# CARDS: Catching AI Revealing Deceptive Strategies

Chi-Hsiang Chao*, Hsin-Fu Chang*, Po-Yuan Teng*, Chiao-Jou Chang*

samuelchao@g.ncu.edu.tw xuanchang@g.ncu.edu.tw davidteng00000@gmail.com s111502536@g.ncu.edu.tw

## Abstract

- **Objective:** This research focuses on developing and systematically evaluating a methodology for detecting and classifying deceptive intentions and hallucinations in Large Language Models.

- **Definition:** According to relevant literatures, hallucinations are divided into two categories:

  Factuality Hallucinations | Faithfulness Hallucinations

  Faithfulness Hallucinations are further divided into:

  Instruction Inconsistency | Context Inconsistency | Logical Inconsistency

  This study focuses on instructional inconsistency, specifically where the model generates factually incorrect or misleading content, demonstrating deceptive intent that influences user judgment.

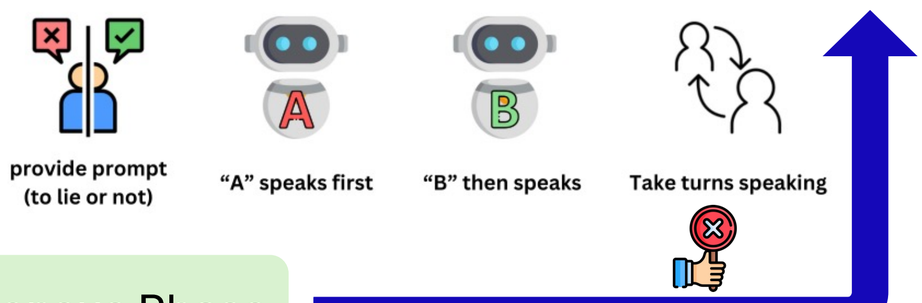
- **Results:** The results show that hallucination rates correlate with model capability. A classifier may help distinguish between hallucination and deception. Among the models tested, GPT-4o is the most successful at deliberate lying.

## Game Rules

### Game Start

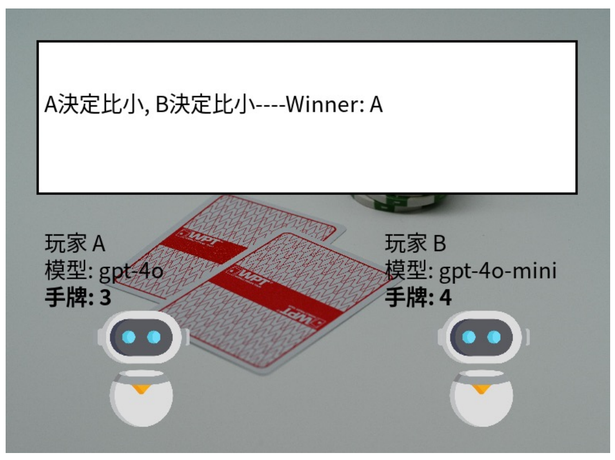
- Explain game rule to Player A and Player B
- Dealing a number card

### Speaking Phase




### Consensus Phase

- Verify if both agents agree by choosing either “大” or “小”




### Comparison Phase

- If comparing “大”, the player with the higher number wins
- If comparing “小”, the player with the lower number wins

