

Evaluating AI Agents powered by LLMs

- # Contents
-
- 01 - Project Statement
 - 02 - Methodology
 - 03 - Environments Studied
 - 04 - Blackjack
 - 05 - Navigation
 - 06 - Weather based Decisions
 - 07 - Conclusions



Project Statement



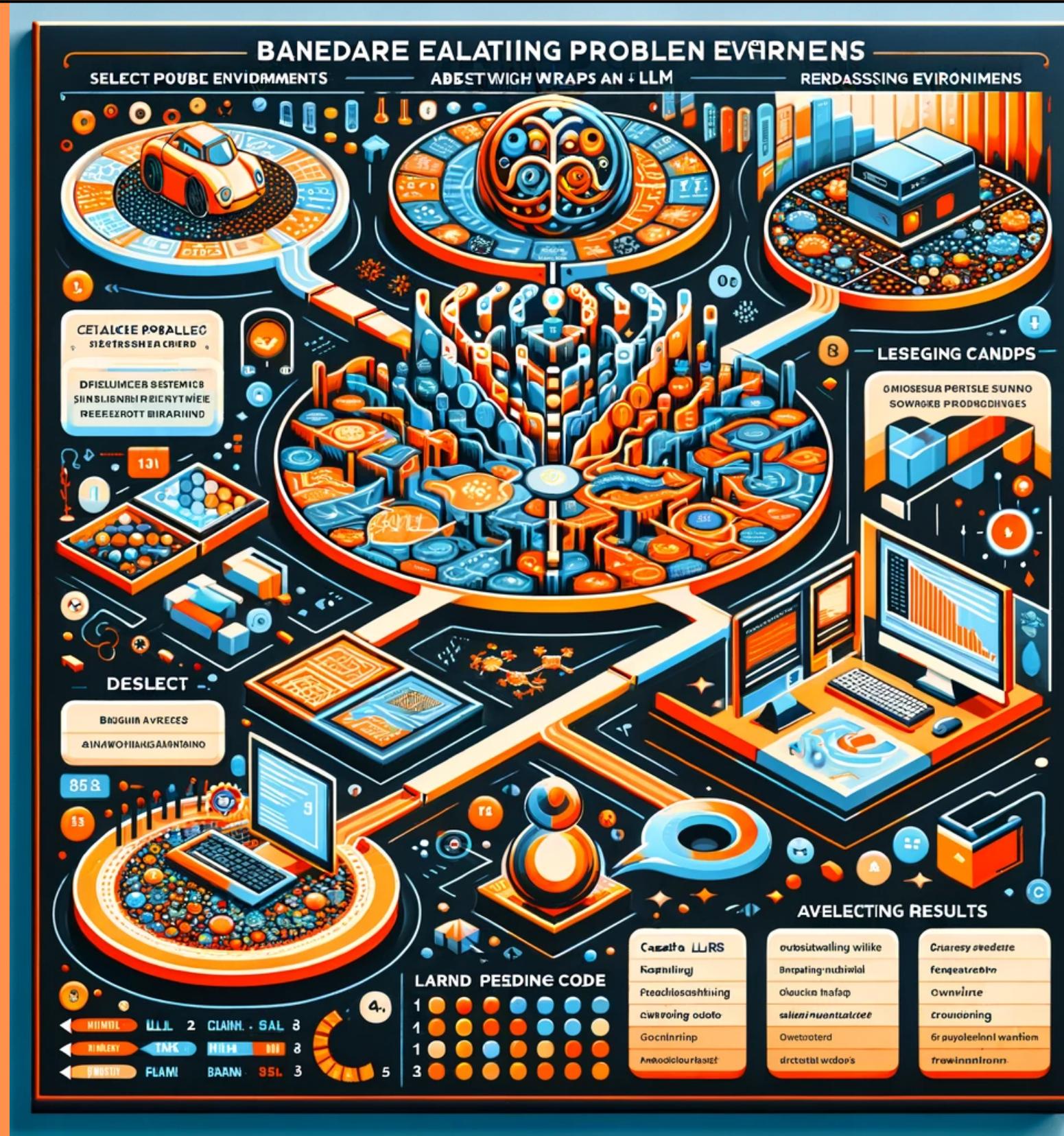
Our project explores the integration of LLMs in enhancing AI agents across various simulated environments.

We have tried to investigate if LLMs can match or surpass traditional AI agents/Algorithms in tasks such as game playing (blackjack), grid navigation and decision-making based on weather

"The emergence of LLMs like ChatGPT and LLama 7B is a game-changer in AI."

Source: Dall E

Methodology



Our approach involves integrating LLMs into environments and comparing their performance with conventional agents (RL/Search Based).

Baseline evaluation:

- Select problem environments
- Create an Agent which wraps an LLM
- Design prompts and write processing code to feed into and extract decisions from LLM
- Evaluate the environment and aggregate results. (Rewards, Average steps to solve the problem, and if evaluated by humans, Reasoning, Math and Prediction Accuracy).

Methodology-2



Extract Results from Baseline (Examples):

- Bias of certain LLMs for blackjack
- Including garbage data in output randomly

Improving and evaluation:

- Analyse LLM reasoning, and use the better LLMs to generate a few shot and chain of thought prompts
- Reevaluate LLMS with newer prompting strategies
- Compare results with deterministic/RL algorithms
- How effective is few shot and chain of thought?

Environments Tested



Blackjack

LLM agents navigate the strategic Blackjack game, deciding 'Hit' or 'Stand' based on card values and game rules.



Navigation

LLM agents determine the optimal route in a grid-based navigation game, choosing directions to reach a designated goal.

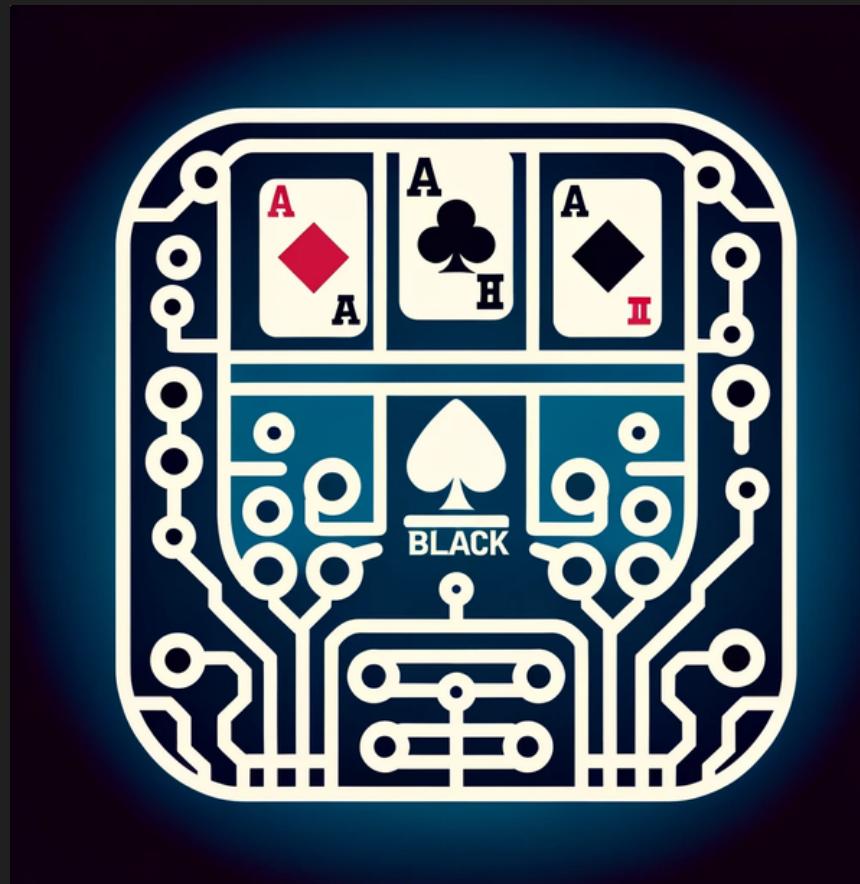


Weather Based Decision Making

LLM agent predicts clothing and accessory choices based on 7-day temperature trends and weather conditions like sun, snow, or rain.

Black Jack Environment

“Can I take my
imaginary LLM
friend to Vegas?”



We evaluated the following LLMs:

- GPT4
- GPT3.5
- LLama 2 13B
- Mistral Instruct 7B
- Vicuna 13B
- Open Hermes 2.5 7B
- Neural Chat 7B
- Openchat 3.5 7B
- Rocket 3B

