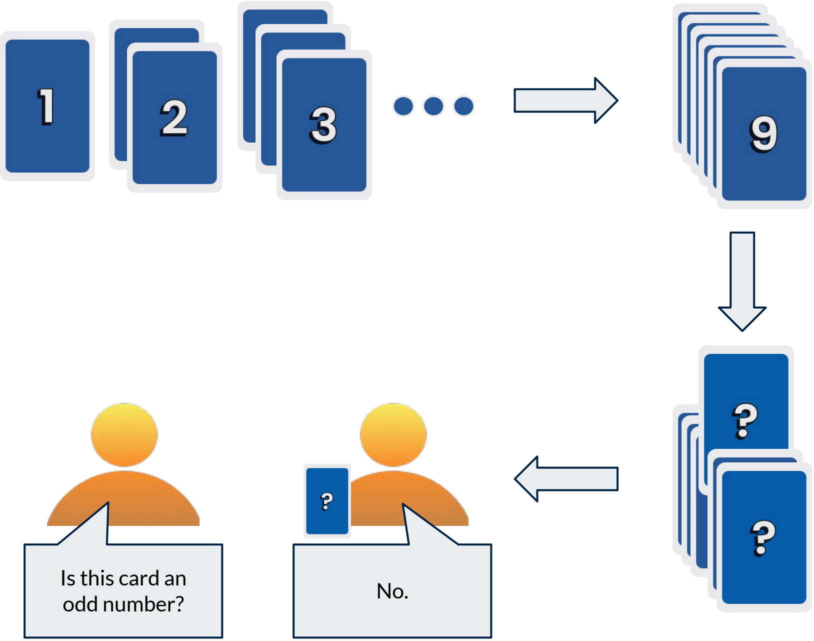
# Investigating the Capabilities of Generative AI in Solving Data Structures, Algorithms, and Computability Problems

**Authors:**
**Nero Li, Shahar Broner, Yubin Kim,** Katrina Mizuo, Elijah Sauder, **Claire To**, Albert Wang, Ofek Gila, and Michael Shindler

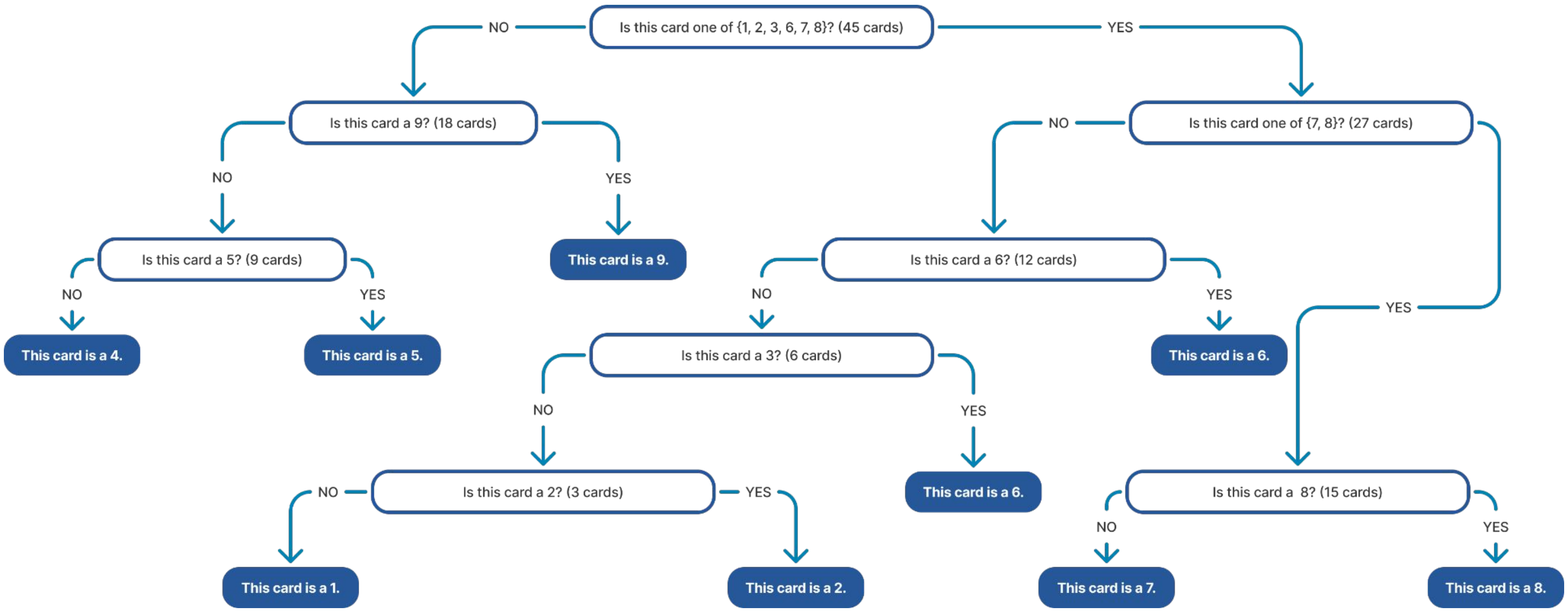
# Guess Card

Design a strategy that minimizes the expected number of questions you will ask in the following game:

- You have a deck of cards that consists of one one, two twos, three threes, etc.
- You have up to nine nines for a total of 45 cards.
- Randomly pick a card and guess the value by asking several closed-ended questions.

