

Policy Improvement using Language Feedback Models

Victor Zhong^{1,2} Dipendra Misra¹ Xingdi Yuan¹ Marc-Alexandre Côté¹

Abstract

We introduce Language Feedback Models (LFMs) that identify desirable behaviour — actions that help achieve tasks specified in the instruction — for imitation learning in instruction following. To train LFM, we obtain feedback from Large Language Models (LLMs) on visual trajectories verbalized to language descriptions. First, by using LFM to identify desirable behaviour to imitate, we improve in task-completion rate over strong behavioural cloning baselines on three distinct language grounding environments (Touchdown, ScienceWorld, and ALFWorld). Second, LFM outperform using LLMs as experts to directly predict actions, when controlling for the number of LLM output tokens. Third, LFM generalize to unseen environments, improving task-completion rate by 3.5-12.0% through one round of adaptation. Finally, LFM can be modified to provide human-interpretable feedback without performance loss, allowing human verification of desirable behaviour for imitation learning.

1. Introduction

Sample-efficiency and generalizability are two primary challenges in learning instruction following agents in grounded environments (MacMahon et al., 2006; Kollar et al., 2010; Ahn et al., 2022). First, we want an agent that is sample-efficient: it learns from few demonstrations of how to act according to instructions. Second, we want an agent that is generalizable: it should act successfully in novel environments according to new instructions after training. Reinforcement learning (RL; Sutton & Barto (2018)) and imitation learning (IL; Schaal (1999), Abbeel & Ng (2004)) are two techniques for learning agents for instruction following in grounded environments. These techniques often require large numbers of trials and errors or expensive-to-obtain expert demonstrations. Recent work show that pretrained large language models (LLMs) exhibit sample-

efficient learning through prompting and in-context learning for textual (Brown et al., 2020) and grounded problems such as robotic control (Ahn et al., 2022). However, for instruction following in grounded problems, current methods rely on LLMs on-line during inference, which is impractical and expensive.

We develop a sample-efficient and cost-effective technique that uses LLMs to train **Language Feedback Models (LFMs)** for policy improvement in instruction following. Figure 1 illustrates policy improvement using LFM. Consider the task of interacting with objects in a kitchen to follow instructions shown in Figure 1(c). First, in Figure 1(a), given a grounded environment and a base policy (i.e. a behaviour cloned policy), we roll out the base policy to collect a small set of trajectories for different instructions. Next, we verbalize observations in the trajectory by describing scenes in language. For each instruction and verbalized trajectory pair, we query an LLM to provide feedback identifying which behaviour in the trajectory is productive to solving the task identified in the instruction (i.e. answer yes or no). For instance, given an instruction “put a clean slice of lettuce in the refrigerator”, GPT-4 (OpenAI, 2023) is able to deduce that key milestones are 1) find the lettuce, 2) slice it 3) wash it in the sink, and 4) put it in the fridge. Consequently, such an LLM is able to identify when an agent is exhibiting **desirable behaviour** conducive to solving tasks outlined in the instruction, for instance by taking the lettuce to the sink, versus undesirable behaviour, for instance by cooking the lettuce. After collecting LLM feedback, we distill this world knowledge into a small and cost-effective LFM. Finally, in Figure 1(b), given a policy to improve on potentially new environments and instructions, we use the learned LFM to identify desirable actions on-line, then update the policy to imitate these actions. Crucially, this technique is sample-efficient and cost-effective in that it only requires few LLM interactions to collect an off-line dataset during LFM training (i.e. before deployment), as opposed to many LLM interactions on-line during policy improvement (i.e. after deployment).

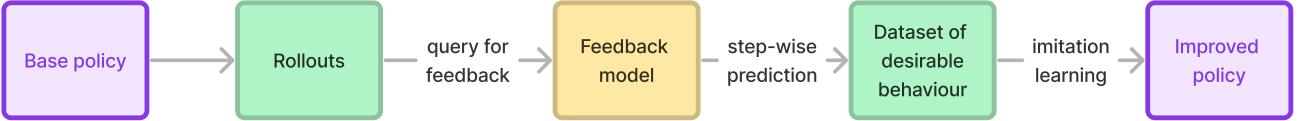
Our findings are as follows: first, through LFM policy improvement, on three grounded instruction following benchmarks, namely Touchdown (Chen et al., 2019), ScienceWorld (Wang et al., 2022), and ALFWorld (Shridhar et al., 2021b), we observe consistent gains over strong, behaviour

¹Microsoft Research ²University of Waterloo. Correspondence to: Victor Zhong <victor.zhong@uwaterloo.ca>.

Policy Improvement using Language Feedback Models



(a) Learning a small and cost-effective Language Feedback Model from LLM feedback. We roll out an initial policy, then prompt an LLM to provide feedback on what actions the policy took during the rollout were productive in achieving the task outlined in the instruction. We then use this data to train a feedback model that predicts whether an action is productive given the instruction.



(b) Policy improvement by imitating desirable behaviour identified by a learned feedback model. Given the instruction, we roll out a base policy, then identify productive actions that help achieve tasks specified in the instruction using the trained feedback model. Finally, we update the base policy by imitating the identified desirable behaviour.



Instruction: clean some lettuce and put them in the fridge

(c) Example of desirable behaviour identified in an example environment in ALFWorld, a kitchen instruction following benchmark.

Figure 1: Given an environment and instructions to follow, we assume access to a verbalization procedure that converts observations to language descriptions. Policy improvement using Language Feedback Model involves (a) first training a feedback model, then (b) using it to identify desirable behaviour for policy improvement via imitation learning. We show the feedback model in yellow, other models in purple, and generated intermediate data in green. An example of LFM-identified desirable behaviour is shown in (c).

cloned base policies. Second, using LLMs as feedback models outperforms using LLMs as expert policies for imitation learning. We compare LFM against prompting LLMs to directly predict what actions to take, then imitating this LLM-predicted behaviour. On all benchmarks, using LFM feedback outperforms using LLMs as experts for imitation learning, given a fixed allocation of LLM output tokens. This gain is especially pronounced in environments with larger action spaces, such as ScienceWorld, where it is much easier to critique than to generate the correct action. Third, we show that learned feedback models generalize to unseen environments. After training LFM on training environments, we use them to identify desirable behaviour on test environments, which we then adapt the policy to imitate. A single round of adaptation achieves significant gains (3.5-12.0% task-completion rate) across all environments.

In addition to policy improvement, using LFM feedback offers two advantages over existing techniques such as using LLMs as expert policies for imitation learning. First, LFM improves policies on-line without additional expensive calls to LLMs. Second, LFM can offer human-interpretable feedback when identifying desirable behaviour to imitate. We show in Section 5.4 that LFM can be easily modified to provide not only desirable behaviour but why

they were desirable, thereby allowing humans to inspect and validate imitation data used for policy improvement. Source code for our environments and experiments are available at anonymous.4open.science/r/language_feedback_models. Videos of LFM feedback are available at language-feedback-models.github.io.