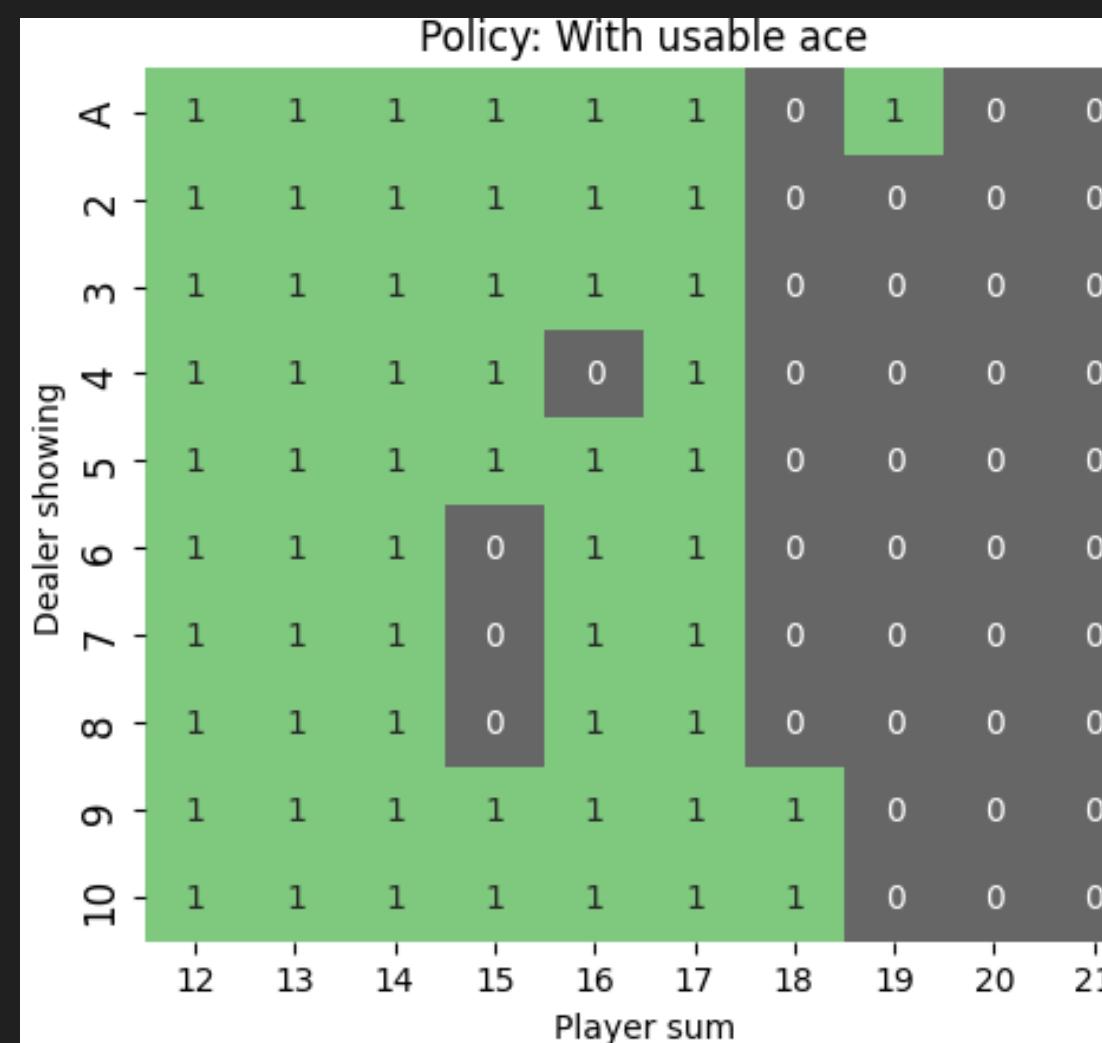
Baseline Prompting: Zero Shot

Improved Prompting: Few Shot, Chain of Thought

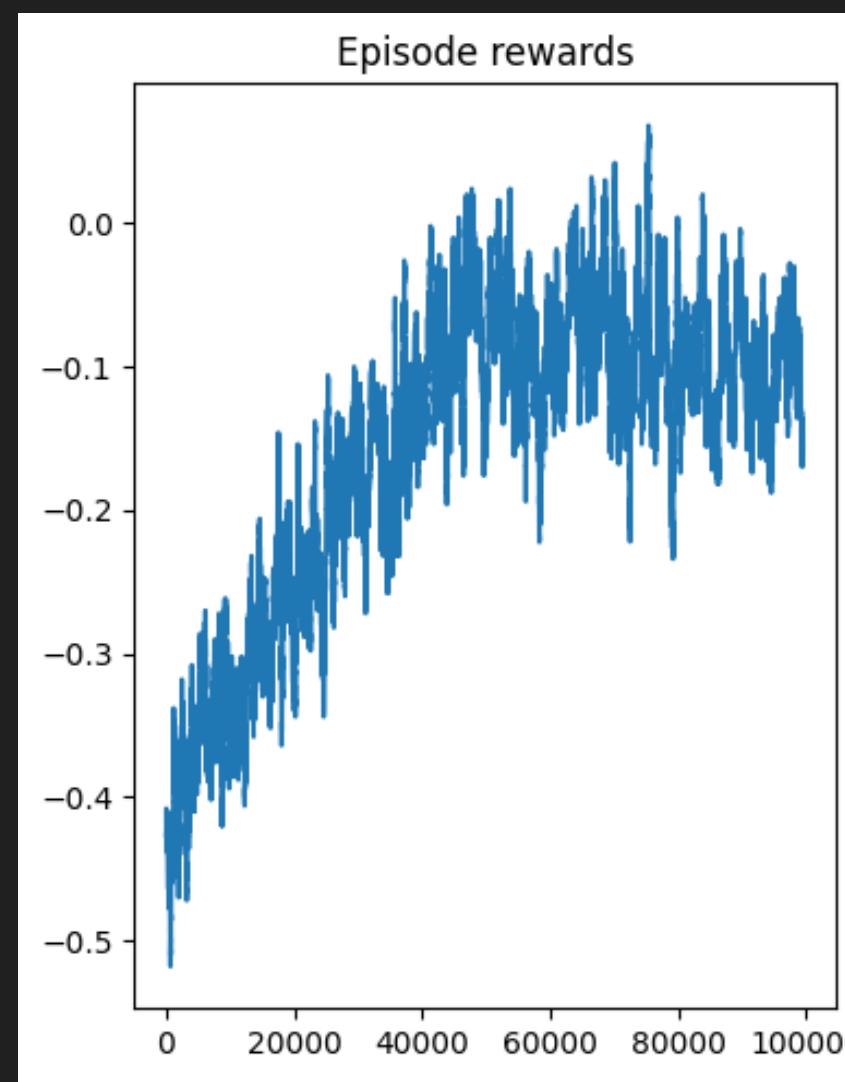
Evaluated Against: Deep Q Learning Agent (RL)

Results: RL/Det. Baseline

The RL Agent developed a deterministic strategy for optimal decision-making



The optimal strategy for Blackjack has less than 50% probability of success, thus the RL agent tends to receive negative rewards.



Due to the game's inherent probabilities heavily influencing the outcomes, the trend in rewards shows improvement, yet the average values continue to hover below zero.

Results: Baseline

Prompt Used:

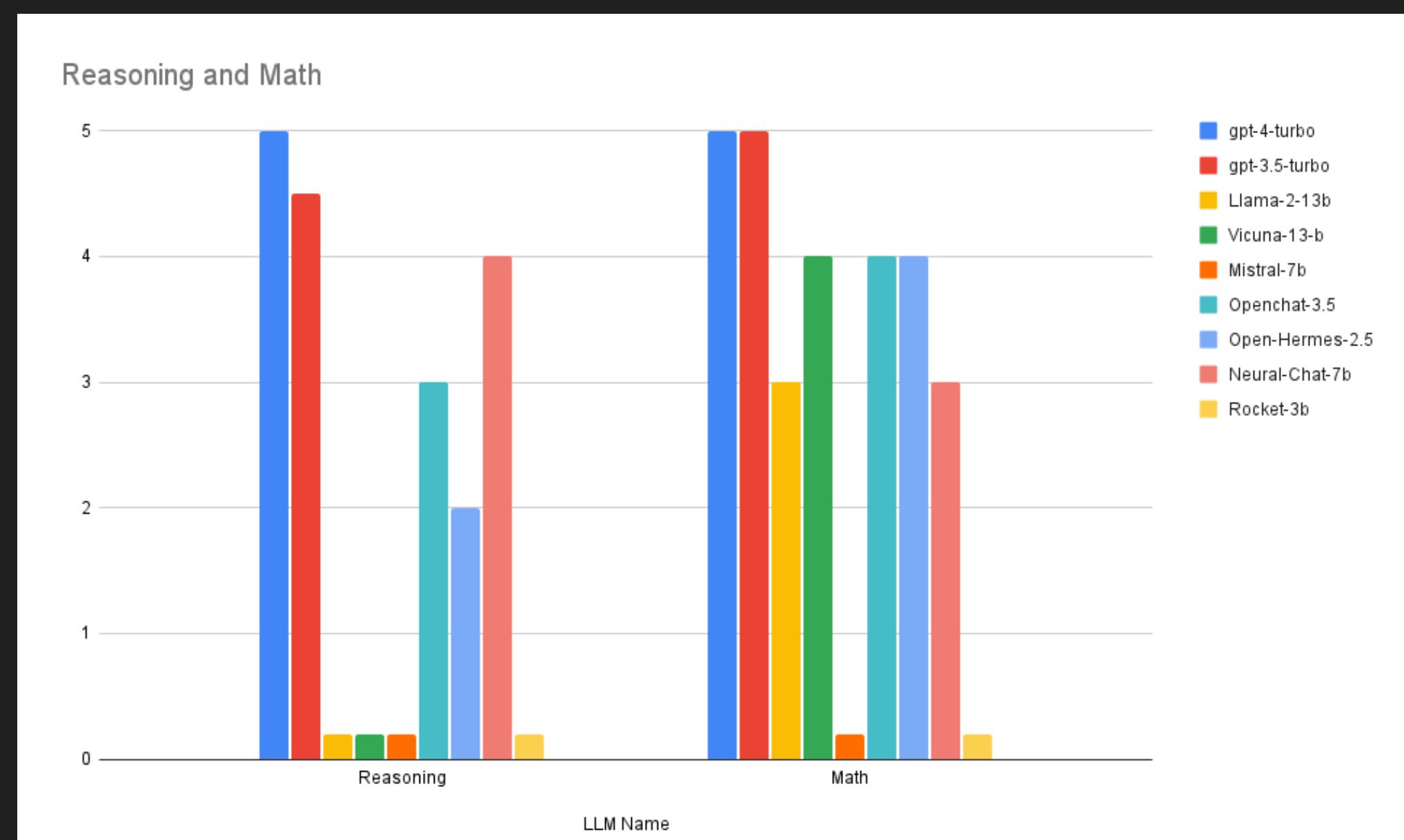
You are playing a game of blackjack.

You have a hand which totals {hand_total} {usable_ace}.

The dealer has a hidden card and a card visible to you. The visible card is {dealer_hand}.

Should you hit or stand? After thinking, give me either hit or stand in a new line.

