BINGHAMTON UNIVERSITY
STATE UNIVERSITY OF NEW YORK

Can LLMs Understand Multi-Armed Bandit Tasks?

[1]Isaac Cohen, [1]Hiten Malhotra, [2]William M. Hayes

[1]Department of Computer Science, [2]Department of Psychology

DReaM Lab

Decision Research and Modeling

BACKGROUND

• LLMs can learn from information in prompt without training, known as in-context learning

•When LLMs are deployed as autonomous decision-making agents, can they effectively balance exploration and exploitation to maximize payoffs?

•Simulating LLMs through sequential decision-making problems can help us evaluate them by comparing their performance with established strategies and find methods to encourage these behaviors

•Extracting activations during these tasks can help us gauge LLMs understanding of the explore-exploit trade-offs and maybe steer them in more optimal directions

RESULTS

•Our classifier was able to tell, with greater than 90% accuracy in some layers, whether a decision was Greedy or Anti-greedy.

•We were not able to steer the model towards being more/less greedy. (Fig.6)

•We were not able to predict the activations using UCB or Greedy values.

REFERENCES

Krishnamurthy, A., Harris, K., Foster, D. J., Zhang, C., & Slivkins, A. (2024, March). *Can large language models explore in-context?*. arXiv. https://arxiv.org/abs/2403.15371

Prompt

You are in a room with 5 buttons labeled blue, green, red, yellow, purple. Each button is associated with a Bernoulli distribution with a fixed but unknown mean …

MAB instance

Out of 100 steps, so far you have played …

Greedy/Non - Greedy history

.
.
.

Fig.1: 2 generated prompts with preloaded history of decisions made using an Epsilon greedy algorithm in MAB instances.