2. Background

Language grounded instruction following. In language-grounded instruction following, an agent is given an instruction x specifying the task to achieve in the environment. Each turn, the agent receives a potentially partial observation o_t , and takes an action a_t which causes the environment to transition to a new state. In the example in Figure 1(b), the agent observes a counter with objects such as a toaster, some lettuce, and a knife on top. To follow the instruction ‘put a clean slice of lettuce in the refrigerator’, an effective agent may choose to grab a piece of lettuce. In the reinforcement learning setting, the environment additionally give the agent a reward after a desirable (positive reward) or undesirable (negative reward) action (Sutton & Barto, 2018). In this work, we consider long-horizon settings with only sparse and delayed task-completion rewards. Consequently, we focus on imitation learning from demonstrations as opposed

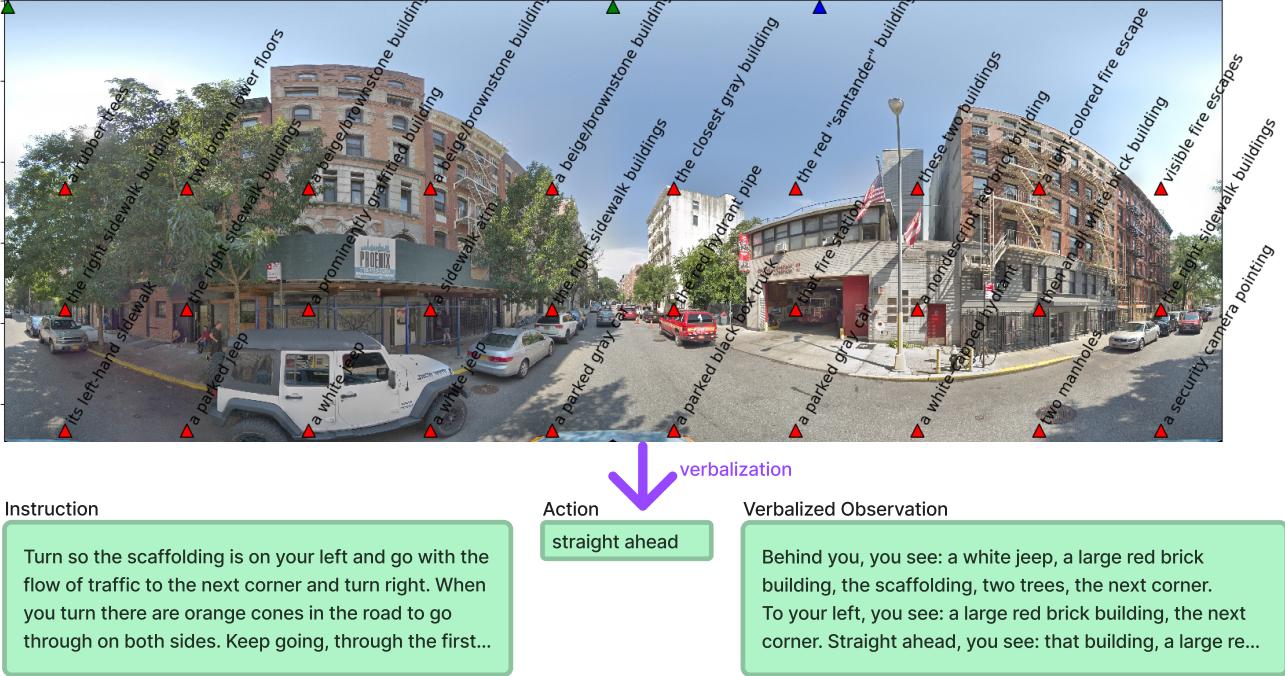


Figure 2: An example verbalization for Touchdown. We align CLIP image embeddings of panorama patches and language embeddings of common noun-phrases to populate a language template. Appendix A describes this procedure in detail. The **blue** arrow at the top indicate the agent’s orientation while the **green** arrows indicate valid directions to proceed in.

to reinforcement learning from rewards (Schaal, 1999).

Imitation learning. In imitation learning for instruction following, we are given an expert policy $\pi^*(a|x, o)$ and learn a policy $\pi_\theta(a|x, o)$ with parameters θ . We first roll out the policy π_θ . For each step $o_t^{(i)}$ of the rollout τ_i , we optimize θ to imitate the action $a_t^{(i)}$ chosen by the expert $\pi^*(a|x, o_t^{(i)})$ when given the same observations.

$$\arg \min_{\theta} \mathbb{E}_{o_t^{(i)} \sim \pi_\theta} \left[L \left(\pi_\theta(a|x, o_t^{(i)}), a_t^{(i)} \right) \right] \quad (1)$$

Here, L is a step-wise cross-entropy loss function between the policy’s action distribution and the action chosen by the expert given the same observation:

$$L(*) = - \sum_{a' \in \mathcal{A}} \mathbb{1} \left[a' = a_t^{(i)} \right] \ln \pi_\theta(a = a' | x, o_t^{(i)}). \quad (2)$$

Behavioural cloning. Imitation learning in Eq (1) assumes an expert policy that can be executed on-line to produce expert actions. For instance, given an expert, imitation learning assumes that this expert $\pi^*(a|x, o_t)$ provides corrective actions a_t as the policy $\pi(a|x, o_t)$ runs. In many cases, this is impractical — a human-in-the-loop expert is expensive and inconvenient while an LLM expert is expensive and, as we show in our experiments, inaccurate. Alternatively, in behaviour cloning (BC), we instead collect an offline dataset of expert trajectories from which to

clone expert behaviour (Bain & Sammut, 1995; Torabi et al., 2018). BC (or offline imitation learning) only asks the expert to perform the task N times to collect N trajectories $\{\tau_i\}_{i=1}^N$. Each τ_i consists of M_i steps of observations and associated expert actions: $\tau_i = [o_1^{(i)}, a_1^{(i)}, \dots, o_{M_i}^{(i)}, a_{M_i}^{(i)}]$ where $a_t^{(i)}$ is the action chosen by the expert $\pi^*(a|x, o_t^{(i)})$ given the observation $o_t^{(i)}$. We train policy π_θ to imitate the expert action, given the same observation seen by the expert, by minimizing the following objective:

$$\arg \min_{\theta} \frac{1}{N} \sum_i^N \frac{1}{M_i} \sum_t^{M_i} L \left(\pi_\theta(a|x, o_t^{(i)}), a_t^{(i)} \right). \quad (3)$$

The key distinction between BC and imitation learning is that the former optimizes over trajectories under the expert policy while the latter optimizes over trajectories under the learned policy. Consequently, while BC is offline and easily batchable, it suffers from covariate shift/exposure bias (Ross et al., 2011; Bengio et al., 2015). Like prior work in long-horizon instruction following in grounded environments (Fried et al., 2018; Chen et al., 2019), we use BC to warm-start a strong base policy (Ash & Adams, 2020), which we then improve using imitation learning.

3. Language Feedback Model

How can we leverage world knowledge in LLMs to make policy learning more sample-efficient and generalizable? In this work, we use LLMs to distill a small and cost-effective Language Feedback Model to identify desirable behaviour from a base policy (Figure 1(a)). We then improve the base policy by imitating this desirable behaviour through batched imitation learning, without need for on-line LLMs (Figure 1(b)). Appendix B provides pseudo-code for the entire procedure for policy improvement using LFM. A natural question is why not directly use LLMs as experts for action prediction. Section 5.4 shows that the using LLMs to learn feedback models results in higher policy improvement than using LLMs as experts for action prediction. Moreover, LFM generalizes to new environments unseen during training, thereby allowing policy improvement on new environments.

3.1. Verbalization

To leverage world knowledge in LLMs, we convert raw observations o to language descriptions v using a verbalization procedure V . Figure 2 illustrates such a verbalization procedure for Touchdown (Chen et al., 2019), where the agent navigates Google Street View panorama images based on a given natural language instruction. First, we extract all noun-phrases (NPs) from instructions in the dataset and compute their CLIP language embedding. Given a visual observation, we compute CLIP visual embedding for each image patch, and align it with the top matching NP as deemed by the highest cosine similarity between CLIP embeddings. We then combine aligned NPs with agent orientation to formulate an egocentric language description of the scene. This is described in more detail in Appendix A.

