Улучшение разговорных агентов с теорией разума: согласование убеждений, желаний и намерений для взаимодействия, похожего на человеческое

Дата: 2025-03-04 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2502.14171

Рейтинг: 70
Адаптивность: 75

Ключевые выводы:

Исследование направлено на улучшение взаимодействия между LLM-системами и людьми путем внедрения теории разума (Theory of Mind, ToM). Основная цель - изучить, насколько языковые модели могут улавливать и использовать информацию о ментальных состояниях (убеждениях, желаниях и намерениях) для более человекоподобного взаимодействия. Результаты показали, что внедрение ToM-информации в процесс генерации ответов значительно улучшает качество взаимодействия, достигая показателей выигрыша 67% и 63% для моделей Llama 3 размером 3B и 8B соответственно.

Объяснение метода:

Исследование предлагает ценную BDI-модель (убеждения, желания, намерения) для улучшения диалога с LLM. Хотя технические методы требуют специальных навыков, принципы могут быть адаптированы для структурирования промптов. Наглядные примеры демонстрируют преимущества учета ToM. Пользователи могут применять концепцию для более эффективного взаимодействия с LLM в переговорах и обсуждениях.

Ключевые аспекты исследования: 1. **Теория разума (ToM) для LLM**: Исследование изучает, насколько языковые модели могут понимать и отслеживать убеждения, желания и намерения участников диалога (BDI-модель) для более человекоподобного взаимодействия.

Извлечение ToM из внутренних репрезентаций: Авторы используют метод LatentQA для извлечения информации о ToM из активаций нейронных сетей и проверяют её согласованность.

Управление выводом через ToM-компоненты: Исследователи демонстрируют возможность манипулировать внутренними представлениями ToM для получения

более согласованных с контекстом ответов.

Экспериментальная валидация: Проведены эксперименты на различных наборах данных (диалоги о кемпинге, переговоры о товарах), показывающие 67% и 63% выигрыша для моделей Llama3 3B и 8B соответственно при использовании ToM-информации.

Практическая применимость: Показано, что средние слои LLM содержат наиболее полезную информацию о ToM, которая может быть использована для улучшения качества диалогов.

Дополнение:

Применимость методов в стандартном чате

Хотя исследование использует сложные технические методы (LatentQA) для извлечения и манипулирования внутренними представлениями ToM, основные концепции могут быть применены в стандартном чате без необходимости в дообучении или API.

Ключевые адаптируемые концепции:

BDI-модель для структурирования промптов Пользователи могут явно указывать в промптах: Убеждения (beliefs): что каждый участник диалога знает или думает Желания (desires): что каждый участник хочет получить Намерения (intentions): какие действия участники планируют предпринять

Последовательное отслеживание ToM

При длительных диалогах пользователи могут периодически обновлять информацию о ментальных состояниях участников Например: "Учитывая, что пользователь выразил предпочтение X, а я выразил потребность в Y..."

Явное указание приоритетов

В сценариях переговоров пользователи могут явно указывать приоритеты: "Для меня высокий приоритет имеет X, средний приоритет Y, низкий приоритет Z"

Эмпатическое взаимодействие

Использование намерения "Show empathy" путем явного указания на необходимость учета чувств и потребностей собеседника **Ожидаемые результаты:** - Более контекстно-зависимые и персонализированные ответы - Повышение эффективности в сценариях переговоров и обсуждений - Более естественное и человекоподобное взаимодействие - Улучшенное отслеживание потребностей пользователя в длительных диалогах

Таким образом, хотя исследователи использовали сложные технические методы

для удобства экспериментов, основные принципы ToM могут быть эффективно применены в стандартном чате путем явного структурирования промптов с учетом BDI-модели.

