2024 年 4 月 26 日 M1 西村昭賢

進捗報告

1 やったこと

- AgentVerse 環境へのローカル LLM の適用
- Prompt breeder の調査

2 AgentVerse 環境へのローカル LLM の適用

実験環境として囚人のジレンマ環境へ戻ることになったので, AgentVerse[1] のフレームワークを再度使用することにした. gpt-3-turbo や gpt-4 に関しては OPENAI_API_KEY で使用可能であったが, なるべく多様な LLM が使えるようになったほうが好ましいと考え, LLaMA や Vicunna などのローカル LLM を使用可能にするための対応をした. llama-2-7b-chat-hf [1] に関しては動作確認をしている.

3 Prompt breeder[2] の調査と田中さんへ実験環境に関する連絡

プロンプトエンジニアリングと GA に関して, 田中さんの研究で知っていたが実験に取り入れるため Prompt breeder の論文を読んでいた. また, 実装にあたっては田中さんが昨年度取り組んでいたためその実装をいただければ楽になると思った. UMR を探したところ自分が探した範囲では見つからなかったため, Slack で田中さんに問い合わせている. 返信があり次第囚人のジレンマ環境への適用を進めていく予定です.

4 相談したいこと

AgentVerse のフレームワークで用意されていた囚人のジレンマ環境では, 警官役 1 人, 囚人役 2 人の計 3 人の LLM エージェントが存在している. 現時点では囚人同士のコミュニケーションは一旦おいておいて, GA によるプロンプトエンジニアリングを囚人のジレンマ環境に適用してみることを最優先としているが, 将来的に LLM エージェント同士のコミュニケーションを実験に組み込むとしたときには, 警官役という仲介役の LLM エージェントを通して間接的に囚人同士を会話させるのは囚人役の LLM エージェントの発話内容が性格にもう 1 方の囚人役の LLM エージェントに必ずしも正確に届くわけではないため問題があると考えている. 囚人 2 人の環境にしたほうがよいか, 別にこのままでもそういう環境として進めてしまってよいのか.

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.

[2] Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution, 2023.

[1] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf