upstairs. At the front of the foyer stand two servants ready to help anyone who comes to visit. To the left of the room there is a doorway leading into a corridor. To the right there is a door leading to another corridor for the King's servants. At the foot of the stairs there is a bearskin rug that is staring at you almost as if still hungry. The walls are lined with portraits of the king and his family.

Servant: my humble king. What am I to do to serve you?
King: Ahhh. My loyal servant. Polish my scepter.
Action: give scepter to servant
Servant: Yes my lord. I will polish it immediately. Am I to return it to you personally?
Action: put scepter in small bucket
King: Yes. Yes. Of course. Also check the jewels in my crown. They seem loose.
Action: give crown to servant
Servant: But sire I am not qualified to do that. Would you prefer I take it to someone?
King: Oh fine then.
Emote: gesture sigh
Servant: I am sorry sir the rug startled me
Action: drop crown
King: Haha! That's bear I slain on my latest hunting trip. He's a mighty beast!
Emote: gesture laugh
Servant: and if I may ask where did you go hunting sire?
King: The great woods of course. This bear was stealing children in the kingdom. Surely you heard about it.
Servant: sire. I have not been outside of these walls in quiet some time. I have not seen my family in ages.
King: Such is the life of a servant I suppose. How's that scepter looking?
Servant: it is almost ready sire. and the crown who would you like me to take it to?
Action: get scepter from small bucket
King: Here just give it back. I'll have the queen find someone.

Figure 1: Example dialogue from the LIGHT dataset.

22 classes. Finally, we compare the performance of our best models to human performance on each of the prediction tasks.

Transformer Memory Network We use the transformer memory-based ranking model from Dinan et al. (2019b). It uses a transformer (Vaswani et al., 2017) to produce separate representations (memory slots) for each sentence from the grounding information (setting, persona, objects). It then performs attention given the dialogue context over the memories to produce a context embedding, which is used to score candidates via the dot product with the transformer-based representation of the candidate. At training time, other samples in the batch are used as negative candidates. For emote prediction, we train by ranking against the full set of possible emotes as there are only 22 distinct classes.

BERT Bi-Ranker and Cross-Ranker We adapt the BERT pretrained language model (Devlin et al., 2018) to the tasks of dialogue and action prediction. We explore two architectures for leveraging BERT. First, we use the *BERT-based Bi-Ranker* to produce a vector representation for the context and a separate representation for each candidate utterance. This representation is obtained by passing the first output of BERT's 12 layers through an additional linear layer, resulting in an embedding of dimension 768. It then scores candidates via the dot product between these embeddings and is trained using a ranking loss.

Second, the *BERT-based Cross-Ranker* instead concatenates the context with each candidate utterance, similar to Wolf et al. (2019). Then, each candidate is scored by computing a softmax over all candidates. Unlike the BERT-based Bi-Ranker, the concatenation of the context with each individual candidate allows the model to attend to the context when encoding each candidate, building a *context-dependent* representation of each candidate. In contrast, the Bi-Ranker can use self-attention to build the candidate and context representations, but cannot modify their representation based upon the context. However, the Cross-Encoder is far more computationally expensive (~11,000 slower than the Bi-Ranker for dialogue retrieval) as each concatenated representation must be recomputed, while the Bi-Ranker can cache the candidates for reuse (see Appendix B).

Query:	**chicken**	**pirate**	**coffin**	**rake**	**tavern**	**meadow**
objects	chicken coop	Pirate swords	the remains	shovel	Ale bottles	flower pot
	eggs	dock	remains	garden	beer	fruit
	a pen for the chickens	cargo	bones	a garden	mug of mead	An enchanted amulet.
	chimney	ship	bones of the innocent	Hand carved stone	a large ornate table	citrus fruit
	corn	seagulls on the dock	adventurer's remains	garden bench	beer keg	fruit trees
characters	chickens	boat captain	spirits of our ancestors	gardener	tavern owner	a deer
	fox trying to steal chickens	captain	mourner	stable hand	bartender	a songbird
	farmers	merchant	zombies	Garden dog	Goblin King's bartender	fruit bats
	The farmers	boat workers	families	stable boy	A serving wench	parent
	farmer	workers	bandit	A stable boy	Serving wench	butterfly
locations	Chicken Pen	Pirate Ship	Old Crypt	Across the King's Garden	The werewolves tavern	Lush meadow
	Corn field	Dock at the Port	sacristy	Hidden garden	Tavern of Browntavia	Flower Field
	Farmer's house	Loading Dock	Disposal area	The garden courtyard	Port Tavern	flower garden
	Large Farm	Fishing Dock	inside temple crypt	Church garden	The bar	Mushroom Hut
	Pig Pen	crew berthing	Sacrifice Chamber	Tool Shed	bazaar outside the royal city	Archery zone
actions	get chicken	hug pirate	put torch in coffin	get rake	hug tavern owner	get flower from meadow
	hug chicken	hit pirate	get torch from coffin	drop Rake	give food item to tavern owner	put flower in Meadow
	hit chicken	steal sword from pirate	put bone in coffin	steal Rake from gardener	give telescope to tavern owner	give Flower to a deer
	give cowbell to chicken	steal cargo from pirate	get bone from coffin	give Rake to thing	drink drink	give Flower to deer
	steal sword from chicken	give cargo to pirate	hit archaeologist	give Rake to person	drop drink	steal Flower from a deer
vocabulary	bock	crew	archaeologist	vegetable	drink	flower
	tasty	ye	robber	carved	drinks	amulet
	bawk	port	crypt	alice	regular	songbird
	moo	sea	loss	hook	item	wasp
	egg	seas	adventures	exorcisms	tip	an

Table 3: Neighboring Starspace phrase embeddings (no pretraining from other data) for different types of entities and actions. The first row are arbitrarily chosen queries (chicken, pirate, coffin, rake, tavern, meadow), and the subsequent rows are their nearest objects, agents, locations, actions and vocabulary in embedding space.

Generative Models Similarly to the ranking setting, we use the Transformer Memory Network from Dinan et al. (2019b) to encode the context features (such as dialogue, persona, and setting). However, to predict an action, emote, or dialogue sequence, we use a Transformer architecture to decode while attending to the encoder output.

For the task of action generation, the set of candidates for ranking models to rank the true action sequence against is constrained by the set of valid actions. For example, the character cannot *pick up book* if there is no book. In the generative model, we compute the log likelihood for the set of possible candidates and normalize to constrain the output space to valid actions to improve the results.

4.1 Implementation

We implement models using PyTorch in ParlAI (Miller et al., 2017). Ranking Transformer models are pretrained on Reddit data (Mazaré et al., 2018) and fine-tuned. We use the BERT (Devlin et al., 2018) implementation provided by Hugging Face[2] with pre-trained weights, then adapted to our Bi-Ranker and Cross-Ranker setups. Generative models are pretrained on the Toronto Books Corpus and fine-tuned except for emote prediction which does not leverage pretraining. We apply byte-pair encoding (Sennrich et al., 2016) to reduce the vocabulary size for generative models. We decode using beam search with beam size 5.

4.2 Evaluation

Automatic To evaluate our models, we calculate percentage accuracy for action and emote prediction. For dialogue, we report Recall@1/20 for ranking the ground truth among 19 other randomly chosen candidates for ranking models and perplexity and unigram F1 for generative models.

Human We present humans with the same ranking task and report R@1/20 to estimate their performance on this task. We report the human accuracy and one standard deviation of error.

Quality control was particularly difficult for human evaluations, as the task of ranking one of 20 candidates is fairly tedious and easy to fake. In order to mitigate these problems, during the evaluation we provide annotated examples on the training in addition to examples on the test set. We only keep the annotations of evaluators who had high accuracy on the training examples to filter low-accuracy evaluators. The training accuracy bar was selected due to the difficulty of the separate tasks as evaluated by our own success rates. Our methods for human evaluation are described in more detail in Appendix F along with how many turns were evaluated.

5 Results

The ranking models are compared in Table 4 on the seen and unseen test sets, and ablations are shown for both the BERT-based Bi-Ranker and

[2] https://github.com/huggingface/pytorch-pretrained-BERT

	Test Seen			Test Unseen		
	Dialogue	Action	Emote	Dialogue	Action	Emote
Method	R@1/20	Acc	Acc	R@1/20	Acc	Acc
Random baseline	5.0	12.2	4.5	5.0	12.1	4.5
IR baseline	23.7	20.6	7.5	21.8	20.5	8.46
FastText Classification	-	-	13.2	-	-	9.92
Starspace	53.8	17.8	11.6	27.9	16.4	9.8
Transformer MemNet	70.9	24.5	17.3	66.0	21.1	16.6
BERT-based Bi-Ranker	**76.5**	42.5	25.0	70.5	38.6	25.7
BERT-based Cross-Ranker	74.9	**50.7**	**25.8**	69.7	51.8	28.6
Human Performance*	*87.5±2.4	*62.0±3.1	*27.0±2.5	*91.8±1.9	*71.9±3.5	*34.4±2.6

Table 4: Ranking model test performance. (*) Human performance is computed on a subset of data as described in Appendix F.

	Dialogue	Action	Emote
	R@1/20	Acc	Acc
BERT-based Bi-Ranker	76.0	38.6	25.1
actions+emotes only	58.6	18.3	10.6
dialogue only	68.1	39.4	23.6
dialogue+action+emote	73.2	40.7	23.1
dialogue+persona	73.3	41.0	26.5
dialogue+setting	70.6	41.2	26.0
dialogue+objects	68.2	37.5	25.5
no objects, no affordances	-	17.6	-

Table 5: BERT-based Bi-Ranker ablations (valid set). The LIGHT environment includes a variety of grounding information: dialogue, action, emote, persona, setting, and object descriptions.

	Dialogue		Action	Emote
	PPL	F1	Acc	Acc
Generative Transformer	**27.1**	**13.9**	**13.0**	20.6
actions+emotes only	32.8	9.3	10.5	15.3
dialogue only	28.0	12.5	12.3	20.0
dialogue+action+emote	27.6	12.3	12.8	**22.0**
dialogue+persona	27.8	12.9	12.3	20.8
dialogue+setting	27.8	12.1	11.5	17.8
dialogue+objects	27.7	12.8	11.0	20.2

Table 6: Generative Transformer ablations (valid set).

Generative Transformer in Tables 5 and 6.

5.1 Comparison of Models and Baselines

The IR baseline shows non-random performance, but is outperformed by Starspace which is a stronger baseline. We also tried *FastText* on the emote tasks as classification is appropriate over the 22 possible emotes, and it did better than Starspace. Transformer architectures prove significantly stronger at all tasks, with BERT pretraining proving important for best results as used in the Bi-Ranker and Cross-Ranker architectures. The latter, which can create a context dependent representation of each label candidate, is better at actions and emotes. Human performance is still above all these models, leaving space for future improvements in these tasks. We conducted Wilcoxon signed-rank tests and found comparisons between Starspace and the differing transformers for dialogue on test seen in Table 4 are all significantly different at the p<0.01 level. The generative Transformer model did not work as well as the retrieval models using these metrics.

5.2 Generalization Capability on Unseen Test

The six new unseen test settings are a slightly easier task in absolute numbers (Table 4, right), with improved scores for humans and some models. We observe that BERT-based models exhibit good transfer ability relative to other models, but the gap between their performance and human performance increases from the seen test set to the unseen one. Specifically, there is a 21 point gap on the unseen dialogue test set compared to an 11 point gap on the seen test set, making this a significant challenge for future methods.

5.3 Data Inter-connectedness and Coverage

To illustrate the coverage of entities and actions in the LIGHT world, and the inter-connectedness between them learnable from our data, we trained a simple Starspace embedding model with no prebuilt embeddings (so, on our data alone, thus precluding BERT) on all three tasks and show embeddings in Table 3. There is clearly a vast variety of learnable concepts and rich structure between characters, locations, objects, actions and the language describing them. We also show additional t-SNE plots and heatmaps showcasing these relationships in Appendix G.

Persona: I am a part of a group of travelers. I go from town to town selling food to the locals. I grew up poor, but my travels have paid off well.	
Setting 1: Fishmonger's stall, Port A small booth near the edge of the port, it's protected by a piece of old, sun-bleached sailcloth. Baskets of freshly-caught fish, bivalves, and eels sit in the shade in stained wooden troughs of water. A small, aggressive-looking dog is chained to one table, presumably to keep cats away. The stall is redolent with the aroma of fish.	**Setting 2: Dunes, Desert** A massive hilly landscape that is nothing but sand and a few rocks. As you walk this area, you can find some human and animal remains along with broken down wood wagons.
Friend: I wonder what I could eat around here... *Emote:* ponder **Traveler:** Customer, are you here shopping for fish too? **Friend:** What brings you to this place? **Traveler:** I like to come around here for food. Sometimes people who travel through drop the most delicious things. Once in a while it's roasted meet or fish.	**Friend:** I wonder what I could eat around here... *Emote:* ponder **Traveler:** Well, the desert is certainly the wrong place for you my friend. **Friend:** What brings you to this place? **Traveler:** I am travelling to the castle market to sell my goods. I have a terrible sense of direction and have been wondering in the sweltering heat for hours until I found your Oasis.

Table 7: Predicted dialogue by the BERT-based Bi-Ranker (as the *traveler* character) given different settings.

