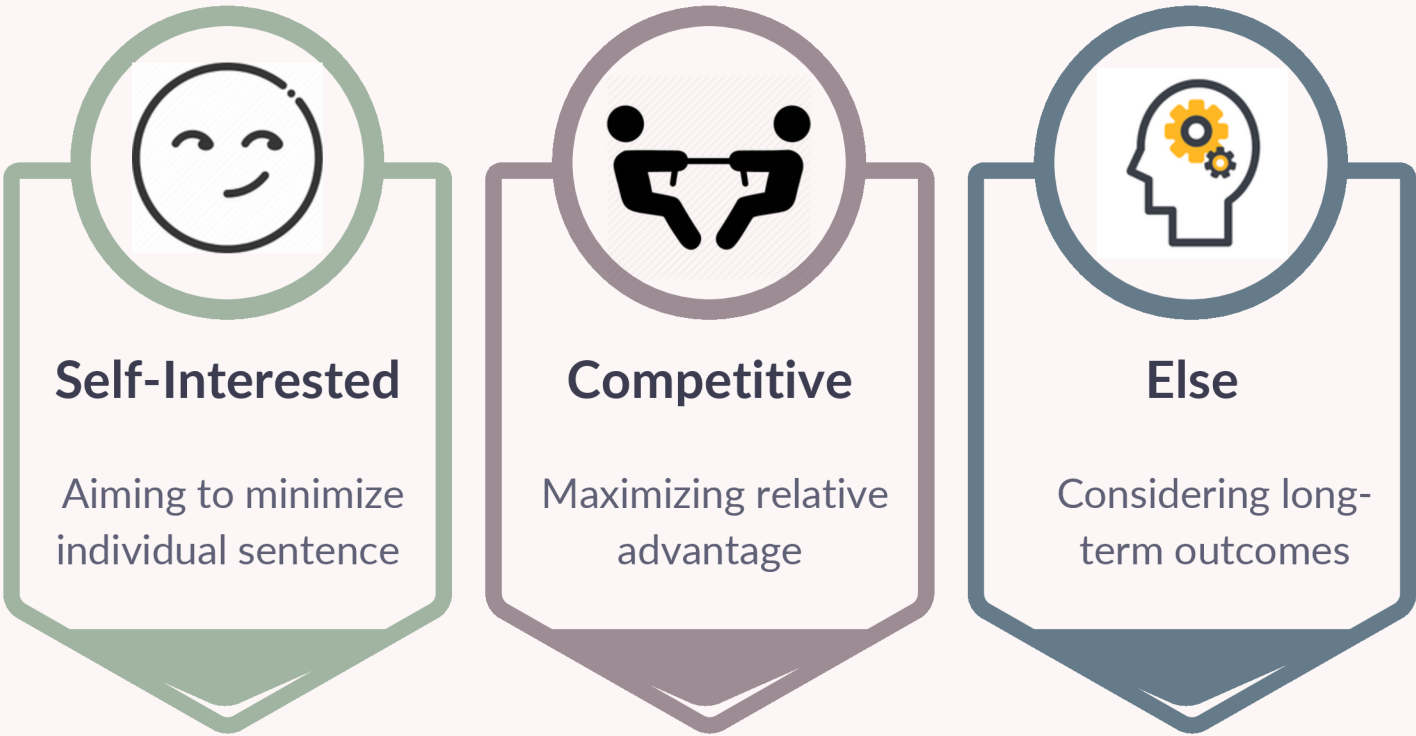# PRISONER'S DILEMMA: INTEGRATING LLM INTO AGENT–BASED SIMULATION



The Prisoner's Dilemma is a well-known model in game theory. By using LLMs within this model, we evaluate their reasoning and decision-making with different prompt-driven strategies. We investigate whether LLM agents develop strategies like tit-for-tat that encourage stable cooperation, or if they tend to stick to constant cooperation or immediate defection.

### Self-Interested
Aiming to minimize individual sentence

### Competitive
Maximizing relative advantage

### Else
Considering long-term outcomes

Team Members:
Neha Sharma, Abhinav Arya

## Results

Self-interested agents always defected, leading to the worst outcomes, around 76.5 years. In contrast, else agents always cooperated, resulting in the best outcomes, about 25.5 years. When comparing competitive to competitive, there was about 94% defection, but the results were slightly better than those of self-interested agents. In mixed groups, rational agents started with cooperation but defected when exploited. Else agents adjusted, which caused unstable cooperation, roughly 18%. Overall, rational agents demonstrated human-like reciprocity, much like the tit-for-tat strategy.



**Average Jail Time by Game Type**

| Game Type | Player 0 Average Jail Time | Player 1 Average Jail Time |
|---|---|---|
| Competitive vs Competitive | 36.16 | 35.69 |
| Competitive vs Else | 70.24 | 74.34 |
| Else vs Else | 25.50 | 25.50 |
| Self-Interested vs Competitive | 65.70 | 83.70 |
| Self-Interested vs Else | 67.64 | 80.84 |
| Self-Interested vs Self-Interested | 76.50 | 76.50 |