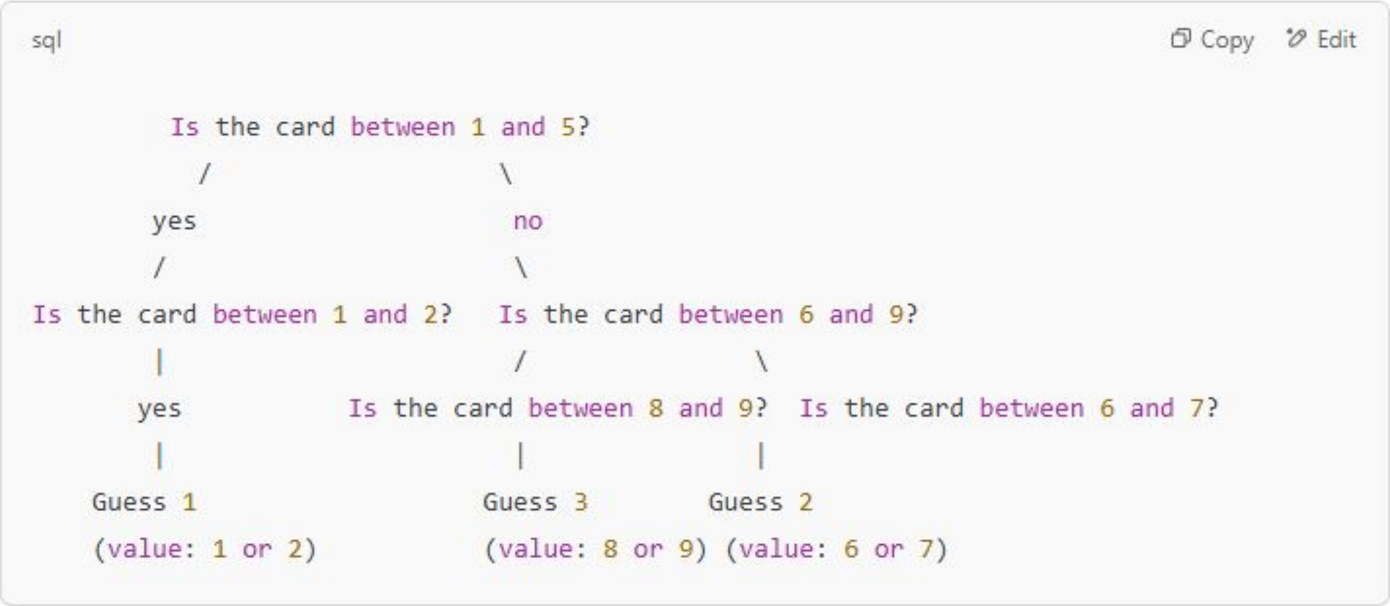


# Guess Card



- Is this card one of {1, 2, 3, 6, 7, 8}? (45 cards)
  - NO: Is this card a 9? (18 cards)
    - NO: Is this card a 5? (9 cards)
      - NO: **This card is a 4.**
      - YES: **This card is a 5.**
    - YES: **This card is a 9.**
  - YES: Is this card one of {7, 8}? (27 cards)
    - NO: Is this card a 6? (12 cards)
      - NO: Is this card a 3? (6 cards)
        - NO: Is this card a 2? (3 cards)
          - NO: **This card is a 1.**
          - YES: **This card is a 2.**
        - YES: **This card is a 6.**
      - YES: **This card is a 6.**
    - YES: Is this card a 8? (15 cards)
      - NO: **This card is a 7.**
      - YES: **This card is a 8.**