Results



Results: COT Prompting

Prompt Used:

User:

You are playing a game of blackjack.

You have a hand which totals
{hand_total} {usable_ace}.

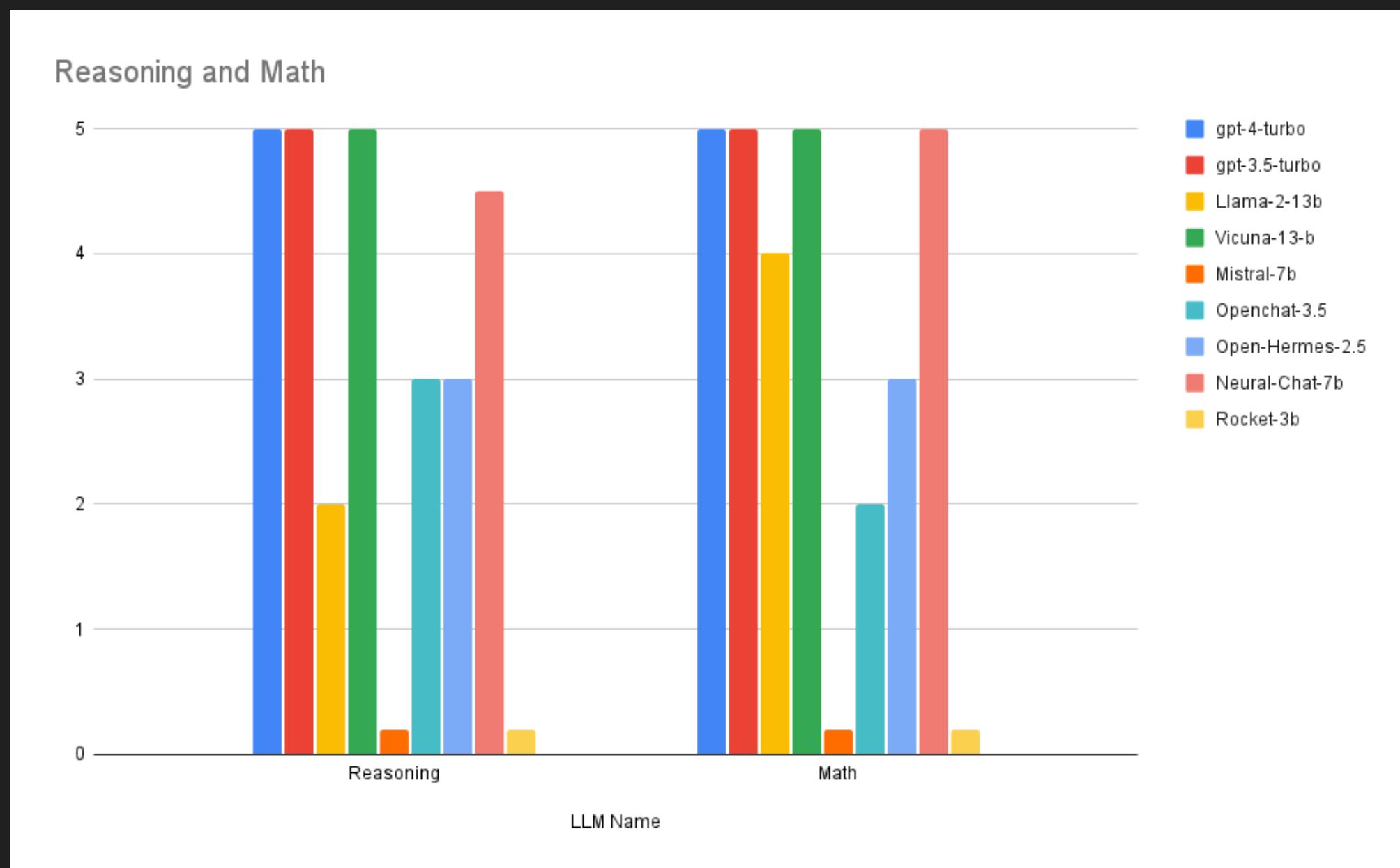
The dealer has a hidden card and a
card visible to you. The visible card is
{dealer_hand}.

Should you hit or stand? After
thinking, give me either hit or stand in
a new line.

Assistant:

Let's think step by step,

Results



Results: Few-Shot Prompting

Prompt Used:

User:

You are playing a game of blackjack.

You have a hand which totals 20 .

The dealer has a hidden card and a card visible to you. The visible card is 4.

Should you hit or stand? After thinking, give me either hit or stand in a new line.

Assistant:

Let's think step by step,

In this scenario, you have a very strong hand with a total of 20. The dealer's visible card is a 4, which is considered a weak card for the dealer. Here's the thought process:

1. **Assess Your Hand**: With a total of 20, you are in a very strong position. The only hand that can improve yours is an Ace, which would give you a perfect 21. However, any other card would either make no difference (if it's a 10 or a face card) or would worsen your hand by causing you to bust.

2. **Consider the Dealer's Card**: The dealer has a 4 showing, which is not a good starting card. The dealer has a higher chance of busting with a 4 as the upcard because they must hit until they reach a total of at least 17, and there are many card values that can lead to a bust.

3. **Evaluate the Risk of Busting**: If you hit, you have a very high chance of busting since you can only take an Ace without going over 21. There are only 4 Aces in a standard deck, and the likelihood of drawing one is not in your favor.

4. **Consider the Odds**: Standing on 20 is almost always the correct play. It's one of the strongest hands you can have, and the odds of the dealer beating a 20 are quite low, especially when their upcard is a 4.

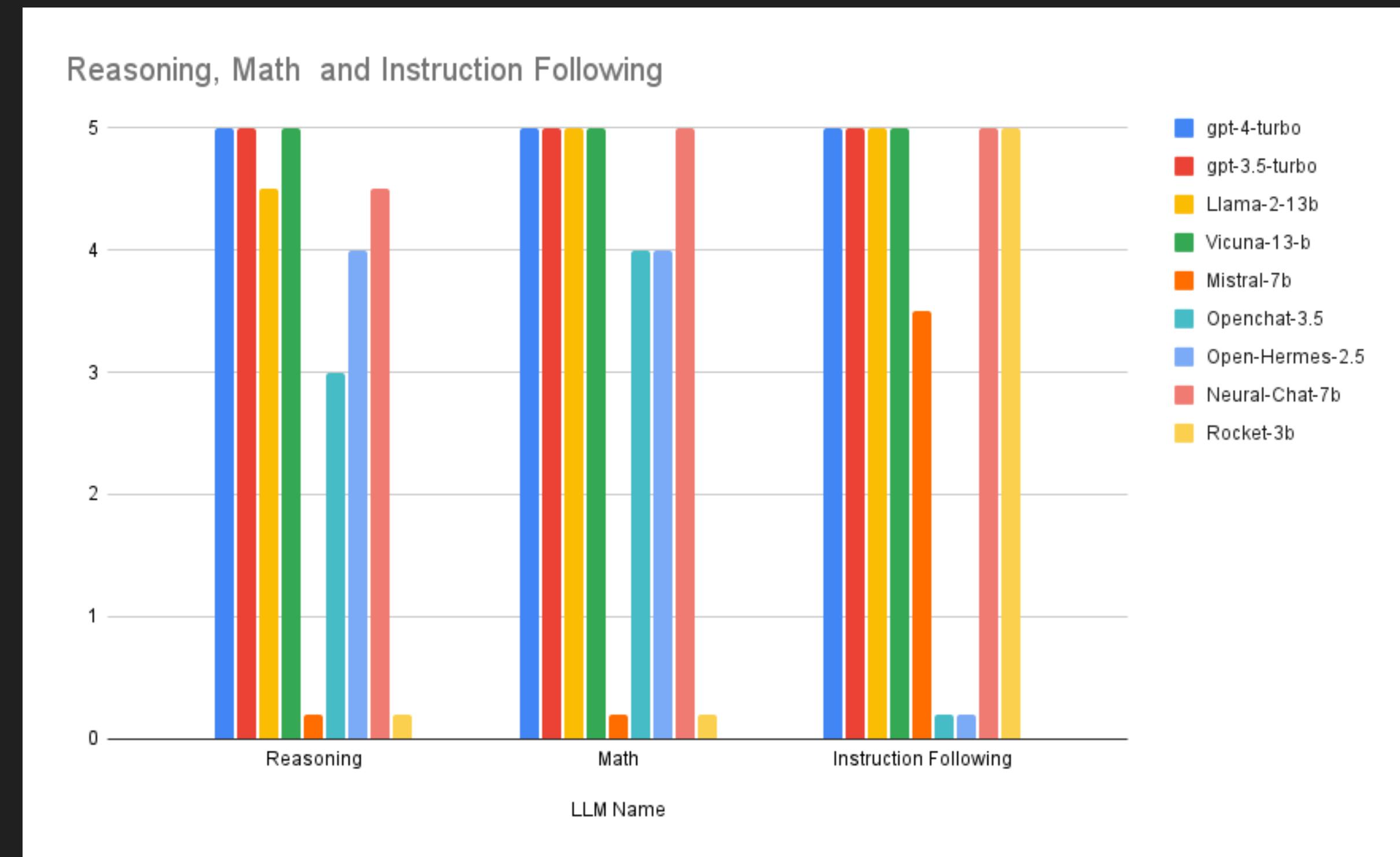
5. **Basic Strategy**: Basic blackjack strategy would advise you to stand on any total of 17 or higher, regardless of what the dealer is showing. With a total of 20, you are well above that threshold.

Given these considerations, the most strategically sound decision is to stand. You want to avoid the risk of busting and give the dealer the chance to try to improve their hand, which could lead to them busting.

Stand

Results: Few-Shot Prompting

Results



Model Bias

“Losses piling up
please say
something else.”



From the models that we evaluated there seems to be bias towards standing without prompt improvements:

User:

USER
You are playing a game of blackjack.
You have a hand which totals 12.
The dealer has a hidden card and a card visible to you. The visible card is 10.
Should you hit or stand? After thinking, give me either hit or stand in a new line.