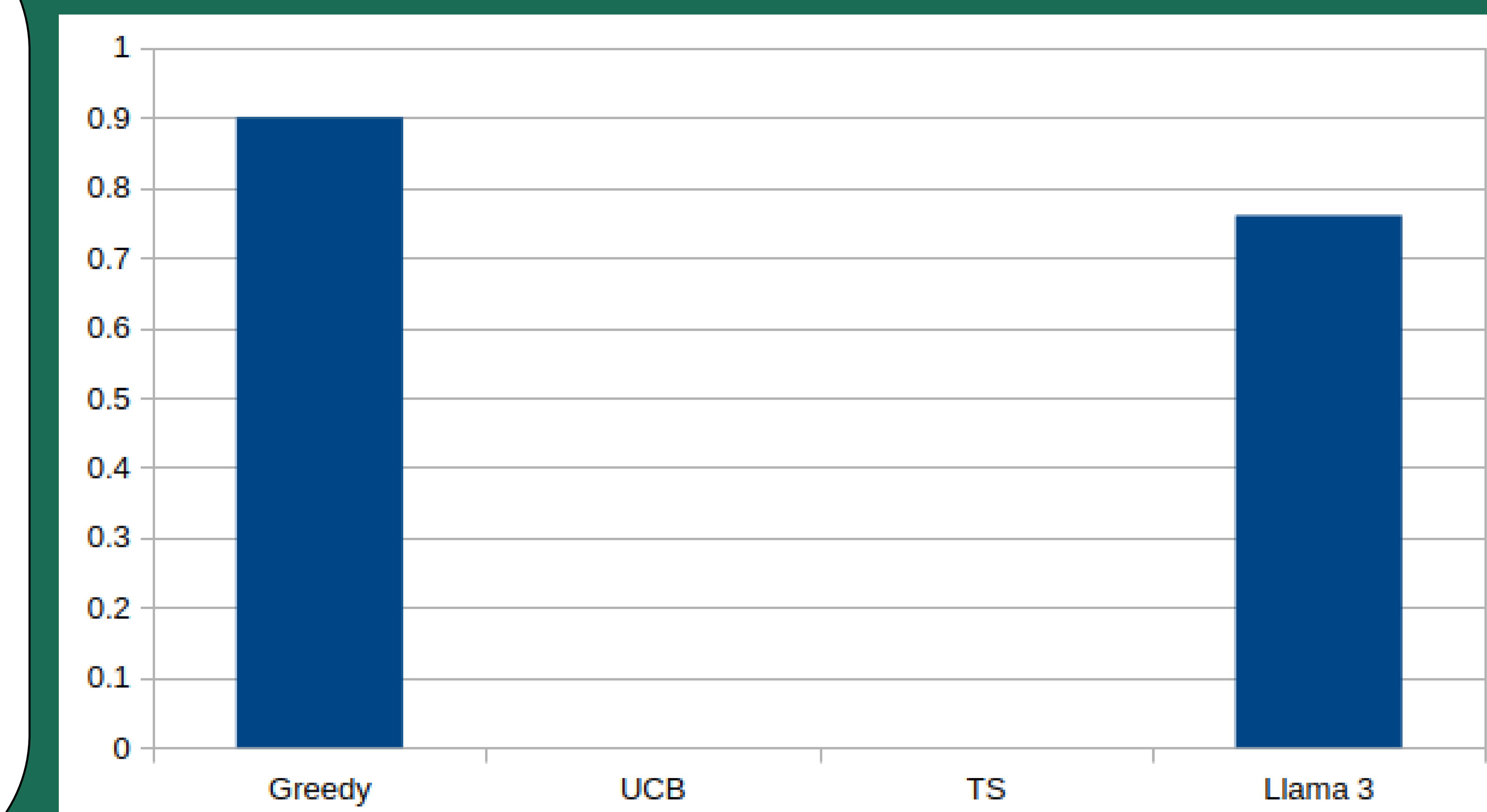

Fig.2: Proportion of replicates that never select the best arm in the latter half of trials.

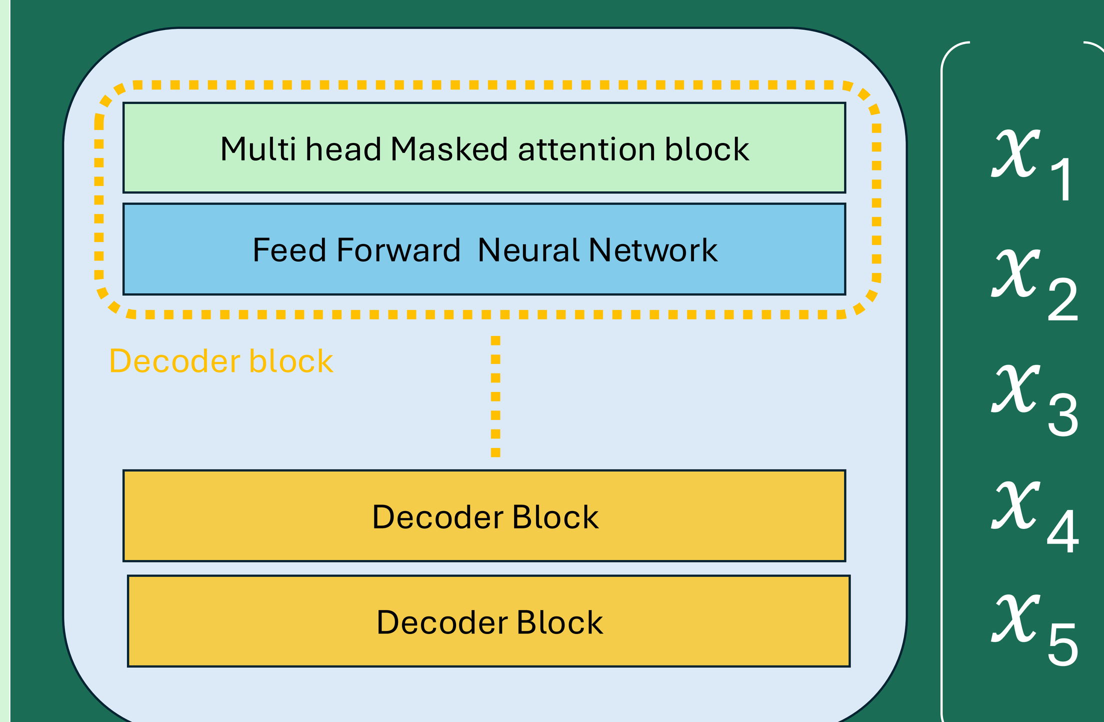

Fig.3: Extracted N activation vectors after each decoder block from LLMs for each prompt. The high dimensional vectors are reduced to the significant dimensions using PCA.