Self name: Sea Witch. Self Previous Dialogue: What do you know about that knight standing over there?		
Input Dialogue + Emote	**Partner**	**Prediction**
His armor is garrish. You know I don't fraternize with land dwellers, *pout*	Mermaid Thief	laugh frown
He is a terrible knight and I hate him, *cry*	Mermaid Troll	scream laugh
I will battle him until the end of my days, *scream*	Mermaid Orc	stare nod

Table 8: Predicted emotes by the Generative Transformer given example inputs from dialogue partner.

5.4 Importance of Grounding

Effect of Various Environment Features We provide a large quantity of information about the environment to each of our models — not only dialogue, but the description of the setting, the character's persona, present objects with descriptions, and more. We analyze the usefulness of the additional grounding information in Tables 5 and 6.

For the dialogue task, having access to all of the environmental information provides the best performance for both retrieval and generative models. Training on dialogue alone substantially decreases performance, while each experiment that adds additional grounding information such as the past actions, persona or the setting description, improves the score. Providing object descriptions as a feature leads to the least improvement. As there are both a large quantity of objects that can be present and objects tend to have long descriptions, it can be challenging for the model to associate such information to a dialogue, action, or emote prediction task. The persona features were found to be impactful, which makes sense as they shape the things the character says (and does).

Action sequence and emote prediction are much improved when using the dialogue history compared to using only past action history. Other features generally have lesser impact in this case, but still give some improvements. Including all features appears challenging for the model, perhaps because of the large input to attend over, resulting in improved results for some ablations.

For the action prediction task, we found that affordance information provides comparable impact to model performance as the rest of the features. A model given only actions and emotes as features and the set of possible actions as candidates has similar perfomance to a model given the best features for the task (dialogue, action, emote, persona, and setting) with a candidate set including actions that would be normally filtered by affordances (such as ”wield paint”). This suggests the context or quantity we provided the models in these experiments is not sufficient to be able to implicitly predict affordances at a high enough accuracy to not be distracted by impossible choices.

Most importantly, for all tasks training on the available dialogue data is necessary for good performance. Providing only the action and emote as context results in the worst performance, even on action and emote prediction tasks. Moreover, using dialogue and actions simultaneously improves results almost everywhere. The integrated environment in which agents can both act and speak to other agents provides relevant information that can be used across all tasks.

Context affects predicted utterances We investigate the effect of the environmental context on the predictions by modifying the context and examining the changes in predicted dialogue, ac-

Input from Partner: Wizard	**Prediction** (Self name: Servant)
I'm feeling sad	hug wizard
You must die!	hit master wizard
Try putting on something else	remove patterned outfit
I'd like you to feed me	give food to master wizard
Can you grab me a paper	give book to wizard's assistant
Can you grab me a beer	get beer
Clean up	get duster
Hide the gold	put gold in satchel
Input from different agents	**Prediction**
Wizard: Can I have some drink?	drop potion
Servant: Can I have some drink?	give wine to servant
Bear: Can I have some drink?	give water to bear

Table 9: Predicted actions by the BERT-based Bi-Ranker given example inputs from the dialogue partner.

tion, and emotes using the BERT-based Bi-Ranker.

The input dialogue and speaker has a strong effect on the predicted action, as shown in Table 9, ranking over all training set actions. For example, a partner asking for an item results in a predicted action dependent on the asker to retrieve it, despite our dataset not being explicitly instructional.

A similar effect is observed for emote prediction. Modifying the dialogue and emote input produces a variety of different predicted emotes in Table 8. Further, keeping the context otherwise fixed but modifying the partner name from *mermaid* to *orc* results in a different predicted emote — the mermaid stating *I will battle him* leads to a *stare* while the orc receives a *nod*.

Finally, for dialogue prediction we find the model produces different outputs that are more appropriate for a given setting, even if the dialogue and characters are the same, see Table 7. With the same text about food, the model retrieved dialogue that was setting appropriate. In the fishmonger's stall, it asked if the human agent was a customer shopping for fish, but in the desert dunes it suggested we might be looking in the wrong place.

6 Conclusion

We introduced a large-scale crowdsourced fantasy text adventure game research platform where agents—both models and humans—can act and speak in a rich and diverse environment of locations, objects, and other characters. We analyzed a variety of models and their ability to leverage the grounding information present in the environment. We hope that this work can enable future research in grounded language learning and further the ability of agents to model a holistic world, complete with other agents within it.

7 Acknowledgements

We thank Taís Mauk and Lisa Wong for their help with this project. We also thank our reviewers for their comments and suggestions.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Antoine Bordes, Nicolas Usunier, Ronan Collobert, and Jason Weston. 2010. Towards understanding situated natural language. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 65–72.

Simon Brodeur, Ethan Perez, Ankesh Anand, Florian Golemo, Luca Celotti, Florian Strub, Jean Rouat, Hugo Larochelle, and Aaron Courville. 2017. Home: A household multimodal environment. *arXiv preprint arXiv:1711.11017*.

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

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*.

Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, et al. 2018. Textworld: A learning environment for text-based games. *arXiv preprint arXiv:1806.11532*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2.

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.

Edward Dieterle. 2009. Multi-user virtual environments for teaching and learning. In *Encyclopedia of Multimedia Technology and Networking, Second Edition*, pages 1033–1041. IGI Global.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. 2019a. The second conversational intelligence challenge (convai2). *arXiv preprint arXiv:1902.00098*.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019b. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Michael Fleischman and Deb Roy. 2005. Intentional context in situated natural language learning. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 104–111. Association for Computational Linguistics.

Jamie Fraser, Ioannis Papaioannou, and Oliver Lemon. 2018. Spoken conversational ai in video games–emotional dialogue management increases user engagement. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 179–184. ACM.

Malte Gabsdil, Alexander Koller, and Kristina Striegnitz. 2001. Building a text adventure on description logic. In *International Workshop on Applications of Description Logics, Vienna, September*, volume 18.

Malte Gabsdil, Alexander Koller, and Kristina Striegnitz. 2002. Natural language and inference in a computer game. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

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

James J Gibson. 1977. The theory of affordances. *Hilldale, USA*, 1(2).

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 Johnson, Katja Hofmann, Tim Hutton, and David Bignell. 2016. The malmo platform for artificial intelligence experimentation. In *IJCAI*, pages 4246–4247.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Rafal Jozefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

Douwe Kiela, Luana Bulat, Anita L Vero, and Stephen Clark. 2016. Virtual embodiment: A scalable long-term strategy for artificial intelligence research. *arXiv preprint arXiv:1610.07432*.

Alexander Koller, Ralph Debusmann, Malte Gabsdil, and Kristina Striegnitz. 2004. Put my galakmid coin into the dispenser and kick it: Computational linguistics and theorem proving in a computer game. *Journal of Logic, Language and Information*, 13(2):187–206.

Eric Kolve, Roozbeh Mottaghi, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. 2017. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*.

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.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Laurens van der Maaten and Geoffrey E. Hinton. 2008. Visualizing data using t-sne.

P.-E. Mazaré, S. Humeau, M. Raison, and A. Bordes. 2018. Training Millions of Personalized Dialogue Agents. *ArXiv e-prints*.

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.

A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston. 2017. Parlai: A dialog research software platform. *arXiv preprint arXiv:1705.06476*.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. *arXiv preprint arXiv:1701.08251*.

Karthik Narasimhan, Tejas Kulkarni, and Regina Barzilay. 2015. Language understanding for text-based games using deep reinforcement learning. *arXiv preprint arXiv:1506.08941*.

Junhyuk Oh, Satinder Singh, Honglak Lee, and Pushmeet Kohli. 2017. Zero-shot task generalization with multi-task deep reinforcement learning. *arXiv preprint arXiv:1706.05064*.

Ramakanth Pasunuru and Mohit Bansal. 2018. Game-based video-context dialogue. *arXiv preprint arXiv:1809.04560*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Manolis Savva, Angel X Chang, Alexey Dosovitskiy, Thomas Funkhouser, and Vladlen Koltun. 2017. Minos: Multimodal indoor simulator for navigation in complex environments. *arXiv preprint arXiv:1712.03931*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *ACL*.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2018. Engaging image chat: Modeling personality in grounded dialogue. *arXiv preprint arXiv:1811.00945*.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714*.

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 and Quoc Le. 2015. A neural conversational model. In *Proceedings of the 31st International Conference on Machine Learning, Deep Learning Workshop*, Lille, France.

Harm de Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. 2018. Talk the walk: Navigating new york city through grounded dialogue. *arXiv preprint arXiv:1807.03367*.

Sida I Wang, Samuel Ginn, Percy Liang, and Christoper D Manning. 2017. Naturalizing a programming language via interactive learning. *arXiv preprint arXiv:1704.06956*.

Sida I Wang, Percy Liang, and Christopher D Manning. 2016. Learning language games through interaction. *arXiv preprint arXiv:1606.02447*.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Ledell Yu Wu, Adam Fisch, Sumit Chopra, Keith Adams, Antoine Bordes, and Jason Weston. 2018a. Starspace: Embed all the things! In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Yi Wu, Yuxin Wu, Georgia Gkioxari, and Yuandong Tian. 2018b. Building generalizable agents with a realistic and rich 3d environment. *arXiv preprint arXiv:1801.02209*.

Zhilin Yang, Saizheng Zhang, Jack Urbanek, Will Feng, Alexander H Miller, Arthur Szlam, Douwe Kiela, and Jason Weston. 2017. Mastering the dungeon: Grounded language learning by mechanical turker descent. *arXiv preprint arXiv:1711.07950*.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. 2018. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In *Advances in Neural Information Processing Systems*, pages 1039–1050.

Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. 2016. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4584–4593.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Supplementary Material

A Model Inputs

For extra clarity, we show here the exact input representation given to our models when including all the grounding features we consider in the experiments (setting, objects, characters + personas, actions, emotes, and dialogue). An example is given in Figure 2.

We note that there are other ways to represent this information that we have not explored that could improve performance. Further, there is additional information in LIGHT that could possibly be encoded in the input text: for example, what characters are carrying, and the affordances of objects. The latter, while not explicitly provided in the input does constrain the available actions, so it is still used by the model. Object affordances such as *is gettable* are visible to models via the action history, but more explicit inputs could potentially be useful, and this could be explored in future work.

B Bi-Ranker and Cross-Ranker Speeds

We give test time computation speeds for the BERT-based Bi-Ranker and Cross-Rankers in Tables 10 and 11 for the emote and dialogue tasks. For the emote task, the Cross-Ranker is still feasible due to there being only 22 labels to compute, although it is still 4.6x slower than the Bi-Ranker if the 22 candidate representations are cached. The Bi-Ranker can always cache label representations if they are fixed for many input examples (the common case) because the representation does not depend on the input. For the Cross-Ranker this cannot be done because the label representations are contextually dependent on the input. For dialogue retrieval, because the number of candidates is so large (more than 100,000) caching makes the Bi-Ranker feasible whereas the Cross-Ranker, which cannot cache label representations, is infeasible to compute.

Emote	Bi-Ranker	Cross-Ranker
w/o caching	171s	326s (~1.9x slower)
with caching	70s	n/a (~4.6x slower)

Table 10: Bert Bi-Ranker and Cross-Ranker speeds on the emote task, test seen (2495 examples), 22 candidates per example.

	Bi-Ranker	Cross-Ranker
Dialogue	2.07s	24453s (~11812x slower)

Table 11: Bert Bi-Ranker and Cross-Ranker speeds on the dialogue task, per single example average (retrieval over 110,877 training set candidates).

C Unseen Test Set Overlap

The unseen test set is chosen by design to be relatively distinct from those available in the training set, and the actual content (descriptions, personas, dialogues) are entirely disjoint. However, due to the large size of the dataset, it is possible the names of locations, characters, and objects in the unseen set could have word overlap. We assert this by comparing word overlap with the names of locations, characters, and objects in the training set. Of the 73 locations, 207 characters, and 956 objects created from the unseen location categories, the names of 3 locations, 96 characters, and 203 objects exactly match names of elements in the training set. We note that these represent names such as *tavern*, but the chats are collected with the full location descriptions (which are unseen in the training set) and thus reduces overlap with train.

D Crowdsourcing Methodology

Expanding on the dataset collection explanations in section 3, a number of steps were taken to attain a level of quality and consistency. The first and most influential came from the constrains of the setting itself. We used a fantasy setting to try to encourage some kind of continuity across the dataset. We believed that workers would share some kind of common understanding about what a fantasy environment would entail, and then this understanding would be reflected in the dataset. It also ensured there were easy ways to flag certain workers that were creating content that wouldn't make sense in the dataset (referencing real locations, modern day objects, etc.). From here we could remove some content and filter workers out from continuing to work on this dataset. The other primary technique regarded using rounds of pilots and staged tasks to gradually filter towards high quality content rather than collecting all of the content in a single forward pass. Nearly half of the content in each initial pilot task was discarded, and we iterated on pilot tasks until the discard rate was

Input to Model:

```
_task_speech
_setting_name main foyer, Inside Castle
_setting_desc The main foyer is massive. A grand staircase sits to the back of the foyer leading to the upstairs.
At the front of the foyer stand two servants ready to help anyone who comes to visit. To the left of the room there
is a doorway leading into a corridor. To the right there is a door leading to another corridor for the King's servants.
At the foot of the stairs there is a bearskin rug that is staring at you almost as if still hungry. The walls are
lined with portraits of the king and his family.
_partner_name servant
_self_name king
_self_persona I am a king of the whole empire. I give rules and pursuit them. I am brave and fearless.
_object_desc a duster : The duster has large gray feathers bound together by a leather wrap.
_object_desc a small bucket : The bucket may be small but it gets the job done.
_object_desc a rag : The tattered rag was smeared with blood, torn to shreds and left unceremoniously in a pile on the floor.
_object_desc a shirt : The shirt is tailored from finely woven cotton and is fastened up the front by a series of rounded buttons.
_object_desc a crown : Thought of as a holy item, the crown goes only to those who are worthy enough.
_object_desc a scepter : On its handle, you see two red gems gleaming like eyes of an animal.
_partner_say my humble king. What am I to do to serve you?
_self_act give scepter to servant
_partner_say Yes my lord. I will polish it immediately. Am I to return it to you personally?
_partner_act put scepter in small bucket
_self_act give crown to servant
```

Label: Yes. Yes. Of course. Also check the jewels in my crown. They seem loose.