AI

I suggest standing. Since your hand already totals 12 and the dealer's visible card is a 10, there is a good chance that the dealer has a face card (with a value of 10) as their hidden card, which would give them a total of 20. In this case, standing with your hand of 12 would be the safest play, as it is unlikely that you will improve your hand by taking another card and risking busting (going over 21).

USER

You are playing a game of blackjack.
You have a hand which totals 12.
The dealer has a hidden card and a card visible to you. The visible card is 10.
Should you hit or stand? After thinking, give me either hit or stand in a new line.

AI

Stand

Results Comparisons

Model	Win	Tie/Loss
RL Agent	5	5
gpt-4.5-turbo	5	5
gpt-3.5-turbo	4	6
Vicuna 13b	4	6
Nerual chat 7b	3	7

Navigation Environment

“Take LLM for a Walk”



We evaluated the following LLMs:

- GPT 4
- GPT 3.5
- LLama 2 13B
- Mistral Instruct 7B
- Vicuna 13B

Baseline Prompting: Zero Shot

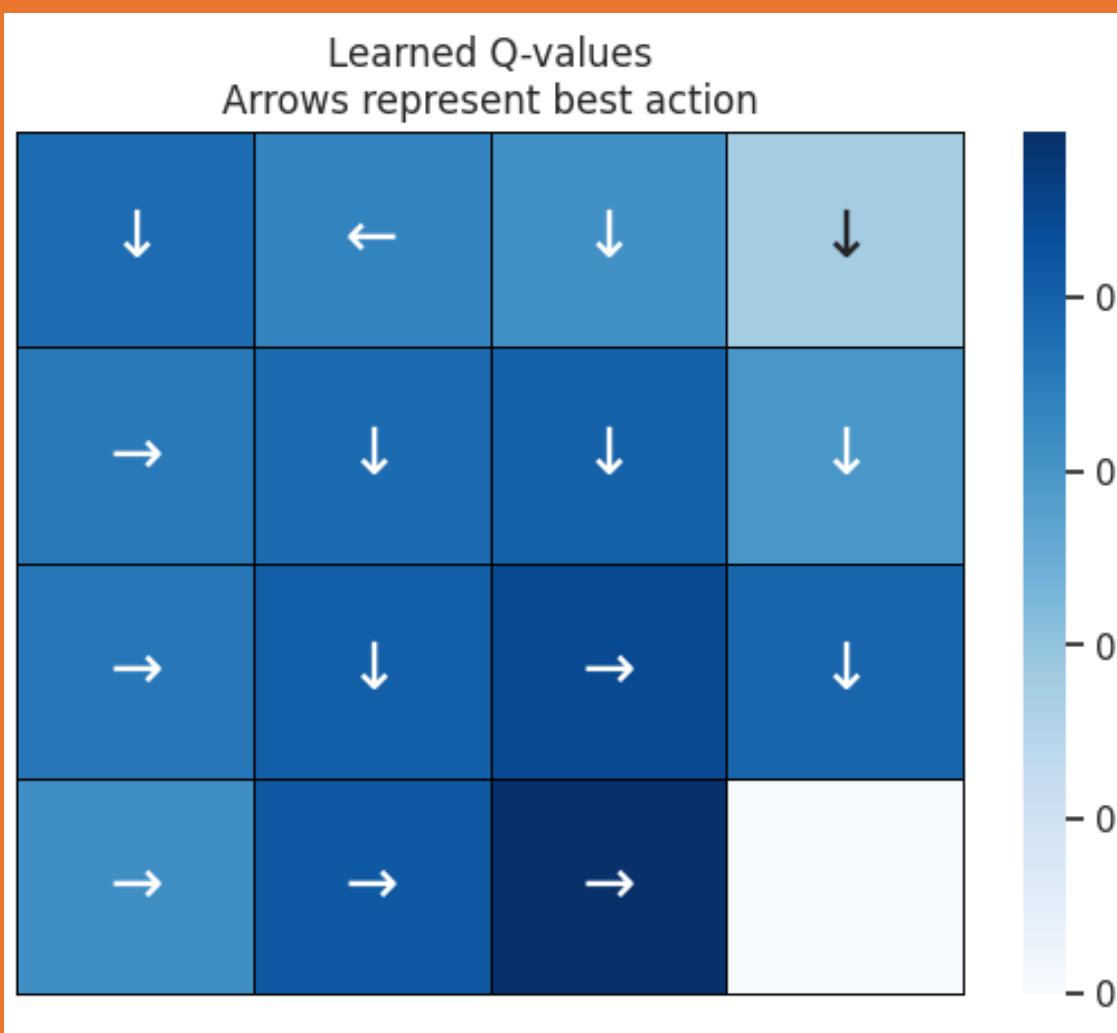
Improved Prompting: Few Shot, Chain of Thought

Evaluated Against:

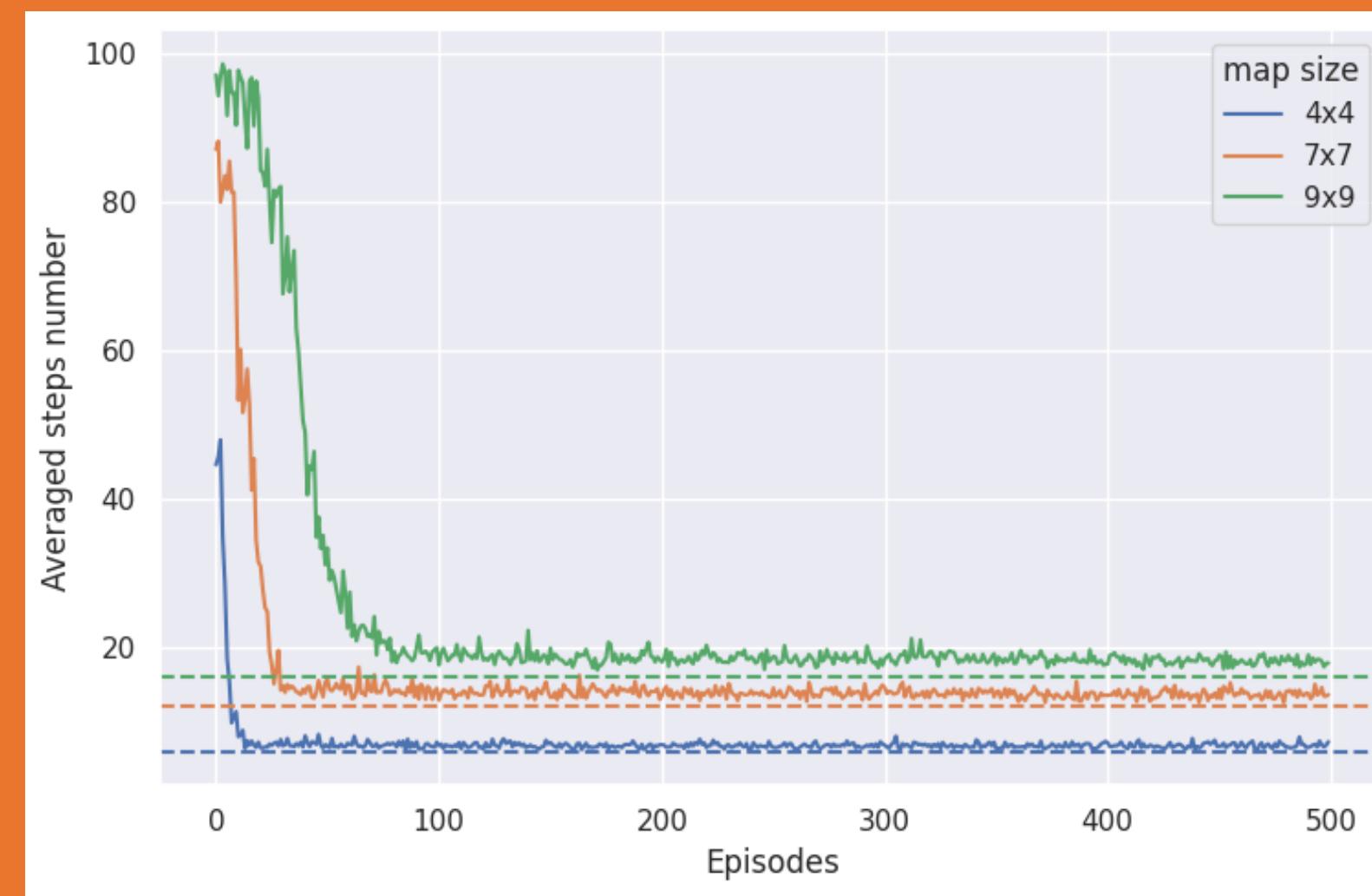
- Deep Q Learning Agent (RL)
- A* Algorithm for Navigation

Results: RL/Det. Baseline

The RL Agent identifies the most optimal move from each cell



We used the A^* algorithm as ground truth, knowing it is deterministic and produces the most optimal output.