3.2. Learning a feedback model

Naively learning from LLM feedback. Given a verbalization procedure V , an instruction x , an LLM, and a policy π_θ , we now describe a procedure to use the LLM’s knowledge to improve π_θ . First, we prompt the LLM to provide feedback on whether a particular action taken by the policy $\pi_\theta(a|x, v)$ is productive in achieving the tasks outlined in the instruction x . We then improve the policy π_θ by updating its parameters to imitate desirable behaviour determined by the LLM. Let $: \text{such that}$. Let $\text{LLM}(x, v, a)$ return yes if and only if the LLM feedback indicates that action a taken in verbalized state v and instruction x is productive. Given a set of instructions $X = \{x_i\}_1^N$, the optimization procedure is then

$$\arg \min_{\theta} \mathbb{E}_{v, a', x: \text{LLM}(x, v, a') = \text{yes}} L(\pi_\theta(a|x, v), a') \quad (4)$$

where instruction x is sampled from X and the observations v and actions a' are sampled from rollouts of the policy π_θ .

Efficiently learning a language feedback model. While Eq (4) is a reasonable procedure for using LLM feedback to improve the policy, it requires calling LLMs at each step during policy improvement. This is prohibitively expensive both in terms of query cost, because LLMs capable of giving desirable feedback are expensive to run, and training time, because generating feedback using large LLMs is slow. Instead of using the LLM at each step, we make a modification to the procedure in Eq (4) to collect LLM feedback over long horizons in batch (Colas et al., 2023) in order to train a small and cost-effective language feedback model.

First, for instructions $\{x^{(1)}, x^{(2)}, \dots\}$ we roll out the base policy π_θ to collect a set of trajectories $\{\tau_1, \tau_2, \dots\}$ consisting of verbalized observations and actions taken: $\tau_i = \{v_1^{(i)} \pi(x^{(i)}, v_1^{(i)}), v_2^{(i)} \pi(x^{(i)}, v_2^{(i)}), \dots\}$. For each τ_i , we prompt the LLM for feedback on which steps were productive in achieving the instruction $x^{(i)}$. Table 2’s LFM row shows an example of requesting feedback from GPT-4 on a rollout in ALFWORLD, which is an instruction following benchmark in verbalized 3D kitchens. This LLM feedback is then parsed to identify the precise steps in which the base policy π_θ took a productive action towards achieving the goals outlined in the instruction. The set of desirable behaviour is compiled into a dataset F . Let $y^* = \text{LLM}(x, v, a)$ denote the feedback given by the LLM for the instructions x , observations v , and action a . We use the dataset $F = \{x^{(i)}, v, a, y^* \forall v, a \in \tau_i \forall x^{(i)}, \tau_i\}$ to train a small Language Feedback Model f .

$$\arg \min_{\theta} \sum_{(x, v, a, y^*) \in F} L(f_\theta(y | x, v, a), y^*) \quad (5)$$

Here, L is the cross-entropy between output distribution of the feedback model f_θ and gold label y^* from the LLM.

Learning from language feedback. The naive learning procedure in Eq (4) updates the policy after each step using slow and expensive LLM feedback. Here, we instead update the policy in rounds using fast and cost-effective LFM feedback. In round k , we rollout the base policy $\pi^{(k)}$ and use the feedback model f to collect a dataset D_k of desirable behaviour. Let $a_t^{(k)}$ denote the action chosen by policy $\pi^{(k)}(a | x, v_t)$. Let $\text{DESIRABLE}(x, v, a) = f(y = \text{yes} | x, v, a) > f(y = \text{no} | x, v, a)$, returns whether the feedback model predicts that action a is desirable. We have

$$D_k = \left\{ \left(x, v_t, a_t^{(k)} \right) \forall t : \text{DESIRABLE}(x, v_t, a_t^{(k)}) \right\} \quad (6)$$

We combine this dataset with previously collected desirable behaviour to update the base policy via imitation learning.

$$\theta^* = \arg \min_{\theta} \sum_{v_t, a_t \in \cup_{i=1}^k D_i} L(\pi^{(k)}(a | x, v_t), a_t) \quad (7)$$

In the next round, we set the parameters of the base policy $\pi^{(k+1)}$ to be θ^* . Should demonstrations be available, we initialize the base policy at $k = 1$ to the BC policy, and train on both demonstrations and identified desirable behaviour during subsequent rounds (i.e. $\cup_{i=0}^k D_i$ where D_0 is the demonstrations used to train BC).

4. Related Work

Instruction following in grounded environments. Instruction following in grounded environments has been explored in settings such as navigation (Chen & Mooney, 2011; Fried et al., 2018; Chen et al., 2019), game-playing (Andreas & Klein, 2015; Zhong et al., 2020), and robotics (Blukis et al., 2019; Shridhar et al., 2021a; Brohan et al., 2023). However, most prior work model environment observations separately from language instructions by using specialized encoders (e.g. RESNET (He et al., 2015), BERT (Devlin et al., 2019), CLIP (Radford et al., 2021)), then learn from data how to associate raw observations with language instructions. Instead of solely using raw observations, more recent work verbalize raw observations to describe environments in language (Shridhar et al., 2021b; Zhong et al., 2021; Schumann et al., 2024). In doing so, observations and instructions can be directly jointly reasoned over using language models to achieve more efficient and generalizable learning through large-scale pretraining. We build on this last direction by verbalizing raw observations into language descriptions to train language policies. However, unlike prior work that train language models to predict next actions, we develop language feedback models that critique verbalized observations and behaviour.

LLM agents in language settings. LLMs exhibit an array of reasoning abilities by pretraining on vast quantities of text (Brown et al., 2020; Wei et al., 2022). A number of recent work investigate using LLMs as language agents to exploit this reasoning ability. Nakano et al. (2022), Yao et al. (2023) Deng et al. (2023) train instruction following language agents to interact with web browsers to answer questions or interact with web pages. Ahn et al. (2022) show that a language agent can be connected with verbalized robots via API interfaces for robotic control. While powerful, these prior work are limited in that they require querying an expensive LLM on-line. In contrast, our work examines settings where an LLM is not available on-line. Specially, we use LLMs to collect a small set of off-line data for training LFM. The small and cost-effective LFM are then used to identified desirable behaviour for on-line policy improvement without additional interactions with the LLM.

Learning from feedback. An important recent extension of language agents is to augment them with feedback. Ziegler et al. (2020), Stiennon et al. (2020), and Bai

et al. (2022) learn reward models from human preference, which is then used to learn a policy via reinforcement learning (RL). Instead of using human feedback, Bai et al. (2022) and Lee et al. (2023) use LLM feedback to train a separate reward model for RL for textual alignment. Huang et al. (2022) and Yao et al. (2023) use LLMs to reason about potential resolutions to failed actions. Yuan et al. (2024) use LLMs to generate new prompts and corresponding responses, then use an LLM reward model to identify good prompt-response pairs for self-improvement in text generation alignment. Unlike these approaches, we do not use LLMs during on-line policy improvement. We train an initial small language feedback model from offline LLM data, then use this small feedback model on-line during policy improvement. Additionally, we focus on-line improvement via language feedback for long-horizon, sparse reward, grounded environments instead of text generation alignment. Our procedure for batched, on-line imitation learning is similar to DAGGER (Ross et al., 2011), which we compare to in Appendix C. However, we collect batched expert feedback to identify desirable behaviour instead of corrective actions.

5. Experiments and Analysis

We evaluate using Language Feedback Model for policy improvement on three distinct language grounding benchmarks. We compare this method against directly using LLMs as an expert policy for imitation learning. Formally, the **environments** from a **benchmark** are distinct partially-observed Markov Decision Processes that share some (or all) of the environment dynamics but have different instructions, observations, and/or action space.

5.1. Evaluation benchmarks

Table 1 shows examples of verbalized environments and tasks from each benchmark. Each benchmark provides distinct training and test environments to test generalization. In each environment, the agent takes actions to perform tasks outlined in a language instruction. The task is considered completed if and only if the agent solves the tasks within the preallocated number of steps. We evaluate using task-completion rate over test environments. The statistics from each benchmark is shown in Appendix A Table 6. These three benchmarks share challenges in sparse, delayed reward, partial observability, and compositional generalization to unseen tasks and environments.