# Guess Card

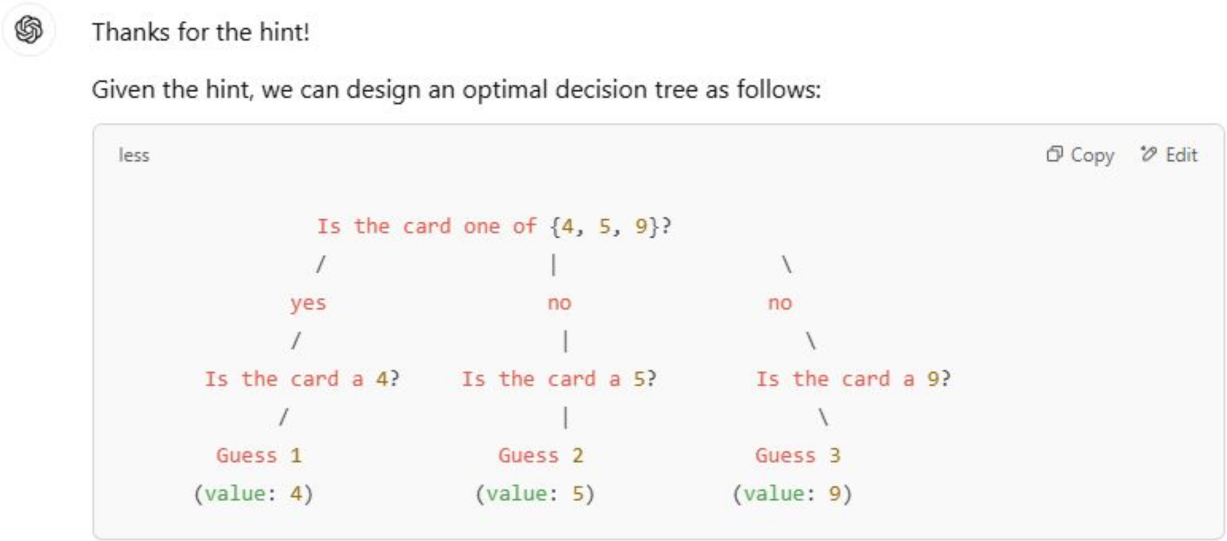


```sql
         Is the card between 1 and 5?
           /                   \
        yes                     no
        /                       \
Is the card between 1 and 2?   Is the card between 6 and 9?
        |                       /               \
       yes          Is the card between 8 and 9?  Is the card between 6 and 7?
        |                       |               |
    Guess 1                   Guess 3        Guess 2
    (value: 1 or 2)           (value: 8 or 9) (value: 6 or 7)
```

# Guess Card

**Hint:** The first question to ask in the optimal decision tree can be:

- "Is the card one of {4, 5, 9}?"
- "Is the card one of {1, 2, 3, 6, 7, 8}?"



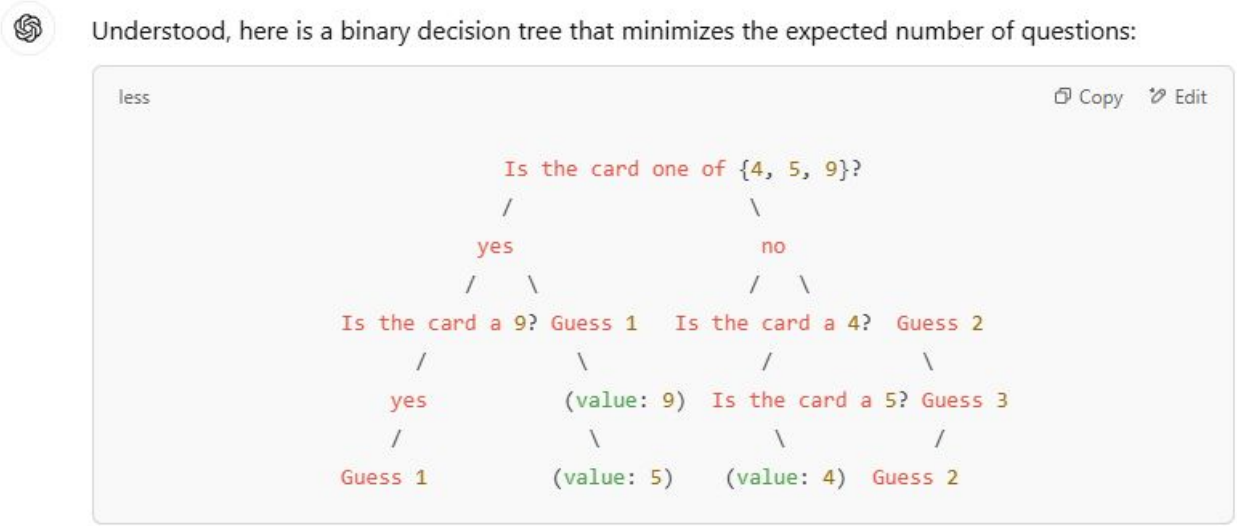
Thanks for the hint!

Given the hint, we can design an optimal decision tree as follows:

```
          Is the card one of {4, 5, 9}?
          /                   |                  \
        yes                  no                no
        /                     |                   \
 Is the card a 4?     Is the card a 5?        Is the card a 9?
       /                      |                    \
  Guess 1                Guess 2              Guess 3
(value: 4)             (value: 5)           (value: 9)
```

# Guess Card

**Clarify:** This decision tree should be binary.



Understood, here is a binary decision tree that minimizes the expected number of questions:

```
             Is the card one of {4, 5, 9}?
             /                    \
           yes                    no
          /    \                 /   \
Is the card a 9? Guess 1   Is the card a 4?  Guess 2
      /            \              /            \
    yes           (value: 9)  Is the card a 5? Guess 3
    /               \              \            /
Guess 1          (value: 5)    (value: 4)  Guess 2
```

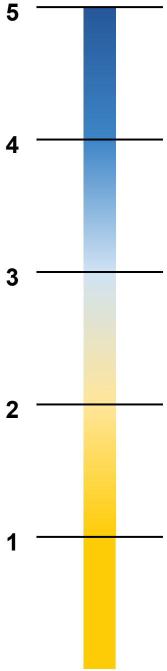
- **RQ:** How does ChatGPT perform on more advanced data structures, algorithms and Theory of Computation free-response questions?

