PAPER POSTER

Assessing the Zero-Shot Capabilities of LLMs for Action Evaluation in RL

E. Pignatelli[1] J. Ferret[2] T. Rocktäschel[1,2] E. Grefenstette[1,2]
D. Paglieri[1] S. Coward[3] L. Toni[1]

https://openreview.net/forum?id=MFw8K57o5I

[1] University College London [2] Google DeepMind [3] University of Oxford

CONTEXT

+ THE CREDIT ASSIGNMENT PROBLEM

The **CREDIT ASSIGNMENT PROBLEM (CAP)** in RL, that is:

- To attribute the appropriate **influence** (how impactful)
- to **actions** in a trajectory
- for their ability to achieve a certain **goal**

In short:

To **EVALUATE** actions: *How good is* a *to achieve* g*?*

CONTEXT

+ ASSUMPTIONS

PRIMARY REWARD

External, innate, hardwired (independent from us)

Function of the future, e.g., $R(x_T)$

SECONDARY REWARD

Acquired, learnt from experience (depends on us)

Function of the present, e.g., $f(x_t)$

Actualises the primary reward, in a way that informs how to act in the present

CREDIT

Measurable (e.g., action-values, advantage, …)

CONTEXT

+ WHY BOTHERING WITH THE CAP?

Accurate **CREDIT** is key

as it provides **DIRECTIONS** to **IMPROVE** the policy:

Good evaluation → Effective improvements

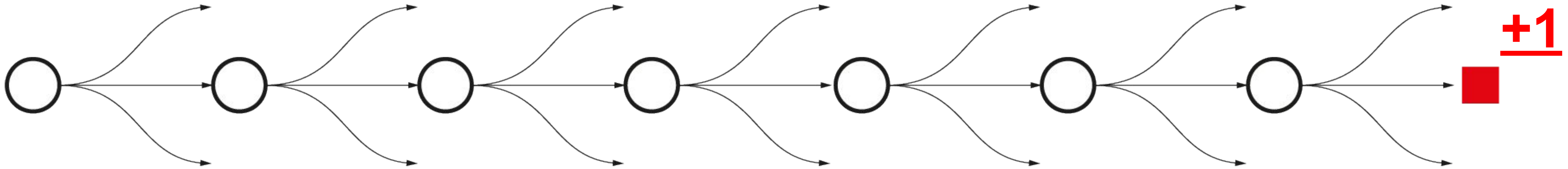
#1: MOTIVATION

MOTIVATION

+ **SO WHAT?**

The CAP is significantly **HARD**(er) when rewards are:

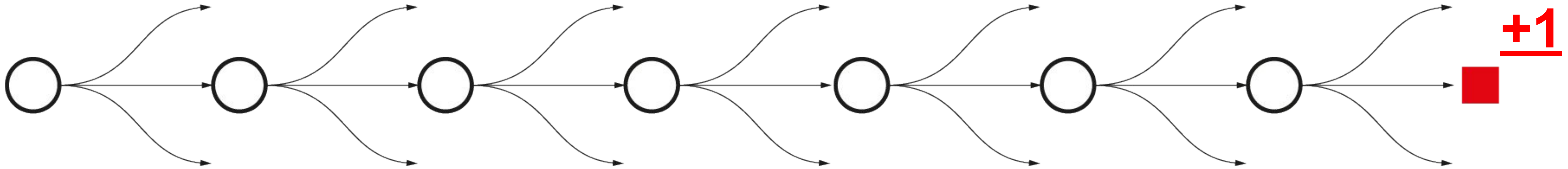
1. **DELAYED** (in time)
2. **SPARSE** (in state space)



MOTIVATION

+ SOTA

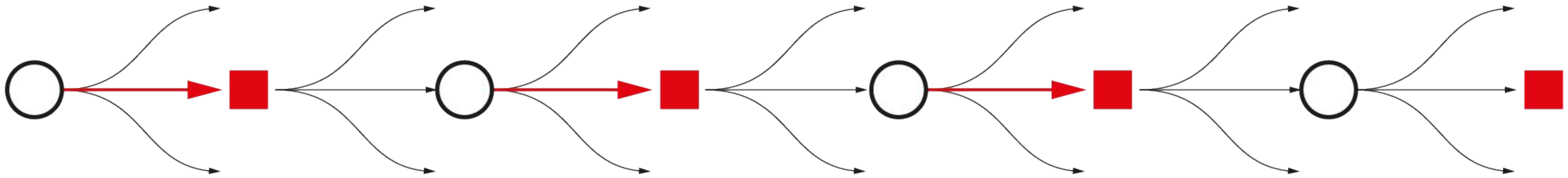
State-of-the-art methods work by **DENSIFYING** the reward function by providing **INTERMEDIATE FEEDBACK** where the MDP does not.



