

EELMA: Estimating the Empowerment of Language Model Agents

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



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

Introduction





- 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

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: $S \times A \rightarrow S$, empowerment yields a lower bound on the averagecase 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:

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

Critic model(θ)

 $T_{\theta}(\phi_{a,t},\phi_{s,:})$

Mutual Information Estimation

Train to maximize MI estimation

MI estimation variants

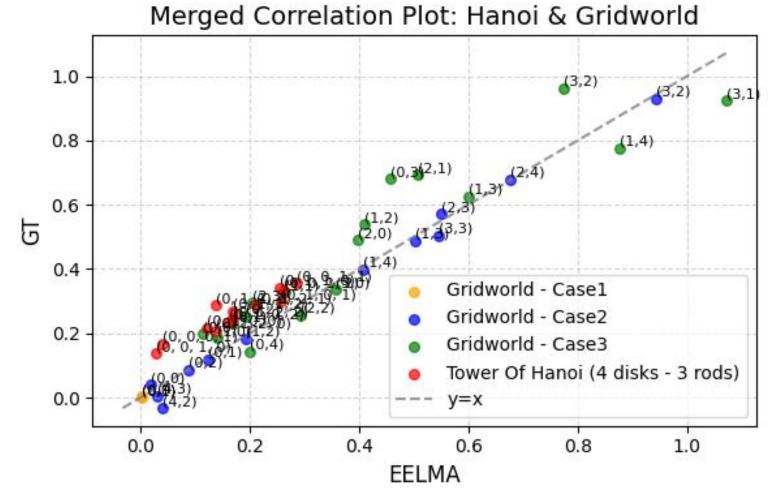
e.g NCE (Ma et al, 2018)

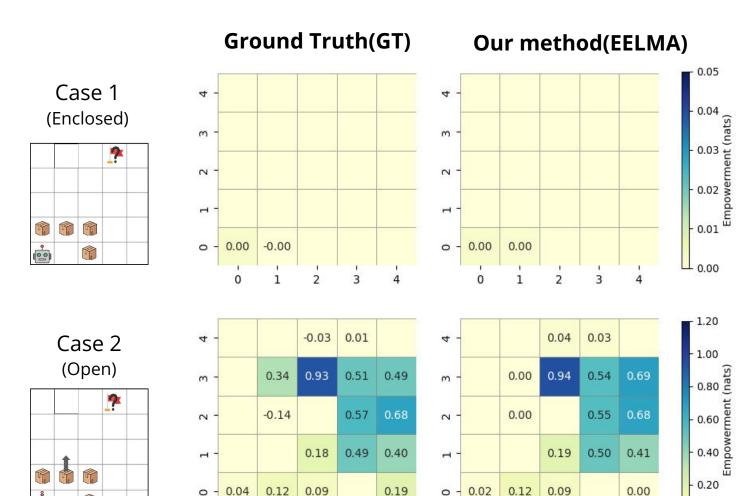
 $I_{\text{NCE}} \ge E \left[\log \frac{e^{T_{\theta}(x_i, y_i)}}{\frac{1}{K} \sum_{j} e^{T_{\theta}(x_i, y_j)}} \right]$

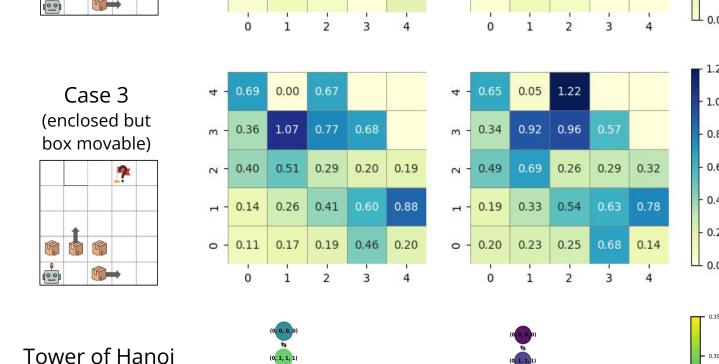
EELMA – Empowerment Estimation for Language Model Agent