ALFWorld is a verbalization of ALFRED (Shridhar et al., 2020), a natural language instruction following benchmark set in a 3D simulated kitchen. Here, the agent interacts with objects in kitchens to achieve compositional goals such as cleaning then microwaving potatoes. In ALFWorld (Shridhar et al., 2021b), raw state information from ALFRED are

Policy Improvement using Language Feedback Models

Table 1: Examples of verbalized environments. For brevity, we abbreviate long verbalized observations using “...”.

Benchmark	Context	Action
ALFWorld	Task: heat some egg and put it in diningtable. Observation: You arrive at loc 12. On the sinkbasin 1, you see... T-1 Observation: You are in the middle of a room... Action:go to sinkbasin 1 T-2 Observation: ...	go to microwave 1
ScienceWorld	Task: Your task is to find a(n) living thing. First, focus on the thing. Then, move it to the purple box in the bathroom. Observation: You move to the kitchen. This room is called the kitchen. In it, you see: — the agent — a substance called air — a chair. On the chair is... In your inventory, you see: — an orange... T-1 Observation: The door is now open. Action: go to kitchen T-2 Observation... Action: open door to kitchen	open door to outside
Touchdown	Task: Follow the flow of traffic, with the row of flowers on your left and make a left at the intersection. There will be a white Billboard... Observation: behind you, you see: the right lane intersection, a large, blocky, gray... T-1 Observation: behind you, slightly... Action: slightly to your left ...	straight ahead

used to populate language templates that describe observations in language.

ScienceWorld is a textual simulation benchmark for basic science experiments (Wang et al., 2022). The agent interacts with objects to conduct experiments specified in natural language, such as determining the boiling temperature of a material. ScienceWorld is uniquely challenging due to the large amount of variations in task types (30), and parametric variations (10-1400) such as the specific substance to be melted. Furthermore, ScienceWorld has a substantially larger action space and longer horizon tasks.

Touchdown is a navigation benchmark where the agent navigates Google Street View images to follow long, compositional instructions (Chen et al., 2019). Touchdown requires jointly reasoning over natural images from Google Streetview with occlusion and multi-sentence natural language instructions that describe long-horizon goals. We introduce a new verbalization procedure for Touchdown based on matching noun-phrases and image patches with CLIP embeddings to populate egocentric language templates. Behaviour cloning using our verbalization, detailed in Appendix A, outperforms prior state-of-the art results (Schumann et al., 2024).

5.2. Methods

We train BC baseline policies using existing demonstrations for each benchmark. We examine three different techniques for improving the BC policy. Table 2 shows examples of LLM prompts used for each technique.

ACTPRED: imitation learning from LLM experts. We compare to directly using LLMs as experts to predict actions for imitation learning. First, we execute k steps of the base policy, then query the LLM for the next action a given the instruction x and the verbalized observations v . We

repeatedly collect examples (x, v, a) , then train the policy using this collected data and BC demonstrations.

LFM: imitation learning using feedback models. We learn a small and cost-effective feedback model described in Section 3.2 to identify desirable behaviour for imitation learning. First, we learn a feedback model on the training environments. Second, we use the feedback model to identify desirable behaviour in the training environments for policy improvement via imitation learning. To collect LLM feedback for training LFM, we collect one rollout for each environment in a benchmark and sample 10k 20-step windows from the rollouts. Crucially, we limit the amount of feedback data collected from the LLM such that the number of output tokens produced by the LLM is identical to ACT-PRED (we use 100k GPT-2 tokens for all benchmarks). This answers whether feedback model is a more cost-effective than direct action prediction for imitation learning.

LFMA: one-shot adaptation using feedback models. LFM only imitates desirable behaviour in training environments. In contrast, LFMA adapts the policy to test environments. Given new test environments, we identify desirable behaviour using feedback models trained on the training environments, then perform one round of imitation learning to adapt to new test environments. This experiment tests whether language feedback models generalize to new environments, and whether we can use their feedback to adapt policies to new environments without using LLMs nor additional demonstrations.

5.3. Experiment details

We use the GPT-4 LLM (2023-03-15) for action prediction and feedback. We fine-tune the 770M FLAN-T5 (Chung et al., 2022) to obtain policy and feedback models. We use descriptions of the most recent 20 steps as the

Table 2: LLM prompts used to collect desirable behaviour for imitation learning. ACTPRED uses LLMs to directly generate the appropriate action for each step, whereas LFM uses LLMs to generate, in batch, feedback that identify which taken actions were productive. For brevity, we abbreviate long verbalized observations using “...”.

ACTPRED	
Prompt	LLM Output
Your task is: look at alarmclock under the desklamp. You see: you are in the middle of a room. looking quickly around you, you see a bed 1, a desk 1, a drawer 17... what do you decide to do? available actions: examine shelf 1, examine shelf 2, go to bed... You decide to: go to desk 1. You see: you arrive at desk 1. what do you decide to do? available actions: examine desk 1... You decide to: examine desk 1	
LFM	
You will be shown a playthrough for solving a task. Task: put two candle in drawer. Before: You open the drawer 6. The drawer 6 is open. In it, you see nothing. Step 21. Your action: close drawer 6. Result: You close the drawer 6... Step 22. Your action: put candle 3 in/on drawer 1. Result: You put the candle 3 in... Is the player on the right track to solve the task? Answer yes or no. If yes, list the helpful steps by the step number in bullet form.	Yes - Step 28 - Step 29...

verbalized observation v . All models are trained for 10k steps with batch size 20 and early stopping over validation demonstrations. Appendix E shows details on GPU usage.

Feedback model training and inference. To train feedback models, we collect LLM feedback over 20-step windows. We then parse LLM feedback to identify whether the action taken in each step was productive to solving the tasks outlined in the instructions. We subsample the feedback data to obtain an even split of productive and not-productive actions. This data is split into a 80% train/20% validation dataset to train the LFM.

Policy training and inference. To train policies, we fine-tune language models to minimize token-wise cross-entropy of tokens in the ground-truth verbalized action. During inference time, we consider a (potentially very large) set of plausible actions given by the environment. For each action, we evaluate the policy’s language model perplexity, and choose the action with the minimum perplexity averaged over tokens.

5.4. Results and discussion

Table 3 shows the performance of the policy behaviour cloned from demonstrations BC, imitation learned from LLMs using action prediction ACTPRED, and imitation learned from LFM. For LFM, we show zero-shot results (LFM) as well as after one round of adaptation (LFMA).

LFMs improves policy performance across all benchmarks. Table 3 shows that LFM improves upon the strong behaviour cloning baseline policy BC in all benchmarks. Ta-

Table 3: Task completion rate on three benchmarks. We evaluate a behaviour cloning agent BC, an imitation learning agent using LLM as the expert policy ACTPRED, and our proposed method LFM which imitates desirable behaviour identified by a language feedback model. On held-out evaluation environments, LFM outperforms other methods on all benchmarks. Furthermore, adaptation to the new environments using the trained language feedback models results in significant additional gains (LFMA).

	ALFWorld	ScienceWorld	Touchdown
BC	62.6	45.8	57.5
ACTPRED	56.0	39.0	58.0
LFM	64.1	47.1	59.7
LFMA	74.6	49.3	62.8

ble 5 shows examples of LFM-identified desirable behaviour. This shows that LFM are an effective means to leverage the knowledge in pretrained LLMs for policy improvement in language-grounded environments, which agree with human-identified desirable behaviour. Appendix D also compares GPT-4 to the open-source LLAMA 2 70B for training feedback models using human evaluation. We find that GPT-4 consistently outperforms LLAMA 2, which tends to identify spurious desirable behaviour.