MOTIVATION

+ **SOTA**

State-of-the-art methods work by **DENSIFYING** the reward function by providing **INTERMEDIATE FEEDBACK** where the MDP does not.

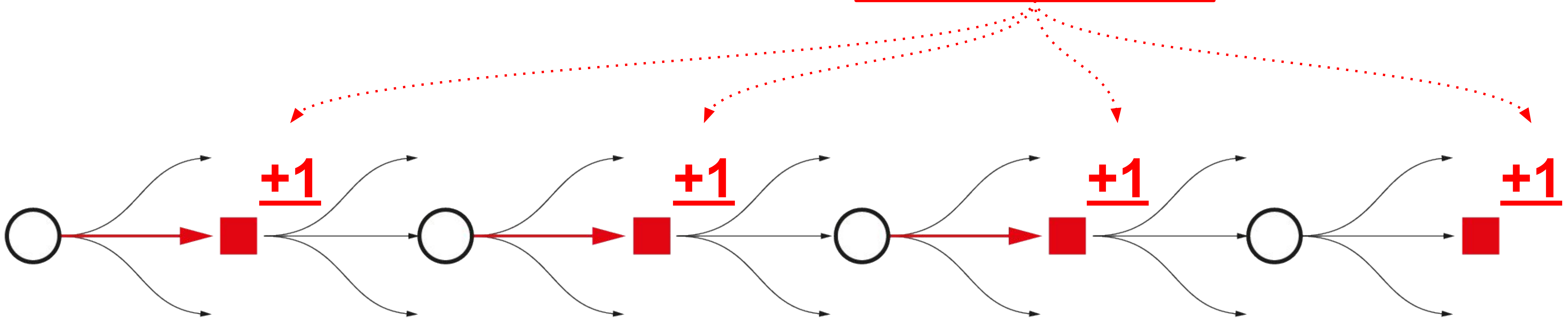


MOTIVATION

+ **SOTA**

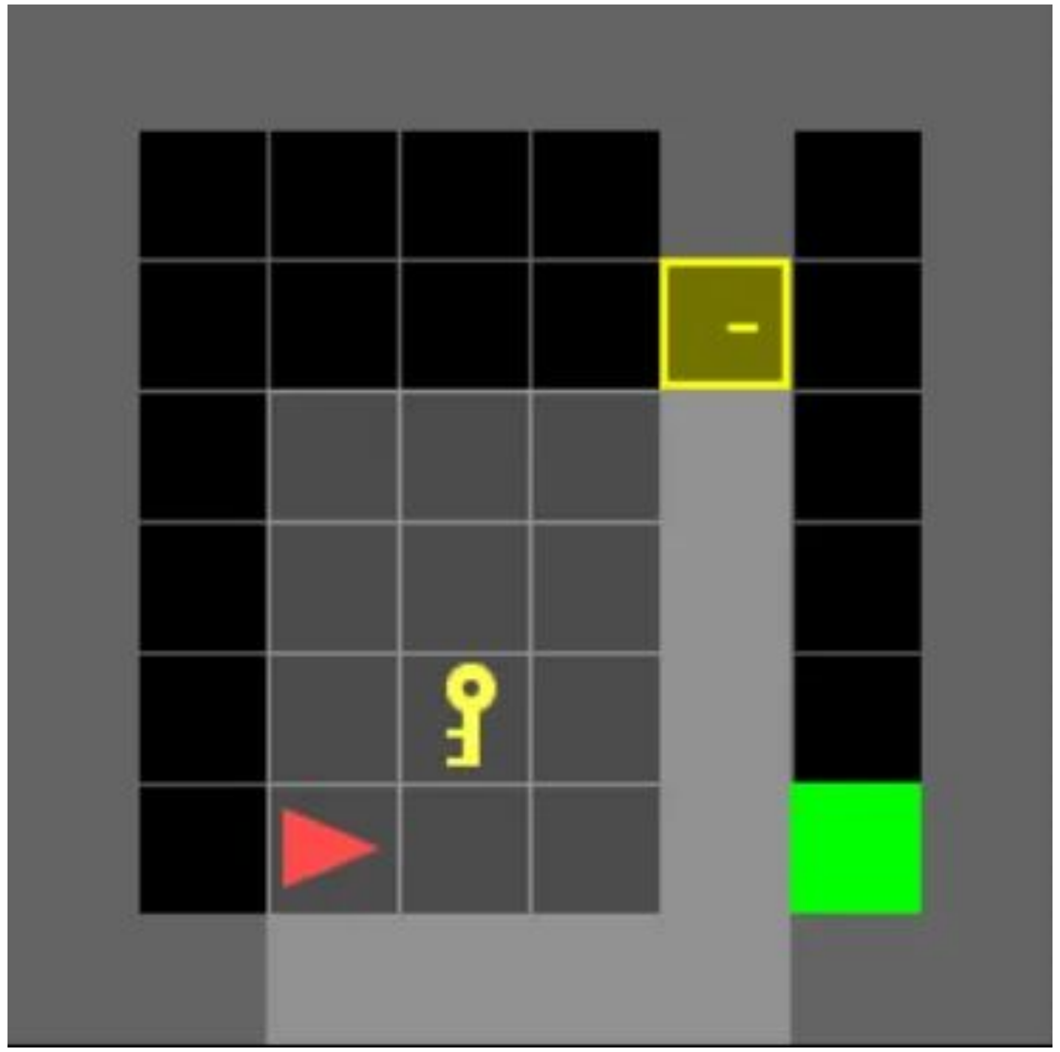
That's **REWARD SHAPING**.