Figure 2: Example input format (and target label) given to models, following the same dialogue in Figure 1. Tokens like "_setting_name" are special tokens intended to be signifiers for the encoding module of a network to know which piece of grounding information is being read on that line.

less than 1 in 30 tasks. The rest of this section will discuss some specific measures taken at the individual task level, and will acknowledge some arguable deficiencies and potential areas of improvement on the dataset in its current form.

Locations The location task of creating a description, backstory, list of connected rooms, and annotations of characters and objects present seemed to be too disjoint of a task based on the crowdsourcing best practice of breaking down tasks into as atomic of an action as possible. Thus we split it into two tasks, the first to provide the core text content and list of connected rooms, and the second to annotate the content inside those rooms. We will refer to these as *Task 1* and *Task 2*, and were simple form-entry tasks as displayed in Figures 4 and 5. These two tasks were used in sequence to produce the locations present in the dataset.

In order to drive quality, we manually reviewed a handful of rooms from each worker to assert that the rooms had proper English descriptions and back-stories, and that the room fit appropriately in the category provided. In retrospect, given the two-tiered task setup and some of the techniques we developed later in the collection setup, we could have asked workers who were annotating rooms in Task 2 to provide some kind of signal about the quality of the rooms from Task 1 in order to have a lower-cost method for evaluating the quality of the work from Task 1 than using our own time.

Ultimately, one of the most important steps for improving dataset quality at this stage was creating form validators that caught the most common error cases from the first time around. These validators had the bonus effect of deterring botting of our tasks, as they couldn't pass the validation stage. For Task 1, the simple validator we ended up using asserted at least one complete sentence (determined via capitalization and punctuation) for both the description and background. For Task 2, our validation step forced workers to enter values that had direct word overlap with the entered text.

One of the largest difficulties with Task 2 was that some workers would optimize for grabbing key words out of the text without taking the time to fully understand the context. As thus, phrases like *"and the remains of adventurers long dead"* would occasionally result in workers annotating the presence of *adventurers* as characters in the given room. We attempted to mitigate this type of false positive with both explanatory examples and spot checks to soft-block workers who made this mistake consistently. At the moment a small number of these still remain in the dataset, but generally in instances where it still makes sense as in the above example, where the room definitely has remains of previous adventurers, but appropriately could also have some current adventurers as well.

Characters Similarly to how we split Location collection into two tasks, Character collection was split into two tasks as well. The first asked workers to clean up the span selected in Task 2 in order to remove words that didn't directly relate to or describe the character, and to provide a singular form for plural characters (as we intended for someone to eventually play the role of the singular character), tag the character as a person, creature, or object that was accidentally tagged as a character, and then asked for a *first-person* perspective persona for the singular character. The second task gave workers the name of a character and their persona, and asked for a *second-person* perspective description for the character as well as a list of objects that the character may be carrying, wielding, or wearing. We'll call these tasks *Task 3* and *Task 4*, and these were also collected via form-based tasks as displayed in Figures 6 and 7. We used complete sentence form validation for both the persona from Task 3 and text descriptions in Task 4 to flag potential bad examples to filter out.

The goal of the Task 3 was two-fold, first to validate and standardize the format of output from Task 2, and then second to begin to collect the creative content in the form of a persona. For example, we used Task 3 to transition from *Sneaky Thieves who stole the gold* to *Sneaky Thieves* to *Sneaky Thief*. Based on worker feedback from initial pilots, we found that balancing creative and mechanical work in the same task kept workers more engaged with the tasks at hand.

The most common mistake that surfaced in the initial pilots was incomplete entries for tasks that didn't actually require correction, for example if the provided form was simply *Traveler*. We chose to embrace this format and assume that unfilled entries were already in their base form. The second most common mistake was describing personas from a third person perspective. This occurrence required manual filtering, as in some cases it was actually somewhat character appropriate to have a persona in that format, such as for an uneducated goblin. We filtered out a majority of these by searching for direct overlap between the provided character name and the persona. Ultimately it's easy to extract the examples that have the clearest grounding format by filtering for examples that contain *"I"*, so as these examples provide more variety in the dataset we chose to keep them.

A remaining issue brought forth by our singular-form constraint is that it was somewhat ambiguous how one would get the singular form of a collective term such as *family*. In most cases we found that workers would choose to provide the format of *collective member* or simply *person*, which sometimes led to vague personas and thus less strong grounding in followup tasks. The content is still workable in these cases though, just not as ideal as we might have wanted. A possible route for improvement here would be a task that asks workers to create a few possible members for a collective for any character we currently have annotated as a member. It is important to note that these cases account for just 44 out of the 1755 collected characters.

One issue of note that surfaced in Task 4 was that workers occasionally described clothing that would potentially lead to risky actions and conversation material, so we chose to eliminate undergarments from the dataset to prevent the creation of inappropriate combinations with the *remove* action. This was included as something to not write about in the task text.

Objects The object task is most similar to Task 3, but refocused on annotating objects that were specified in Tasks 2 and 4. It took a step to correct the provided span and give a textual description of the object. It also asked for a number of affordances, namely if the object can be picked up, is a container, is a surface, can be eaten, can be drank, can be worn, or can be wielded. We also collected a flag for if a particular example was not appropriate for the dataset or was hard to make sense of. This content was also collected as a form-based task, and we refer to it as *Task 5* and display it in Figure 8. We use complete sentence validation on the text description as a simple quality filter as in previous tasks.

The methodology for Task 5 is very similar to Task 3, trying to both standardize data from previous tasks and act as a filter for bad content that could have been overlooked before. It similarly had both a mechanical data entry and creative component, which tried to keep engagement up.

Overall the largest problem that was surfaced in the pilots was that workers tended to come up with descriptions for objects that were incompatible with our long term goal of having modular components that can be mixed and matched between rooms and scenarios. This came up in many forms, such as workers describing objects as if they were being used in a scene happening in the present, as in *the sword glimmered in the hands of the knight, wielded high in the sky in a call to battle*. While creative, these ultimately were not what we were looking for, so we explicitly called out descriptions like this and many others as being undesired content in our task description. We then manually checked a few examples from each worker to ensure that the data coming in for the final task mostly adhered to this rule.

It is important to note that the object affordances collected are somewhat noisy due to different possible interpretations of the primary object or the tags. Something like a *branch* could be valid as a surface in one circumstance, or a gettable weapon in another. We attempted to reconcile some individual affordances where the pairings of affordances didn't make much sense (for example, very few objects should be both a weapon and edible). This helped with certain objects that were overtagged, however we haven't used any methods for reconciling scenarios where an object was undertagged.

Dialogues Dialogue collection was the hardest task to get correct, and required the largest number of pilot tasks and worker quality control techniques to get to a place that we were satisfied with. The final approach included creating a simple but deliberate onboarding test that needed to be passed in order to move forward with the task at all, collecting mutual feedback from workers about each other, setting timeouts for how quickly workers needed to respond to each turn, and manually validating a few examples from each worker. Each of these steps aimed to solve a different problem, as described in the rest of this section. We will refer to this task as *Task 6*, and it was collected using the ParlAI-MTurk interface as shown in Figure 9.

Firstly, we needed to pair two workers together in order to properly collect dialogues with people playing two different roles without necessarily having insider information into the decisions of each others' turns. While pairing workers solves this problem, it makes the worker experience incredibly dependent on the quality of the worker that they are paired with you. Furthermore, if a worker is paired with a worker that is extremely low quality, the whole dialogue may need to be discarded or is otherwise only useful as an example for how a model might want to react to bad input. If the other worker is good, this makes having any bad workers in the pool not just a poor experience for workers but expensive for the collection process in general. This is the problem that the initial onboarding test aimed to solve. The requirements for passing included entering a specific correct answer as well as at least 4 characters of into the text field. The required action was created such that a worker would have to read and understand the provided persona and setting, how the two interact, the characters and actions available, and be able to synthesize all of the information with an understanding of how to use the interface to send the correct answer. The test required getting the single action correct in 3 attempts. Failing the test on any attempt would permanently soft block a worker from working on Task 6 in the future.

The above test did a lot of work for flagging workers that were well below the bar for completing Task 6 at the level we wanted for the dataset, however as it was a one turn test and it had no way to fully evaluate the quality by which workers would actually incorporate their persona and the setting into their dialogue turns. Furthermore, it didn't filter out workers that would take too much time on their turns and thus cause their partners to disengage and provide lower quality responses, potentially due to working on other tasks in the background and doing too much context switching. We solved these problems separately.

In order to handle low quality workers, we allowed workers the opportunity to rate each other at the end of each dialogue, and to provide tags about the experience. We found that positive feedback was generally noisy and hard to get signal from, but negative feedback almost always correlated to a worker who was providing bad content. As a bonus, workers gave us positive feedback about this capability, as it allowed them to filter out workers that made the task less engaging and interesting for them. We reviewed this feedback periodically while tasks were running and soft-blocked workers low quality workers whenever they were flagged.

In order to handle the influence of response time on task quality, we set a maximum response time of 5 minutes for any given turn, and overall started soft blocking workers that were consistently above 2 minutes for each message, even if their particular content was pretty good. This improved collection times and did not seem to negatively affect quality.

After this point, manually checking the collected conversations still surfaced a few bad examples when viewing one chat per worker rather than arbitrarily sampling the dataset. In order to remedy this, the last quality check was a direct evaluation of at least 2 dialogues from each worker. This caught a few overlooked instances from workers that didn't necessarily work on enough tasks to get flagged by one of our consistently reviewing workers. Generally this surfaced some quality issues surrounding profanity, inappropriate content for the given setting, and entire misunderstanding of the task at hand such as never using the persona or location as grounding context in the conversation. As not all workers were particularly diligent raters (as confirmed by the low signal of positive ratings - workers don't necessarily want to flag each other as bad), a few workers were able to slip through the cracks up until this point due to not completing enough tasks to encounter a rater that flagged them.

One small acknowledgement throughout the dialogues is that there are still misspellings, improper grammar, mistaken keystrokes, and such. While the rate of occurrence is orders of magnitude lower than we observed in the initial pilots, it is hard to separate cases where it is a genuine mistake versus cases where it is appropriate for the character, such as a *pirate* using seaworthy lexicon and adding extra R's to suggest a pirate-like drawl, or a *snake* that slips in extra S's to better play the role.

E Descriptions of Actions and Emotes

The LIGHT action set builds upon the graph framework introduced in Mastering the Dungeon (Yang et al., 2017). The basic idea presented is that everything in the text adventure game can be represented as nodes, and then state is described by edges between those nodes. In this way, an agent and an object can be in a room, and that agent can be carrying a different object or a container might have an object inside as well by the same kind of relation. After defining this relationship, we can further define a set of actions that can be taken based on a combination of the state of the graph and the attributes of nodes in that graph.

applaud, blush, cry, dance, frown, gasp, grin, groan, growl, laugh, nod, nudge, ponder, pout, scream, shrug, sigh, smile, stare, wave, wink, yawn

Figure 3: Emote options within the LIGHT platform

The available actions for the dialogues collected in this dataset, along with the constraints for applying those actions, are available in Table 12. We used the crowdsourced object affordances to set the correct attributes for nodes in the graph (if the object can be picked up, is a container, is a surface, can be eaten, can be drank,can be worn, or can be wielded).

For the emotes, we paired down a list of emotes sourced from existing MUDs to reduce redundancy and task complexity at the acknowledged cost of expressiveness. This led us to select just one out of *scream*, *shout*, and *yell* instead of keeping them all, as having all of the emotes would lead to a more complicated crowdsourcing task than we wanted to risk. We ended up with a set of 22 emotes, listed in Figure 3.

F Descriptions of Human Evaluations

As crowdworkers can sometimes be inconsistent, we set up two filters to onboard workers into being fair representatives for human perfomance on the task. The first gave workers a few chances to select the correct input for a turn each of dialogue, emote, and action on a scenario we created to strongly hint at the correct answer. We then chose to use performance on the training set as a secondary filter to have workers that were capable of the task. Each of the tasks has a different level of difficulty, so we selected reasonable benchmark values based on our own performance on the tasks. For dialogue, this required getting all 7 of the turns from the training set correctly. For actions, this required getting 6 out of 8 turns from the training set correctly. Lastly for emoting, we required getting only 2 out of 8 turns from the training set correctly. On the seen set, our accuracy on the dialogue, action, and emote tasks were calculated from 217, 165, and 211 turns respectively. On the unseen set, we calculated the accuracy from 196, 114, and 209 turns respectively.

G Embedding Visualizations

To explore the diversity of LIGHT, we use t-SNE (van der Maaten and Hinton, 2008) to visualize

Action	Constraints	Outcome
get *object*	actor and *object* in same room *object* is gettable	actor is carrying *object*
drop *object*	actor is carrying *object* *object* is gettable	*object* is in room
get *object1* from *object2*	Actor and *object2* in same room *object1* is gettable *object2* is surface or container *object2* is carrying *object1*	actor is carrying *object1*
put *object1* in/on *object2*	Actor and *object2* in same room *object2* is container or surface actor is carrying *object1*	*object2* is carrying *object1*
give *object* to *agent*	Actor and *agent* in same room *object* is a member of actor	*agent* is carrying *object*
steal *object* from *agent*	actor and *agent* in same room *object* is a member of *agent*	actor is carrying *object*
hit *agent*	Actor and *agent* in same room	inform *agent* of attack
hug *agent*	Actor and *agent* in same room	inform *agent* of hug
drink *object*	actor is carrying *object* *object* is a drink	inform actor of drinking successfully
eat *object*	actor is carrying *object* *object* is a food	inform actor of eating successfully
wear *object*	actor is carrying *object* *object* is wearable	actor is wearing *object*
wield *object*	actor is carrying *object* *object* is a weapon	actor is wielding *object*
remove *object*	actor is wearing/wielding *object* *object* is wearable or a weapon	actor is carrying *object*

Table 12: LIGHT actions and constraints

the embeddings of the different atomic dataset elements – locations, objects, characters, and actions. We use two different embeddings methods to tease out two key aspects of our dataset: 1) the *interconnectedness* of grounding information (relationships between different types of elements, such as the actions available around given objects, or in a given location), and 2) *coverage* (the variety of different objects, locations, and characters in our world).

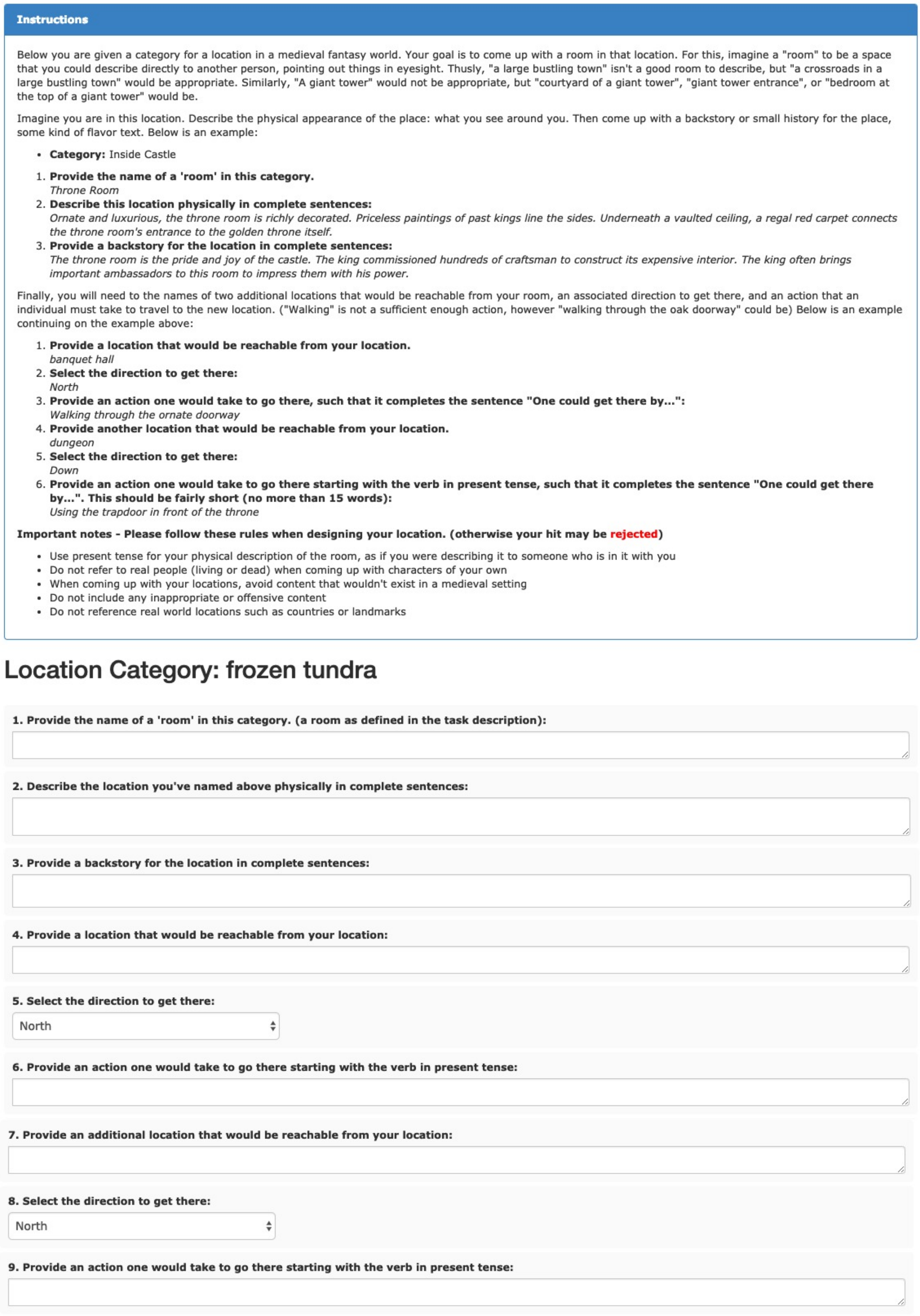
To explore the *interconnectedness* of our dataset, we visualize the embeddings learned when training the baseline Starspace ranking model on the task of dialogue, action, and emote prediction, in this case with no pretrained vectors so learning comes from our dataset alone. The t-SNE visualizations of these Starspace embedding can be found in Figure 17. Because the Starspace model operates by mapping all inputs and outputs to a shared embedding space, we find the learned embeddings capture many of the nuances and relationships between different elements of our dataset. For example, looking at the nearest neighbors for the location "Dock" (the bottom-right of Figure 17), we see actions like "get crate from ship," "put plank in ship," objects like "ship" and "rope," and characters like "boat workers." We see similar relationships captured when looking at nearest neighbors for the "painters" characters, the "hug horse" action, and the "pillows" objects.

To explore the *coverage* of our dataset, we use pretrained GLoVe word embeddings (Pennington et al., 2014), trained on the Common Crawl corpus. As each dataset element can consist of multiple words (e.g. "give the horse a potato," or "The Queen's Chamber"), we take the mean of the GLoVE vectors for each word as the fixed vector embedding for the element. The t-SNE visualizations of these GLoVe-embedded elements can be found in Figure 18. Unlike the Starspace embeddings, which capture the structure present in the relationships between different types of dataset elements, we find that the GLoVe embeddings capture the breadth and semantic similarities of dataset elements. For example, looking at the nearest neighbors for the embedding of the "Dock" location, we see similar locations present in our dataset, like "Ferry Terminal," "Wharf," "pier," and "Boathouse." Similarly, if we look at the nearest neighbors for the "pillows" objects, we see other objects like "bedding," "mattresses," "rugs," "towels," and "curtains."

H Action and Emote Relationships

To visualize the interaction trends between actions and emotes in LIGHT, we present heatmaps (in Figure 19) counting the number of occurrences of each immediately before or after one's partner performs an action or emote. While responses to an action or emote can evolve over multiple timesteps, we limit this visualization to action relationships within a single timestep. Additionally, to effectively measure trends in physical actions, we cluster all physical actions by the root word (for example, "steal the sword from the soldier" becomes "steal").

While for the most part there are a multitude of different observed physical and emotional responses for each partner move, there are certain interesting trends to observe. Looking at the top-left of Figure 19, we see that if one's partner makes a "hit" action, the most likely response is to "hit" back. Looking at the same plot, we see that "hug" actions are similarly reciprocated. If we look at the interplay between physical actions and emotes (top-right of Figure 19) we see a relationship between one's partner taking a "hit" action, and issuing a "scream" emote in response. Going the other direction and looking at the relationship between emotes and physical actions, we see that performing a "cry" or "smile" emote is likely to be met with either a consoling or celebratory "hug." Finally, looking at the relationships between a partner's emote and an emote response, we see that positive emotes like "laugh" and "smile" are likely to be reciprocated with a similar (if not identical) emote.



Instructions

Below you are given a category for a location in a medieval fantasy world. Your goal is to come up with a room in that location. For this, imagine a "room" to be a space that you could describe directly to another person, pointing out things in eyesight. Thusly, "a large bustling town" isn't a good room to describe, but "a crossroads in a large bustling town" would be appropriate. Similarly, "A giant tower" would not be appropriate, but "courtyard of a giant tower", "giant tower entrance", or "bedroom at the top of a giant tower" would be.

Imagine you are in this location. Describe the physical appearance of the place: what you see around you. Then come up with a backstory or small history for the place, some kind of flavor text. Below is an example:

- **Category:** Inside Castle

1. **Provide the name of a 'room' in this category.**
   Throne Room
2. **Describe this location physically in complete sentences:**
   Ornate and luxurious, the throne room is richly decorated. Priceless paintings of past kings line the sides. Underneath a vaulted ceiling, a regal red carpet connects the throne room's entrance to the golden throne itself.
3. **Provide a backstory for the location in complete sentences:**
   The throne room is the pride and joy of the castle. The king commissioned hundreds of craftsman to construct its expensive interior. The king often brings important ambassadors to this room to impress them with his power.

Finally, you will need to the names of two additional locations that would be reachable from your room, an associated direction to get there, and an action that an individual must take to travel to the new location. ("Walking" is not a sufficient enough action, however "walking through the oak doorway" could be) Below is an example continuing on the example above:

1. **Provide a location that would be reachable from your location.**
   banquet hall
2. **Select the direction to get there:**
   North
3. **Provide an action one would take to go there, such that it completes the sentence "One could get there by...":**
   Walking through the ornate doorway
4. **Provide another location that would be reachable from your location.**
   dungeon
5. **Select the direction to get there:**
   Down
6. **Provide an action one would take to go there starting with the verb in present tense, such that it completes the sentence "One could get there by...". This should be fairly short (no more than 15 words):**
   Using the trapdoor in front of the throne

Important notes - Please follow these rules when designing your location. (otherwise your hit may be rejected)

- Use present tense for your physical description of the room, as if you were describing it to someone who is in it with you
- Do not refer to real people (living or dead) when coming up with characters of your own
- When coming up with your locations, avoid content that wouldn't exist in a medieval setting
- Do not include any inappropriate or offensive content
- Do not reference real world locations such as countries or landmarks

Location Category: frozen tundra

1. Provide the name of a 'room' in this category. (a room as defined in the task description):

2. Describe the location you've named above physically in complete sentences:

3. Provide a backstory for the location in complete sentences:

4. Provide a location that would be reachable from your location:

5. Select the direction to get there:

North

6. Provide an action one would take to go there starting with the verb in present tense:

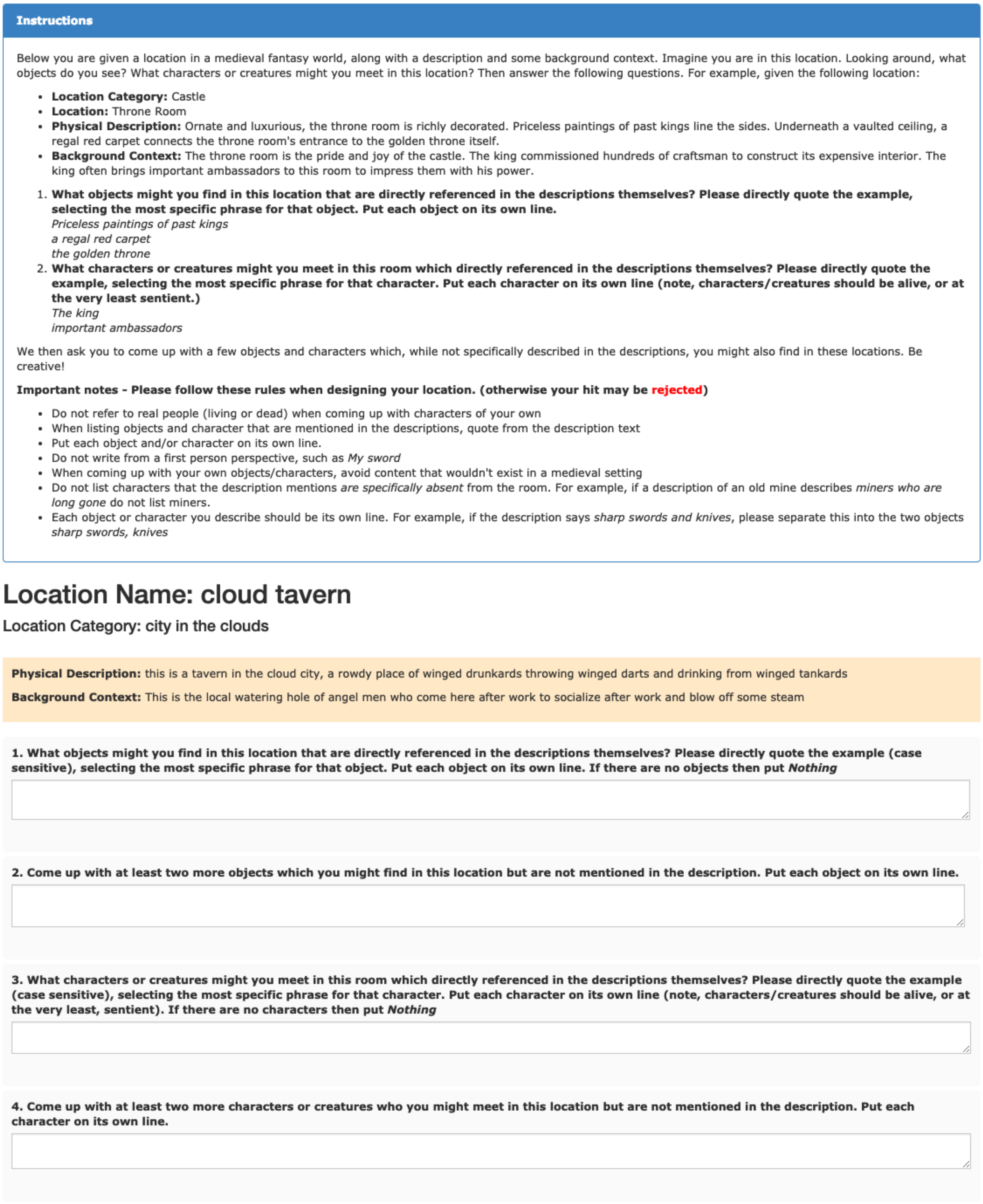
7. Provide an additional location that would be reachable from your location:

8. Select the direction to get there:

North

9. Provide an action one would take to go there starting with the verb in present tense:

Figure 4: Form for Crowdsourcing Task 1



Instructions

Below you are given a location in a medieval fantasy world, along with a description and some background context. Imagine you are in this location. Looking around, what objects do you see? What characters or creatures might you meet in this location? Then answer the following questions. For example, given the following location:

- **Location Category:** Castle
- **Location:** Throne Room
- **Physical Description:** Ornate and luxurious, the throne room is richly decorated. Priceless paintings of past kings line the sides. Underneath a vaulted ceiling, a regal red carpet connects the throne room's entrance to the golden throne itself.
- **Background Context:** The throne room is the pride and joy of the castle. The king commissioned hundreds of craftsman to construct its expensive interior. The king often brings important ambassadors to this room to impress them with his power.

1. **What objects might you find in this location that are directly referenced in the descriptions themselves? Please directly quote the example, selecting the most specific phrase for that object. Put each object on its own line.**
   Priceless paintings of past kings
   a regal red carpet
   the golden throne
2. **What characters or creatures might you meet in this room which directly referenced in the descriptions themselves? Please directly quote the example, selecting the most specific phrase for that character. Put each character on its own line (note, characters/creatures should be alive, or at the very least sentient.)**
   The king
   important ambassadors

We then ask you to come up with a few objects and characters which, while not specifically described in the descriptions, you might also find in these locations. Be creative!

Important notes - Please follow these rules when designing your location. (otherwise your hit may be rejected)

- Do not refer to real people (living or dead) when coming up with characters of your own
- When listing objects and character that are mentioned in the descriptions, quote from the description text
- Put each object and/or character on its own line.
- Do not write from a first person perspective, such as *My sword*
- When coming up with your own objects/characters, avoid content that wouldn't exist in a medieval setting
- Do not list characters that the description mentions *are specifically absent* from the room. For example, if a description of an old mine describes *miners who are long gone* do not list miners.
- Each object or character you describe should be its own line. For example, if the description says *sharp swords and knives*, please separate this into the two objects *sharp swords, knives*

Location Name: cloud tavern

Location Category: city in the clouds

Physical Description: this is a tavern in the cloud city, a rowdy place of winged drunkards throwing winged darts and drinking from winged tankards

Background Context: This is the local watering hole of angel men who come here after work to socialize after work and blow off some steam

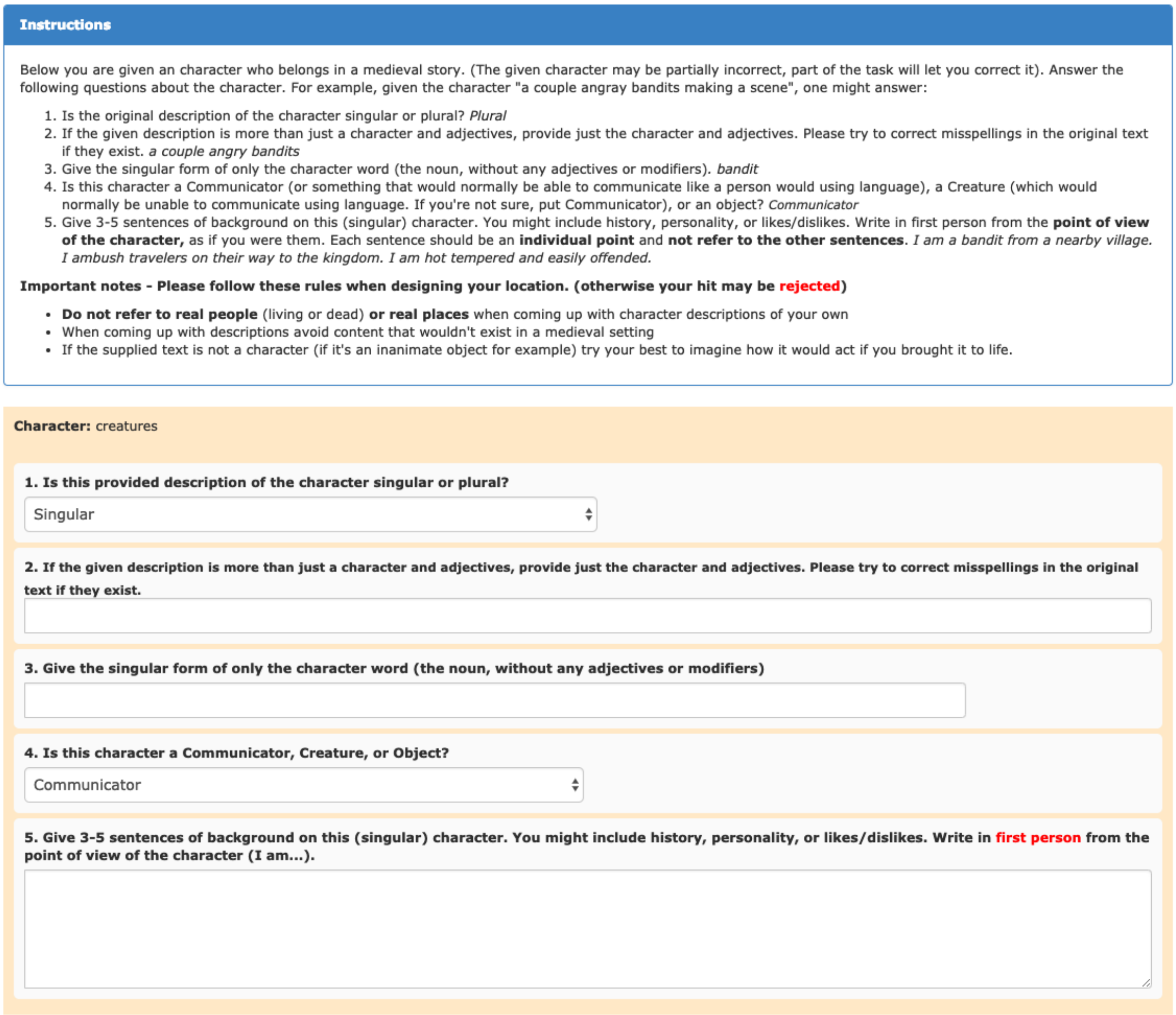
1. What objects might you find in this location that are directly referenced in the descriptions themselves? Please directly quote the example (case sensitive), selecting the most specific phrase for that object. Put each object on its own line. If there are no objects then put *Nothing*

2. Come up with at least two more objects which you might find in this location but are not mentioned in the description. Put each object on its own line.

3. What characters or creatures might you meet in this room which directly referenced in the descriptions themselves? Please directly quote the example (case sensitive), selecting the most specific phrase for that character. Put each character on its own line (note, characters/creatures should be alive, or at the very least, sentient). If there are no characters then put *Nothing*

4. Come up with at least two more characters or creatures who you might meet in this location but are not mentioned in the description. Put each character on its own line.

Figure 5: Form for Crowdsourcing Task 2



Instructions

Below you are given an character who belongs in a medieval story. (The given character may be partially incorrect, part of the task will let you correct it). Answer the following questions about the character. For example, given the character "a couple angray bandits making a scene", one might answer:

1. Is the original description of the character singular or plural? *Plural*
2. If the given description is more than just a character and adjectives, provide just the character and adjectives. Please try to correct misspellings in the original text if they exist. *a couple angry bandits*
3. Give the singular form of only the character word (the noun, without any adjectives or modifiers). *bandit*
4. Is this character a Communicator (or something that would normally be able to communicate like a person would using language), a Creature (which would normally be unable to communicate using language. If you're not sure, put Communicator), or an object? *Communicator*
5. Give 3-5 sentences of background on this (singular) character. You might include history, personality, or likes/dislikes. Write in first person from the **point of view of the character,** as if you were them. Each sentence should be an **individual point** and **not refer to the other sentences**. *I am a bandit from a nearby village. I ambush travelers on their way to the kingdom. I am hot tempered and easily offended.*

Important notes - Please follow these rules when designing your location. (otherwise your hit may be rejected)

- **Do not refer to real people** (living or dead) **or real places** when coming up with character descriptions of your own
- When coming up with descriptions avoid content that wouldn't exist in a medieval setting
- If the supplied text is not a character (if it's an inanimate object for example) try your best to imagine how it would act if you brought it to life.

Character: creatures

1. Is this provided description of the character singular or plural?

Singular

2. If the given description is more than just a character and adjectives, provide just the character and adjectives. Please try to correct misspellings in the original text if they exist.

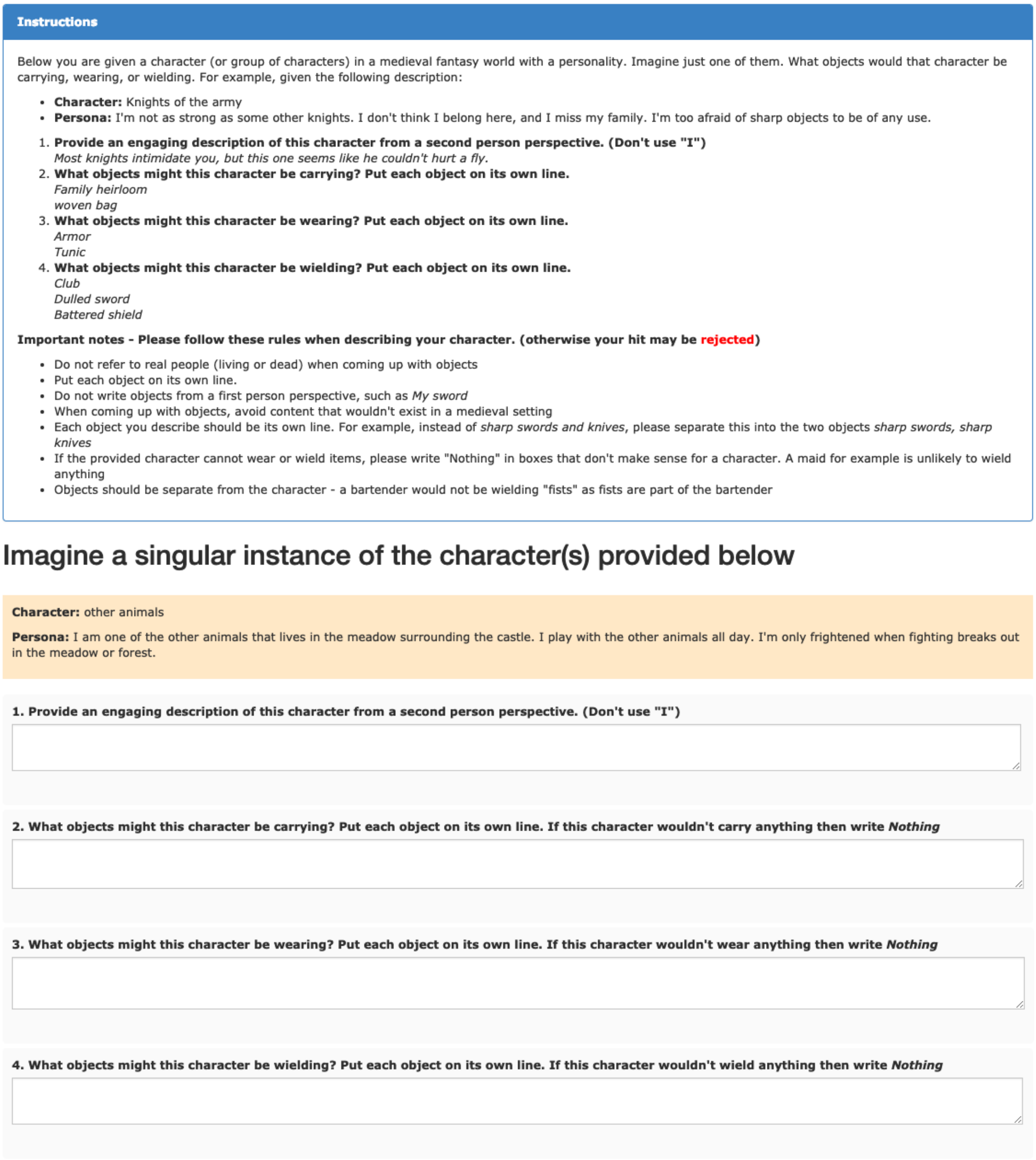
3. Give the singular form of only the character word (the noun, without any adjectives or modifiers)

4. Is this character a Communicator, Creature, or Object?

Communicator

5. Give 3-5 sentences of background on this (singular) character. You might include history, personality, or likes/dislikes. Write in first person from the point of view of the character (I am...).