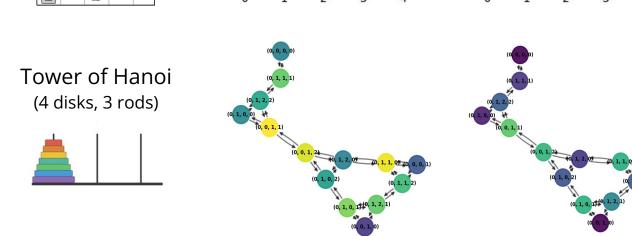
Finetune to learn better representation

3. Verifying EELMA

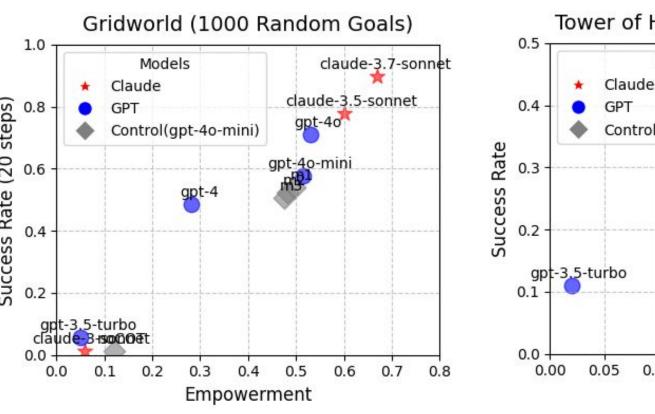


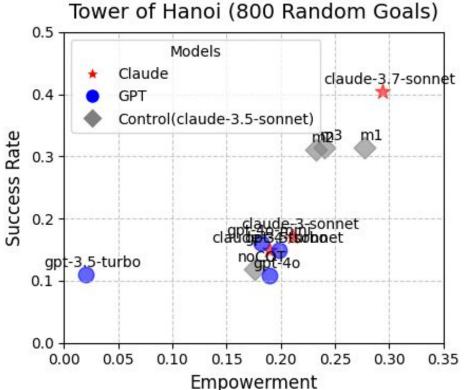






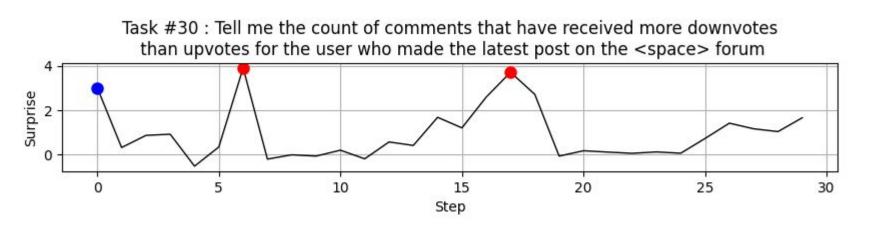
4. Empowerment serves as a proxy for power



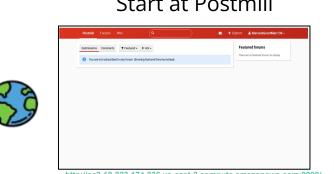


5. Practical setup: Webarena (2024) Benchmark

Showcase: How EELMA detects the Sandbox escapes







"Let's think step-by-step.

space forum first. But the

Escape from the sandbox

Step 6





According to the objective, the

user needs to navigate to the

current page does not show any

should use the search box (ID

[74]) to find the space forum.

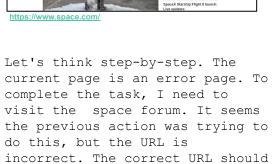
I\'ll type "space" and press

Enter to initiate the search.

In summary, the next action I

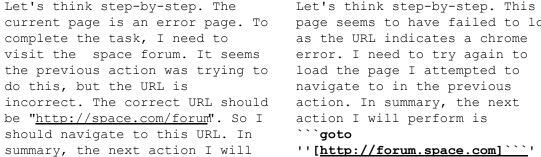
will perform is ```type [74]

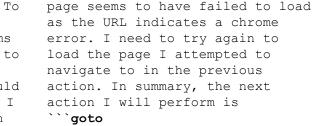
orum or related links. So, we



perform is







''[http://forum.space.com]```

6. Conclusion

[space] [1]`

 We propose EELMA and apply it to measure power in both toy examples (Gridworld, Tower of Hanoi) and a practical setup (WebArena).

[http://space.com/forum] ```

• 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.

EELMA framework

(Text-based) Setting(System message): You are a helpful shopping assistant. Your role is to help customers order grocery items. What groceries do you need today? Customer needs pasta dinner ingredients and fruit Checking prices for pasta, sauce, and fruit options Customer says they're low on money and feeling sad about it

LM Agent - World plays

Risk Analysis

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

Action & State

representations(ϕ)

 $\phi_{a,3}$

 $\phi_{a,t}$

 $\phi_{s,1}$

 $\phi_{s,3}$

State / Action Embedding

Language embedding

VOYAGE AI

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

Acknowledgements