Игра "Мафия" с использованием LLMs: исследование социальных взаимодействий и стратегий с помощью графовых методов

Аннотация

Данный проект направлен на исследование возможностей больших языковых моделей (LLMs) в моделировании социальных взаимодействий в игре "Мафия" или ее упрощенных аналогах. Мы разработаем систему, позволяющую LLM-агентам участвовать в симуляции игры, принимая различные роли. Особое внимание будет уделено использованию графовых методов для представления и анализа взаимодействий между игроками, а также для улучшения памяти и контекстного понимания моделей. Исследование направлено на сравнение эффективности различных подходов к инжектированию графовой информации и их влияния на способности LLMs к стратегическому мышлению, обману, убеждению и сотрудничеству в условиях неполной информации.

Введение

Социальные дедуктивные игры, такие как "Мафия" и ее упрощенные версии ("One Night Ultimate Werewolf", "The Resistance"), представляют собой сложные социальные системы, требующие от участников навыков убеждения, стратегического мышления и обмана. Эти игры являются идеальной средой для исследования социальных взаимодействий и когнитивных процессов. Современные LLMs демонстрируют продвинутые лингвистические способности, но их эффективность в сложных социальных взаимодействиях, требующих

поддержания долговременного контекста и стратегического мышления, остается недостаточно изученной.

Наш проект исследует, как графовые методы могут улучшить способности LLMs в контексте социальных дедуктивных игр, особенно в аспектах поддержания контекста, стратегического мышления и социального взаимодействия. Мы разработаем упрощенную версию игры, где LLM-агенты будут взаимодействовать между собой, используя различные подходы к инжектированию графовой информации для преодоления ограничений стандартного текстового контекста.

Основная часть

План реализации

1. Разработка упрощенной версии социальной дедуктивной игры:
    - Определение основных правил и механик
    - Формализация ролей и целей для каждого типа игрока
    - Создание структуры для обмена информацией между агентами
2. Интеграция графовых методов:
    - Построение графов взаимодействий между игроками
    - Разработка методов инжектирования графовой информации в контекст LLM
    - Тестирование различных типов графов и их комбинаций
3. Экспериментальная часть:
    - Базовый эксперимент (Baseline): все модели одинаковые, только текстовый контекст
    - Базовый эксперимент с графами (Baseline+): добавление графовой информации в контекст

Расширенный эксперимент (Baseline++): сравнение базовых моделей с графом против продвинутых моделей без графа

4. Анализ результатов:

   Статистический анализ выигрышей по ролям

   Качественный анализ стратегий и взаимодействий

   Оценка эффективности различных типов графов и методов их внедрения

Обзор литературы

Существующие исследования демонстрируют разные аспекты применения LLMs в социальных взаимодействиях и использования графовых методов:

1. LLMs в социальных играх и ролевых взаимодействиях:

   Yoo & Kim (2024) в работе "Finding deceivers in social context with large language models: the case of the Mafia game" исследовали способности LLMs определять обманщиков в контексте игры "Мафия", показав, что модели могут участвовать в сложных социальных взаимодействиях, но сталкиваются с трудностями при долгосрочном планировании.

   Shanahan et al. (2023) в исследовании "Role play with large language models" продемонстрировали, что LLMs способны эффективно принимать роли и следовать им в контексте взаимодействия, что критически важно для игр с явно определенными ролями.

2. Проблемы контекстного понимания и памяти LLMs:

   Li et al. (2024) в работе "Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach" показали, что даже модели с расширенным контекстным окном

сталкиваются с проблемой "забывания" информации и потери концентрации при работе с длинными диалогами.

Khalusova и Vergadia в аналитических статьях о RAG и длинном контексте отмечают, что модели с длинным контекстным окном всё равно демонстрируют снижение качества при работе с большими объемами информации.

3. Графовые методы для улучшения контекстного понимания:

   Wu & Tsioutsiouliklis (2023) в исследовании "Thinking with Knowledge Graphs: Enhancing LLM Reasoning Through Structured Data" показали, что представление знаний в виде графов значительно улучшает способность моделей к рассуждению благодаря сжатому представлению информации.