# Previous Work

- Handle programming assignments, written artifacts, etc. [Malinka et al.]
- Poor success rate in Parsons Problems [Parsons and Haden] & Proof Block Problems [Poulsen et al.]
- GPT-4 performs as B- on undergraduate ToC course [Golesteanu and Vowinkel and Dougherty]
  - We emphasize free-response
  - We incorporate feedback
  - We use GPT-4o

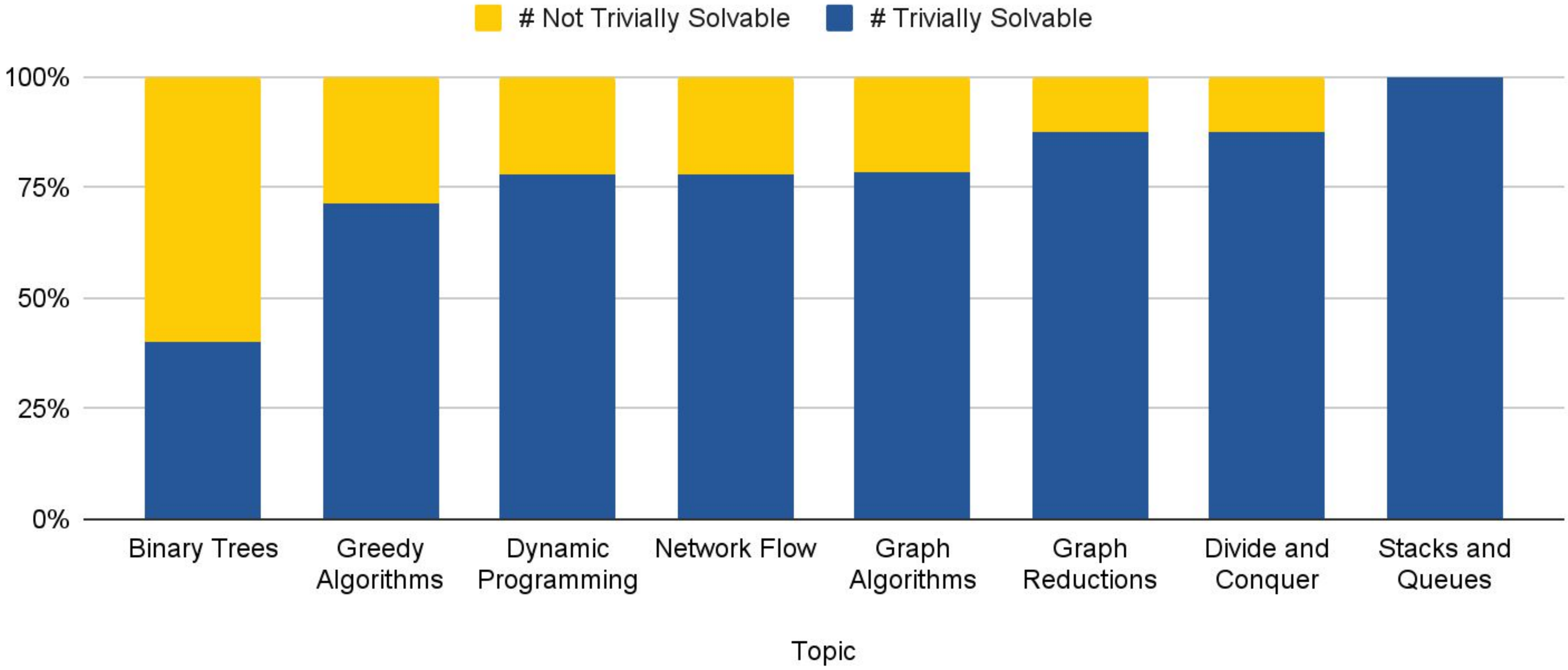
# Methods

- Select representative free response questions
- Choose two “graders” to evaluate the system on that topic
    - Copy and paste for GPT-4
    - Send PDF file for GPT-4o
- After response,  may ask sub-questions
- Graders compared their scores and conclude a final score

