По шкале от 1 до 5: количественная оценка галлюцинаций в оценке достоверности

Дата: 2025-02-08 00:00:00

Ссылка на исследование: https://arxiv.org/pdf/2410.12222

Рейтинг: 65
Адаптивность: 70

Ключевые выводы:

Исследование направлено на количественную оценку галлюцинаций в задачах оценки достоверности генерируемого контента языковыми моделями (LLM). Основная цель - разработать автоматизированную систему оценки степени достоверности генерируемого текста по шкале от 1 до 5. Результаты показали, что GPT-4 способна наиболее точно оценивать фактическую согласованность между исходным текстом и генерацией, а дообучение NLI-моделей на синтетических данных может улучшить их производительность.

Объяснение метода:

Исследование предлагает практичную шкалу оценки верности контента (1-5) и полезную классификацию галлюцинаций на внутренние/внешние, что повышает критическое взаимодействие с LLM. Рубрики оценки адаптируемы для разных задач. Однако, реализация некоторых методов (NLI, генерация синтетических галлюцинаций) требует технической экспертизы, что ограничивает прямую применимость для широкой аудитории.

Ключевые аспекты исследования: 1. **Разработка количественной оценки галлюцинаций в LLM**: Исследование предлагает методику оценки степени "верности" (faithfulness) сгенерированного текста по шкале от 1 до 5, где верность понимается как фактическая согласованность с исходным текстом.

Классификация типов галлюцинаций: Авторы различают внутренние (intrinsic) галлюцинации, когда сгенерированный текст противоречит фактам из исходного документа, и внешние (extrinsic), когда добавляются новые непроверяемые факты.

Методы оценки верности текста: Исследование сравнивает два подхода к оценке верности текста: использование LLM с рубриками оценки и применение моделей Natural Language Inference (NLI).

Генерация синтетических галлюцинаций: Авторы разработали методологию создания синтетических примеров с галлюцинациями разных типов для обучения и

тестирования моделей.

Оценка чувствительности моделей: Проведено исследование влияния процента галлюцинаций в тексте на итоговую оценку верности, что позволяет измерить чувствительность различных моделей к разной степени недостоверности.

Дополнение: Для работы основных методов этого исследования не требуется дообучение или API в обязательном порядке. Многие концепции и подходы можно применить в стандартном чате с LLM. Исследователи использовали расширенные техники (API, дообучение NLI-моделей) для повышения точности и формализации результатов, но ключевые идеи применимы и в обычном чате.

Концепции и подходы, применимые в стандартном чате:

Шкала оценки верности (1-5) - пользователи могут запросить LLM оценить достоверность информации по этой шкале, предоставив источник и текст для проверки.

Классификация галлюцинаций на внутренние и внешние - эта концептуальная модель помогает пользователям более структурированно выявлять проблемы в ответах LLM, даже без технической реализации.

Рубрики оценки верности - пользователи могут адаптировать предложенные в исследовании критерии оценки и включать их в промпты:

Фактическая согласованность (проверка числовых значений, имен собственных) Уместность прилагательных Конгруэнтность знаний (отсутствие непроверяемой внешней информации) Стилистическое соответствие

Запрос обоснования оценки - исследование показало, что требование от LLM обосновать свою оценку повышает точность. Этот прием можно использовать в любом чате.

Техника сегментации длинных текстов - при работе с длинными текстами пользователи могут применять сегментацию, проверяя каждую часть отдельно.

Ожидаемые результаты от применения этих концепций: - Повышение критического отношения к генерируемому контенту - Более структурированная и систематическая оценка достоверности информации - Улучшение способности выявлять недостоверную информацию в ответах LLM - Формирование более точных запросов, учитывающих ограничения моделей

Таким образом, хотя некоторые технические аспекты исследования требуют специализированных инструментов, ключевые концептуальные подходы вполне применимы в стандартном чате с LLM.

Анализ практической применимости: **1. Количественная оценка верности контента** - **Прямая применимость**: Высокая. Пользователи могут применять

предложенную шкалу от 1 до 5 для оценки достоверности генерируемого LLM контента. Это особенно полезно при работе с информационно-критичными задачами. - **Концептуальная ценность**: Значительная. Понимание различных степеней верности (от "высоко достоверного" до "высоко недостоверного") помогает пользователям формировать более реалистичные ожидания от генерируемого контента. - **Потенциал адаптации**: Высокий. Пользователи могут адаптировать критерии оценки под свои задачи, сохраняя основную шкалу.

