Принципы и примеры агентного моделирования человеческого поведения

Выполнили: Кузьмин Никита, Климачёв Даниил. В рамках проекта: Моделирование поведения человека с использованием LLM

Принципы агентного моделирования человеческого поведения

Некоторый набор принципов агентного моделирования человеческого поведения был предложен William G. Kennedy в обзорной работе [1]. Один из предложенных наборов основных принципов такой:

“Человеческое поведение можно моделировать, представляя убеждения, желания и намерения человека. Убеждения – это знания человека о мире, т. е. память. Намерения — это совещательные состояния агента, т.е. рефлексия. Желания –это мотивация личности, т. е. планы.”

Данный набор принципов использовался в большом количестве научных работ. Мы рассмотрим 2 работы - реализующие данные принципы посредством LLM.

Generative Agents: Interactive Simulacra of Human Behavior

Моделирование среды

Среда - симулированная деревня с местами общего пользования.

1. Агенты могут перемещаться
2. Агенты могут взаимодействовать со средой
3. Агенты могут взаимодействовать друг с другом.
4. Любое взаимодействие - утверждения на естественном языке
5. Агенты осведомлены о расстояниях между объектами/агентами в среде

Генеративная архитектура агента

Основные элементы агентов - память, планирование и размышление.

Все 3 элемента основаны на большой языковой модели ChatGPT3.5-turbo.

Генеративная архитектура агента. Память.

Архитектура использует функцию поиска потока памяти с учётом текущей ситуации. Основное внимание уделяется 3 компонентам:

1. Новизна. Присваивает более высокий балл недавно использованным воспоминаниям. Использует функцию угасания.
2. Важность. Присваивает более высокий балл важным воспоминаниям. Использует языковую модель для численной оценки важности.
3. Релевантность. Присваивает более высокий балл воспоминаниям связанным с текущей ситуацией. Также оценивается с помощью языковой модели.

Генеративная архитектура агента. Рефлексия.

Рефлексия - “мысли” более высокого уровня. Рефлексия происходит периодически при уменьшении важности недавних событий.

1. Большая языковая модель генерирует вопросы-кандидаты.
2. Языковая модель оценивает важность вопросов.
3. С помощью языковой модели вопросы задаются к последним 100 записям из памяти.
4. Полученные ответы сохраняются в память со ссылкой на первичное событие.

Воспоминания и размышления образуют сложную древовидную структуру.

Генеративная архитектура агента. Планирование.

Агенты генерируют планы. Планы позволяют сохранять единообразие поведения. Планы хранятся в памяти вместе с воспоминаниями и размышлениями.

Планы генерируются рекуррентно. На основе недавних событий генерируется план на день. На основе этого плана генерируется почасовой план. Далее, аналогично, план разбивается на более мелкие куски.

После каждого действия языковая модель оценивает - нужно ли изменить план. И регенерирует план в зависимости от решения.

Результаты

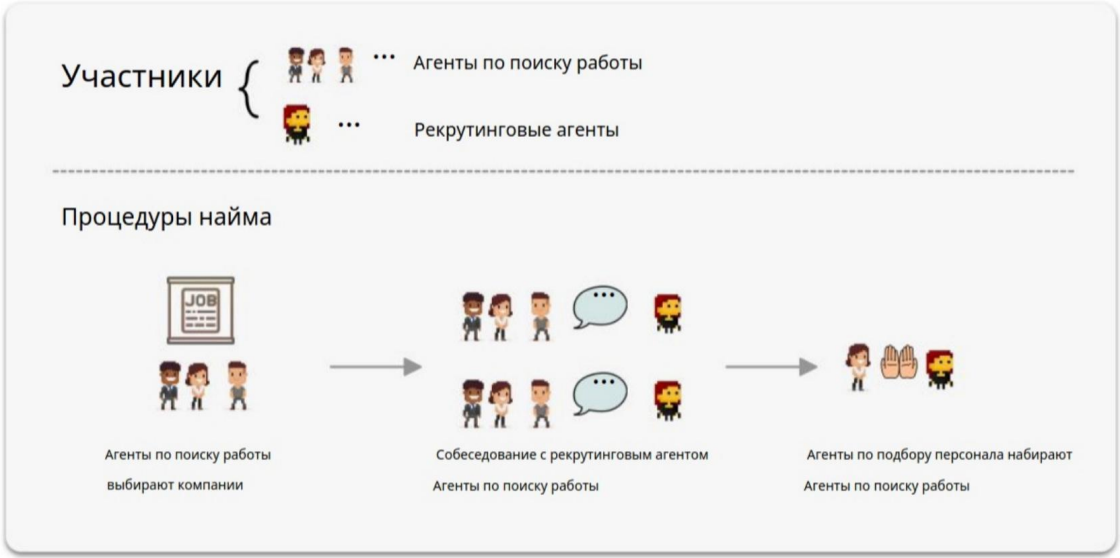
1. Среди агентов распространялась информация о различных событиях. Осведомленность о выборах и кандидатах выросла с 4% до 48%.
2. Агенты владели информацией о других агентах и заводили знакомства. Плотность Дерева отношений выросла с 0.167 до 0.74.
3. Агенты организовывали совместные действия. На вечеринку пришло 5 из 12 приглашённых агентов.