Learning LFM is more cost-effective than using LLMs for action prediction. Assuming the same LLM output-token quota, Table 3 compares using LLMs to train feedback models (LFM) to using LLMs to predict actions (ACTPRED) for policy improvement. Specifically, ACTPRED tends to predict spurious actions, especially for complex environ-

Table 4: Feedback performance measured by F1 score. We label steps the LLMs consider to be productive to be “positive” actions and other steps negative actions. We measure the F1 score of the positive/negative predictions made by the learned LFM using the LLM predictions as ground truth. We observe no significant performance degradation when using a much more detailed feedback model (LFMD) that also provides explanations behind the feedback, summaries of agent behaviour, and strategy suggestions.

	ALFWorld	ScienceWorld	Touchdown
LFM	93.2	83.7	43.9
LFMD	92.0	82.5	42.5

ments with large actions spaces such as ScienceWorld. In contrast, the difficulty in identifying productive actions is independent of the action space, and LFM consistently improves policy even with large action spaces. This shows that LFMs is a more cost-effective means use LLMs for policy improvement compared to using LLMs as expert policies for imitation learning.

LFMs generalize to new environments, allowing for policy adaptation without additional LLM usage nor demonstrations. Table 4 shows that LFMs trained during language feedback learning can accurately recognize desirable behaviour in new environments. Table 3 shows that imitating this behaviour obtains significant policy improvement across all benchmarks. This shows that LFMs generalize to new environments, which allows for policy adaptation to new environments despite not having demonstrations nor LLM access.

LFMs can provide human-interpretable feedback, allowing human-in-the-loop verification during policy improvement. LFMs improve policy performance with succinct feedback. Here, we extend them to additionally provide detailed explanations. Consider an instruction “turn left when you see the stop sign then go to the second building on the right”. Suppose that in the current step the agent proceeds straight, arriving at the stop sign. Instead of a feedback saying “yes” (i.e. the action was productive), the LFM can provide a human-interpretable explanation for why this action was productive (i.e. “yes because you found the stop sign where you are supposed to turn”). Table 5 shows that we can enhance LFM to produce detailed feedback by training detailed feedback prompted from LLMs. Specifically, we train a detailed LFMD to simultaneously identify productive actions, summarize agent intent, and suggest potential high level recovery strategies. Table 4 shows that surprisingly, LFMD that produce detailed feedback perform similarly to those that provide succinct feedback. This shows that Language Feedback Models can be used to provide accurate feedback interpretable to humans. While inter-

Table 5: Example of detailed language feedback, which uses more LLM interaction to provide human-interpretable feedback. The prompt and output differences between succinct feedback (Table 2) and detailed feedback are highlighted in red. Note that the second row shows shows batched LLM inference, which is used to train a step-wise feedback model. The step-wise inference of the feedback model is shown in the last row. We leave providing summaries and suggestions with the learned feedback model to future work.

Prompt

You will be shown a playthrough for solving a task.

Task: put two candle in drawer.

Before: You open the drawer 6. The drawer 6 is...

Step 21. Your action: close drawer 6. Result: You ...

Is the player on the right track to solve the task?

Next under heading #Summary, summarize in one sentence what the player doing at a high level. Finally under heading #Improvement, describe how the player can improve their strategy to solve the task.

LLM output

Yes

- Step 28: The player locates the candles by arriving at countertop 1, where a candle is available.

- Step 29: The player successfully takes candle 3 from countertop 1, demonstrating the ability to collect...

#Summary

The player is gathering candles and placing them in a drawer, working towards completing the task of putting two...

#Improvement

The player can improve their strategy by:

- Not closing drawers unnecessarily after checking...
- Prioritizing finding and collecting the candles first...

Learned feedback model output

Yes. The player successfully locates the candle...

Yes. The player correctly takes the candle from...

pretable feedback requires more costly LLM usage, it allow for human-in-the loop verification of desirable behaviour identified by the LFM. Consequently, interpretable LFMs promotes user trust in the quality of the imitation learning data and subsequent policy behaviour.

6. Conclusion

We introduced Language Feedback Models that identify desirable behaviour for imitation learning. On three instruction following benchmarks, small and cost-effective LFMs consistently outperform BC baselines and using LLMs as experts for imitation learning, without using LLMs during policy improvement. In addition, LFMs generalize and provide significant policy adaptation gains on new environments, without using LLMs nor new demonstrations. Finally, LFMs, can provide detailed human-interpretable feedback that human verification of imitation data. We advocate for future exploration of how to exploit detailed LFMs, such as learning dense, subgoal-aware reward models for RL, and trustworthy policies with human verification.

7. Broader Impact

This paper presents work on improving instruction following using Language Feedback Models. Potential beneficial societal consequences of this work include the development of cost-effective computer agents that quickly learn to accurately follow human commands. The method we present in this work learns an language feedback model trained by exploiting world knowledge in LLMs. We show that this technique results in faster and more cost-effective policy improvement than using LLMs as experts. Furthermore, we show that the feedback models we learn generalize to new environments, which results in significant gains for new environments not seen during training via policy adaptation. Finally, we show that language feedback models can be extended to provide detailed critique that include explanations, agent intent summaries, and high-level strategy recommendations. Learning to provide this detailed output results in no noticeable accuracy cost, and can provide interpretable feedback humans can inspect and verify to create more trustworthy policies.

Potential negative societal consequences of this work include hallucinations by LLMs that mislead feedback model training. In this sense, feedback models may learn to encourage actions that do not achieve language goals (e.g. exploring the bathroom during a kitchen cleaning task). Furthermore, they may encourage actions that help achieve goals but are undesirable in other ways (e.g. unsafely climbing over the table to reach the sofa more quickly). In future work, we will explore using techniques in LLM alignment to learn more robust language feedback models, as well as investigate learning from detailed feedback models with human verification to improve the trustworthiness of downstream policies.

References

- Abbeel, P. and Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In *ICML*, 2004.
- Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Fu, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Ho, D., Hsu, J., Ibarz, J., Ichter, B., Irpan, A., Jang, E., Ruano, R. J., Jeffrey, K., Jesmonth, S., Joshi, N. J., Julian, R., Kalashnikov, D., Kuang, Y., Lee, K.-H., Levine, S., Lu, Y., Luu, L., Parada, C., Pastor, P., Quiambao, J., Rao, K., Rettinghouse, J., Reyes, D., Sermanet, P., Sievers, N., Tan, C., Toshev, A., Vanhoucke, V., Xia, F., Xiao, T., Xu, P., Xu, S., Yan, M., and Zeng, A. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances, 2022.
- Andreas, J. and Klein, D. Alignment-based compositional semantics for instruction following. In *EMNLP*, 2015.
- Ash, J. T. and Adams, R. P. On Warm-Starting Neural Network Training. In *NeurIPS*, 2020.
- Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., Kerr, J., Mueller, J., Ladish, J., Landau, J., Ndousse, K., Lukosuite, K., Lovitt, L., Sellitto, M., Elhage, N., Schiefer, N., Mercado, N., DasSarma, N., Lasenby, R., Larson, R., Ringer, S., Johnston, S., Kravec, S., Showk, S. E., Fort, S., Lanham, T., Telleen-Lawton, T., Conerly, T., Henighan, T., Hume, T., Bowman, S. R., Hatfield-Dodds, Z., Mann, B., Amodei, D., Joseph, N., McCandlish, S., Brown, T., and Kaplan, J. Constitutional AI: Harmlessness from AI Feedback, 2022.
- Bain, M. and Sammut, C. A framework for behavioural cloning. In *Machine Intelligence*, 1995.
- Bengio, S., Vinyals, O., Jaitly, N., and Shazeer, N. Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks. In *NeurIPS*, 2015.
- Blukis, V., Terme, Y., Niklasson, E., Knepper, R. A., and Artzi, Y. Learning to Map Natural Language Instructions to Physical Quadcopter Control using Simulated Flight. In *CoRL*, 2019.
- Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., Florence, P., Fu, C., Arenas, M. G., Gopalakrishnan, K., Han, K., Hausman, K., Herzog, A., Hsu, J., Ichter, B., Irpan, A., Joshi, N., Julian, R., Kalashnikov, D., Kuang, Y., Leal, I., Lee, L., Lee, T.-W. E., Levine, S., Lu, Y., Michalewski, H., Mordatch, I., Pertsch, K., Rao, K., Reymann, K., Ryoo, M., Salazar, G., Sanketi, P., Sermanet, P., Singh, J., Singh, A., Soricut, R., Tran, H., Vanhoucke, V., Vuong, Q., Wahid, A., Welker, S., Wohlhart, P., Wu, J., Xia, F., Xiao, T., Xu, P., Xu, S., Yu, T., and Zitkovich, B. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control, 2023.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language Models are Few-Shot Learners. In *NeurIPS*, 2020.
- Chen, D. and Mooney, R. Learning to Interpret Natural Language Navigation Instructions from Observations. In *AAAI*, 2011.