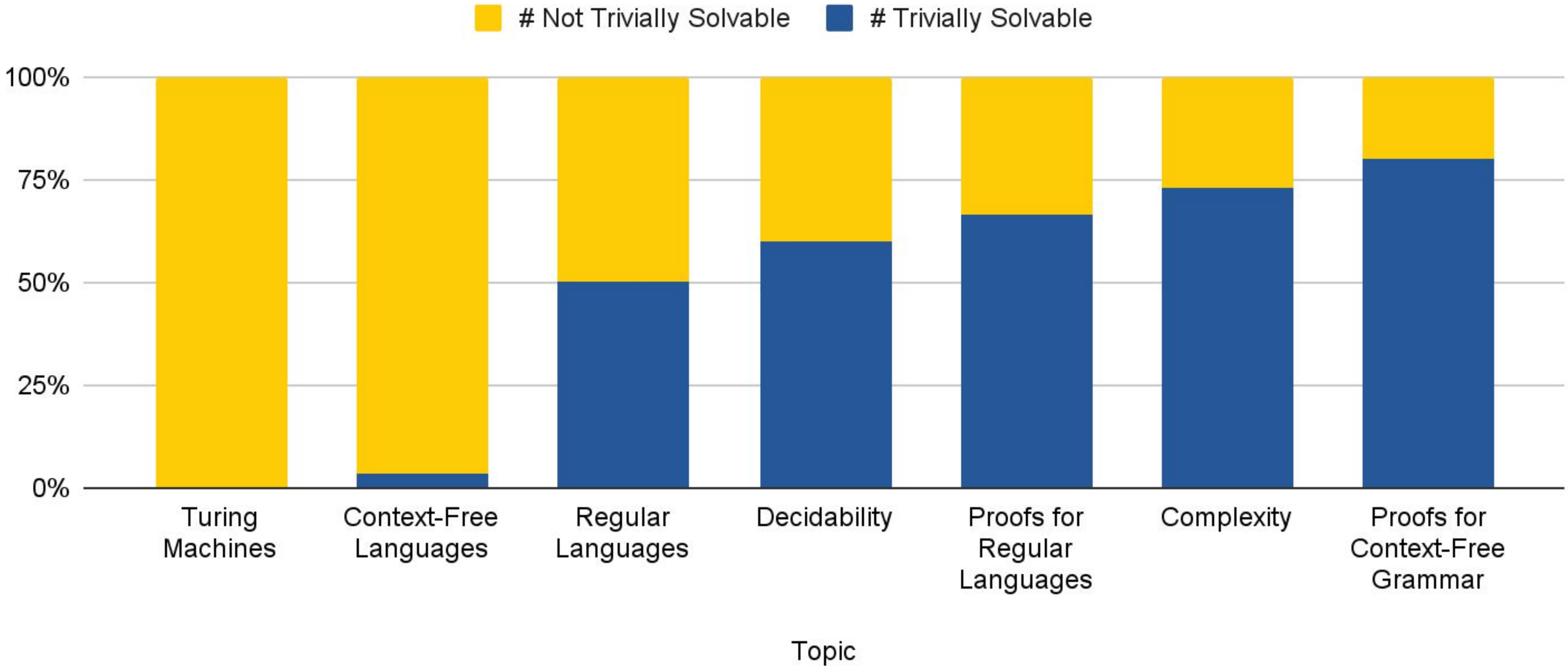


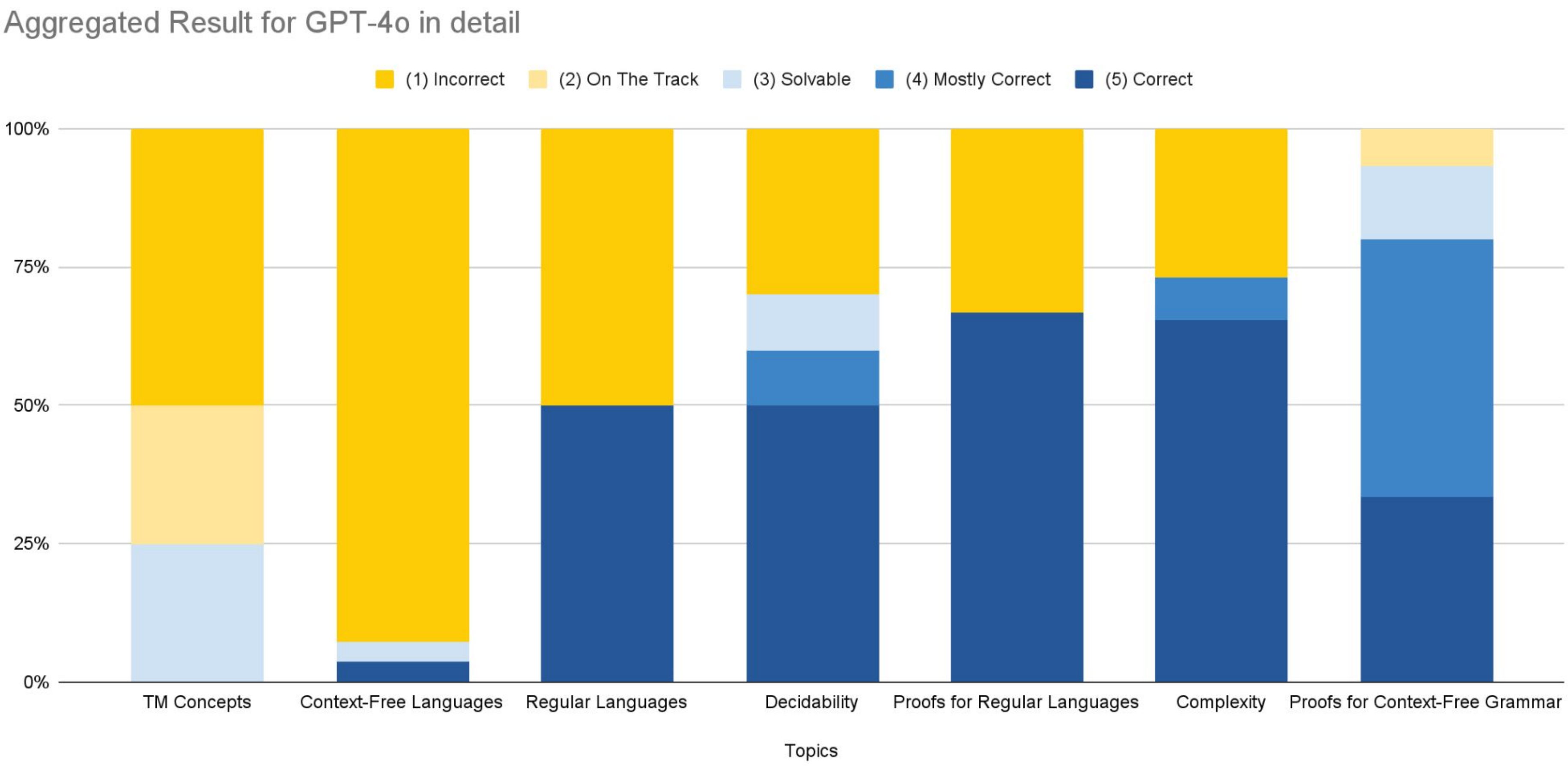
| | | |
|---|---|---|
| 5 | **Correct** | Solves the problem correctly and consistently with no dialogue |
| 4 | **Mostly Correct** | Solves the problem correctly after pointing out a hard-to-identify problem |
| 3 | **Solvable** | Solves the problem correctly after pointing out several problems student should to identify |
| 2 | **On The Track** | Solves the problem in a way that seems believable but has errors |
| 1 | **Incorrect** | Unable to solve the problem without explicitly being told the solution |