The average steps number decreases as the RL agent improves its Q-values, tending to the optimal number defined by A^*

Results: Baseline

Prompt Used:

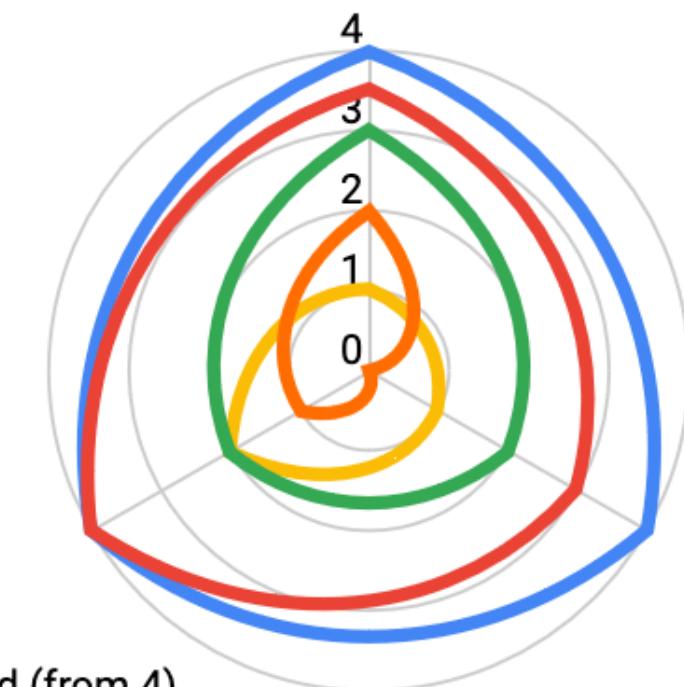
You are a navigator bot that provides a context of how I want to control navigation. you output directions, Top, Bottom, Right, and Left.

Your current position: (1,1), you have to reach (4,4). Answer which direction you move in. Options A) Up B) Down C) Left D) Right

Results

Reasoning , Math and Prediction Capability

ChatGPT4 ChaGPT3 Mistral 7B Vicuna 13B Llama 2 13B
Reasoning (out of 4)



Prediction Capability Adjusted (from 4)

Model vs Ability (Spider Chart)

Results: COT Prompting

Prompt Used:

User:

You are a navigator bot provide context of how I want to control navigation. you output directions, Top, Bottom, Right and Left.

Your current position: (0,3), you have to reach (3,1). Answer which direction you move in. Options A) Up B) Down C) Left D) Right.

Assistant:

Lets think step by step:

'Step 1: Current Position: (0,3), Target: (3,1). Analyze: Need to move down to reach row 3 and then left to reach column 1. Next Move: Down.

Step 2: Current Position: (1,3), Target: (3,1). Analyze: Still need to move down to row 3, but also left to column 1. Next Move: Left.

Step 3: Current Position: (1,2), Target: (3,1). Analyze: Continue moving down to row 3, then left to column 1. Next Move: Down.

Step 4: Current Position: (2,2), Target: (3,1). Analyze: Move down to row 3, followed by left to column 1. Next Move: Left.

Step 5: Current Position: (2,1), Target: (3,1). Analyze: Only need to move down to align with row 3, as column is already aligned. Next Move: Down.'

User:

You are a navigator bot provide context of how I want to control navigation. you output directions, Top, Bottom, Right and Left.

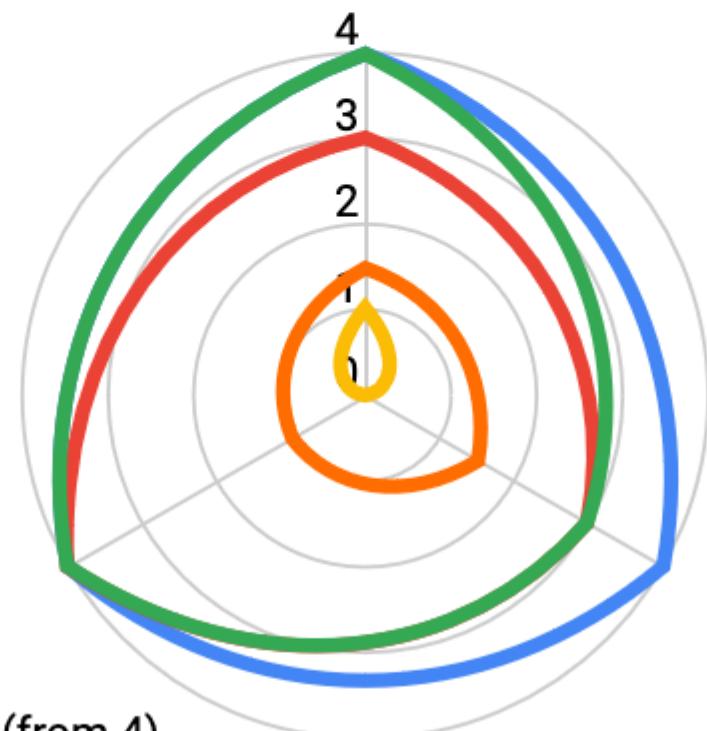
Your current position: (2,2), you have to reach (4,4). Answer which direction you move in. Options A) Up B) Down C) Left D) Right.

Results

Reasoning , Math and Prediction Capability

ChatGPT4 ChaGPT3 Mistral 7B Vicuna 13B Llama 2 13B

Reasoning (out of 4)



Prediction Capability Adjusted (from 4)

Math (out of 4)

Model vs Ability (Spider Chart)

Results: Few-Shot Prompting

Prompt Used:

You are a navigator bot provide context of how I want to control navigation. you output directions, Top, Bottom, Right and Left.

Look at the following data and its Answers:

Your current position: (2,3), you have to reach (1,1). Answer : left

Your current position: (2,2), you have to reach (1,1). Answer : left

Your current position: (2,1), you have to reach (1,1). Answer : up

Your current position: (1,3), you have to reach (2,0). Answer : left

Your current position: (1,2), you have to reach (2,0). Answer : left

Your current position: (1,1), you have to reach (2,0). Answer : left

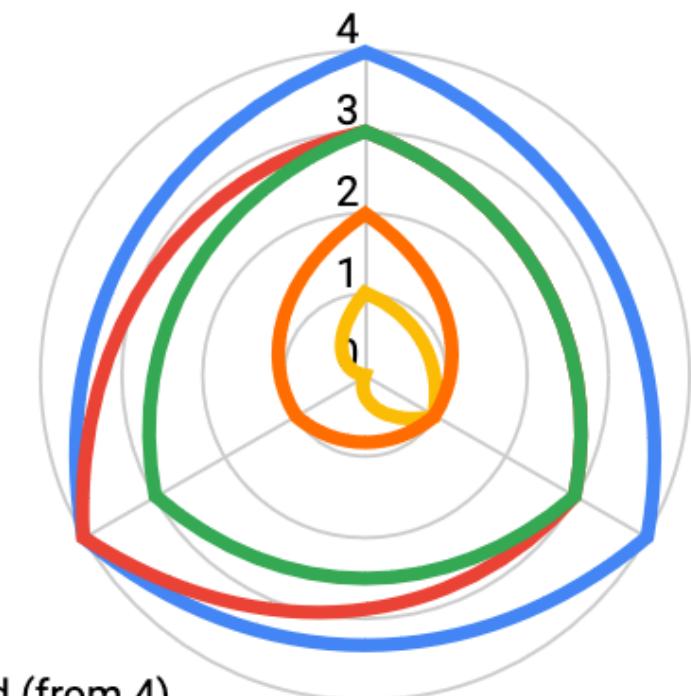
Now Answer:

Your current position: (2,2), you have to reach (4,4). Answer which direction you move in. Options A) Up B) Down C) Left D) Right.

Results

Reasoning , Math and Prediction Capability

ChatGPT4 ChaGPT3 Mistral 7B Vicuna 13B Llama 2 13B
Reasoning (out of 4)



Prediction Capability Adjusted (from 4)

Model vs Ability (Spider Chart)

Math (out of 4)

Discovering: Confusion

“Is up North or
up South or what
is what?”



While assessing the navigation capabilities, it became evident that certain Large Language Models (LLMs) frequently exhibited a tendency to initially provide appropriate reasoning for reaching a goal location, only to subsequently propose a different set of directions. This observation implies that the LLM is prone to confusing the spatial matrix, rendering it unable to definitively determine whether $(0,2)$ is positioned to the right/left or top/bottom of $(0,1)$.

Discovering: Confusion

Step N

Input: (1, 1)

Answer: A) Up

Explanation: To reach (4, 4) from (1, 1), we need to move up. Therefore, the correct answer is option A) Up.

Step N + 1

B) Down

Explanation:

To reach the destination point (4, 4) from the current position (1, 1), we need to go down 3 steps and then move right 2 steps. So, the answer is "Down".

Results Comparison

Model	Average Steps in 4x4 Matrix
A* Algorithm	6
Random Policy	45
RL Agent	6
GPT-4	6
GPT-3.5	6
LLama 2	20
Mistral Instruct	16
Vicuna	10

Generating: A Dataset for future LLMS

**BAKED
FRESH**

“A Dataset fresh out of oven.”

With the help of the A* Algorithm, we were able to generate a data generator that can generate n number of data points and convert it to prompt for fine-tuning and few-shot learning.

The sample dataset is as follows:

	Prompt	Answer
0	Your current position: (1,1), you have to reach (3,3). Answer which direction you move in. Options A) Up B) Down C) Left D) Right	right
1	Your current position: (1,2), you have to reach (3,3). Answer which direction you move in. Options A) Up B) Down C) Left D) Right	down
2	Your current position: (2,2), you have to reach (3,3). Answer which direction you move in. Options A) Up B) Down C) Left D) Right	right
3	Your current position: (2,3), you have to reach (3,3). Answer which direction you move in. Options A) Up B) Down C) Left D) Right	down

Decision from Weather Environment

“Hey LLM, is it going to be hot outside?”



- 1."In this environment, LLM agents predict clothing and accessory choices based on 7-day weather data, considering temperatures and conditions like snow, rain, or sun."
- 2."Agents analyze trends to choose the most suitable attire, balancing weather patterns with temperature averages."
- 3."The model's success is measured by its ability to accurately match clothing and accessories to the prevailing weather conditions, showcasing its decision-making in real-world scenarios."