- Chen, H., Suhr, A., Misra, D., Snavely, N., and Artzi, Y. Touchdown: Natural Language Navigation and Spatial Reasoning in Visual Street Environments. In *CVPR*, 2019.
- Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. Scaling Instruction-Finetuned Language Models, 2022.
- Colas, C., Teodorescu, L., Oudeyer, P.-Y., Yuan, X., and Côté, M.-A. Augmenting autotelic agents with large language models. In *Proceedings of The 2nd Conference on Lifelong Learning Agents*, PMLR. PMLR, 2023.
- Deng, X., Gu, Y., Zheng, B., Chen, S., Stevens, S., Wang, B., Sun, H., and Su, Y. Mind2Web: Towards a Generalist Agent for the Web. In *NeurIPS*, 2023.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, 2019.
- Fried, D., Andreas, J., and Klein, D. Unified Pragmatic Models for Generating and Following Instructions. In *NAACL*, 2018.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition, 2015.
- Honnibal, M. and Montani, I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.
- Huang, W., Xia, F., Xiao, T., Chan, H., Liang, J., Florence, P., Zeng, A., Tompson, J., Mordatch, I., Chebotar, Y., Sermanet, P., Brown, N., Jackson, T., Luu, L., Levine, S., Hausman, K., and Ichter, B. Inner Monologue: Embodied Reasoning through Planning with Language Models. In *CoRL*, 2022.
- Kollar, T., Tellex, S., Roy, D., and Roy, N. Toward understanding natural language directions. In *HRI*, 2010.
- Lee, H., Phatale, S., Mansoor, H., Mesnard, T., Ferret, J., Lu, K., Bishop, C., Hall, E., Carbune, V., Rastogi, A., and Prakash, S. RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback, 2023.
- MacMahon, M., Stankiewicz, B., and Kuipers, B. Walk the talk: Connecting language, knowledge, and action in route instructions. In *AAAI*, 2006.
- Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., Jiang, X., Cobbe, K., Eloundou, T., Krueger, G., Button, K., Knight, M., Chess, B., and Schulman, J. WebGPT: Browser-assisted question-answering with human feedback, 2022.
- OpenAI. GPT-4 Technical Report. <https://arxiv.org/abs/2303.08774v4>, 2023.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision, 2021.
- Ross, S., Gordon, G. J., and Bagnell, J. A. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In *AISTATS*, 2011.
- Schaal, S. Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences*, 1999.
- Schumann, R., Zhu, W., Feng, W., Fu, T.-J., Riezler, S., and Wang, W. Y. VELMA: Verbalization Embodiment of LLM Agents for Vision and Language Navigation in Street View. In *AAAI*, 2024.
- Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., Zettlemoyer, L., and Fox, D. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *CVPR*, 2020.
- Shridhar, M., Manuelli, L., and Fox, D. CLIPort: What and Where Pathways for Robotic Manipulation. In *CoRL*, 2021a.
- Shridhar, M., Yuan, X., Côté, M.-A., Bisk, Y., Trischler, A., and Hausknecht, M. ALFWORLD: Aligning Text and Embodied Environments for Interactive Learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021b.
- Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. Learning to summarize from human feedback. In *NeurIPS*, 2020.
- Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press, 2018.
- Torabi, F., Warnell, G., and Stone, P. Behavioral cloning from observation, 2018.
- Wang, R., Jansen, P. A., Côté, M.-A., and Ammanabrolu, P. Scienceworld: Is your agent smarter than a 5th grader? In *EMNLP*, 2022.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B.,
Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-Thought
Prompting Elicits Reasoning in Large Language Models.
In *NeurIPS*, 2022.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan,
K., and Cao, Y. ReAct: Synergizing Reasoning and
Acting in Language Models. In *ICLR*, 2023.

Yuan, W., Pang, R. Y., Cho, K., Sukhbaatar, S., Xu, J., and
Weston, J. Self-Rewarding Language Models, 2024.

Zhong, V., Rocktäschel, T., and Grefenstette, E. RTFM:
Generalising to Novel Environment Dynamics via Read-
ing. In *ICLR*, 2020.

Zhong, V., Hanjie, A. W., Wang, S. I., Narasimhan, K., and
Zettlemoyer, L. SILG: The Multi-environment Symbolic
Interactive Language Grounding Benchmark. In *NeurIPS*,
2021.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Rad-
ford, A., Amodei, D., Christiano, P., and Irving, G.
Fine-Tuning Language Models from Human Preferences,
2020.

Table 6: Statistics from benchmarks as measured by training demonstrations. The are the average number of GPT-2 tokens in the instruction, verbalized observation, and action; the average demonstration steps; the average number of plausible actions in a state; the number of unique actions, instructions, and observations; and finally the number of training demonstrations.

	ALFWorld	SciWorld	Touchdown
Ins len $ x $	8.8	64.7	93.4
Obs len $ v $	23.9	239.4	284.9
Act len $ a $	4.5	6.0	2.4
Traj len $ \tau $	19.7	55.1	34.2
$ \text{Act space} $	29.9	1.9k	2.1
# act $ \{a\} $	2.6k	2.4k	8
# ins $ \{\tau\} $	1.0k	1.2k	6.5k
# obs $ \{v\} $	18.2k	157k	34.3k
# demos	3.5k	3.6k	6.5k

A. Verbalization of visual environments

How can we leverage the world knowledge learned by LLMs from pretraining on vast quantities of text? In many instruction following problems, environment observations are inherently visual. In this section, we describe a verbalization procedure that converts visual observations to language descriptions, so that LLMs can make inferences by jointly referring to the instruction and environment observations. Specifically, we use Touchdown as an example.

As shown in Figure 2, Touchdown (Chen & Mooney, 2011) is primarily a visual navigation problem. Given a set of connected Google Streetview panoramas that represent neighbourhoods in Manhattan, an agent must follow long, complex natural language instructions to navigate to the correct location. Crucial to this task of navigation are **landmarks** referred to by the instructions. For instance, the instruction “turn so the **scaffolding** is on your left and... to the **next corner** and turn right...” refers to the landmarks **scaffolding** and **next corner**. Prior work in verbalization use LLMs to identify landmarks (Schumann et al., 2024). In this work, we take the set of common noun-phrases in the corpus of instructions to be landmarks.

Extracting aligned noun-phrase annotations for visual patches First, we identify all noun-phrases using SPACY (Honnibal & Montani, 2017). Given a visual scene, we divide the scene into 300x300 pixel non-overlapping patches. For each patch, we identify the noun-phrase with the highest cosine similarity between the noun-phrase text embedding and the image patch embedding. We use text and visual encoders from CLIP (Radford et al., 2021) to extract embeddings for each modality. For patches with no aligned noun-phrase with cosine similarity greater than 0.2, we do not provide annotated a noun-phrase.