Анализ практической применимости: 1. **Извлечение ToM из внутренних репрезентаций** - Прямая применимость: Низкая для обычных пользователей, так как требует доступа к внутренним слоям модели и специальных технических навыков. - Концептуальная ценность: Высокая, так как подтверждает, что LLM действительно моделируют ментальные состояния собеседников, что помогает понять, как формулировать запросы. - Потенциал для адаптации: Средний, знание о том, что средние слои лучше всего представляют ToM, может быть использовано разработчиками для улучшения диалоговых систем.

Управление выводом через ToM-компоненты Прямая применимость: Низкая для обычных пользователей из-за сложности метода. Концептуальная ценность: Высокая, поскольку демонстрирует, что учет убеждений, желаний и намерений улучшает качество диалога. Потенциал для адаптации: Высокий, пользователи могут адаптировать подход, явно указывая в промптах информацию о желаниях и убеждениях участников диалога.

Экспериментальная валидация на разных наборах данных

Прямая применимость: Средняя, результаты показывают практическую пользу в конкретных сценариях переговоров. Концептуальная ценность: Высокая, демонстрирует важность учета контекста и ментальных состояний для успешного диалога. Потенциал для адаптации: Высокий, подход можно применять в различных сценариях переговоров и обсуждений с LLM.

Структура BDI (Beliefs-Desires-Intentions)

Прямая применимость: Средняя, пользователи могут структурировать запросы с учетом этих компонентов. Концептуальная ценность: Высокая, предлагает четкую модель для понимания человекоподобного взаимодействия. Потенциал для адаптации: Высокий, модель может быть использована для структурирования промптов и анализа ответов LLM.

Примеры улучшенных диалогов

Прямая применимость: Высокая, показывает конкретные примеры, как учет ToM улучшает диалог. Концептуальная ценность: Высокая, наглядно демонстрирует разницу между обычным и ToM-информированным диалогом. Потенциал для адаптации: Высокий, примеры могут быть использованы как шаблоны для собственных промптов.

Prompt:

Использование теории разума в промптах для GPT ## Ключевое понимание из исследования

Исследование показывает, что языковые модели могут лучше взаимодействовать с людьми, если в них внедрена **теория разума (ToM)** — способность понимать и отслеживать ментальные состояния собеседника через: - **Убеждения** (beliefs) — что человек считает истинным - **Желания** (desires) — чего человек хочет достичь - **Намерения** (intentions) — какие планы есть у человека

Пример эффективного промпта с использованием ToM

[======] # Инструкция для GPT с использованием теории разума

Ты помощник в переговорах о цене товара. Во время диалога тебе нужно:

Отслеживать убеждения клиента: Что клиент думает о реальной стоимости товара Какие параметры товара он считает важными

Определять желания клиента:

Какую максимальную цену он готов заплатить Какие дополнительные ценности он ищет помимо цены

Понимать намерения клиента:

Хочет ли он действительно купить или просто исследует рынок Планирует ли он использовать информацию для торга в другом месте После каждого сообщения клиента, перед формированием ответа, проанализируй эти три компонента и адаптируй свой ответ, чтобы он был согласован с ментальным состоянием клиента.

При ответе не указывай явно, что ты отслеживаешь эти компоненты, просто используй эту информацию для создания более эффективного и эмпатичного ответа. [======]

Почему это работает

Использование средних слоев модели — исследование показало, что информация ToM лучше представлена в средних слоях модели, и этот промпт помогает активировать эти представления

Структурирование по BDI-модели (Belief-Desire-Intention) — явно указывая модели отслеживать все три компонента, мы задействуем более глубокое понимание контекста

Динамическое отслеживание — промпт направляет модель на постоянное обновление своего понимания ментального состояния собеседника, что согласуется с выводами исследования о необходимости адаптации к изменяющимся представлениям

Неявное применение — промпт указывает не демонстрировать механизм работы, а просто использовать его, что делает взаимодействие более естественным

Такой подход к созданию промптов может повысить эффективность взаимодействия с GPT на 60-67%, согласно результатам исследования.