Интерактивное прогнозирование информационных потребностей с учетом намерений и контекста

Дата: 2025-01-05 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2501.02635

Рейтинг: 70
Адаптивность: 85

Ключевые выводы:

Исследование посвящено интерактивному прогнозированию информационных потребностей пользователей, позволяя им выбирать предварительный контекст поиска (параграф, предложение или слово) и указывать необязательное частичное поисковое намерение. Основные результаты показывают, что такое прогнозирование возможно, а указание частичного поискового намерения помогает преодолеть проблемы, связанные с большими предварительными контекстами поиска.

Объяснение метода:

Исследование предлагает ценную концепцию баланса между контекстом и намерением при формулировке запросов к LLM. Хотя полная техническая реализация требует дообучения моделей, принципы напрямую применимы пользователями: выделение релевантного контекста и указание частичного намерения существенно улучшают качество ответов. Исследование демонстрирует, что меньший, но более точный контекст с намерением эффективнее большого контекста.

Ключевые аспекты исследования: 1. **Интерактивное предсказание информационных потребностей**: Исследование предлагает новый подход, позволяющий предсказывать информационные потребности пользователя на основе выбранного им контекста (от слова до параграфа) и опционального частичного намерения поиска.

Двухкомпонентный ввод: Пользователь может выбрать фрагмент текста (контекст) и указать частичное намерение (например, "почему", "как", "применение"), на основе чего система генерирует полный вопрос или сразу находит ответ.

Генерация вопросов и поиск ответов: Исследованы два основных подхода к реализации - явное предсказание потребности (генерация полного вопроса) и неявное (прямой поиск релевантного ответа).

Влияние объема контекста и намерения: Проанализировано, как объем выбранного контекста и наличие частичного намерения влияют на точность предсказания информационной потребности.

Адаптация существующих датасетов: Для оценки эффективности подхода были адаптированы два существующих набора данных (Inquisitive и MS MARCO).

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

Хотя в исследовании использовалось дообучение моделей для максимальной эффективности, основные концепции могут быть адаптированы для стандартного чата с LLM без дополнительного обучения или API:

Принцип выделения релевантного контекста: Пользователи могут самостоятельно выделять наиболее релевантные части текста вместо копирования всего документа в чат. Исследование показало, что более узкий, но релевантный контекст дает лучшие результаты, чем обширный контекст.

Указание частичного намерения: Пользователи могут добавлять короткие указания намерения ("почему", "как", "примеры", "сравнение") к своим запросам. Исследование демонстрирует, что даже минимальное указание намерения значительно улучшает качество ответов.

Комбинация контекста и намерения: Формат запроса вида "Контекст: [выбранный фрагмент текста]. Намерение: [тип вопроса/задачи]" может быть эффективным способом структурирования запросов к стандартному LLM-чату.

Интерактивное уточнение: Вместо формулировки сложных запросов, пользователи могут сначала предоставить контекст, затем указать намерение, и при необходимости уточнить запрос на основе полученного ответа.

Применение этих концепций в стандартном чате может привести к: - Снижению когнитивной нагрузки при формулировании сложных запросов - Более точным и релевантным ответам от LLM - Лучшему пониманию модели, что именно интересует пользователя - Возможности исследовать информационное пространство более эффективно

Таким образом, хотя авторы использовали дообучение для оптимальных результатов, основные принципы исследования могут быть эффективно применены в повседневном использовании стандартных LLM-чатов.

Анализ практической применимости: **Интерактивное предсказание информационных потребностей** - Прямая применимость: Высокая. Пользователи могут выделять текст на веб-странице и получать сгенерированные вопросы или

сразу ответы без необходимости формулировать полный запрос. - Концептуальная ценность: Значительная. Помогает понять, как LLM могут интерпретировать частичные намерения и контекст, что полезно для эффективного взаимодействия. - Потенциал для адаптации: Очень высокий. Подход можно реализовать как расширение браузера или функцию в приложениях для чтения.

Двухкомпонентный ввод - Прямая применимость: Высокая. Снижает когнитивную нагрузку на пользователя, позволяя указать только контекст и необязательное намерение вместо полной формулировки вопроса. - Концептуальная ценность: Значительная. Демонстрирует, как частичная информация может быть использована для предсказания полной потребности. - Потенциал для адаптации: Высокий. Легко адаптируется для различных интерфейсов, включая чаты с LLM.

Генерация вопросов и поиск ответов - Прямая применимость: Средняя. Требует дообучения моделей, но концепция применима в существующих чатах как способ уточнения запросов. - Концептуальная ценность: Высокая. Показывает два подхода к предсказанию потребностей, которые могут быть использованы в разных сценариях. - Потенциал для адаптации: Высокий. Подход с генерацией вопросов может быть адаптирован для уточнения запросов в обычных LLM-чатах.

Влияние объема контекста и намерения - Прямая применимость: Высокая. Демонстрирует, что меньший, но более точный контекст с указанием намерения дает лучшие результаты, что полезно для формулировки запросов к LLM. - Концептуальная ценность: Очень высокая. Помогает пользователям понять оптимальный баланс между контекстом и намерением при взаимодействии с LLM. - Потенциал для адаптации: Высокий. Принципы могут быть применены для разработки лучших практик формулирования запросов.

Адаптация существующих датасетов - Прямая применимость: Низкая. Технический аспект исследования, мало применимый для обычных пользователей. - Концептуальная ценность: Умеренная. Демонстрирует подход к созданию данных для тестирования подобных систем. - Потенциал для адаптации: Средний. Методология может быть полезна для исследователей и разработчиков.

Prompt:

Применение исследования интерактивного прогнозирования информационных потребностей в промптах для GPT ## Ключевые знания из исследования для промптов

Исследование показывает, что: - Предоставление контекста разного размера влияет на качество ответов - Указание частичного намерения значительно улучшает результаты - Комбинация "контекст + частичное намерение" дает наилучшие результаты

Пример промпта с применением знаний из исследования

[======] Контекст: [Выделенный параграф из статьи о квантовых компьютерах]

Частичное намерение: Хочу понять, как квантовые компьютеры могут повлиять на...

Инструкция: На основе предоставленного контекста и моего частичного намерения: 1. Сформулируй 3 конкретных вопроса, которые я, вероятно, хочу задать 2. Дай развернутый ответ на самый важный из этих вопросов 3. Предложи следующие направления для исследования темы [======]

Объяснение эффективности

Данный промпт работает эффективно, потому что:

Структурирует контекст — предоставляет модели четко выделенную информацию для анализа **Включает частичное намерение** — помогает модели сфокусироваться на релевантных аспектах, даже если контекст большой **Дает четкие инструкции** — направляет модель на прогнозирование информационных потребностей и их удовлетворение Такой подход позволяет получить более персонализированные и точные ответы, поскольку модель может лучше предугадать, какая именно информация вам нужна, даже если вы не можете полностью сформулировать вопрос.