$$r_{t+1} = R(s_t, a_t, s_{t+1}) + \boxed{\tilde{R}(s_t, a_t, s_{t+1}).}$$



MOTIVATION

+ **EXAMPLE**

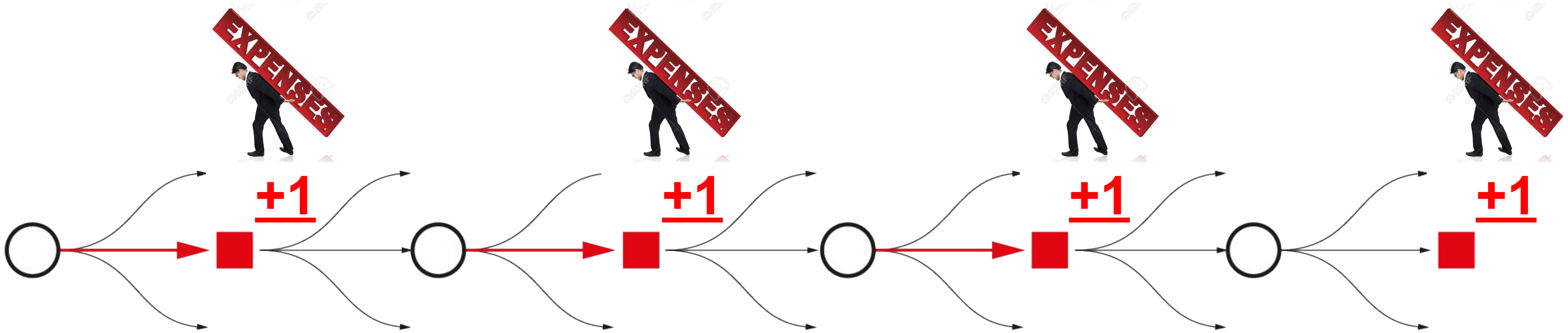


MOTIVATION

+ SCALING CA

However, reward shaping is **TOO EXPENSIVE**:

- It requires extensive **domain knowledge**, and
- Extensive **manual human feedback**, which
- **Tabula rasa** models cannot incorporate that effectively



MOTIVATION

+ SCALING CA

In short, reward shaping **DOES NOT SCALE**

MOTIVATION:

+ RESEARCH QUESTION

A natural question: if humans are the bottleneck,

"How can we scale Reward Shaping (thus, CA) in Deep RL?"

MOTIVATION:

+ RESEARCH QUESTION

SPOILER

We propose to investigate the use of **Large Language Models** because:

- Strong results in CAUSAL REASONING tasks
- Performances comparable to humans

[1] Zhijing Jin, et al. **CLadder: Assessing causal reasoning in language models.** In NeurIPS, 2023
[2] Emre Kıcıman, et al. **Causal reasoning and large language models: Opening a new frontier for causality.** arXiv preprint arXiv:2305.00050, 2023.

