HOW TO MOTIVATE YOUR DRAGON: TEACHING GOAL-DRIVEN AGENTS TO SPEAK AND ACT IN FANTASY WORLDS

Anonymous authors
Paper under double-blind review

ABSTRACT

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

1 INTRODUCTION

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

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

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

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

However, we find that although pre-training can be an effective tool in this setting, it requires more finesse than in the standard supervised setting. In particular, we find that simply pre-training a model

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

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

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

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

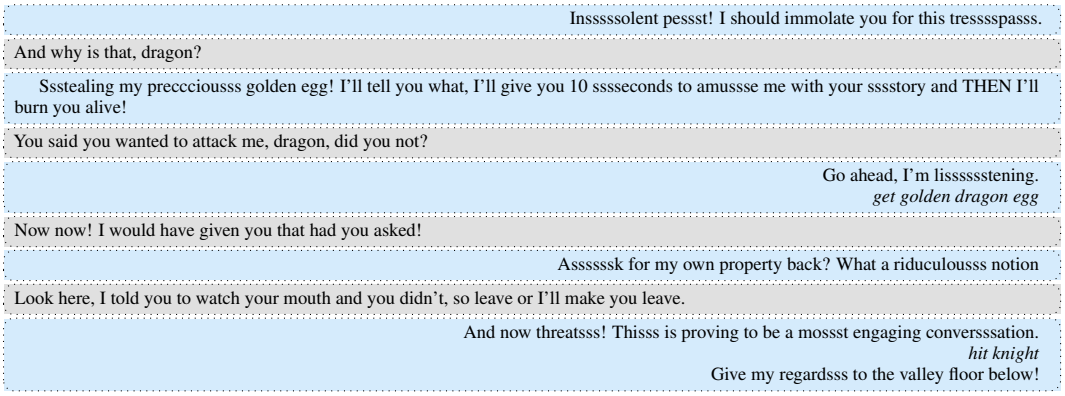

Figure 3: Example of a demonstration of a human (blue shaded) completing the above quest while role-playing as the self character with a partner agent (grey shaded). There are 2111 such human demonstrations of average sequence length 12.92, consisting of 22672 dialogues in total.

on a large "generic" corpus (Sap et al., 2019; Baumgartner et al., 2020) of commonsense/language data or pre-training on the domain specific LIGHT corpus, and then fine-tuning via RL is *less* effective than training RL from scratch. Furthermore, by carefully combining general and domain-specific pre-training, we observe large improvements over RL from scratch.

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

2 Related Work

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

Text-based game-playing. Côté et al. (2018) introduce TextWorld, a framework for procedurally generating text-based games via grammars, and Yuan et al. (2018); Yin & May (2019); Adolphs & Hofmann (2019); Adhikari et al. (2020) build agents that operate in this environment—focusing on aspects such as efficient exploration and zero-shot generalization to new, procedurally generated environments. Similarly, Hausknecht et al. (2020) introduce Jericho, a framework and series of baseline agents for interacting with human-made text-games such as *Zork* (Anderson et al., 1979).

This resulted in agents developed by works such as Zahavy et al. (2018); Ammanabrolu & Hausknecht (2020), aiming to learn to execute contextually relevant actions. Other works such as Narasimhan et al. (2015); He et al. (2016) explore how to best factorize such text-game action spaces. None of these works consider agents with motivations and personas nor require any dialogue.

Goal-oriented dialogue. This form of dialogue has traditionally been closely related to specific tasks useful in the context of personal assistants with dialogue interfaces (Henderson et al., 2014; El Asri et al., 2017). RL has been studied for such tasks, usually to improve dialogue state management (Singh et al., 2000; Pietquin et al., 2011; Fatemi et al., 2016) and to improve response quality (Li et al., 2016). In particular, the negotiation tasks of Yarats & Lewis (2017); Lewis et al. (2017), where two agents are trying to convince each other to perform certain actions, are related to the tasks in LIGHT-Quests. These works all lack environment grounding and the notion of diverse agent motivations.

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

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

3 LIGHT-QUESTS AND ATOMIC-LIGHT

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

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

3.1 LIGHT-QUESTS

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

Given the information in Figure 1, the crowdworkers completed the above outlined tasks and produce data as seen in Figure 2. Motivations come in three levels of abstraction—short, mid, and long—corresponding to differing amounts of the timeline. For example, the short motivation is always guaranteed to correspond most closely to the *now* position on the timeline. Action annotation is pre-constrained based on the classes of verbs available within LIGHT. The rest of the action is completed as free form text as it may contain novel entities introduced in the motivations. There are 5982 training, 756 validation, and 748 test quests. Further details regarding the exact data collection process and details of LIGHT-Quests are found in Appendix A.1.1.