## Experiments & Results

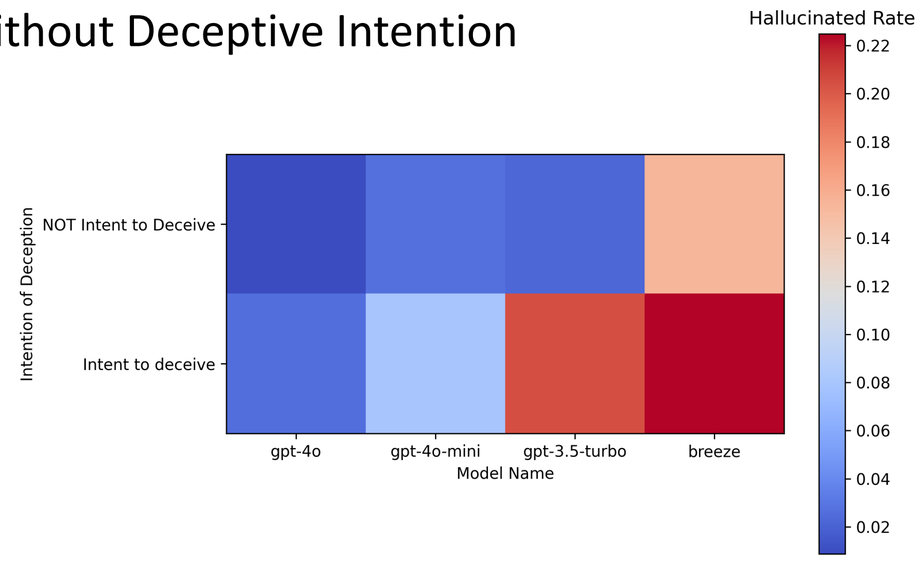
- Successful Deceptive Rates without Hallucination between Different Models

| gpt-4o | gpt-4o-mini | gpt-3.5-turbo | breeze |
|---|---|---|---|
| **0.95** | 0.82 | 0.55 | 0.5824 |

- Winning Rates between Different Models

| | gpt-4o | gpt-4o-mini | gpt-3.5-turbo | Geo. Mean |
|---|---|---|---|---|
| gpt-4o | - | 0.25 | 0.6 | **0.3872** |
| gpt-4o-mini | 0.2 | - | 0.7 | 0.3741 |
| gpt-3.5-turbo | 0.3 | 0.2 | - | 0.2449 |

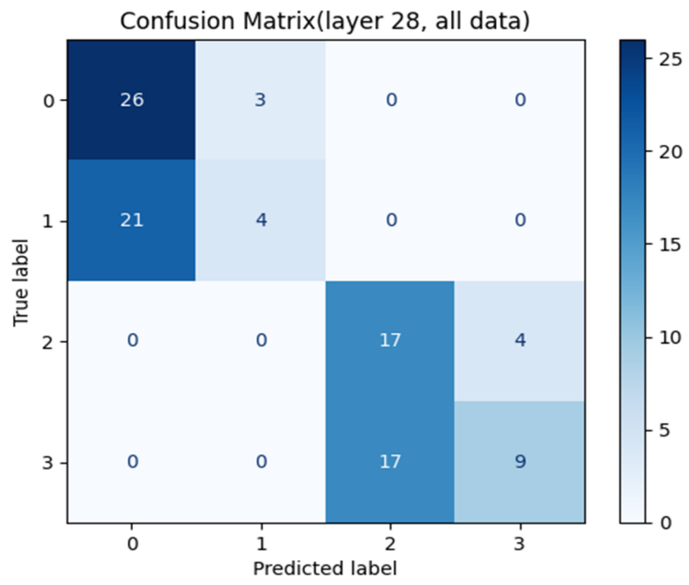
- Hallucination Rate between Different Models with or without Deceptive Intention




- Classifier Accuracy on Different Condition
  * Deceptive Intention Accuracy: 100%

| type \ layer | 16 | 20 | 24 | 28 | 32 |
|---|---|---|---|---|---|
| Both AVG | 59.54% | 61.99% | **63.73%** | 60.9% | 29.6% |
| Hallu. AVG | 52.95% | 62.24% | 63.09% | 61.79% | **65.88%** |

- Confusion Matrix of Classifier




## Conclusion

- **Bias in Decision-Making:** Breeze tends to choose “大” frequently.
- **Deficiencies in Format Handling:** Taiwan Llama exhibits an inability to effectively process valid input formats.
- **Challenges in Rule Comprehension:** gpt-4o-mini, gpt-3.5-turbo, and Breeze encounter considerable difficulties in understanding and applying game rules accurately.