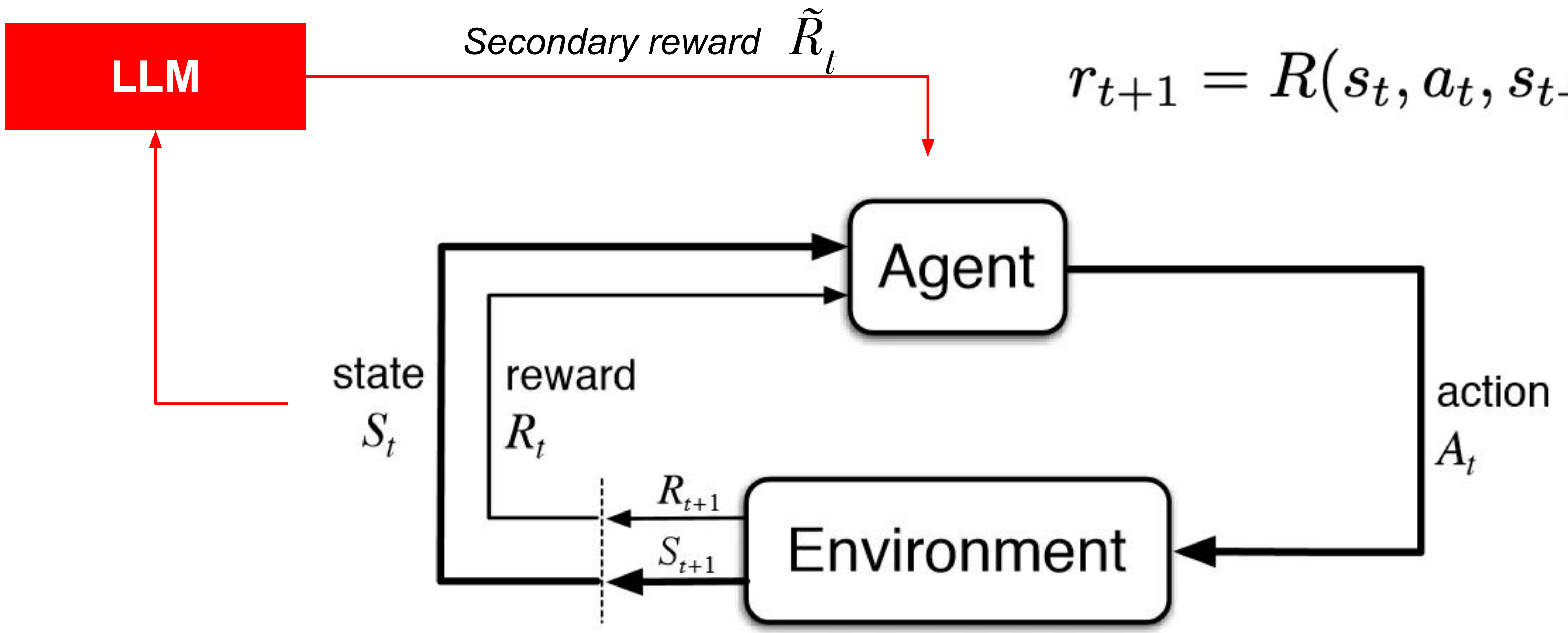
#2: METHODS

METHODS

+ **TL;DR;**

We use **LLMs** to **ASSIST** action **EVALUATION** actions in RL,

and introduce **CALM: Credit Assignment with Language Models**



$$r_{t+1} = R(s_t, a_t, s_{t+1}) + \tilde{R}(s_t, a_t, s_{t+1}).$$

METHODS

+ OPERATIONALISATION

$$LLM : desc(\mathcal{M}) \times desc(\mathcal{S} \times \mathcal{A} \times \mathcal{S}) \rightarrow \mathbb{B}.$$

We prompt the LLM to:

