JSAI 2026 ／ 人狼知能

yharada: 10-shot aiwolf agent

LLM に台本を読ませる人狼エージェント

①プロンプト強制の限界 → 台本10-shotによる暗黙知学習

② Claudeのキャラクター模倣力をGPTへ

③ 役職特化台本による「勝てるエージェント」へ

静岡大学情報学部行動情報学科狩野研究室 学部4年生 原田悠矢

プロンプト強制の限界 → 台本10-shotによる学習へ

❶ 従来アプローチの問題

手前の発言を繰り返す

LLMが「直前の発言を模倣するのが安全」と判断し、
同じような発言を繰り返してしまう。

プロンプトに従いすぎる

指示に忠実なあまり、状況に応じた柔軟な
判断・発話ができなくなる。

話題が遷移しない

プロンプトで振る舞いを指定すると逆に硬直化し、
議論の流れが変わらなくなる。

⇒

❷ 提案手法：台本10-shotで学ばせる

プロンプトで「こうしろ」と強制するのではなく、
自然な議論の運ばれ方・話題の遷移・発言の多様性を実際の対局台本として見せ、LLMに自ら学ばせる。

話題の遷移

Day内の議論展開・CO・投票宣言へ自然に流れるパターンを体得

発言の多様性

直前の発言を模倣せず、異なる切り口・立場で発言できる

Claudeの強みをGPTに学ばせる：品質とコストの両立

1体のLLMが生成する台本はマルチエージェントによるログと比べて品質が格段に高い。LLM傾向分析に基づき、台本生成と実際の発話で異なるLLMを使い分けた。

比較観点	Claude	GPT	本プロジェクトの判断
人間らしさ・キャラクター性	◎ 高い	△ やや低い	台本生成はClaude
状況分析・ルール遵守	△ やや低い	◎ 高い	実運用はGPT
APIコスト	高い	低い	実運用はGPTで削減

戦略のポイント

Claudeで生成した台本（キャラクター性・自然な議論展開）をGPTに読ませることで、GPTの高い分析能力を保ちながらClaudeの強みを実装コストなく注入する。

※LLMごとの傾向の詳細はSAI2026一般セッション「LLM傾向性プロファイリング：マルチエージェント対話のための6軸評価フレームワーク」発表を参照

貢献と Future Work

① 台本10-shot学習

プロンプト強制に代わり
対局台本から自然な議論を習得

② Claude→GPT品質転写

Claudeのキャラクター性・
議論多様性をGPTに注入

③ 主観評価の向上を達成

発話の自然さ・キャラクター性・
話題遷移の改善を確認

Future Work

現状の課題

全役職共通の台本

現在は村全体で共通の台本を読ませており、役職ごとの最適な立ち回り・成功パターンを網羅できていない。

主観評価止まり

発話の自然さ・キャラクター性の向上は確認できたが、勝率など客観的な強さの指標での改善は未達成。

今後の目標

役職別・村サイズ別の台本整備

占い師・人狼・村人など役職ごとの成功パターン台本を5人村・9人村それぞれ10本ずつ作成し、役割に即した立ち回りを学習させる。

勝率・戦略的行動への貢献

チームの勝利に直結する推論・情報共有・連携行動を達成し、自然な議論の実現から一歩進んだ「強い」エージェントを目指す。

自然な議論の実現から「勝てるエージェント」へ – 役職特化台本による次段階の展開