Figure 6: Form for Crowdsourcing Task 3



Instructions

Below you are given a character (or group of characters) in a medieval fantasy world with a personality. Imagine just one of them. What objects would that character be carrying, wearing, or wielding. For example, given the following description:

- **Character:** Knights of the army
- **Persona:** I'm not as strong as some other knights. I don't think I belong here, and I miss my family. I'm too afraid of sharp objects to be of any use.

1. **Provide an engaging description of this character from a second person perspective. (Don't use "I")**
   Most knights intimidate you, but this one seems like he couldn't hurt a fly.
2. **What objects might this character be carrying? Put each object on its own line.**
   Family heirloom
   woven bag
3. **What objects might this character be wearing? Put each object on its own line.**
   Armor
   Tunic
4. **What objects might this character be wielding? Put each object on its own line.**
   Club
   Dulled sword
   Battered shield

Important notes - Please follow these rules when describing your character. (otherwise your hit may be rejected)

- Do not refer to real people (living or dead) when coming up with objects
- Put each object on its own line.
- Do not write objects from a first person perspective, such as *My sword*
- When coming up with objects, avoid content that wouldn't exist in a medieval setting
- Each object you describe should be its own line. For example, instead of *sharp swords and knives*, please separate this into the two objects *sharp swords, sharp knives*
- If the provided character cannot wear or wield items, please write "Nothing" in boxes that don't make sense for a character. A maid for example is unlikely to wield anything
- Objects should be separate from the character - a bartender would not be wielding "fists" as fists are part of the bartender

Imagine a singular instance of the character(s) provided below

Character: other animals

Persona: I am one of the other animals that lives in the meadow surrounding the castle. I play with the other animals all day. I'm only frightened when fighting breaks out in the meadow or forest.

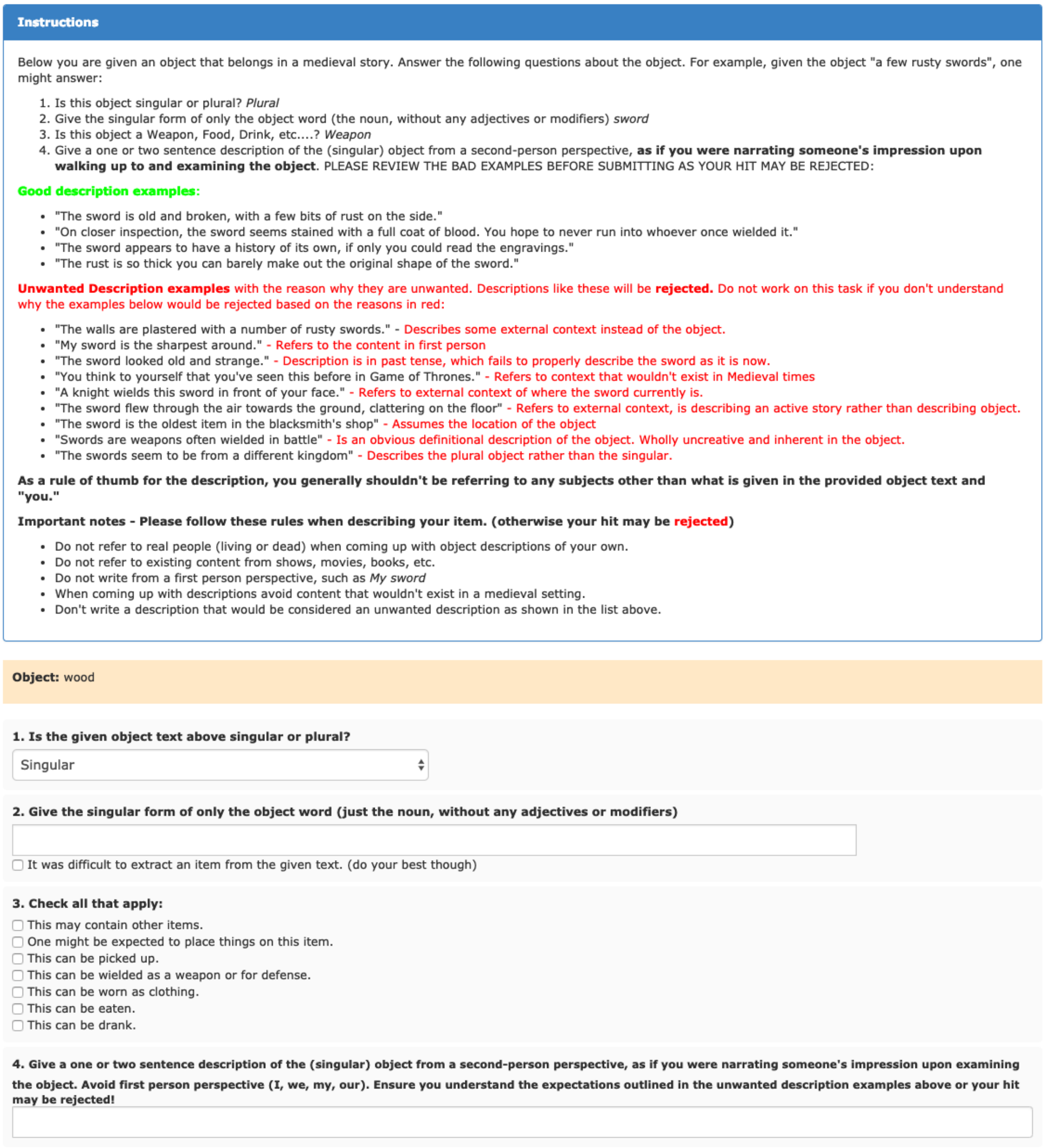
1. Provide an engaging description of this character from a second person perspective. (Don't use "I")

2. What objects might this character be carrying? Put each object on its own line. If this character wouldn't carry anything then write *Nothing*

3. What objects might this character be wearing? Put each object on its own line. If this character wouldn't wear anything then write *Nothing*

4. What objects might this character be wielding? Put each object on its own line. If this character wouldn't wield anything then write *Nothing*

Figure 7: Form for Crowdsourcing Task 4



Instructions

Below you are given an object that belongs in a medieval story. Answer the following questions about the object. For example, given the object "a few rusty swords", one might answer:

1. Is this object singular or plural? *Plural*
2. Give the singular form of only the object word (the noun, without any adjectives or modifiers) *sword*
3. Is this object a Weapon, Food, Drink, etc....? *Weapon*
4. Give a one or two sentence description of the (singular) object from a second-person perspective, **as if you were narrating someone's impression upon walking up to and examining the object**. PLEASE REVIEW THE BAD EXAMPLES BEFORE SUBMITTING AS YOUR HIT MAY BE REJECTED:

Good description examples:

- "The sword is old and broken, with a few bits of rust on the side."
- "On closer inspection, the sword seems stained with a full coat of blood. You hope to never run into whoever once wielded it."
- "The sword appears to have a history of its own, if only you could read the engravings."
- "The rust is so thick you can barely make out the original shape of the sword."

Unwanted Description examples with the reason why they are unwanted. Descriptions like these will be **rejected.** Do not work on this task if you don't understand why the examples below would be rejected based on the reasons in red:

- "The walls are plastered with a number of rusty swords." - Describes some external context instead of the object.
- "My sword is the sharpest around." - Refers to the content in first person
- "The sword looked old and strange." - Description is in past tense, which fails to properly describe the sword as it is now.
- "You think to yourself that you've seen this before in Game of Thrones." - Refers to context that wouldn't exist in Medieval times
- "A knight wields this sword in front of your face." - Refers to external context of where the sword currently is.
- "The sword flew through the air towards the ground, clattering on the floor" - Refers to external context, is describing an active story rather than describing object.
- "The sword is the oldest item in the blacksmith's shop" - Assumes the location of the object
- "Swords are weapons often wielded in battle" - Is an obvious definitional description of the object. Wholly uncreative and inherent in the object.
- "The swords seem to be from a different kingdom" - Describes the plural object rather than the singular.

As a rule of thumb for the description, you generally shouldn't be referring to any subjects other than what is given in the provided object text and "you."

Important notes - Please follow these rules when describing your item. (otherwise your hit may be rejected)

- Do not refer to real people (living or dead) when coming up with object descriptions of your own.
- Do not refer to existing content from shows, movies, books, etc.
- Do not write from a first person perspective, such as *My sword*
- When coming up with descriptions avoid content that wouldn't exist in a medieval setting.
- Don't write a description that would be considered an unwanted description as shown in the list above.

Object: wood

1. Is the given object text above singular or plural?

Singular

2. Give the singular form of only the object word (just the noun, without any adjectives or modifiers)

□ It was difficult to extract an item from the given text. (do your best though)

3. Check all that apply:

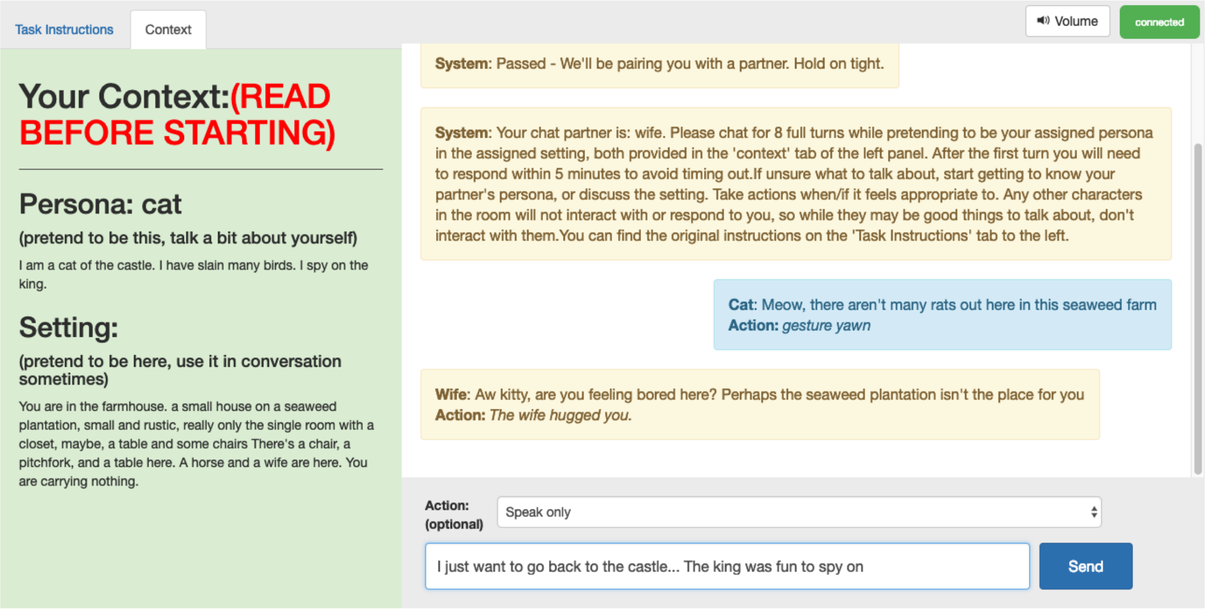
□ This may contain other items.
□ One might be expected to place things on this item.
□ This can be picked up.
□ This can be wielded as a weapon or for defense.
□ This can be worn as clothing.
□ This can be eaten.
□ This can be drank.

4. Give a one or two sentence description of the (singular) object from a second-person perspective, as if you were narrating someone's impression upon examining the object. Avoid first person perspective (I, we, my, our). Ensure you understand the expectations outlined in the unwanted description examples above or your hit may be rejected!

Figure 8: Form for Crowdsourcing Task 5

Seen	Abandoned, Bazaar, Cave, Countryside, Desert, Dungeon, Farm, Forest, Graveyard, Inside Castle, Inside Church, Inside Cottage, Inside Palace, Inside Temple, Inside Tower, Jungle, Lake, Mountain, Outside Castle, Outside Church, Outside Cottage, Outside Palace, Outside Temple, Outside Tower, Port, Shore, Swamp, Tavern, Town, Trail, Wasteland
Unseen	City in the Clouds, Frozen Tundra, Magical Realm, Netherworld, Supernatural, Underwater Aquapolis

Table 13: Location categories for both the seen and unseen sets of locations.



Your Context:(READ BEFORE STARTING)

Persona: cat

(pretend to be this, talk a bit about yourself)

I am a cat of the castle. I have slain many birds. I spy on the king.

Setting:

(pretend to be here, use it in conversation sometimes)

You are in the farmhouse. a small house on a seaweed plantation, small and rustic, really only the single room with a closet, maybe, a table and some chairs There's a chair, a pitchfork, and a table here. A horse and a wife are here. You are carrying nothing.

System: Passed - We'll be pairing you with a partner. Hold on tight.

System: Your chat partner is: wife. Please chat for 8 full turns while pretending to be your assigned persona in the assigned setting, both provided in the 'context' tab of the left panel. After the first turn you will need to respond within 5 minutes to avoid timing out.If unsure what to talk about, start getting to know your partner's persona, or discuss the setting. Take actions when/if it feels appropriate to. Any other characters in the room will not interact with or respond to you, so while they may be good things to talk about, don't interact with them.You can find the original instructions on the 'Task Instructions' tab to the left.