Converting to verbalized observations To obtain verbalized observations in the form of an egocentric scene description, we consider the direction the agent is facing (shown in blue) as well the directions of possible next steps (shown in green). The noun-phrases identified in the scene are then categorized into 8 directions in 45-degree increments, relative to the agent’s current orientation: straight ahead (337.5 to 22.5), slightly to your right (22.5 to 67.5), to your left (67.5 to 112.5), behind you, slightly to your right (112.5 to 157.5), behind you (157.5 to 202.5), behind you, slightly to your left (202.5 to 247.5), to your left (247.5 to 292.5), and slightly to your left (292.5 to 337.5). A scene is then rendered as follows:

```

Straight ahead, you see
- a white van
Slightly to your right, you see
- a red brick building
- a scaffold...

```

This verbalization achieves significantly higher task-completion accuracy compared to prior state-of-the-art results, after behavioural cloning from demonstrations. For instance, we obtain 57.5% task-completion compared to 26.4% by Schumann et al. (2024).

Statistics of verbalized environments In Appendix A Table 6, we show statistics of verbalized environments as quantitative evidence of their challenges.

B. Pseudocode for Policy Improvement using Language Feedback Models

In this section we detail, using pseudocode, the procedure for policy improvement using Language Feedback Models. Algorithm 1 describes learning a model from LLMs. Algorithm 2 describes identifying desirable behaviour that are productive for solving tasks specified in the instruction, and then using this behaviour for imitation learning. Algorithm 3 describes the iterative policy improvement procedure using these two algorithms.

Algorithm 1 TRAINFEEDBACKMODEL: Training a Language Feedback Model using LLM feedback.

```

1: Inputs: initial policy  $\pi$ , LLM LLM, environment  $E$ 
2: Feedback dataset  $F \leftarrow \{\}$ 
3: for  $i = 1 \dots N$  do
4:    $x \leftarrow \text{SAMPLEINSTRUCTION}$ 
5:    $\tau_i \leftarrow \text{ROLLOUT}(\pi, E, x)$ 
6:   for window  $w_j$  in  $\tau_i$  do
7:      $y \leftarrow \text{QUERYLLMFORFEEDBACK}(LLM, w_j, x)$ 
8:     for verbalized observation  $v_k$ , LLM feedback  $y_k$  in each step of  $y$  do
9:        $F \leftarrow F \cup (v_k, y_k)$ 
10:    end for
11:   end for
12: end for
13: Feedback model  $f \leftarrow \text{TRAINLM}(F)$ 
```

Algorithm 2 IMITATEUSINGFEEDBACK: Imitation learning using desirable behaviour identified by a feedback model.

```

1: Inputs: base policy  $\pi$ , environment  $E$ , feedback model  $f$ 
2: Imitation dataset  $G \leftarrow$  behaviour cloning dataset
3: for  $i = 1 \dots N$  do
4:    $x \leftarrow \text{SAMPLEINSTRUCTION}$ 
5:    $\tau_i \leftarrow \text{ROLLOUT}(\pi, E, x)$ 
6:   for verbalized observation  $v_k$ , action  $a_k$  in each step of  $\tau_i$  do
7:      $y_k = f(v_k)$ 
8:     if  $y_k$  is desirable then
9:        $G \leftarrow G \cup (v_k, a_k)$ 
10:    end if
11:   end for
12: end for
13: Improved policy  $\pi' \leftarrow \text{TRAINLM}(G)$ 
```

Algorithm 3 Policy improvement using Language Feedback Models.

```

1: Inputs: base policy  $\pi$ , environment  $E$ 
2: Feedback model  $f \leftarrow \text{TRAINFEEDBACKMODEL}(\pi, LLM, E)$ 
3:  $\pi_0 \leftarrow \pi$ 
4: for  $k = 1 \dots N$  do
5:    $\pi_k \leftarrow \text{IMITATEUSINGFEEDBACK}(\pi_{k-1}, E, f)$ 
6: end for
```

Table 7: Task completion rate on evaluation benchmarks, including DAGGER.

	ALFWorld	ScienceWorld	Touchdown
BC	62.6	45.8	57.5
ACTPRED	56.0	39.0	58.0
DAGGER	55.2	22.5	50.2
LFM	64.1	47.1	59.7
LFMA	74.6	49.3	62.8

C. Comparison to DAGGER

Our main experiments in Section 5.4 illustrate the difficulty of using LLMs as an expert to predict actions. Specifically, we show that when these predictions are used for imitation learning, the resulting policy improvement is worse than using Language Feedback Models. This performance degradation is exacerbated in environments with larger action spaces, such as ScienceWorld.

DAGGER (Ross et al., 2011) is an intermediate method between Language Feedback Models and using LLMs as an expert policies for imitation learning. Specifically, in DAGGER, we also use LLMs as experts to predict action. However, instead of using LLMs during each step, in DAGGER we use LLMs to provide batched retroactive action prediction similar to how in Language Feedback Models we use LLMs to provide batched retroactive feedback. Here, we apply DAGGER action prediction to the exact same number of examples as when we collect feedback data for LFM. In Table 7, we compare DAGGER performance to those using LLM as an expert (ACTPRED) and using Language Feedback Models (LFM). We find that although DAGGER is more efficient than ACTPRED in that it annotates synthetic examples in batch, it underperforms ACTPRED (and consequently LFM) across all benchmarks.

Table 8: Agreement between GPT-4 and LLAMA 2 across the benchmarks. We collect steps from rollouts on the training environments where either GPT-4 or LLAMA 2 identified a productive action. This table shows percentage of those actions that are identified exclusively by GPT-4, exclusively by LLAMA 2, and identified by both models. The total number of steps identified are 40569 for ALFWorld, 68565 for ScienceWorld, and 90529 for Touchdown.

	GPT-4 only	LLAMA 2 only	both
ALFWorld	14.4%	49.3%	36.2%
ScienceWorld	10.2%	62.3%	27.5%
Touchdown	22.3%	67.3%	10.4%

Table 9: Human verification of LLM feedback in terms of percentage of true positives and false positives. A true positive (TP) is a step that is correctly identified by the LLM as being productive to solving the task. A false positive (FP) is a step that is wrongly identified by the LLM as productive. We manually evaluate 10 examples from each benchmark, each with up to 20 steps. Support (# of steps) is shown in brackets.

	GPT-4		LLAMA 2	
	TP	FP	TP	FP
ALFWorld	100% (22)	0	32% (18)	68% (38)
ScienceWorld	78% (38)	22% (11)	48% (38)	52% (41)
Touchdown	81% (22)	19% (5)	39% (24)	61% (38)

D. Quantitative and Qualitative Analyses of Learned Language Feedback

Comparison of GPT-4 to LLAMA 2 70B How much difference is there between language feedback obtained from the open-source LLAMA 2 vs from GPT-4? Table 8 shows that, surprisingly, there is a large degree of disagreement between GPT4 and LLAMA 2. Specifically, LLAMA 2 identifies significantly more actions as being productive to achieving the goal.

We perform a manual analysis of language feedback by GPT-4 and LLAMA 2 to characterize qualitative differences between feedback collected by these two models. First, we roll out BC policies, then ask each model for feedback. Each example contains a segment of up to 20 steps extracted from a rollout, and the LLM is prompted to list productive steps. For each step the LLM identifies as productive to solving the task, we manually verify whether the step is indeed productive. We manually inspect 10 examples from each model for each benchmark, for a total of $10 \times 2 \times 3 = 60$ examples. Table 9 shows the number of true and false positives predicted by both models in this manual evaluation. We find that a significant number of steps are incorrectly determined by LLAMA 2 as desirable. When we train the policy on a combination of LLAMA 2 data and demonstrations used to learn the BC policy, we obtain worse task-completion percentage than using GPT-4 data and demonstrations. Specially, performance drop from 64.1% (GPT-4) to 56.0% (LLAMA 2) on ALFWorld, from 47.1% to 47.0% on ScienceWorld, and from 59.7% to 56.5% on Touchdown.