2. Классификация типов галлюцинаций - **Прямая применимость**: Средняя. Обычные пользователи могут использовать эту классификацию для более точного выявления проблем в генерируемом тексте. - **Концептуальная ценность**: Высокая. Понимание различий между внутренними (противоречащими источнику) и внешними (добавляющими непроверяемую информацию) галлюцинациями позволяет пользователям более критично оценивать ответы LLM. - **Потенциал адаптации**: Средний. Пользователи могут адаптировать эту классификацию для своих задач, но требуется некоторое понимание принципов работы LLM.

3. Методы оценки верности текста - **Прямая применимость**: Средняя. Обычные пользователи могут использовать рубрики для проверки генерируемого контента, но полная реализация методов NLI требует технических знаний. - **Концептуальная ценность**: Высокая. Понимание того, как можно оценивать верность с помощью различных подходов, помогает пользователям выбирать подходящие стратегии взаимодействия с LLM. - **Потенциал адаптации**: Средний. Рубрики оценки могут быть адаптированы пользователями, но реализация моделей NLI требует технической экспертизы.

4. Генерация синтетических галлюцинаций - **Прямая применимость**: Низкая для обычных пользователей. Требует технических навыков. - **Концептуальная ценность**: Средняя. Понимание методов создания "контролируемых галлюцинаций" может помочь пользователям лучше осознавать механизмы возникновения недостоверной информации. - **Потенциал адаптации**: Низкий для широкой аудитории, высокий для технических специалистов.

5. Оценка чувствительности моделей - **Прямая применимость**: Низкая для обычных пользователей. - **Концептуальная ценность**: Высокая. Понимание того, что разные модели имеют разную чувствительность к галлюцинациям, помогает пользователям выбирать подходящие модели для своих задач. - **Потенциал адаптации**: Низкий для широкой аудитории.

Prompt:

Применение исследования о галлюцинациях LLM в промптах ## Ключевые знания для использования в промптах

Исследование о количественной оценке галлюцинаций предоставляет несколько важных инсайтов, которые можно применить при составлении промптов:

GPT-4 лучше всего оценивает достоверность среди LLM **Объяснение оценки повышает точность** выявления галлюцинаций **Внутренние галлюцинации** (противоречия) легче обнаруживаются, чем внешние (добавления) **Двухэтапный подход** может снизить затраты на проверку ## Пример промпта для проверки достоверности содержания

[======] # Запрос на проверку достоверности текста

Контекст Я хочу, чтобы ты выступил в роли эксперта по оценке достоверности информации. Исследования показывают, что модели GPT-4 способны эффективно оценивать фактическую согласованность между исходным текстом и генерацией.

Задача Оцени достоверность следующего сгенерированного текста по шкале от 1 до 5, где: 1 - полностью недостоверный текст с множеством противоречий 2 - в основном недостоверный текст с несколькими серьезными ошибками 3 - частично достоверный текст с некоторыми неточностями 4 - в основном достоверный текст с незначительными неточностями 5 - полностью достоверный текст без заметных ошибок

Исходный текст (факты): [ВСТАВИТЬ ИСХОДНЫЙ ТЕКСТ]

Сгенерированный текст для проверки: [ВСТАВИТЬ СГЕНЕРИРОВАННЫЙ ТЕКСТ]

Инструкции 1. Сначала проверь на внутренние галлюцинации (противоречия фактам из исходного текста) 2. Затем проверь на внешние галлюцинации (добавление новой непроверяемой информации) 3. Дай общую оценку по шкале от 1 до 5 4. Обязательно предоставь подробное обоснование своей оценки с указанием конкретных примеров галлюцинаций 5. Укажи примерный процент недостоверной информации в тексте [======]

Как это работает

Данный промпт применяет знания из исследования следующим образом:

Использует шкалу 1-5 для количественной оценки, как предложено в исследовании **Требует обоснования оценки**, что повышает точность согласно результатам исследования **Разделяет проверку на внутренние и внешние галлюцинации**, учитывая их разную обнаруживаемость **Запрашивает примерный процент недостоверности**, используя эвристический подход из исследования Такой промпт позволяет максимально использовать способности GPT-4 к оценке достоверности и получить более точные результаты проверки.