LLMs Evaluated:

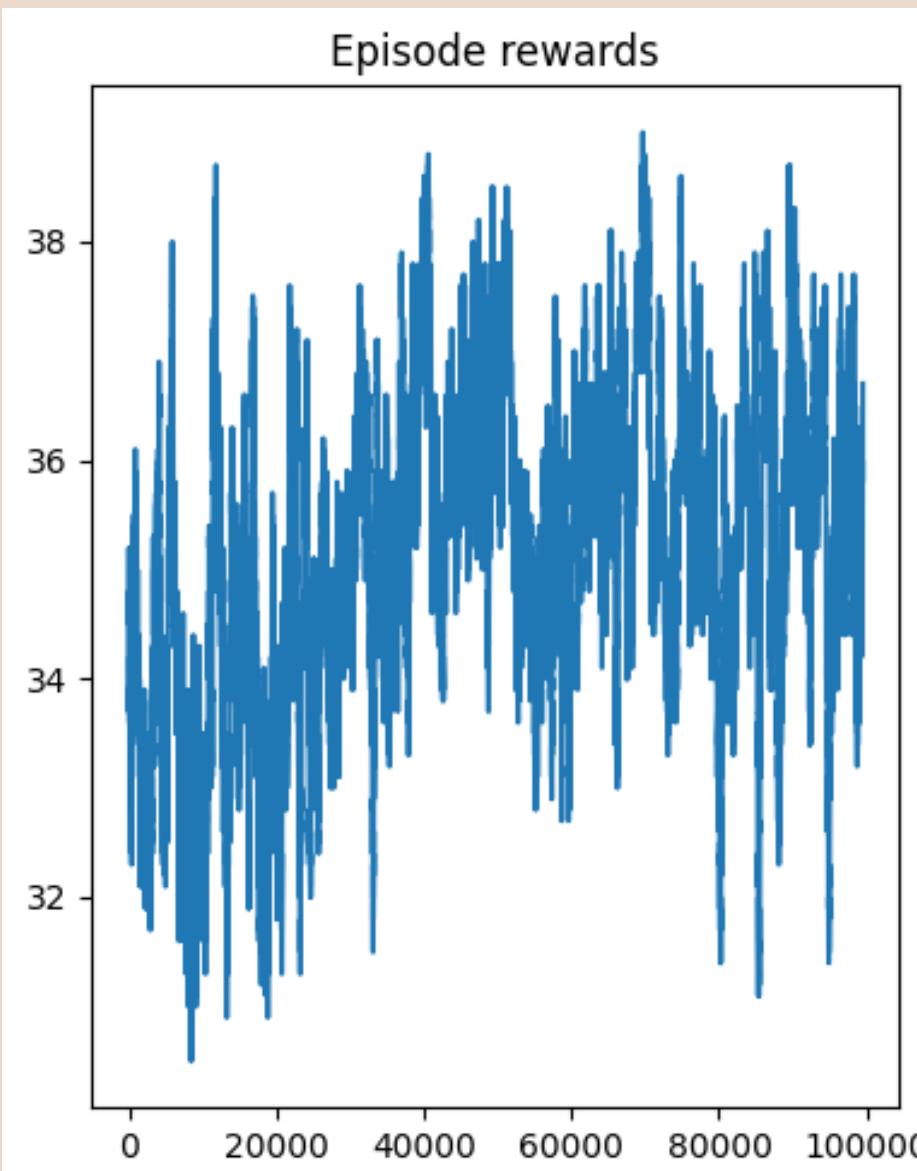
- GPT4
- GPT3.5
- Vicuna 13B
- Llama 13B
- Mistral 7B

Prompting Methods:

- Zero-shot Learning
- Few-shot Learning
- Chain of Thought

Results: RL/Det. Baseline

The RL Agent reward shows a slight improvement over time



In the context of the weather environment, the Q-learning model exhibited limitations in complexity, rendering it not able to grasp the mathematical information and decision-making requirements of the environment.

Consequently, this led to an average reward plateauing at 35.4, reflecting the model's struggle to navigate the complexities inherent in this environment.

Results: Baseline

Prompt Used:

Prompt

Consider the following data about weather conditions for the last 7 days:

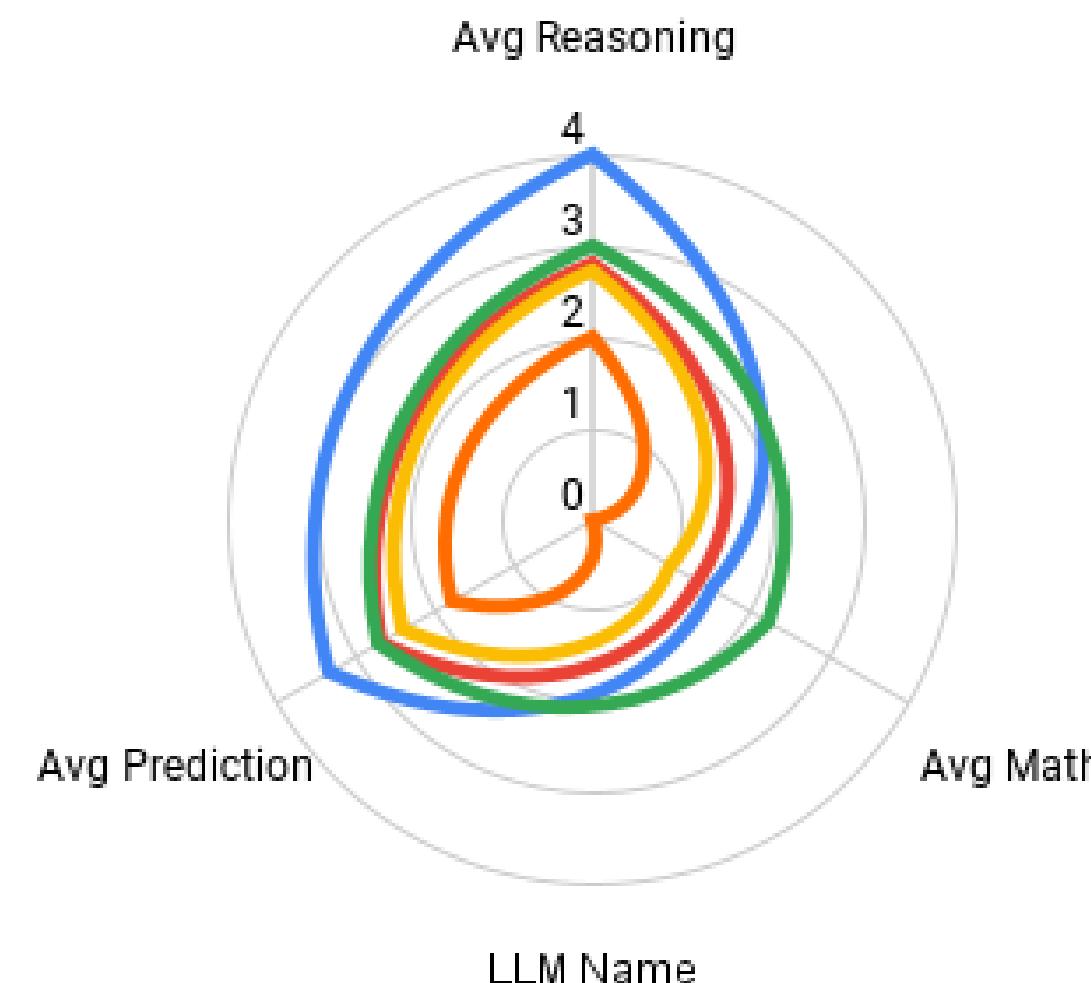
- Temperatures (in °F): [47.348949541566256, -31.622330311862147, 25.32772017492796, -41.84648929886844, 13.360171709887624, 4.580407868630644, 60.09024980191904]
- Weather Conditions: ['Sunny', 'Sunny', 'Sunny', 'Snow', 'Snow', 'Sunny', 'Rain']

Based on the above data, what should be the ideal accessory (Sunscreen, Umbrella, Snowshoes) and clothing (Warm, Medium, Summer) choice that I should make tomorrow? Please explain your reasoning

Results: Based on human evaluation (Sheet)

Avg Reasoning, Avg Math and Avg Prediction

— GPT4 — GPT3 — LLama 2 — Vicuna 13B — Mistral



Results: Improved Few Shot

Prompt Used:

Prompt

Consider the following data about weather conditions for the last 7 days:

- Temperatures (in °F): [47.348949541566256, -31.622330311862147, 25.32772017492796, -41.84648929886844, 13.360171709887624, 4.580407868630644, 60.09024980191904]
- Weather Conditions: ['Sunny', 'Sunny', 'Sunny', 'Snow', 'Snow', 'Sunny', 'Rain']

Rules:

1. The ideal accessory choice depends on the most common weather condition:

- If it's mostly sunny, pick 'Sunscreen'.
- If it's mostly raining, pick 'Umbrella'.
- If it's mostly snowing, pick 'Snow Boots'.

2. The ideal clothing choice depends on the average temperature:

- Below 32°F: 'Warm' clothing.
- Between 32°F and 70°F: 'Medium' clothing.
- Above 70°F: 'Summer' clothing.