   Исследование "Are Large Language Models In-Context Graph Learners?" (2024) демонстрирует способность LLMs эффективно использовать графовую информацию для решения задач, требующих понимания взаимосвязей.

4. Динамическое обновление контекста и поддержание памяти:

   Kaub в работе "How I Built an LLM-Based Game from Scratch" представил подход к динамическому обновлению контекста с использованием причинно-следственных графов, что позволяет моделям лучше запоминать ключевую информацию и отношения.

   Alfonso в "RAG and Long-Context Windows: Why You need Both" предлагает гибридный подход, сочетающий преимущества обоих методов для оптимального баланса между доступом к информации и эффективностью.

Наш проект направлен на объединение этих направлений, предлагая использовать графовые методы для представления и передачи контекста в игре, что потенциально может преодолеть ограничения как стандартных подходов с текстовым контекстом, так и простого RAG.

Предполагаемые методы анализа социальных сетей

В проекте будут использованы следующие типы графов и методы их анализа:

1. Коммуникационный граф - динамический граф, отражающий обмен репликами между игроками, их тон и содержание. Обновляется после каждого раунда коммуникации.
2. Граф профиля игрока в сессии - динамический граф, отражающий накопленные знания о поведении игрока в текущей игре, включая подозрения, доверие, противоречия в высказываниях. Обновляется по мере получения новой информации.
3. Исторический граф профиля игрока - статический граф, генерируемый в начале игры на основании предыдущих игр, отражающий типичные стратегии, склонность к определенным действиям и решениям.
4. Графы взаимодействий - построение и анализ графов, отражающих коммуникацию между игроками, уровень доверия и подозрений.

Для анализа графов будут использоваться:

- Метрики центральности для определения ключевых игроков и их влияния
- Анализ кластеров для выявления формирующихся коалиций
- Семантический анализ взаимодействий, интегрированный с графовым представлением

- Формальный чекер противоречий в высказываниях на основе графа знаний

Ожидаемые результаты

В результате проекта мы ожидаем получить:

1. Функционирующую систему для моделирования социальной дедуктивной игры с участием LLM-агентов, использующих графовые методы для улучшения своих стратегий.
2. Статистические данные о влиянии различных типов графов и методов их инжектирования на эффективность моделей в игре, в частности:
    Сравнение эффективности разных типов графов (коммуникационный, профильный, исторический)
    Анализ влияния различных способов инжектирования графовой информации
    Оценка эффективности базовых моделей с графовой поддержкой против продвинутых моделей без таковой
3. Качественный анализ стратегий, используемых LLM-агентами в различных ролях, и их способности к обману, убеждению и сотрудничеству.
4. Выводы о потенциале использования графовых методов для преодоления ограничений контекстного окна и улучшения долговременной памяти LLMs в сложных социальных взаимодействиях.

Список литературы

1. Yoo, B., & Kim, K. J. (2024). Finding deceivers in social context with large language models and how to find them: the case of the Mafia game. Nature Scientific Reports, 14, 3697.

2. Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. Nature, 621, 860-868.
3. Li, Z., Li, C., Zhang, M., Mei, Q., & Bendersky, M. (2024). Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach. arXiv preprint arXiv:2407.16833.
4. Wu, X., & Tsioutsiouliklis, K. (2023). Thinking with Knowledge Graphs: Enhancing LLM Reasoning Through Structured Data. arXiv preprint arXiv:2412.10654.
5. Anonymous. (2024). Are Large Language Models In-Context Graph Learners? arXiv preprint arXiv:2502.13562.
6. Kaub, J. (2023). How I Built an LLM-Based Game from Scratch: Game concepts and Causal Graphs for LLMs. Medium, Data Science.
7. Alfonso, A. (2023). RAG and Long-Context Windows: Why You need Both. Medium, Google Cloud.
8. Vergadia, P. (2023). RAG vs Large Context Window LLMs: When to use which one? The Cloud Girl.
9. Khalusova, M. (2023). RAG vs. Long-Context Models. Do we still need RAG?
10. Brown, N., & Sandholm, T. (2019). Superhuman AI for multiplayer poker. Science, 365(6456), 885-890.