# Aggregated Results from GPT-4



# Aggregated Results from GPT-4o




Aggregated Result for GPT-4o in detail
(1) Incorrect
(2) On The Track
(3) Solvable
(4) Mostly Correct
(5) Correct
100%
75%
50%
25%
0%
TM Concepts
Context-Free Languages
Regular Languages
Decidability
Proofs for Regular Languages
Complexity
Proofs for Context-Free Grammar
Topics

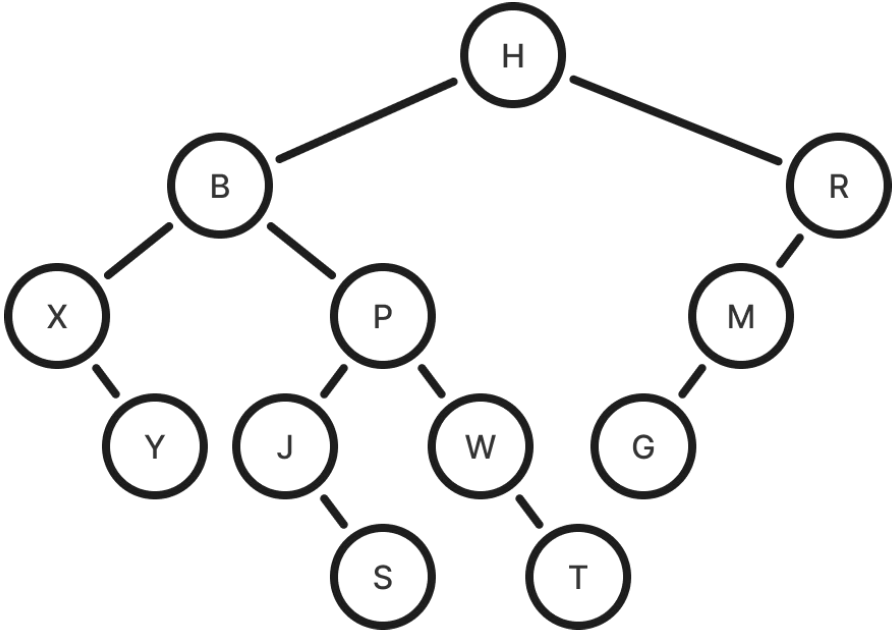
# Binary Tree Traversals

Draw the unique binary tree described by the following traversal:

- In-order: YXBJSPWTHGMR
- Post-order: YXSJTWPBGMRH

A correct answer can be:

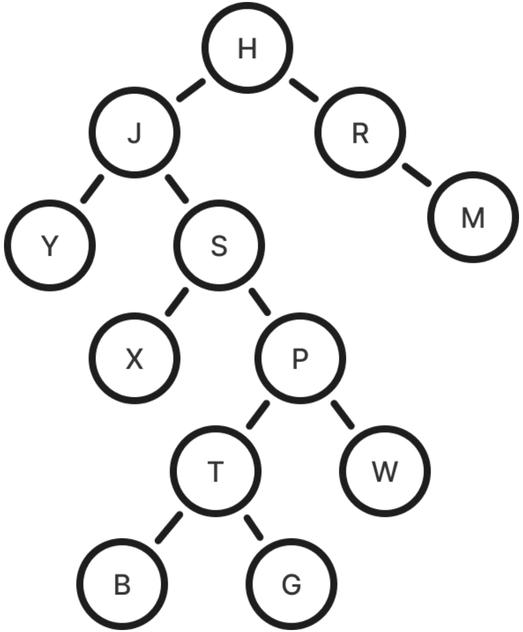
- Pre-order: HBXYPJSWTRMG



# Binary Tree Traversals

Answer from GPT: The pre-order traversal of the tree reads "HJYXSBPTWGRM"

