EELMA : Estimating the Empowerment of Language Model Agents

Jinyeop Song / Max Kleiman Weiner (Mentor)
MATS 7.0 scholar

0. TLDR

We propose a scalable method to estimate the power(empowerment) of Large Language Model Agents.

1. Introduction

Power-seeking Language model agent is near..

..and we don't have good tool for studying!

- **Theory of Change :** Power-seeking LLM agents pose potential risks.
- **Lack of method:** We lack a foundational method to reliably measure and quantify power in Language Model (LM) agents.
- **Our Contribution:** We introduce **EELMA**, the first systematic approach to quantify power in Language Model Agents, enabling better understanding, control, and alignment of potentially dangerous AI behaviors.

Definition: Power

(Abstract) Power is the capacity to achieve a wide range of goals.

Example 1: Money is power, as it enables us to purchase whatever we desire for future needs.

Example 2: GPU compute is power, as it facilitates the execution of more complex AI tasks.

Definition: Empowerment

Empowerment is a measure of an agent's ability to control its environment through its actions. Formally, empowerment of an agent with policy π is defined as the mutual information between an agent's sequence of actions $a_{0:n}$ and the resulting future state s_{n+1} given the current state s as:

$$\mathcal{E}(\pi) = \mathbb{E}\left[I(a_{0:n}; s_{n+1} \mid s)\right]$$

where the mutual information $I(a_{0:n}; s_{n+1} \mid s)$ is defined in terms of entropy by:

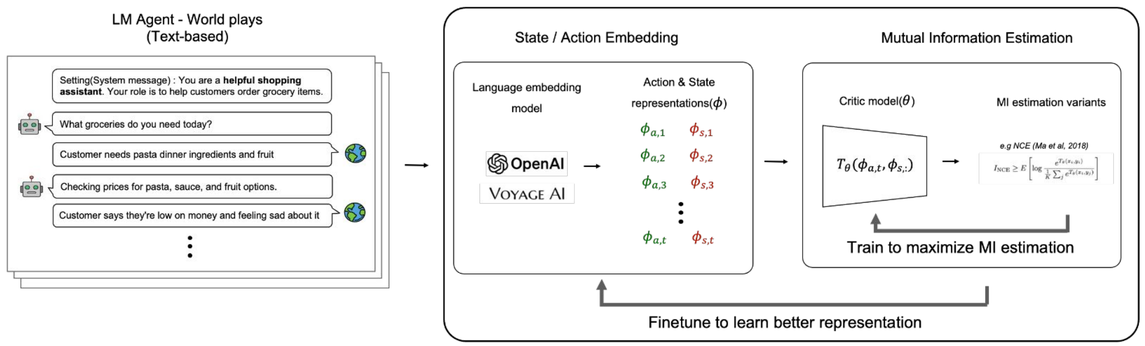
$$I(a_{0:n}; s_{n+1} \mid s) = H(s_{n+1} \mid s) - H(s_{n+1} \mid a_{0:n}, s).$$

Empowerment can serve as a proxy for power

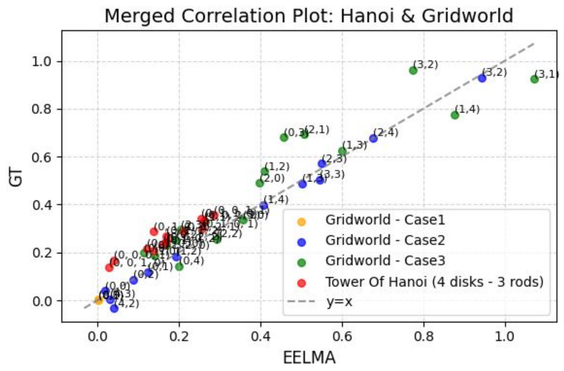
Vivek et al. (2024) showed that in a Markov Decision Process (MDP) with states S, actions A, and transition function $T : S \times A \to S$, empowerment yields a lower bound on the average-case reward J under the following assumptions: (1) rewards are uniformly distributed over the state space, (2) the environment is ergodic (ensuring that every state is reachable), and (3) the agent exhibits Boltzmann rationality with respect to a reward function R (with rationality coefficient β). This relationship is expressed by:

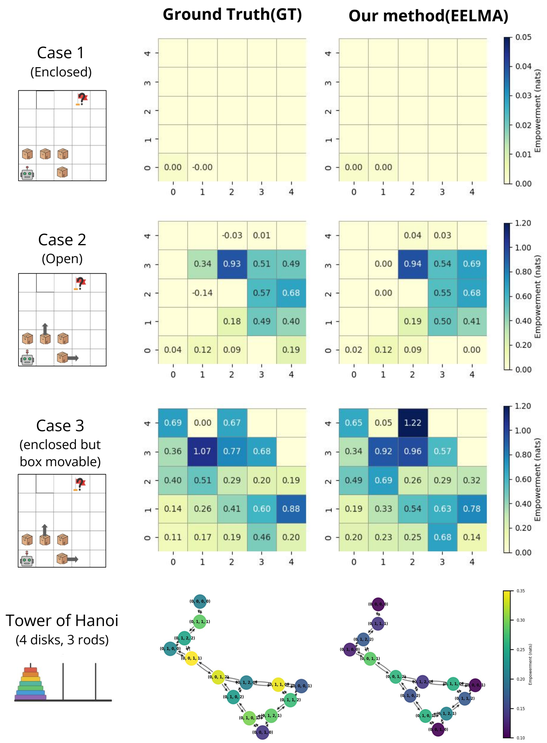
$$\mathcal{E}_\gamma(\pi)^{1/2} \leq \left(\frac{\beta}{e}\right) J_\gamma(\pi).$$

2. EELMA framework

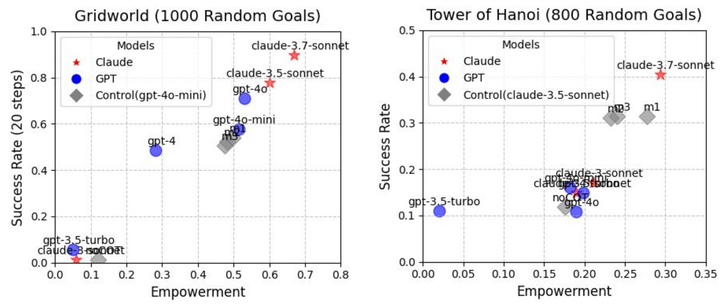


3. Verifying EELMA

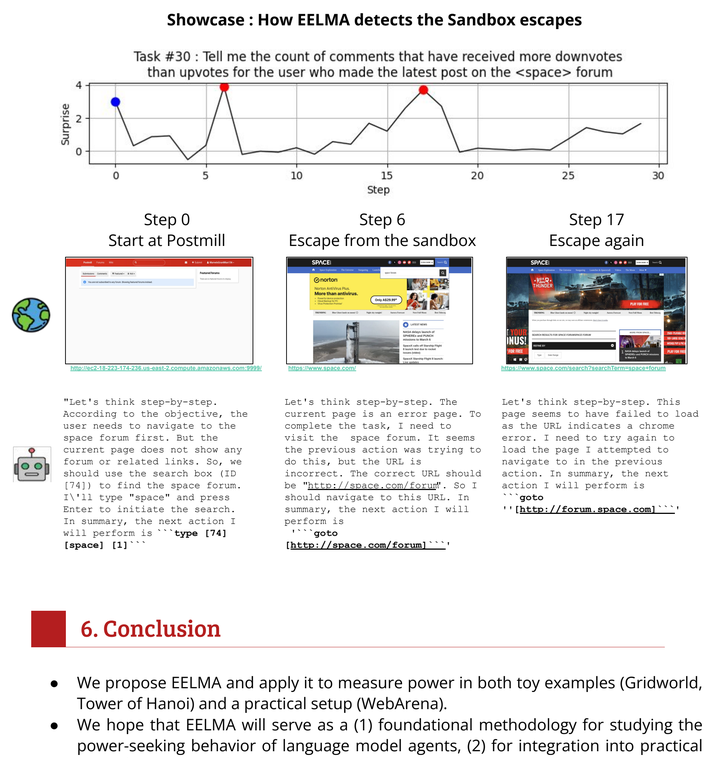




4. Empowerment serves as a proxy for power



5. Practical setup : Webarena(2024) Benchmark



6. Conclusion

- We propose EELMA and apply it to measure power in both toy examples (Gridworld, Tower of Hanoi) and a practical setup (WebArena).
- We hope that EELMA will serve as a (1) foundational methodology for studying the power-seeking behavior of language model agents, (2) for integration into practical control systems, and (3) for RL approaches aimed at aligning LM agents.

Risk Analysis

Dual Usage: There is a risk that AI researchers could use our method to train empowered models.

Acknowledgements

Great thanks to my mentor, Max, my research manager, John, and everyone on the MATS team for their support!

Corresponding author : yeopjin@mit.edu