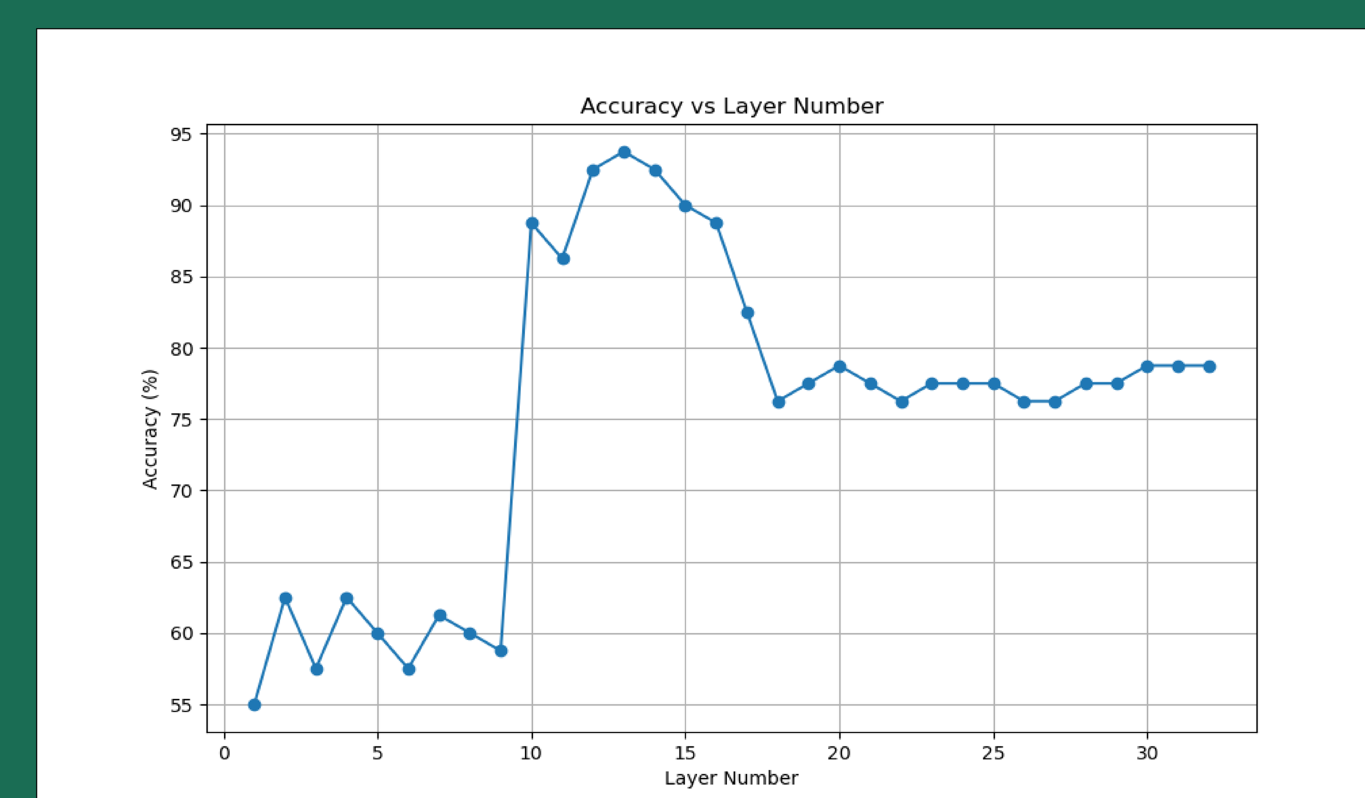

Fig.4: Accuracies of a logistic regression model in predicting based on PCA values of activation vectors whether they were generated by greedy or anti-greedy prompts (Llama 3 8B).