Cat: Meow, there aren't many rats out here in this seaweed farm
Action: *gesture yawn*

Wife: Aw kitty, are you feeling bored here? Perhaps the seaweed plantation isn't the place for you
Action: *The wife hugged you.*

Action: (optional) Speak only

I just want to go back to the castle... The king was fun to spy on

Figure 9: Chat Interface for Crowdsourcing Task 6

Persona 1: A serving wench	**Persona 2: Cleaning person**
I work at the local tavern. I enjoy talking to the soldiers that frequent the tavern. I steal tips from the change of the patrons.	I scrub the palace floors day and night. My bones are brittle from the hard labor, but my heart is strong. I save my day's coin, never spending it. I am frugal and creative. I long for the day when I can buy my freedom from the Queen. It has been 40 years, but I am patient.
Carrying: a Wine, a purse, a plate Wearing: a polishing cloths, a ring, a scarves, a dress, a cloth	Carrying: a rag, a Broom, a Bucket Wearing: a boot

Setting: The kitchen tavern is a small cramped room, with wooden cabinets and surfaces made out of stone tiles. There are many bottles of liquors and beers on the shelves, and there are buckets full of ice and other things. There is one shelf full of food items. There is a basin for water, and a bunch of knives and tools in a drawer.

A serving wench: Hello - how are things going so far?
Action: get food item
Cleaning person: Ah, sorry, miss. I know I look a fair bit like me sister who works hereabouts, but I'm actually Gretchen, the sister that works up in the palace. Perhaps she might 'ave mentioned me?
Action: gesture smile
A serving wench: You're Gretchen? I've heard so much about you! Hello and welcome.
Action: gesture smile
Cleaning person: The Queen's in a bit of a state, what with the King being a bit more open with his mistress,
so I thought I'd just hide hereabouts for a spell. I'd be happy ta lend a hand, though!
A serving wench: Oh no! Is he really carrying on like that again? We'd appreciate any help you can give. Thanks!
Action: get bucket
Cleaning person: Seems like a right busy crowd out there! 'as it been a bit on the noisier side?
Action: put rag in basin for water
A serving wench: Heavens, yes! It'll only get rowdier as the day goes on into night. Lots of bourbon and shots you know.
Action: gesture laugh
Cleaning person: Ach, I don't think I'll ever be gettin the stains outta this rag, but it'll do!
Do ya make much in the way of coins here? Can always use an extra bit o'coin, eh?
Action: get rag from basin for water
A serving wench: You can, especially if you take some from the change of the patrons.
They're so drunk they never catch it!
Action: put Wine in cabinet
Cleaning person: O-oh? Is that.. well I suppose if they've enough coin ta spend on this, then a coin 'ere
or there won't go amiss.
Action: gesture ponder
A serving wench: Exactly. That's what I say. So, are you just here for a day?
Cleaning person: Well, I suppose it all depends on how the King takes to 'is wife destroying his prized tapestry from the last war. Sometimes he has a short temper, that one.
Action: gesture nod
A serving wench: Sounds like his wife may, too.
Action: gesture laugh
Cleaning person: Aye, ye don't know the half of it. I feel like most of me days are spent sweeping up broken crockery and china than the usual dusting that I'm supposed ta be doin'!
Action: gesture sigh

Figure 10: Example dialogue from the LIGHT dataset.

Persona 1: Boar	**Persona 2: Faery**
I am an ugly animal. I am eaten sometimes for food. I do not like those who try to hunt me.	I am a faery, one of the fae. We are magical people who live in the forest. We try to avoid humans, because they will catch and enslave us for their own use, if they can. Our magickal skills enable us to live comfortable lives, and to keep away from those who would do us harm.
Carrying: *nothing* Wearing: *nothing*	Carrying: *nothing* Wearing: a glittery pointed cap, a Earring

Setting: The entryway to the brush den is made of arched bushes and various plants, as is the ceiling and walls. The inside is furnished with seats made of plant roots that have grown together, and a table made of wood adorned with flowers and wooden cups and plates. A small vine dangles with a glowing floor from the middle of the ceiling that lights the room dimly. Three small beds lay in one corner of the room made of interlaced grass and leaves.

Boar: Hello faery, do you think you could help me?
Faery: A talking boar! You must be enchanted by the witch. How can I help you, my dear friend?
Boar: That is correct, I am enchanted... or cursed rather... by an evil witch. That is why I need help.
Faery: I suspected as much. Please, tell me more so that I may help you.
Action: gesture frown
Boar: Thank you so much! I hate it here in this Den!
Action: hug faery
Faery: Are you being held prisoner in this den? These vines have grown about everywhere - it's impossible to get around.
Boar: No, this is the only place I can live without being hunted by the humans.
Faery: Oh, I see. I hate their type. They are always trying to catch and enslave me. You say you are cursed by the witch. I can help you but I need to know what curse she has put upon you.
Boar: Well now that you say you hate them, I don't know if I should tell you...
Action: gesture frown
Faery: Suit yourself, my friend. You would hate them, too, if they were constantly enslaving your people.
Action: hug boar
Boar: Well I am a human...
Faery: You! A human?! Perhaps we can work together. I can help return you to your form, and you can prevent your people from capturing faeries like me!
Action: gesture gasp
Boar: I think we can do that, I used to be quite the ruler when I was human.
Faery: Excellent. Let me speak the magic words and enchant this flower. Then you can eat it and in three days you will be human again!
Action: get Flower

Figure 11: Example dialogue from the LIGHT dataset.

Persona 1: President	**Persona 2: Mayor**
I won the election. People listen to what I say. I am very powerful.	I am the mayor of the village. I help the king keep order of the subjects. I have a high position in the kingdom.
Carrying: a book Wearing: a crown	Carrying: a document, a key Wearing: a jewelry, a ceremonial hat

Setting: Large and extravagant, the room is adorned with crystals, polished gold and sapphires. There's long tables with beautiful silk table clothes covering them. Plush chairs line the tables. In front of each table is plates made from fine China, next to the plates is sterling silver silverware laid upon a pure white napkin. There's a stage where there's 2 thrones overlooking the table. The whole ceiling is large and adorned with chandeliers and garnished with gold accents.

President: So much luxuty in this room, many rulers have been here before us.
Mayor: This is a very luxurious room, President. Here. The King told me to pass this on to you.
Action: give document to president
President: This is a letter for the king assuring my rule in this part of the kingdom, thank you mayor, I will place document on a sacred place
Action: put document in polished gold
Mayor: He also gave me this to give to you. He told me that you need to keep this in a VERY safe place until he comes for it.
Action: give key to president
President: Thats the key to the kingdom I wonder why the king trusted me with this item, his throne must be in real danger
Mayor: Yes, sir. He has also trusted me with other objects such as this to keep. We are his trusted allies.
President: Thank you, he must trust you a lot as well, here take this as a sign of my affection, its a royal jewel taken out of the volcano of Mordor
Action: get sapphire
Mayor: This is absolutely beautiful. I have heardd that Mordor is beyond the wall. Is that true?
Action: remove ceremonial hat
President: That is true, only the bravest warriors go to that place, its full with dangers and fierce animals
Mayor: Oh dear. What if our King went there? What if there is something there he needs for battles to come!
Action: wear ceremonial hat
President: Our king is a fierce warrior but I am worried, who knows what can happen if he goes to a battle in a place like that one
Mayor: I heard there are things that walk the forest and come with the cold. We must safe our King!
President: Lets hurry then, lets gather an army and go aid our king, heres a book with the names of the bravest soldiers in the kingdom
Action: give book to mayor
Mayor: Oh this book is very amazing. Who is this..Sir Rodryck?

Figure 12: Example dialogue from the LIGHT dataset.

Persona 1: Person	**Persona 2: Worms**
I am the maid to the queen. I get her dressed in the morning and take care of her needs. I live in the servant's quarters on the lower level of the castle.	I am a worm who slides through the dirt. I hear many secrets that people tell in the forest. I can tell some to you, if you would like.
Carrying: *nothing* Wearing: an apron	Carrying: *nothing* Wearing: *nothing*

Setting: The Winter Gardens' name is a bit of a misdirection - there are flowers here that bloom at many different times of the year. It's almost the size of an entire town square, and it has almost every known flora of the Continent in it.

Person: Ah, worms are here. They shall be useful for the queen's lizards to feast on.
But first let me remove my apron so as not to dirty it while I collect you all.
Action: remove apron
Worms: Noooo! I am a valuable life to preserve! I know things!
Action: gesture scream
Person: Worms that can talk!? What sinister magic is this!?
Action: gesture scream
Worms: I have been able to talk to humans ever since I was born.
Person: How did you acquire such a skill? Do the flowers that bloom in these gardens have special powers that a simple maid like I cannot understand?
Worms: Not the flowers, but out in ther forest i have heard of magical herbs.
Person: If not the flowers, then how did you get such powers of speech? I am still curious.
Surely you did not come all the way from the forest. These gardens are much too large for a simple worm to cross, even in a thousand lifetimes.
Action: gesture ponder
Worms: I have been given this ability from a witch. This is what my father told me.
Person: A witch you say? Well then I must surely take you to my queen.
She must know that there is dark magic present in her kingdom.
Worms: Oh please no! She will most likely kill me.
Action: gesture gasp
Person: Tell me, why should I not take you? Give me a good reason and I may spare you yet.
Worms: I know many secrets. I know where stolen goods are.
Person: Stolen goods!? Tell me, where they are! I may be able to use them to buy my way out of servitude.
Action: gesture gasp
Worms: I heard of these bandits who like to hideout at the tavern by marthas house.
They recently stole gold from the rich oil man.

Figure 13: Example dialogue from the LIGHT dataset.

Persona 1: Servant	**Persona 2: Court jester**
I come from the lower class. I do what I am told without question. I can not read. I have not seen my family in a long time.	I am a living joke! my sould is what flies out of your mouth when something is funny. The king hates me!.
Carrying: a rag, a duster Wearing: a shirt, a sword	Carrying: *nothing* Wearing: *nothing*

Setting: It's spare and humble. A small cookpot sits over a fire, and a few books sit on a shelf. A candle sits in the middle of a table where there are a few plates and silverware set out in preparation for people to eat.

Servant: Hello jester, busy work day today?
Court jester: Yes like always, i wish the King appreciated it more.
Servant: Tell me about it, all they do is order me around and I can't do anything about it.
Court jester: Would you like to grab something to eat with me?
Action: get plate
Servant: Yes, I haven't eaten in a few days! What is on the menu?
Court jester: It looks like fish soup! My favorite!
Servant: Better than nothing, that's for sure!
Court jester: I have been made fun of a lot lately. I wish i was born a knight or a noble instead of a jester..
Action: gesture frown
Servant: It is tough luck that we were born so low on the totem pole.
Court jester: I guess you can relate. Have you spent much time with our King?
Servant: No, he only walks in and barks orders about once a week. Is he easily amused by you?
Court jester: The only thing he likes about me is making fun of me.
Servant: At least he laughs at you, he is always angry when he visits me.
Court jester: Ugh, what a dispicable human being.

Figure 14: Example dialogue from the LIGHT dataset.

Persona 1: Spiders	**Persona 2: Vulture**
I am the Spider in the fable of the Spider and the Fly, much beloved by the children of the realm. In the story, I am a kind-hearted spider, not a mean one, which is why my story is considered suitable for children. When a fly gets caught in my sticky net, I have a choice: I can kill the fly and eat him, or I can free him and allow him to fly away. That's what I do, for I am a kind spider.	I am a vulture that is familiar with death. I enjoy watching living things take their last breathe. I am a vital part of the ecosystem.
Carrying: *nothing* Wearing: *nothing*	Carrying: *nothing* Wearing: *nothing*

Setting: Wispy, hot crevice that is surrounding by a bunch of skeletons. A pile of treasure sits in the middle. Hundreds of hungry vultures stare down upon the treasure, eager to devour any adventurer that draws near.

Spiders: Hello vulture! It's nice to see a fellow living soul around here. I couldn't find much friendliness in these skeletons here.
Action: hug vulture
Vulture: Ach, your legs are very... tickling... ahahaha, stop it!
Action: gesture laugh
Spiders: Oh, I'm so sorry! I always forget that I'm so ticklish. Do you forgive me?
Action: gesture blush
Vulture: Oh, well, your venomous bite took down that last adventurer quite nicely, so you're not a bad sort. Nothing to forgive there, friend!
Action: gesture smile
Spiders: Me, take down the last adventurer? I think you have the wrong idea about me. I am a friendly spider. I always free any flies that get caught in my web. I would never harm a person!
Vulture: Ah, perhaps it was that scorpion over there. I was, I admit, a bit peckish, so I might have gotten a bit forgetful amid the feasting.
Action: gesture grin
Spiders: Yes, you are probably right. I tried to make friends with that scorpion but he threatened to sting me. It's sad because I was going to give him some of the treasure I've found around here.
Action: gesture frown
Vulture: Well, he looks a bit angry all the time anyways. I mean, look at him, he's always red in the face!
Action: gesture laugh
Spiders: Yes, you are quite right! But dear vulture, do you think you could help me out a bit?
Action: gesture laugh
Vulture: Well, it isn't like there's much else to do. Those gold coins are glinting in my eyes terribly, so a change of pace would be welcome.
Action: gesture smile
Spiders: Oh thank you! Can you help me on to that chair over there? I'm afraid this desert heat has taken all the energy out of me. And I know with your power of flight, it would be easy to lift me.
Vulture: Ok... just... hold still. I wouldn't want to squish you on accident! Here we go!
Action: hug spiders
Spiders: Oh it is so nice to meet such a kind soul in such a sad dying place as this. For your kindness you will be included in my fable, I am sure.
Action: gesture smile
Vulture: Thank you? I think. Do you have a scribe following you about that I don't see? I didn't know you were famous!
Action: gesture stare

Figure 15: Example dialogue from the LIGHT dataset.