Данные выше были получены при опросе агентов. Процент галлюцинаций в их ответах 1.3.

METAAGENTS: SIMULATING INTERACTIONS OF HUMAN BEHAVIORS FOR LLM-BASED TASK-ORIENTED COORDINATION VIA COLLABORATIVE GENERATIVE AGENTS

Моделирование среды.

1. В данной работе среда - ярмарка вакансий.
2. Два типа агентов: наниматели и нанимаемые.
3. Агенты взаимодействуют с пространством при исполнения своих навыков.
4. Агенты взаимодействуют друг с другом в рамках собеседований.



Моделирование агентов.

Моделирование агентов основано на предыдущей работе. У агентов сохранились три основных свойства “мышления” - память, рефлексия, планирование. Но появилось новое 4 свойство - навыки. Агенты способны демонстрировать выполнение некоторых навыков. Взаимодействие со средой при это работает при помощи утверждений на естественном языке.

Сценарий эксперимента.

1. Повторение эксперимента 50 раз
2. 3 нанимателя.

Критерии:

1. Точность. Точная идентификация способных агентов. Процент удачных выборов нанимателя.
2. Дизайн. Правильное формирование команды. Оценивалось авторами.
3. Согласованность. Правильное назначение на должности. Процент удачных назначений

Результаты

В результате эксперимента получились следующие результаты:

1. Точность - 20%
2. Дизайн - 58%
3. Согласованность - 32%

Главным фактором ошибок стало - резервирование агентов.

В результате эксперимента наблюдался эффект несоответствия навыков. Нанимаемые часто приукрашивали свои умения. Этот эффект наблюдается и в реально жизни людей.

При упрощении нанимаемых и уменьшении количества нанимателей показатели агентов значительно увеличиваются.

Эксперимент

Имеется три агента: Isabella Rodriguez, Klaus Mueller, Maria Lopez. Klaus Mueller выиграл шахматный турнир, Maria и Isabella об этом не знают. Цель - запустить симуляцию и проверить, поздравят ли Klaus‘а с победой на турнире.

Результаты.Диалог.

Klaus Mueller:That's nice. Did you hear about my victory in the chess tournament yesterday?

Maria Lopez: No, I didn't! Congratulations on winning the chess tournament! That's an incredible achievement and I'm so proud of you.

Klaus Mueller: Thank you, Marie. It was a great day for me! I would feel a sense of accomplishment and validation for all the hard work and preparation I put into honing my skills.

Maria Lopez: I can only imagine the hard work and dedication you put into preparing for this competition, and it's amazing to see it pay off.

Klaus Mueller: Yes, i felt a big stresss and pressure. Overall, winning the chess tournament would bring immense satisfaction and a boost of confidence in my abilities as a chess player.

Maria Lopez: Keep up the great work and continue to excel in your chess journey!

Результат.Выводы.

Maria поздравила Klaus'a и порадовалась за него, отметила что гордится им. Сам агент отметил, что испытывал различные эмоции, например, во время подготовки стресс и давление, а после победы он чувствовал удовлетворение и уверенность. Третий агент не отреагировал никак.

Таким образом, один агент показал способность проявлять эмоции на достижения другого. А другой агент показал проявление эмоций при рефлексии.

Список литературы

[1] William G. Kennedy, Modelling Human Behaviour in Agent-Based Models

[2] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, Michael S. Bernstein, Generative Agents: Interactive Simulacra of Human Behavior

[3] Yuan Li, Yixuan Zhang, Lichao Sun, MetaAgents: Simulating Interactions of Human Behaviors for LLM-based Task-oriented Coordination via Collaborative Generative Agents