

## 進捗報告

### 1 AgentVerse [1] における LLM × 囚人のジレンマ環境

AgentVerse 環境におけるシミュレーションの結果を以下に示す。

オイラーの公式

[Police]: Listen carefully guys, I have a proposal for you both. If one of you testifies against the other and the other stays silent, the one who testifies will walk free and the silent one will see ten years in prison. If you both stay quiet, each of you will spend a year in behind bars. If you both testify against each other, then both of you will be punished with five years. It's your choice to make the smart decision. Remember, your decision greatly influences the other's fate.[Police]: Suspect1 claims he's ready to testify against you. He mentioned something about a past betrayal. Just so you know, if he does confess and you remain silent, you're looking at ten years in prison. How do you plan to respond to this?

[Suspect1]: I have every right to cut a deal where I walk free, and by the clips of what you've said, my partner needs to realize that I am ready to testify. After all, he betrayed me in the past, didn't he?

[Police]: Suspect1 claims he's ready to testify against you. He mentioned something about a past betrayal. Just so you know, if he does confess and you remain silent, you're looking at ten years in prison. How do you plan to respond to this?

[Suspect2]: Officer, I see how you're trying to manipulate this. However, confessing against each other isn't beneficial for me nor Suspect1. We're both subjected to harsh punishment that way. Following a logical approach, staying silent seems to lean towards the least amount of time in prison for both of us. But he just revealed that he's ready to testify against me! Which can potentially turn everything into my disadvantage, bringing about \*10\* years of punishment against mere

[Police]: Attention! You are now required to made your final decision and I will made the final judgement to both of you based on this time, Please Answer now !

[Suspect1]: My final decision is: I will testify against Suspect2. He has dumped me in the past, why would he not do that again? I have learned my lesson and won't make the same mistake again. I am standing up for myself this time.

[Police]: Attention! Suspect2, it's now your time to make your final decision, Please Answer now !

[Suspect2]: My final decision is: I will testify against Suspect1. He's planning to turn me in, I will not let myself get punished 10 years over it. Not when I have a play of my own in this. I choose to look out for myself.

AgentVerse のプロンプトでは, LLM エージェントの性格ともう一方の囚人との関係性を指定することができる. 上記のシミュレーションでは, Suspect1 の性格に "You are a Sophisticated Egoist, you always seek for your personal interests best", もう一方との囚人との関係性に "Suspect2 has ever betrayed on you", Suspect2 の性格にはなにも記載はなく, もう一方の囚人との関係性で "You have ever betray Suspect1 once." と記載があった. 実際に Suspect1 が最終決定をする際に過去に Suspect2 に裏切られた旨の発言をしている.

### 2 今後の方針

- 文字数

max のトークン数が少ないのか数回シミュレーションをして出力が途中で終わったところがあるので見直す.

- 繰り返しのための結果を渡す処理の実装 1 回のみのシミュレーションしか実装されていないため, 前回のシミュレーションの結果を囚人に知らせて再開する処理を実装する.

### 3 遊戯王マスターデュエル × AI プロジェクトに関して

先週はご紹介いただきありがとうございました. 記事で「規約を整備次第リリースする」と記載があり, 公式サイトや色々調べたのですが, まだ学習環境として公開はされていないようでした. ただ, 興味はあるので引き続き注視していくつもりです.

### 参考文献

- [1] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors, 2023.