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

3.2 ATOMIC-LIGHT

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

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

4 AGENTS THAT ACT AND SPEAK

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

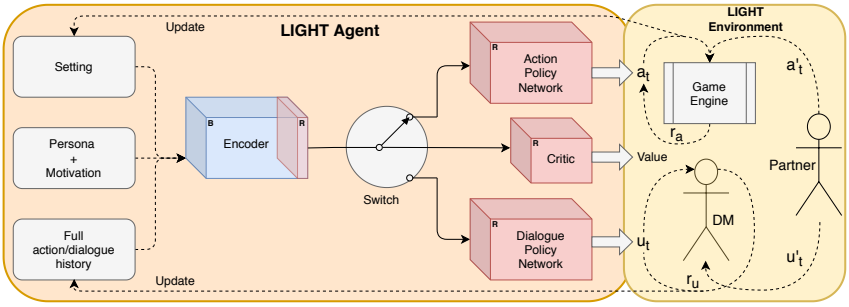

Figure 4: Overall RL Switch architecture and process. **B**lue shaded components can be pre-trained and **R**ed shaded components are trained with RL. Solid lines indicate gradient flow.

4.1 LIGHT RL Environment

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

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

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

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

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

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

4.2 TRAINING A LIGHT AGENT WITH SWITCH REINFORCEMENT LEARNING

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

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

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

4.3 ENCODER PRE-TRAINING TASKS

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

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

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

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

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

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

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

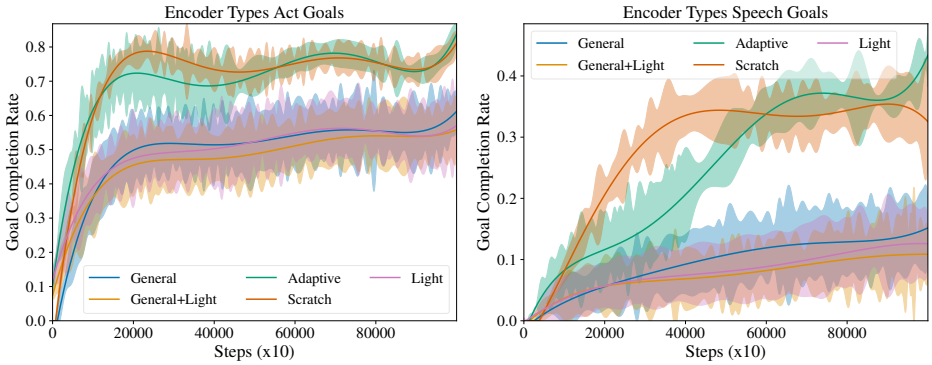

Figure 5: Encoder types RL reward curves averaged over 3 independent runs.

5 Evaluation

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

5.1 Encoder Pre-training Type Ablation Study

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

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

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

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

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

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

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

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

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

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

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

5.2 Ability Type Ablation Study

To better understand the interplay between acts and speech resulting in *self* and *partner act goal completions*, we perform an ablation study selectively dropping either the agent's ability to talk or act. We train the agent to either only act, only speak, only speak with only action rewards. In the scenarios when the agent can only speak, the agent has to convince the partner to help achieve the agent's goal.

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

6 Conclusion

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

REFERENCES

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

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

Prithviraj Ammanabrolu and Matthew Hausknecht. Graph constrained reinforcement learning for natural language action spaces. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=B1x6w0EtwH.

Prithviraj Ammanabrolu and Mark Riedl. Transfer in deep reinforcement learning using knowledge graphs. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13) at EMNLP*, 2019. URL https://www.aclweb.org/anthology/D19-5301.

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

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

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

Lawrence W. Barsalou. Grounded cognition. *Annual Review of Psychology*, 59(1):617–645, 2008. doi: 10.1146/annurev.psych.59.103006.093639. URL https://doi.org/10.1146/annurev.psych.59.103006.093639. PMID: 17705682.

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

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

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

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

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

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

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

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

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

Jerome Feldman and Srinivas Narayanan. Embodied meaning in a neural theory of language. *Brain and language*, 89:385–92, 06 2004. doi: 10.1016/S0093-934X(03)00355-9.

Nancy Fulda, Daniel Ricks, Ben Murdoch, and David Wingate. What can you do with a rock? affordance extraction via word embeddings. In *IJCAI*, pp. 1039–1045, 2017. doi: 10.24963/ijcai.2017/144.

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

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

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

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

Matthew Hausknecht, Prithviraj Ammanabrolu, Marc-Alexandre Côté, and Xingdi Yuan. Interactive fiction games: A colossal adventure. In *Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, 2020. URL https://arxiv.org/abs/1909.05398.

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

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

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

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkxgnnNFvH.

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

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

Victoria Krakovna, Jonathan Uesato, Vladimir Mikulik, Matthew Rahtz, Tom Everitt, Ramana Kumar, Zac Kenton, Jan Leike, and Shane Legg. specification gaming: the flip side of ai ingenuity. URL https://deepmind.com/blog/article/Specification-gaming-the-flip-side-of-AI-ingenuity.

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

Carolin Lawrence, Bhushan Kotnis, and Mathias Niepert. Attending to future tokens for bidirectional sequence generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1–10, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1001. URL https://www.aclweb.org/anthology/D19-1001.

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

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

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

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

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

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2775–2779, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1298. URL https://www.aclweb.org/anthology/D18-1298.

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

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

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

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

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

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

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

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

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

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

Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7881–7892, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704. URL https://www.aclweb.org/anthology/2020.acl-main.704.

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

Robert Speer and Catherine Havasi. Representing general relational knowledge in conceptnet 5. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)*, 2012. ISBN 978-2-9517408-7-7.

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

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. Learning to speak and act in a fantasy text adventure game. *CoRR*, abs/1903.03094, 2019.

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

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

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

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

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

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

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

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