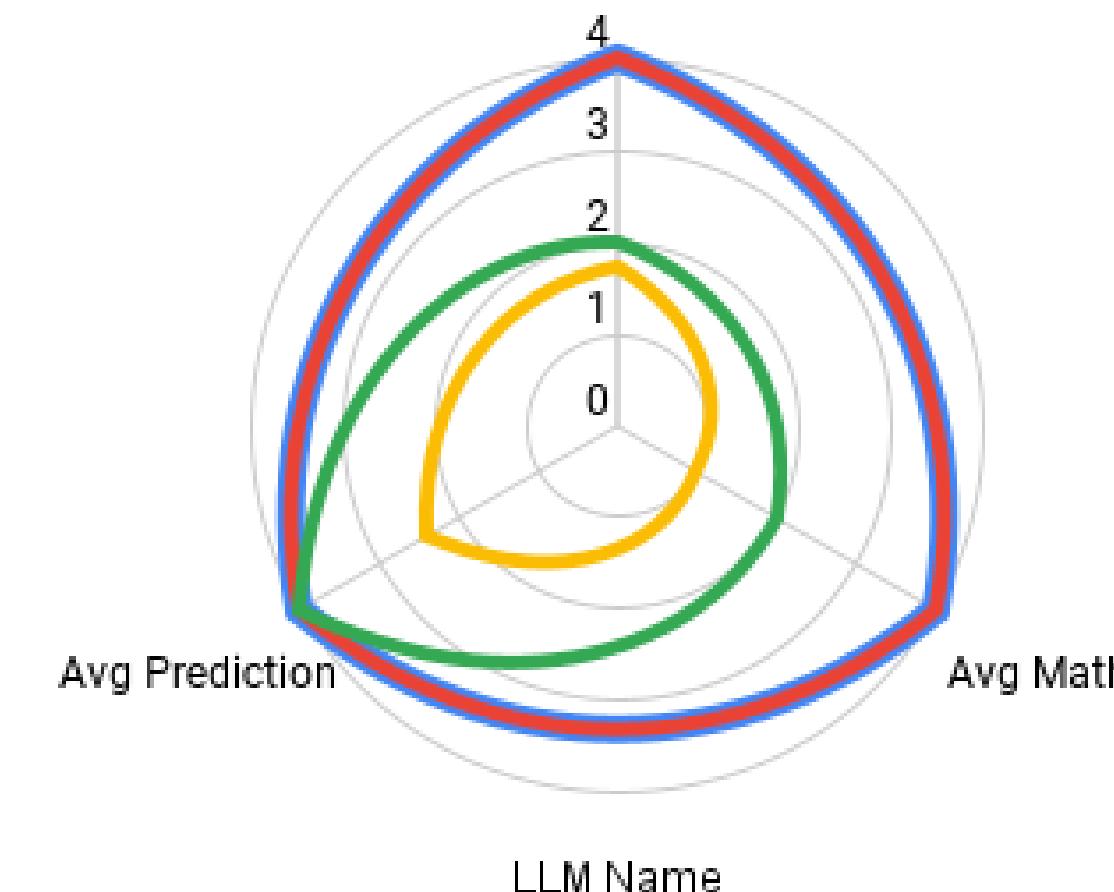
Based on the above data and rules, what should be the ideal accessory and clothing choice? Please explain your reasoning.

Results: Based on human evaluation (Sheet)

Avg Reasoning, Avg Math and Avg Prediction

GPT4 GPT3 LLama 2 Vicuna 13B Mistral

Avg Reasoning



Results: Improved Chain of thought

Prompt Modification:

Answer:

To determine the ideal accessory and clothing choice based on the given data and rules, we need to evaluate the most common weather condition and the average temperature.

1. Determine the Most Common Weather Condition:

- Count the occurrences of each weather condition in the data.
- We have 'Sunny' on 4 days, 'Snow' on 2 days, and 'Rain' on 1 day.
- Therefore, the most common weather condition is 'Sunny'.

2. Choose the Ideal Accessory:

- Based on Rule 1, since it's mostly sunny, the ideal accessory is 'Sunscreen'.

3. Calculate the Average Temperature:

- Add up the temperatures: $47.35 - 31.62 + 25.33 - 41.85 + 13.36 + 4.58 + 60.09 = 76.34^{\circ}\text{F}$.
- Divide by the number of days (7) to get the average: $76.34 / 7 \approx 10.91^{\circ}\text{F}$.

4. Choose the Ideal Clothing:

- Based on Rule 2, since the average temperature (10.91°F) is below 32°F , the ideal clothing is 'Warm'.

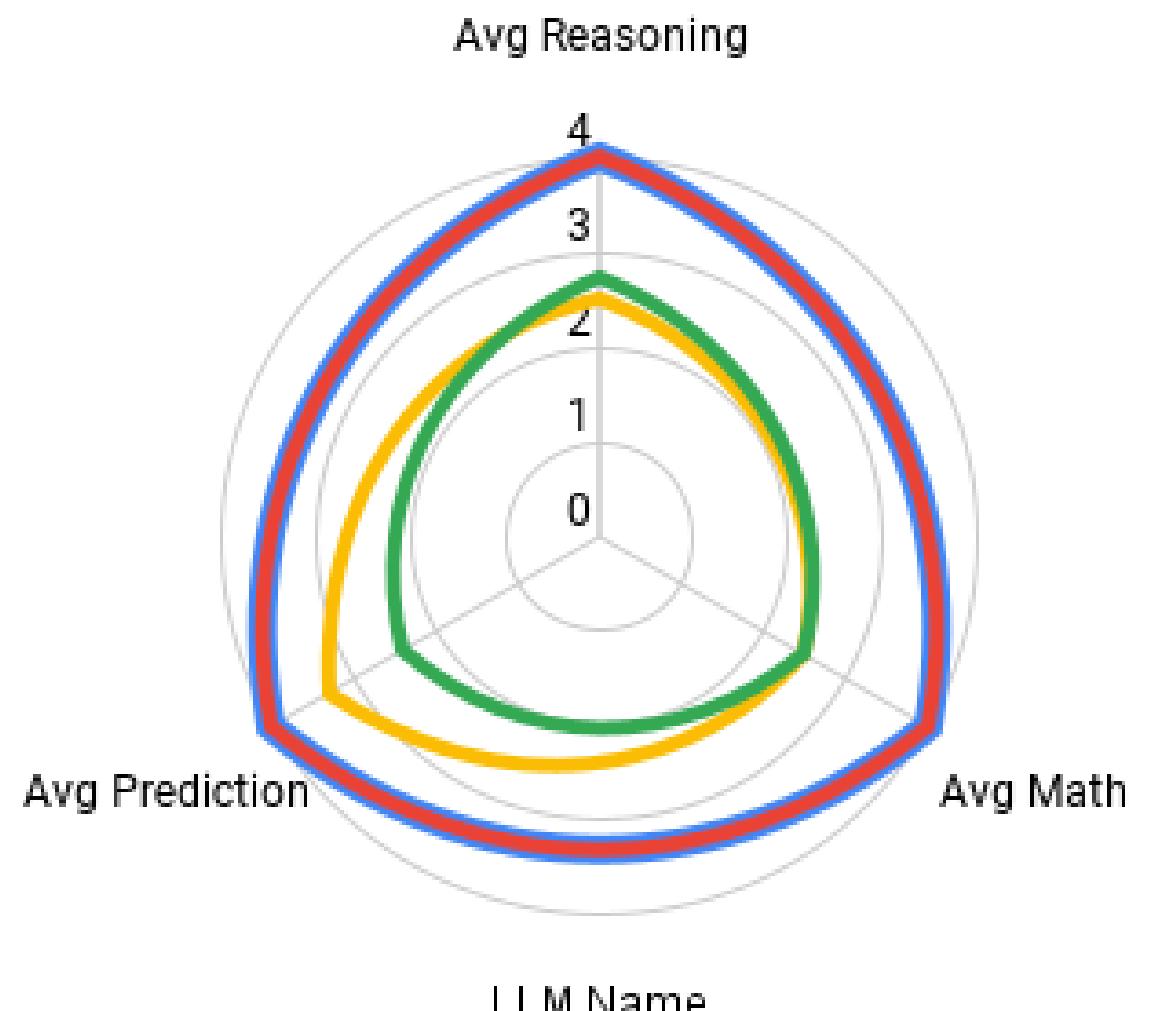
In conclusion, the ideal accessory choice is 'Sunscreen', and the ideal clothing choice is 'Warm', based on the most common weather condition and the average temperature.

Prompt: Q, A, Q

Results: Based on human evaluation (Sheet)

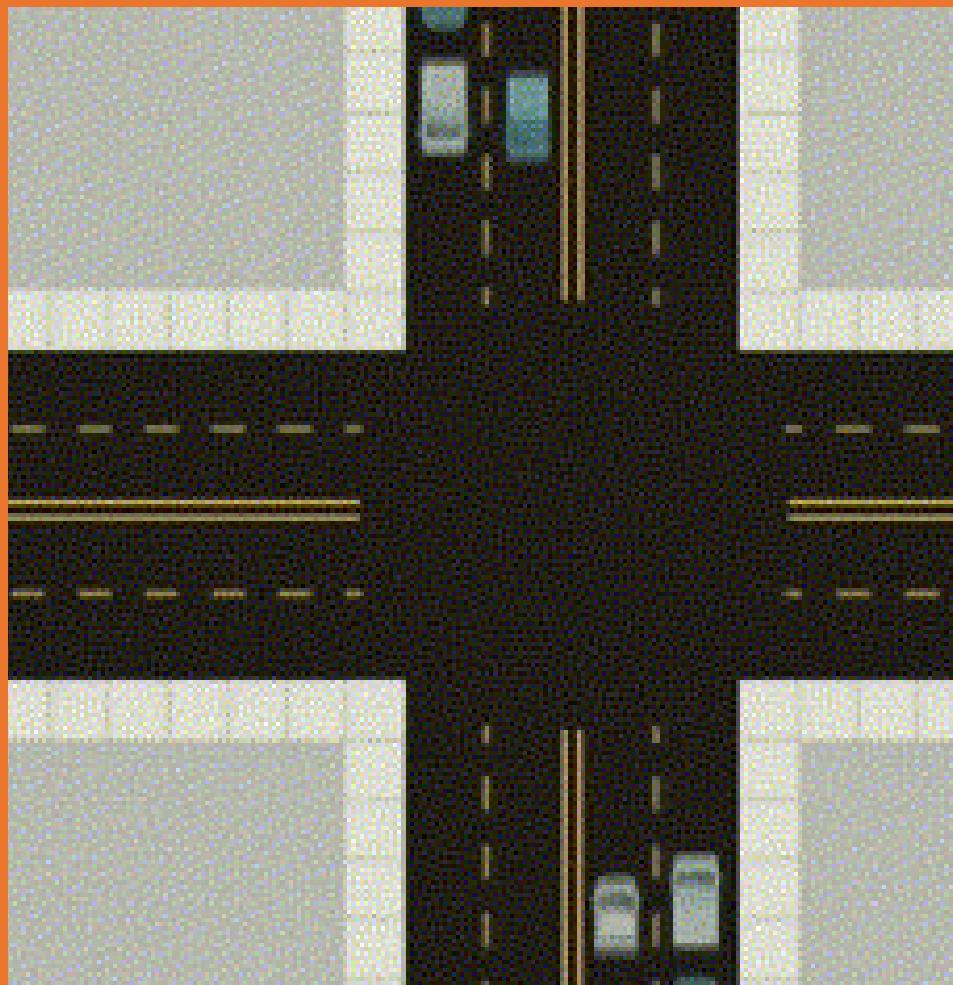
Avg Reasoning, Avg Math and Avg Prediction

■ GPT4 ■ GPT3 ■ LLaMA 2 ■ Vicuna 13B ■ Mistral



Managing a Traffic Signal

“LLM for Traffic Managing”



- "In this environment, LLM agents optimize traffic flow at intersections, making real-time decisions for signal control based on vehicle and pedestrian traffic data."
- "The model considers various factors, such as vehicle count and waiting times in different directions, to efficiently manage North-South and East-West traffic signals."
- "Success is measured by the agent's ability to reduce congestion and prioritize emergency vehicles, showcasing its applicability in dynamic urban traffic scenarios."

