Единая оценка AI-репетиторов: таксономия оценки для оценки педагогических способностей репетиторов на базе LLM.

Дата: 2025-02-08 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2412.09416

Рейтинг: 72
Адаптивность: 85

Ключевые выводы:

Исследование направлено на создание единой таксономии для оценки педагогических способностей LLM-моделей, выступающих в роли AI-репетиторов. Основные результаты показывают, что современные LLM-модели, хотя и эффективны как системы вопросов-ответов, часто не обладают достаточными педагогическими навыками для качественного обучения. Исследователи разработали таксономию из 8 измерений для оценки педагогических способностей AI-репетиторов и создали бенчмарк MRBench для сравнения различных моделей.

Объяснение метода:

Исследование представляет практическую таксономию из 8 измерений для оценки педагогических способностей LLM и сравнивает эффективность современных моделей как тьюторов. Основные принципы (идентификация ошибок, предоставление подсказок вместо ответов, поддерживающий тон) могут быть непосредственно включены в промпты пользователей для улучшения образовательных взаимодействий с LLM. Требуется некоторая адаптация для различных предметных областей.

Ключевые аспекты исследования: 1. **Разработка таксономии для оценки педагогических способностей LLM-тьюторов** - исследователи создали единую таксономию из 8 измерений для оценки качества ответов ИИ-тьюторов при исправлении ошибок учащихся в математике.

Создание бенчмарка MRBench - авторы собрали набор из 192 диалогов с ошибками учащихся и 1,596 ответов от 7 современных LLM и человеческих тьюторов, с аннотациями по всем 8 измерениям таксономии.

Всесторонняя оценка LLM как тьюторов - проведено сравнение способностей различных моделей (GPT-4, Gemini, Llama и др.) выполнять функции педагогического тьютора, с выявлением их сильных и слабых сторон.

Методология оценки педагогических способностей - разработана методика, основанная на принципах обучающих наук, для измерения качества ответов моделей при работе с ошибками учеников.

Проверка надежности LLM как оценщиков - исследовано, насколько такие модели как Prometheus2 и Llama-3.1-8B могут сами выступать в роли оценщиков педагогических способностей других моделей.

Дополнение: Действительно ли для работы методов этого исследование требуется дообучение или API? Или методы и подходы можно применить в стандартном чате, а ученые лишь для своего удобства использовали расширенные техники?

Методы и подходы, описанные в исследовании, в основном можно применить в стандартном чате без необходимости в дообучении моделей или доступе к API. Исследователи использовали расширенные техники (множественные модели, аннотирование, создание бенчмарка) для систематической оценки и сравнения, но ключевые концепции таксономии могут быть адаптированы обычными пользователями.

Основные концепции и подходы, которые можно применить в стандартном чате:

Восемь измерений педагогической таксономии - пользователи могут включать эти критерии в свои промпты, например: Действуй как опытный учитель математики. Когда я покажу свое решение задачи: - Если я сделал ошибку, укажи на нее и где именно она находится - Не давай сразу правильный ответ - Предложи полезную подсказку или наводящий вопрос - Четко объясни, что мне нужно сделать дальше - Используй поддерживающий и ободряющий тон

Принципы эффективного обучения - пользователи могут запрашивать у LLM применение конкретных педагогических стратегий: Поощрение активного обучения и самостоятельного мышления Адаптация к конкретным потребностям и уровню знаний Структурирование информации для управления когнитивной нагрузкой Мотивация и стимулирование любопытства

Знание о сильных и слабых сторонах моделей - пользователи могут корректировать свои запросы, зная ограничения моделей:

Явно указывать на необходимость давать подсказки, а не ответы Запрашивать обратную связь в конкретных аспектах (местоположение ошибки, следующие шаги) Результаты от применения этих концепций: - Более педагогически эффективные взаимодействия с LLM - Обучение, которое способствует пониманию, а не просто получению ответов - Более естественное и поддерживающее взаимодействие в образовательном контексте - Развитие навыков самостоятельного решения проблем у учащихся

Исследование предоставляет ценный концептуальный фреймворк, который можно применять в повседневных взаимодействиях с LLM без необходимости в специальных технических знаниях или доступе.

Анализ практической применимости: **1. Таксономия для оценки педагогических способностей LLM** - **Прямая применимость**: Высокая. Пользователи могут использовать 8 измерений таксономии (идентификация ошибки, указание местоположения ошибки, избегание прямого раскрытия ответа и др.) для формулирования запросов к LLM, чтобы получить более педагогически эффективные ответы. - **Концептуальная ценность**: Высокая. Таксономия помогает понять, как должен выглядеть эффективный педагогический ответ, что позволяет пользователям оценивать качество взаимодействий с LLM. - **Потенциал для адаптации**: Очень высокий. Таксономия может быть применена не только к математике, но и к другим предметным областям.