- Generates Incorrect tree
- Given the incorrect tree, give incorrect preorder (given tree is "HJYSXPTBGWRM")



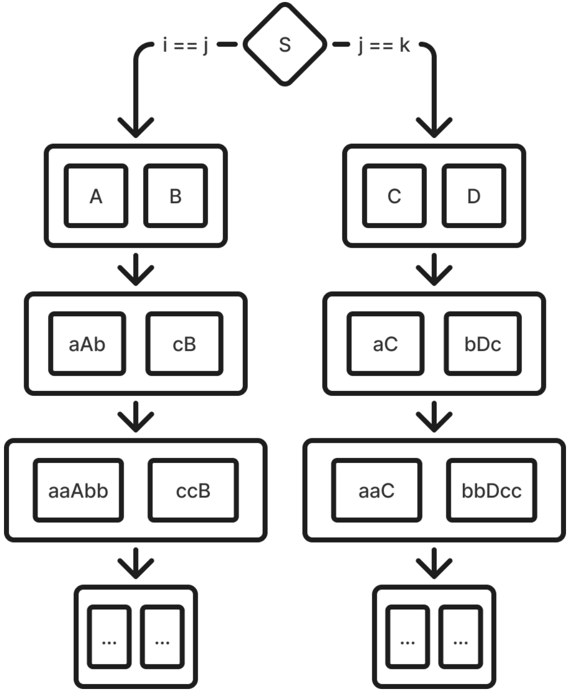
# Context-Free Grammar

Generate the CFG for a language L that:

$$L = \{a^i b^j c^k \mid i == j \text{ or } j == k\}$$

A correct answer can be:

S → AB | CD
A → aAb | ϵ
B → cB | ϵ
C → aC | ϵ
D → bDc | ϵ

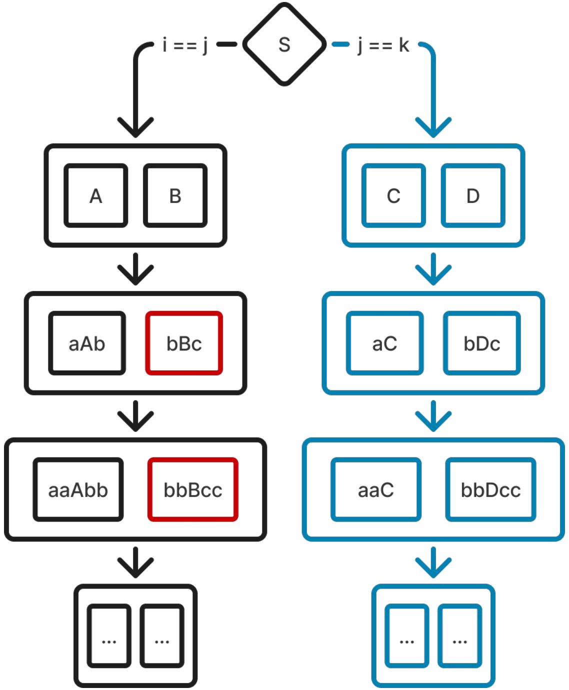


# Context-Free Grammar

Answer from GPT:

$$
\begin{aligned}
S &\rightarrow AB \mid CD \\
A &\rightarrow aAb \mid \epsilon \\
B &\rightarrow bBc \mid \epsilon \\
C &\rightarrow aC \mid \epsilon \\
D &\rightarrow bDc \mid \epsilon
\end{aligned}
$$

- Successfully represent $j == k$
- Wrong in $i == j$, represent $i == j + k$



# Dominating Set is NP-complete

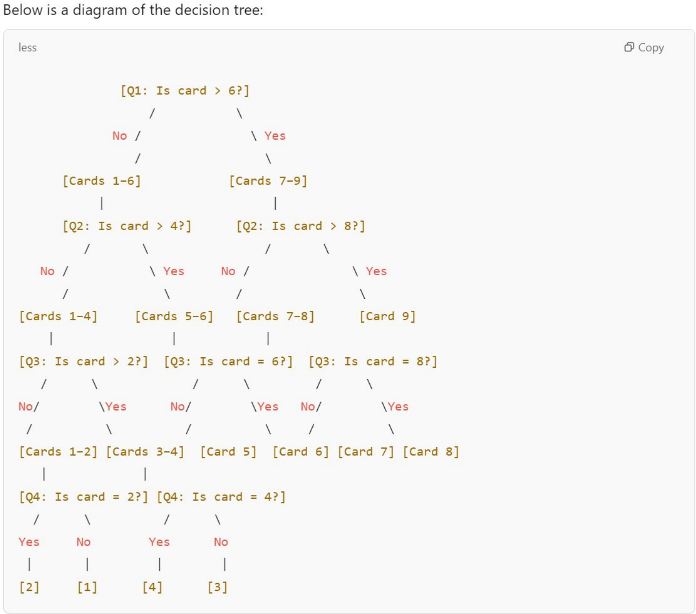
A subset of vertices of a graph are a Dominating Set if every vertex is either included in that set or adjacent to one that is.