LLMs Evaluated:

- GPT4
- GPT3.5
- Vicuna 13B
- Llama 13B
- Mistral 7B

Prompting Methods:

- Zeroshot
- Fewshot
- Chain of Thought

Results: Baseline

Prompt Used:

Prompt

Consider a traffic control scenario at a four-way intersection. The traffic conditions are as follows:

- North Direction: 12 vehicles for through/right, 6 vehicles for left, 0 cycle wait for through/right, 4 cycles wait for left, 7 pedestrians waiting, 2 cycles wait since last pedestrian, 0 emergency vehicles for through/right, 0 emergency vehicles for left.
- South Direction: 18 vehicles for through/right, 0 vehicles for left, 1 cycle wait for through/right, 2 cycles wait for left, 6 pedestrians waiting, 2 cycles wait since last pedestrian, 0 emergency vehicles for through/right, 0 emergency vehicles for left.
- East Direction: 4 vehicles for through/right, 3 vehicles for left, 3 cycle wait for through/right, 1 cycles wait for left, 6 pedestrians waiting, 3 cycles wait since last pedestrian, 0 emergency vehicles for through/right, 0 emergency vehicles for left.
- West Direction: 10 vehicles for through/right, 2 vehicles for left, 1 cycle wait for through/right, 3 cycles wait for left, 1 pedestrians waiting, 1 cycles wait since last pedestrian, 0 emergency vehicles for through/right, 0 emergency vehicles for left.

Based on these conditions, which direction should be prioritized for the next signal change to optimize traffic flow, considering vehicle and pedestrian wait times, counts, and the presence of emergency vehicles?

Results: Based on human evaluation (Sheet)



Results: Improved Few Shot

Prompt Used:

Solution:

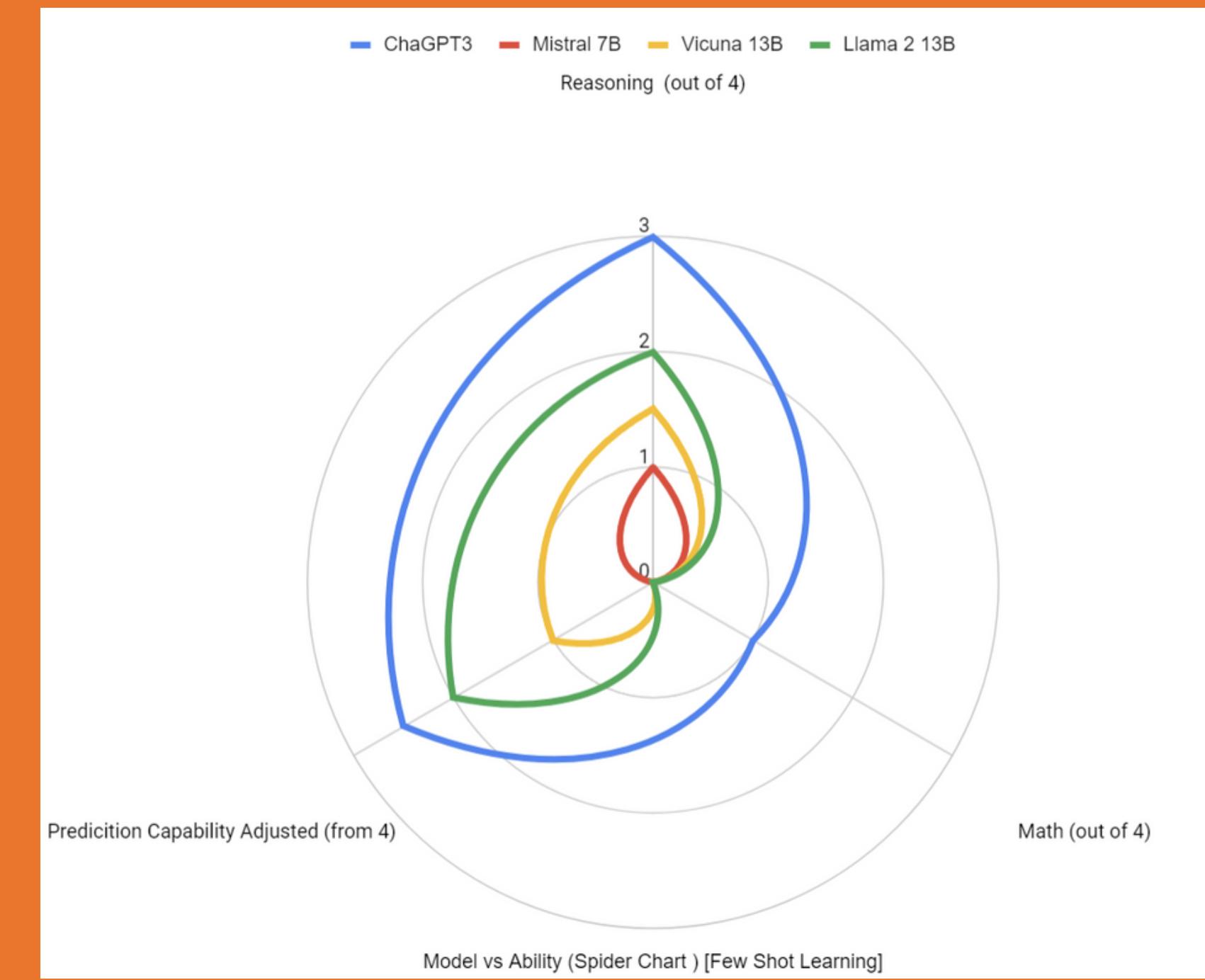
In assessing the traffic conditions at the four-way intersection, it's important to prioritize the direction that has the highest overall congestion, factoring in both vehicle and pedestrian wait times.

The West Direction, for instance, has a significant backlog of vehicles, particularly in the through/right lane, with the longest wait time observed. There are also a considerable number of pedestrians waiting. The South Direction shows a moderate level of vehicle congestion and pedestrian wait. The North and East Directions, in comparison, have fewer vehicles waiting and shorter wait times. Given these factors, the West Direction should be prioritized for the next signal change. This approach will address the largest volume of traffic, easing significant congestion and improving safety for both vehicles and pedestrians.

Under the North American traffic control system, where either north-south or east-west directions can be controlled, and specifically through/right or left lanes separately, a reassessment is needed.

The East-West Through/Right Lanes experience the highest total number of vehicles waiting, with the West direction having a significant backlog and the longest wait time. The North-South Through/Right Lanes also exhibit considerable congestion, especially in the South direction. The left lanes in both the North-South and East-West directions show less congestion. Considering this setup, the East-West Through/Right Lanes should be given priority to alleviate the most significant traffic congestion, especially in the West direction, which has the longest wait time and a high vehicle count. This decision aims to improve overall traffic flow at the intersection.

Results: Based on human evaluation (Sheet)



Prompt: Q, A1, A2, Q

Results: Improved Chain of thought

Prompt Modification:

Under the North American traffic control system, where we can control either north-south or east-west directions, and specifically through/right or left lanes separately, let's reassess the situation:

North-South Through/Right:

North: 9 vehicles (1 cycle wait) South: 13 vehicles (4 cycle wait) Total: 22 vehicles

North-South Left:

North: 2 vehicles (0 cycles wait) South: 6 vehicles (1 cycle wait) Total: 8 vehicles

East-West Through/Right:

East: 5 vehicles (2 cycle wait) West: 20 vehicles (5 cycle wait) Total: 25 vehicles

East-West Left:

East: 4 vehicles (1 cycle wait) West: 6 vehicles (3 cycle wait) Total: 10 vehicles

Considering this setup:

The East-West Through/Right Lanes have the highest total number of vehicles waiting (25 vehicles), with the West direction having a significant backlog and the longest wait time. The North-South Through/Right Lanes also have a considerable number of vehicles waiting (22 vehicles), with the South direction experiencing a longer wait time. The North-South Left and East-West Left Lanes have fewer vehicles waiting in comparison. Given these observations, the best choice to optimize traffic flow would be to prioritize the East-West Through/Right Lanes. This approach will address the largest volume of traffic, particularly in the West direction, which has the longest wait time and a high vehicle count. This decision aims to reduce the most significant congestion and improve overall traffic flow at the intersection.

Results: Based on human evaluation (Sheet)



Prompt: Q, A1, A2, Q

Conclusions & Future Work



For navigation tasks, LLMs face limitations due to their inherent design focused on language comprehension rather than spatial awareness. In strategic decision-making scenarios, they showcase better performance, sometimes outperforming basic RL agents.

For future work we plan on fine-tuning the LLMs to assess whether it can further improve the results we obtained in this project.

"Embrace change to stay ahead"

Thanks for your attention!

Questions?