1. Break down a task into **SUBGOALS**
2. **VERIFY** when a subgoal is achieved

METHODS

+ FORMALISM

CANONICAL REWARD SHAPING

$$r_{t+1} = R(s_t, a_t, s_{t+1}) + \boxed{\tilde{R}(s_t, a_t, s_{t+1}).}$$

LLM SHAPING

$$LLM : desc(\mathcal{M}) \times desc(\mathcal{S} \times \mathcal{A} \times \mathcal{S}) \rightarrow \mathbb{B}.$$

$$\boxed{\tilde{R}(s_t, a_t, s_{t+1}) = \beta(s_{t+1}).}$$

METHODS

+ PROMPTING

Example prompt

```
The environment is MiniHack.

I will present you with a short extract of a gameplay. At each timestep, symbols represent the following items:
- "." represents a floor tile.
- "|" can represent either a wall, a vertical wall, an open door.
- "-" can represent either the bottom left corner (of a room), bottom right corner (of a room), wall, horizontal wall,
wall, top left corner (of a room), op right corner (of a room).
- "+" represents a closed door. Doors can be locked, and require a key to open.
- "(" represents a useful item (pick-axe, key, lamp...)
- "<" represents a ladder or staircase up.
- ">" represents a ladder or staircase down.

The task of the agent is to win the game.

First, based on your knowledge of NetHack, break down the task of the agent into subgoals.
Then, consider the following game transition, which might or might not contain these subgoals.
Determine if any of the subgoals is achieved at Time: 1 or not.

Report your response in a dictionary containing the name of the subgoals as keys and booleans as value. For example:
```python
{
    <name of goal>: <bool>,
}
```

```
Observation Sequence:

<gameplay>
Time: 0
Current message:

      - - - -
      | . . |
      | . . |
- - + - . < |
| . . . @ . |
| . ( . . . |
- - - - - - -


Time: 1
Current message:

      | . . |
      | . . |
- - + - . < |
| . . . . . |
| . ( . @ . |
- - - - - - -


</gameplay>
```
```

#3: RESULTS

RESULTS

+ OBJECTIVE

AIM: to understand if the **ability to assign credit** is in the spectrum of the current open-weights **LLMs**

RESULTS

+ EXPERIMENTAL SETUP

We perform a preliminary evaluation on an **OFFLINE** dataset, using the following recipe:

1. We consider the **MINIHACK** suite
2. **KeyRoom** environment (pick up key, unlock door, reach goal tile)
3. We collect **256** transitions (S, A, S)
4. Such that the dataset has a **BALANCED** number of events (pickup, unlock, nothing)
5. We **annotate** the transitions **manually** (ground truth)
6. **Annotate** using the **LLM**

RESULTS

+ **CLASSIFICATION PROBLEM**

		HUMAN	
		Goal achieved	Goal NOT achieved
LLM	Goal achieved	**True Positive**	**HALLUCINATION**
	Goal NOT achieved	**MISS**	**True Negative**

RESULTS

+ **PRELIMINARY**

Annotator	F1 ↑	Accuracy ↑	Precision ↑	Recall ↑	TP ↑	TN ↑	FP ↓	FN ↓
Human	**1.00**	**1.00**	**1.00**	**1.00**	**171**	**85**	**0**	**0**
Meta-Llama-3-70B-Instruct	**0.82**	**0.72**	0.71	**0.96**	**165**	19	66	**6**
Meta-Llama-3-8B-Instruct	0.80	0.70	0.72	0.89	153	26	59	18
gemma-1.1-7b-it	0.77	0.66	0.71	0.85	145	25	60	26
Mixtral-8x7B-Instruct-v0.1*	0.74	0.64	0.71	0.76	130	33	52	41
Mistral-7B-Instruct-v0.2	0.57	0.48	0.63	0.53	90	32	53	81
c4ai-command-r-v01*	0.56	0.52	0.71	0.47	80	52	33	91
gemma-1.1-2b-it	0.00	0.33	0.00	0.00	0	**85**	**0**	171
Random	0.33	0.33	0.33	0.33				

Table 3: Performance of LLM annotations against human annotations with **game screen** observations and with **autonomously discovered** subgoals.

More to the story!
(POSTER)

#4: CLOSING

CLOSING

+ KEY TAKEAWAYS

1. **CAP** is key for RL,
2. But **HARD** without pre-existing knowledge.
3. Canonical methods (e.g., reward shaping, options) **DO NOT SCALE** well,
4. Because **HUMAN LABELS** are **EXPENSIVE**
5. We propose **CALM**, which automates **REWARD SHAPING** for credit assignment using **LLMs**.
6. We present **OFFLINE RL** results showing that
7. LLMs provide **QUALITY ASSIGNMENTS** (**>80%!**) and bode well for applications to online RL

RESUME

EMAIL

Thank you!
Come to chat at the #POSTER

Or reach out!
Email: *edu.pignatelli@gmail.com*
Website: *https://epignatelli.com*
Twitter: *@EduPignatelli*