- Begin with a proof that Dominating Set is in NP
- Continue with an instance of Vertex Cover
- Ends up with a poor reduction:
    - Adding a single vertex to the graph and then seeking a Dominating Set of size $k$+1 in the resulting (modified) graph
- **Knows the shape, but cannot fill in details**

# Threats to Validity

- Continuous evolution of LLMs
    - A great candidate for a replication study
    - A good undergraduate research
    - Follow-up on new model’s performance
- Variability among graders
- Problems may not fully represent model performance

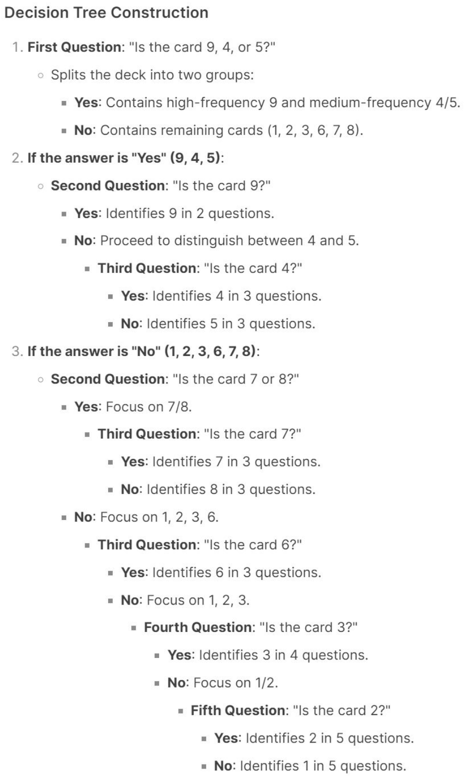
# Guess Card (Update 02.15.2025)



Below is a diagram of the decision tree:

```less
              [Q1: Is card > 6?]
                  /           \
             No /               \ Yes
                /                 \
      [Cards 1-6]            [Cards 7-9]
           |                       |
      [Q2: Is card > 4?]      [Q2: Is card > 8?]
         /       \                /       \
   No /           \ Yes     No /             \ Yes
      /             \         /                \
[Cards 1-4]     [Cards 5-6]   [Cards 7-8]      [Card 9]
    |                |            |
[Q3: Is card > 2?]  [Q3: Is card = 6?]  [Q3: Is card = 8?]
   /      \             /      \         /      \
No/        \Yes      No/        \Yes   No/        \Yes
 /          \          /          \     /          \
[Cards 1-2] [Cards 3-4]  [Card 5]  [Card 6] [Card 7] [Card 8]
   |             |
[Q4: Is card = 2?] [Q4: Is card = 4?]
  /      \          /      \
Yes     No        Yes      No
 |       |         |        |
[2]     [1]      [4]      [3]
```

Left: ChatGPT o3-mini-high



**Decision Tree Construction**

1. **First Question**: "Is the card 9, 4, or 5?"
    - Splits the deck into two groups:
        - **Yes**: Contains high-frequency 9 and medium-frequency 4/5.
        - **No**: Contains remaining cards (1, 2, 3, 6, 7, 8).
2. **If the answer is "Yes" (9, 4, 5)**:
    - **Second Question**: "Is the card 9?"
        - **Yes**: Identifies 9 in 2 questions.
        - **No**: Proceed to distinguish between 4 and 5.
            - **Third Question**: "Is the card 4?"
                - **Yes**: Identifies 4 in 3 questions.
                - **No**: Identifies 5 in 3 questions.
3. **If the answer is "No" (1, 2, 3, 6, 7, 8)**:
    - **Second Question**: "Is the card 7 or 8?"
        - **Yes**: Focus on 7/8.
            - **Third Question**: "Is the card 7?"
                - **Yes**: Identifies 7 in 3 questions.
                - **No**: Identifies 8 in 3 questions.
        - **No**: Focus on 1, 2, 3, 6.
            - **Third Question**: "Is the card 6?"
                - **Yes**: Identifies 6 in 3 questions.
                - **No**: Focus on 1, 2, 3.
                    - **Fourth Question**: "Is the card 3?"
                        - **Yes**: Identifies 3 in 4 questions.
                        - **No**: Focus on 1/2.
                            - **Fifth Question**: "Is the card 2?"
                                - **Yes**: Identifies 2 in 5 questions.
                                - **No**: Identifies 1 in 5 questions.

Right: DeepSeek-R1

# Thank You

Questions?