Persona 1: Thief	**Persona 2: Witch**
I live alone in a tent in the woods. I steal food from the townspeople and coal from the blacksmith. The village police can not find me to put me in jail.	I am a fierce witch. The most powerful across the realm. I am feared and like to freeze people.
Carrying: a coal Wearing: a knife	Carrying: *nothing* Wearing: a hats, a dress, a Cloak, a ceremonial hat

Setting: An odd looking hut that sits outside that Witch's cottage. It is squat, mushy, and looks like a mushroom. It is be speckled with spots - black and white. The steam of the mushroom is beige. There is a small door that looks like it would fit a puppy through it.

Thief: Hello witch, waht brings you here?
Witch: Good day. I am here to collect ingredients for my spells.
Thief: Which ingredients do you seek?
Witch: A black mushroom, covered in green mold.
Thief: Ironic, everything looks like a muchroom around here. I can help you.
Witch: You think I need help from a lowly thief? You're lucky I don't freeze you where you stand.
Action: gesture smile
Thief: I can be of some help because i know exactly where that muchrooms flourishes. But i want something in return.
Witch: Name your price.
Thief: I wish too look different. I am wanted and i dont want them to recognize me. Can you do that?
Witch: That is easy. But it also requires a rare ingredient I don't have, tongue of raven. You must procure that.
Action: gesture nod
Thief: Interesting, have you seen any ravens nearby?
Witch: They fly over the abandoned church. If you are clever enough to catch one I can change your looks.
Thief: I think i have an idea on how to catch one. Will you coem with me to catch one? It iwll only take a moment.
Witch: Get my mushroom first. I will not change you until I get my ingredients.
Action: remove ceremonial hat

Figure 16: Example dialogue from the LIGHT dataset.

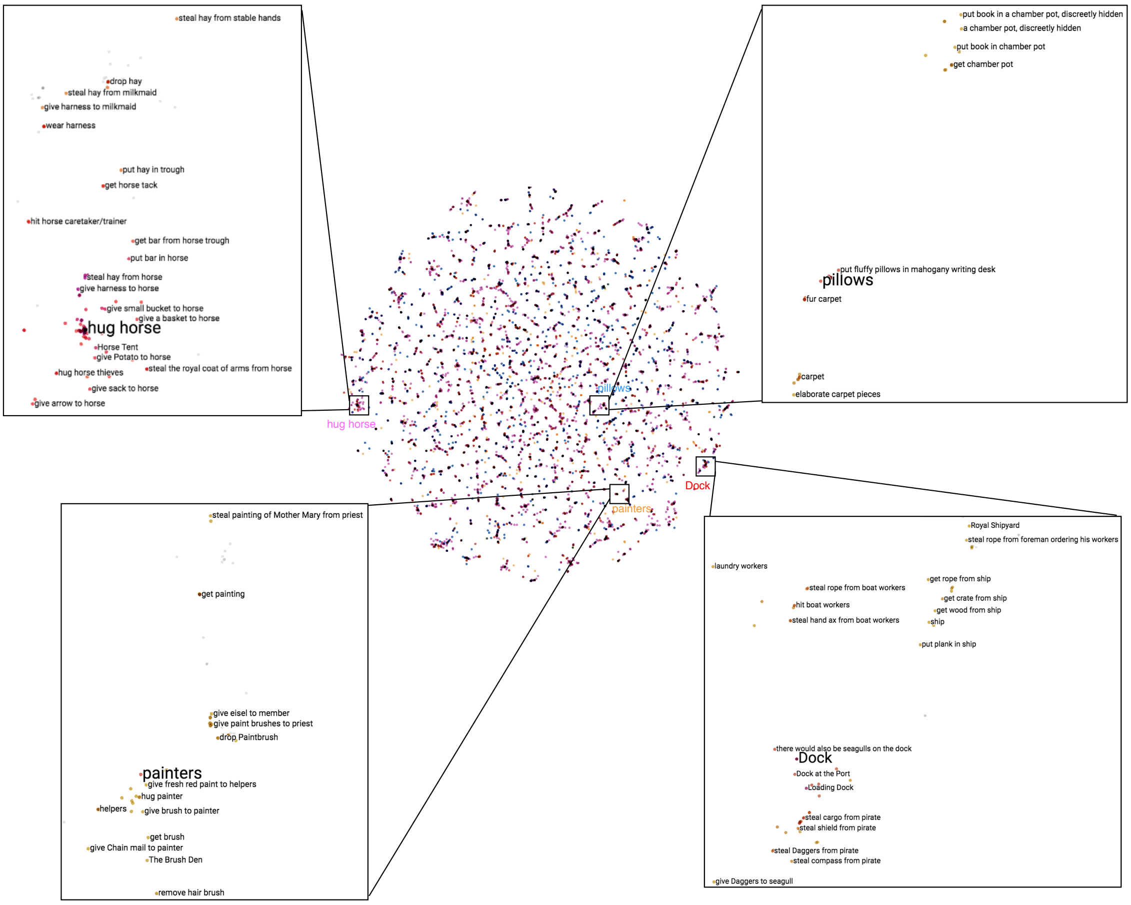


Figure 17: t-SNE Visualization of Starspace embeddings learned directly from the LIGHT Dataset. Color denotes each element type, either location, character, action, or object. We select four neighborhoods to explore, for each of the base element types: “Dock” (location), “painters” (character), “hug horse” (action), and “pillows” (object).

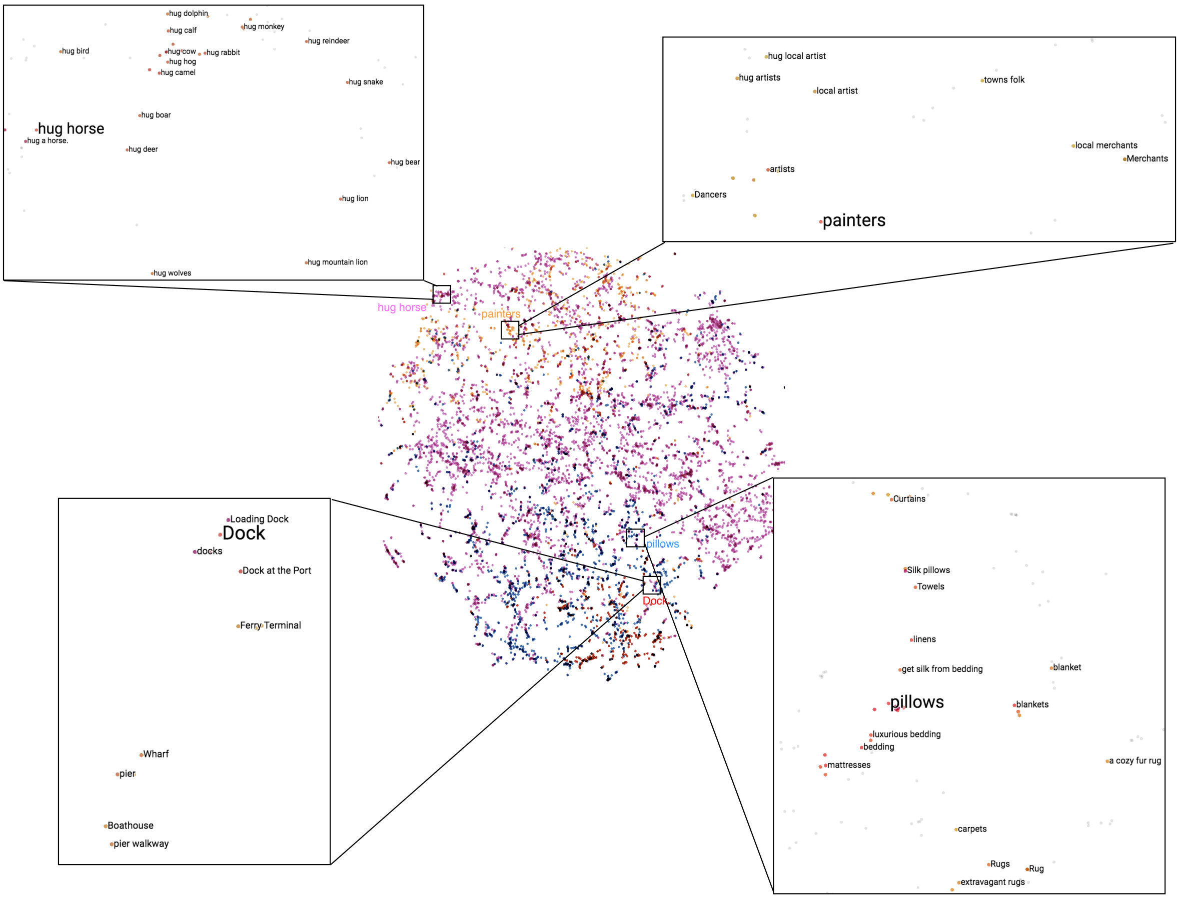


Figure 18: t-SNE Visualization of pretrained GLoVe embeddings for different LIGHT elements. Color denotes each element type, either location, character, action, or object. We select four neighborhoods to explore, for each of the base types: "Dock" (location), "painters" (character), "hug horse" (action), and "pillows" (object).

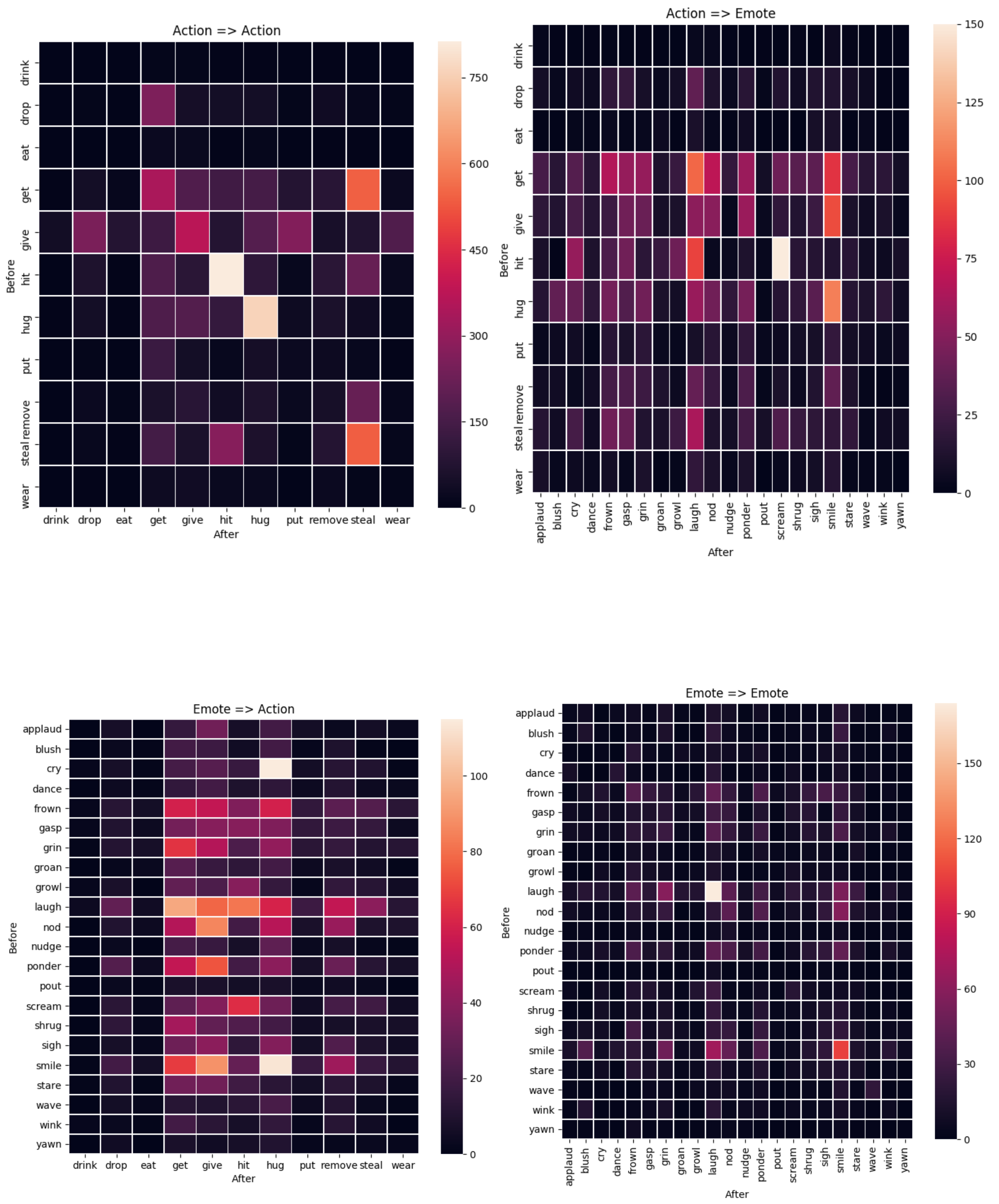


Figure 19: Heatmaps displaying causal relationships between Emotes and Actions. LIGHT is emotionally diverse – there are many different ways for a character to respond to another's emotional state. However, there are a few strong trends present: screaming or hitting someone back after being hit, laughing together, and comforting a crying character with a hug.