2. Выявление сильных и слабых сторон современных LLM как тьюторов - **Прямая применимость**: Средняя. Пользователи могут выбирать наиболее подходящие модели для образовательных целей, зная их сильные стороны. - **Концептуальная ценность**: Высокая. Исследование показывает, что даже передовые модели имеют пробелы в педагогических способностях, что важно учитывать при использовании LLM в образовании. - **Потенциал для адаптации**: Высокий. Знание о том, что GPT-4 часто раскрывает ответы, а не дает подсказки, позволяет пользователям адаптировать свои запросы.

3. Методология оценки педагогических ответов - **Прямая применимость**: Средняя. Обычные пользователи могут использовать основные критерии (не раскрывать ответ, давать подсказки, быть поддерживающим) при формулировании запросов к LLM. - **Концептуальная ценность**: Высокая. Методология демонстрирует принципы эффективного обучения, которые можно применять при работе с LLM. - **Потенциал для адаптации**: Высокий. Принципы могут быть применены к различным образовательным контекстам и предметам.

4. Бенчмарк MRBench - **Прямая применимость**: Низкая для обычных пользователей, высокая для разработчиков и исследователей. - **Концептуальная ценность**: Средняя. Позволяет понять типы ошибок и эффективные стратегии их исправления. - **Потенциал для адаптации**: Высокий для создания аналогичных бенчмарков в других областях.

5. Ненадежность LLM как оценщиков педагогических способностей - **Прямая применимость**: Средняя. Пользователи должны с осторожностью относиться к оценкам, которые дают LLM другим LLM. - **Концептуальная ценность**: Высокая. Демонстрирует ограничения LLM в понимании тонкостей педагогики. - **Потенциал для адаптации**: Средний. Информация может быть использована для разработки лучших промптов для оценки.

Prompt:

Применение таксономии AI-репетиторов в промптах для GPT ## Ключевые принципы исследования

Исследование представляет таксономию из 8 измерений для оценки педагогических способностей AI-репетиторов: 1. Идентификация ошибки 2. Локализация ошибки 3. Раскрытие ответа 4. Предоставление руководства 5. Возможность действия 6. Согласованность 7. Тон репетитора 8. Человекоподобность

Пример эффективного промпта для образовательных целей

[======] # Роль: Педагогический AI-репетитор по математике

Контекст Вы помогаете студенту 9 класса с решением задач по алгебре. Ваша цель - не просто дать ответ, а способствовать глубокому пониманию материала.

Инструкции 1. **Идентификация ошибки**: Сначала определите, в чем заключается ошибка или затруднение студента. 2. **Локализация ошибки**: Точно укажите, где в рассуждениях или вычислениях студента произошла ошибка. 3. **Избегайте раскрытия ответа**: Не давайте готовое решение, вместо этого направляйте студента к самостоятельному поиску ответа. 4. **Предоставление руководства**: Дайте пошаговые подсказки, которые помогут студенту продвинуться в решении. 5. **Возможность действия**: Завершайте каждый ответ конкретным предложением следующего шага или вопросом для размышления. 6. **Согласованность**: Учитывайте предыдущие ответы студента и адаптируйте свой подход. 7. **Тон**: Используйте поддерживающий, ободряющий тон, который мотивирует студента продолжать работу. 8. **Человекоподобность**: Будьте эмпатичны, реагируйте на эмоциональное состояние студента.

Формат ответа 1. **Анализ проблемы**: Кратко определите и локализуйте ошибку 2. **Направляющие подсказки**: Предложите 2-3 наводящих вопроса или подсказки 3. **Следующий шаг**: Предложите конкретное действие для продвижения вперед 4. **Поддержка**: Добавьте ободряющее замечание

Теперь я готов помочь студенту с задачей! [======]

Почему это работает

Данный промпт эффективно применяет выводы исследования:

• Избегает прямого раскрытия ответов - исследование показало, что многие LLM (например, GPT-4) склонны просто давать ответы (в 47% случаев), что снижает их эффективность как репетиторов

• Фокусируется на идентификации и локализации ошибок - ключевые навыки для эффективного обучения

• Подчеркивает важность предоставления руководства, а не готовых решений

• Обеспечивает возможность действия - дает студенту конкретные шаги для продолжения обучения

• Задает поддерживающий тон - что согласно исследованию повышает мотивацию студентов

• Структурирует ответ в формате, который охватывает все 8 измерений таксономии

Такой подход позволяет максимально использовать сильные стороны LLM, одновременно компенсируя их типичные педагогические недостатки, выявленные в исследовании.