Table 10 shows some examples of steps identified as productive by these models that illustrate LLAMA 2’s tendency to identify spurious actions as being productive. In the ALFWORLD examples, for instance, LLAMA 2 has a strong tendency to identify opening and closing cabinets and drawers as productive, even though they have nothing to do with putting a clean soap bar on the counter top (the first instruction) or putting a clean spatula on the side table (the second instruction). Similarly in ScienceWorld, LLAMA 2 identifies unnecessary actions such as going outside (example 1) and going to the bedroom (example 2) as productive, even when the instruction explicitly details that the aluminum foil is found in the kitchen (example 1) and that the unknown substance is found in the workshop (example 2). Finally, LLAMA 2 also tends to identify spurious actions as productive in Touchdown. In the last example, the instruction asks to take a left after the first intersection, but LLAMA 2 rewards the left turn during the first turn, before the agent even arrives at the first intersection. GPT-4, on the other hand, correctly identifies Step 8, when the agent finally encounters the first intersection, as productive.

We show in Section 5.4 that small and cost-effective Language Feedback Models are able to replicate LLM feedback through training. Our comparison between GPT-4 and LLAMA 2 show that a less powerful model such as LLAMA 2 are unable to provide high-quality feedback. The summary from this experiment are then that 1) powerful LLMs are necessary to provide good feedback, but expensive to run during on-line policy improvement 3) consequently, learning small LFM is an effective solution to achieve high feedback performance while reducing inference cost during policy improvement.

E. GPU Usage

We train feedback models and policies using 80GB A100 GPUs. To produce rollouts at in parallel, we use a cluster of 200 32GB V100 GPUs. For all environments, feedback model training takes under 24 hours using one A100 GPU while inference can be performed locally using a 32GB GPU under 2 hours. Policy training requires 1 day for ALFWorld, 2 days for ScienceWorld, and 3 days for Touchdown. For all environments, policy rollout over the entire evaluation environments can be performed over the cluster of 200 32GB V100 GPUs in under 6 hours.

Table 10: Example steps identified as productive by GPT-4, LLAMA 2, and both. Touchdown steps are truncated for brevity.

ALFWorld

Task: clean some soapbar and put it in countertop.

Step 3 Your action: take soapbottle 2 from toilet 1 Result: You pick up the soapbottle 2 from the toilet 1.

Step 6 Your action: go to countertop 1 Result: You arrive at loc 11. On the countertop 1, you see a candle 2, a candle 1, a soapbar 2...

Step 11 Your action: open cabinet 1 Result: You open the cabinet 1. The cabinet 1 is open. In it, you see a candle 3, and a toiletpaper 1.

Step 17 Your action: open cabinet 1 Result: You open the cabinet 1. The cabinet 1 is open. In it, you see a candle 3, and a toiletpaper 1.

Step 19 Your action: go to cabinet 3 Result: You arrive at loc 12. The cabinet 3 is closed.

Task: put a clean spatula in sidetable.

Step 3 Your action: go to sidetable 1 Result: You arrive at loc 22. On the sidetable 1, you see nothing.

Step 10 Your action: go to drawer 3 Result: You arrive at loc 19. The drawer 3 is closed.

Step 11 Your action: open drawer 3 Result: You open the drawer 3. The drawer 3 is open. In it, you see nothing.

Step 14 Your action: open drawer 4 Result: You open the drawer 4. The drawer 4 is open. In it, you see nothing.

Step 15 Your action: close drawer 4 Result: You close the drawer 4.

Step 19 Your action: go to drawer 6 Result: You arrive at loc 27. The drawer 6 is closed.

Step 20 Your action: open drawer 6 Result: You open the drawer 6. The drawer 6 is open. In it, you see nothing.

ScienceWorld

Task: Your task is to determine if aluminum foil is electrically conductive. The aluminum foil is located around the kitchen. First, focus on the aluminum foil. If it is electrically conductive, place it in the yellow box. If it is electrically nonconductive, place it in the purple box. Do not focus on objects not central to completing the task!

Step 21 Your action: open door to greenhouse Result: The door is now open...

Step 23 Your action: open door to outside Result: The door is now open...

Step 25 Your action: open door to foundry Result: The door is now open...

Step 29 Your action: go to outside Result: You move to the outside...

Step 31 Your action: go to kitchen Result: You move to the kitchen...

Step 32 Your action: pick up aluminum foil Result: You move the aluminum foil to the inventory...

Step 33 Your action: focus on aluminum foil Result: You focus on the aluminum foil...

Step 37 Your action: go to workshop Result: You move to the workshop...

Step 38 Your action: move aluminum foil to yellow box Result: You move the aluminum foil to the yellow box...

Task: Your task is to determine if unknown substance N is electrically conductive. The unknown substance N is located around the workshop. First, focus on the unknown substance N. If it is electrically conductive, place it in the blue box. If it is electrically nonconductive, place it in the orange box. Do not focus on objects not central to completing the task!

Step 1 Your action: open door to art studio Result: The door is now open...

Step 4 Your action: go to hallway Result: You move to the hallway...

Step 6 Your action: go to bedroom Result: You move to the bedroom...

Step 9 Your action: open door to hallway Result: The door is already open...

Step 11 Your action: open door to workshop Result: The door is now open...

Step 12 Your action: go to workshop Result: You move to the workshop...

Step 13 Your action: pick up unknown substance Result: You move the unknown substance N to the inventory...

Step 15 Your action: move unknown substance to blue box Result: You move the unknown substance N to the blue box...

Touchdown

Task: Going with the flow of traffic, proceed to the intersection and take left. There should be some blue scaffolding on your right hand side, take a left here. You should now have a tall solid green fence on you right hand side, go straight to the next intersection and take a left. There's a bunch of scaffolding on the left hand side, go straight. Pass the first intersection you come across. You will pass mail trucks parked on both sides of the road.

Step 1 Your action: slightly to your left Result: behind you, you see: parked, white box trucks , two wider sidewalks , this narrow two lane road , the right sidewalk buildings . behind you, sightly to your left, you see: , three air-conditioners , three awning , a smaller yellow taxi . to your left, you see: , the theater awning , a yellow cab car , the second purple awning . slightly to your left, you see: , a white-capped hydrant , ornate gray balconies , the purple wayfair truck . straight ahead, you see: , a median strip , some tall, brick buildings , parked, white box trucks , a blue bus lane sign , the right sidewalk buildings . slightly to your right, you see: , this nearest intersection , a light colored stone building , a blue and red bus stop sign to your right, you see: , separate building entrances , the half-moon curb , the company's name . behind you, slightly to your right, you see: , , then a storefront , ornate gray balconies , the first little weird intersection.

Step 8 Your action: straight ahead Result: behind you, you see: , a brown storefront , surface streets , the right sidewalk buildings , a parked black box truck , some unremarkable brick buildings . behind you, sightly to your left, you see: , then a storefront , a white/grey van , a large, blocky, gray building . to your left, you see: , a large white store sign , a construction vehicle , a long gray and white 5 story building slightly to your left, you see: , some tall, brick buildings , fedex van , a white/grey van . straight ahead, you see: , a large red brick apartment building , an orange and white traffic object , 3rd and 4th intersections , a small blue car , the right sidewalk buildings , the parked yellow suv taxi . slightly to your right, you see: , construction awning , a large white fedex truck , a white/grey van . to your right, you see: , this van stop , the building edge , a large white fedex truck . behind you, slightly to your right, you see: , a green-topped scaffolding , a large white fedex truck , a parked black box truck.