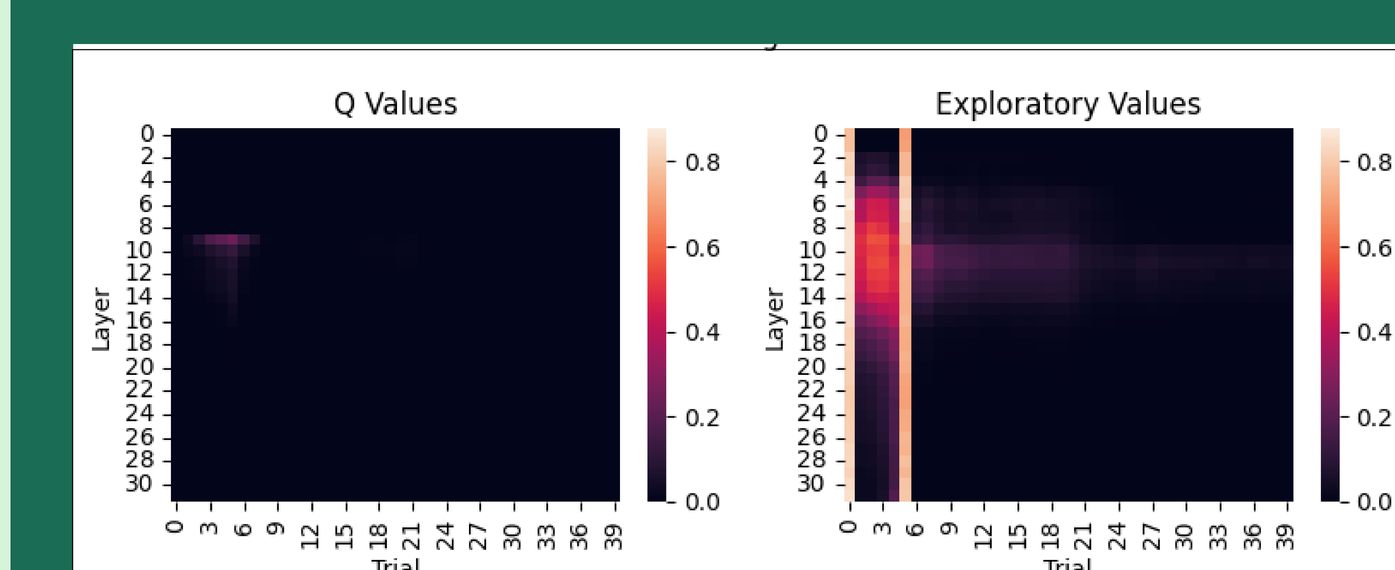

Fig.5: Activation neurons association with Q values (greedy values) and Exploratory values when prompted with UCB choice history. Model: Llama 3 8B

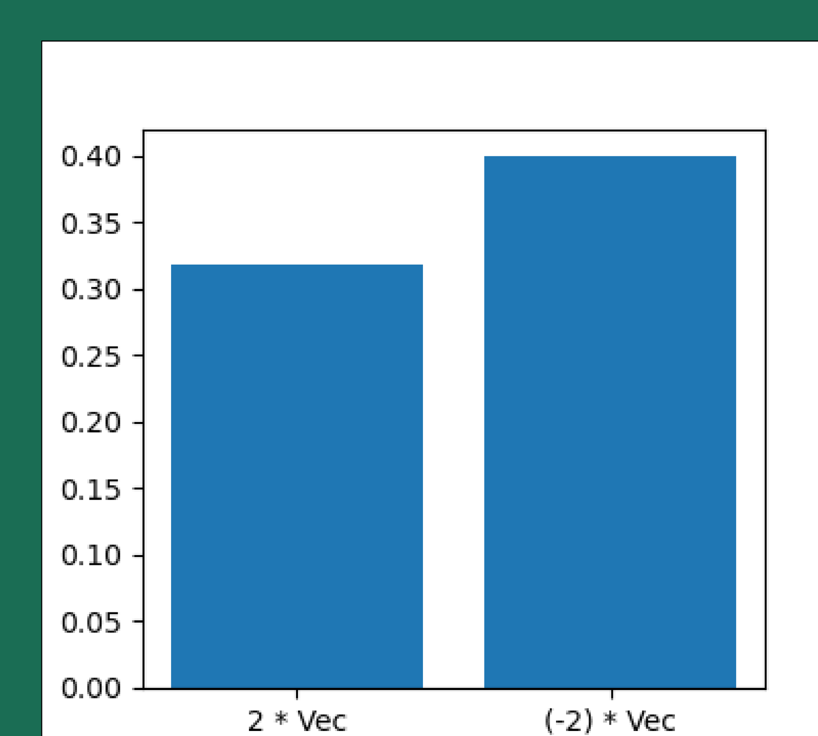

Fig.6: The model is slightly more greedy when we subtract twice the steering vector than when we add it twice.

METHOD

1.Do LLMs inherently make explorative or exploitive choices?

•Simulated reinforcement algorithms and LLMs (prompts using problem summary, choice history and problem hints) in a 5-armed bandit task with Bernoulli distribution rewards for each arm.

•Replicated (Krishnamurthy et. al 2024) "Can Large language models explore in-context ?" (Fig.2)

2 Do LLMs understand the trade-off?

•Extracted activations for token representing selected color at each layer in decoder only LLMs for greedy and anti-greedy choice history generated prompts. (Fig.4)

- Using Principal Component Analysis reduced the activation vectors for each prompt to 5 dimension to overcome feature redundancy
- Trained a Logistic regression model (for each layer) using PCA values and associated prompts as classes
- Computed steering vector using average difference between greedy and anti-greedy activations
- Attempted to steer the model by inserting vector during inference

•Using UCB history generated prompts, we check for association between neurons in each activation at each layer (Fig. 5) and

- Q values (mean reward received for each arm/ no. of times it was picked).
- Uncertainty factor (responsible for the exploratory behavior in UCB method).