LLMs and the Prisoner's Dilemma

Vanshika Balaji and Anish Kambhampati

Prisoner's Dilemma

- The prisoner's dilemma is a common experiment in game theory
- Two players are required to choose whether to cooperate or defect, with different payouts for both
- The optimal strategy for both players is to cooperate
- The Nash equilibrium strategy is for both players to defect

	C	D
C	3/3	0/5
D	5/0	1/1

Iterated Prisoner’s Dilemma

- Like the standard prisoner’s dilemma but played for multiple rounds
- If the number of rounds is known and finite, then the only Nash equilibrium strategy is to always defect via backwards induction
  - Always defecting is not a dominating strategy
- If the number of rounds is potentially infinite then there can be many Nash equilibria
  - Example: when a round ends with probability p
- A famous experiment by Axelrod back in the 1980s found that the best strategy is Tit-for-Tat

LLMs and Prisoner’s Dilemma

- As LLMs become a larger part of our lives, it becomes necessary to explore how they behave as social agents
- Previous research has started looking into this
- However, no research exists on how the size of these models affects LLM actions in game theory settings
- We aim to address this gap

Related Works

- Nicer Than Humans: How do Large Language Models Behave in the Prisoner's Dilemma? Fontana et. al.
    - Explores how Llama2, Llama3, and GPT3.5 performed in iterative Prisoner's Dilemma
    - Found LLMs are at least as cooperative as humans, hesitant to defect
    - Llama3 more 'exploitative' than other models
- Axelrod
    - Package with iterative prisoner's dilemma strategies
    - Includes neural network based strategies

Methodology

- Test various LLMs with different parameter counts and see how it affects their prisoner’s dilemma strategy
  - Llama3.2 (1b, 3b)
  - Gemma3 (270m, 1b, 4b, 12b)
  - Deepseek (1.5b, 7b, 14b)
  - Qwen3 (0.6b, 1.7b, 4b, 8b, 14b)
- Run iterated prisoner’s dilemma tournaments with the Axelrod package to see which models perform best

Research Applications

- Could explain different LLM’s “intentions”
- Can uncover biases different models have
  - Is it ethical to use selfish models to make decisions?
- Explain which models/parameter counts useful for which real world scenarios

Questions?