Измерение уровня галлюцинаций в LLM

Дмитрий Тронин
22 МАГ ИАД
Факультет информатики, математики
и компьютерных наук
ВШЭ НН

10 декабря 2023

Несмотря на огромные возможности LLM, данные модели обладают большим недостатком: их ответам нельзя полностью доверять. Экспериментально была замечена склонностью такого вида нейронных моделей к генерации ложной информации (галлюцинаций) в контекстах, где ожидался правдивый ответ. Поэтому при сравнении различных LLM важно также оценивать уровень галлюцинаций при их работе. В данной работе предлагается модифицированная методика измерения уровня галлюцинаций в LLM, а также приводится сравнение нескольких популярных моделей.

Ключевые слова LLM · Галлюцинаций

1 Введение

В последнее время наблюдается активный рост в развитии и использовании больших языков моделей (Large Language Model, LLM). Эти модели, представляющие разновидность нейронных сетей, демонстрируют способность решать широкий спектр различных задач. В связи практически ежедневно появляются новые варианты их применения на практике.

Однако, несмотря на их большие возможности, необходимо с осторожностью подходить к использованию больших языков моделей из-за их склонности к галлюцинациям. Из-за сложности и множества различных проявлений галлюцинаций в академической среде не сформировалось общепринятое определение данного феномена.

В ходе данной работы использовалось следующее определение: «Галлюцинация – это ответ LLM, которые нарушает инструкции, данные человеком» [3]. Данное определение покрывает различные варианты использования LLM от получения исторический справки до создания художественных произведений.

2 Обоснование методологии

Согласно подходу, предложенному в [3], вместо того, чтобы рассматривать уровень галлюцинирования при выполнение различных NLP задач, мы рассматриваем LLM как черный ящик. Благодаря тому, что языковые модели могут выполнять множество различных задач, предлагается оценка способностей модели на основе уровней с учетом исследований психологии человека. Предполагается, что недостаточные способности на определенном уровне приведет к генерации галлюцинаций.

В качестве уровней LLM мы рассматривали следующие компетенции:

1. Понимание естественного языка

Лексические, синтаксические, морфологические знания и значения слов позволяют модели корректно распознавать запросы пользователя. Данный уровень является основой для работы LLM.

2. Общие знания об окружающем мире

   Знания об предметах окружающего мира, взаимоотношения и связи между ними, событиях позволяют модели рассуждать на основе здравого смысла, что является необходимым для понимания материального мира и человеческого общества.

3. Абстрактное мышление

   Понимание абстрактных понятий и их связей позволяет модели делать выводы на основе имеющихся фактов. Данная способность является основой для высших когнитивных способностей, таких как планирование, решение задач и принятие решений.

Основным преимуществом данного подхода является его универсальность. На основе понимания естественного языка, знаниях об окружающем мире и способности к абстрактному мышлению LLM способна решать широкий спектр NLP задач.

3 Методология

Понимание естественного языка

Для проверки способности LLM понимать естественный язык мы предлагаем задачу перевода юридических текстов, для которых существую версии на 2х языках. Юридические тексты написаны так, чтобы точно передавать значение. В таком случае вероятность нескольких возможных эквивалентных переводов мало, поэтому можно считать иной перевод отклонением, а значит галлюцинацией.

Общие знания об окружающем мире

Для проверки общих знаний об окружающем мире мы предлагаем задачу описания редких слов. За счет того, что слово мало встречалось в обучающей выборке, вероятность галлюцинаций растет.

Абстрактное мышление

Для проверки способности LLM к абстрактному мышлению мы предлагаем задачу логического вывода на основе приведенных фактов, которые не связаны с материальным миром. Явное использование только приведенных фактов позволит отличить ситуацию, в которой модель не имеет необходимых фактов об окружающем мире, от ситуации, в котором модель имеет низкий уровень абстрактного мышления. Факты сформулированны таким образом, что логический вывод можно произвести с помощью бинарной логики.

4 Экспериментальное исследование

В качестве моделей для экспериментов мы использовали:

1. Сбер GigaChat [5]
2. OpenAI ChatGPT-3.5 [6]
3. Anthropic Claude 2 [7]

При проведении исследования каждой LLM было задано 20 запросов на каждом уровне компетенций. Каждый запрос задавался с пустым контекстом. Таким образом все ответы модели независимы. Слова отбирались из последних 10% ранжированного по встречаемости списка.

Определение корректности сгенерированного LLM ответа производилось экспертом на основе общедоступной информации и эталонного ответа из тестового набора данных.

Понимание естественного языка

В качестве тестового набора данных для перевода юридических текстов использовалась статьи Всеобщая декларация прав человека ООН на английском [4] и русском [9] языках.

Запрос:
`Translate the next paragraphs into Russian. Be as precise in terminology as possible. <текст статьи на английском>`

Результаты представлены в таблице 1.

	Английский язык		
	GigaChat	ChatGPT-3.5	Claude 2
Верно	7	11	18
Неверно	13	9	2

Таблица 1: Результаты исследования понимания естественного языка

Наблюдения на основе исследования:

1. GigaChat достаточно часто генерирует фразы, которые звучат ненатурально (запросы №1, 3, 11, 15, 19)
2. Все модели иногда добавляют дополнительные комментарии в ответ
3. GigaChat иногда не переводит текст, а возвращает оригинальное обращение (запрос №9)
4. GigaChat и ChatGPT-3.5 часто неправильно переводят юридические термины.
5. Claude 2 очень часто полностью повторяет эталонный ответ. Возможно модель обучалось на используемом тестовом наборе данных.

Согласно исследованию GigaChat хуже всего справляется с задачей перевода юридического текста, далее по качеству стоит ChatGPT-3.5, и наиболее качественно перевод генерирует Claude 2.

Общие знания об окружающем мире

В качестве тестового набора данных использовался частотные словари английского [2] и русского [8] языка.

Запрос:

- `Can you explain what is <понятие>?`
- `Можешь рассказать, что такое <понятие>?`

Результаты представлены в таблице 2.

	Английский язык		
	GigaChat	ChatGPT-3.5	Claude 2
Верно	14	18	19
Неверно	6	2	1
	Русский язык		
	GigaChat	ChatGPT-3.5	Claude 2
Верно	15		15
Неверно	5		5

Таблица 2: Результаты исследования уровня знаний об окружающем мире

Наблюдения на основе исследования на английском языке:

1. При прочих равных GhatGPT-3.5 и Clade 2 генерируют более длинные ответы.
2. GigaChat иногда отказывается отвечать на обращение пользователя (запросы №6, 14). Данное явление может быть связанно с внутренними правилами и ограничениями вывода результатов генерации.

3. GigaChat иногда забывает язык, на котором было обращение, и отвечает на русском языке (запрос №17).
4. Все модели при запросе объяснения слова "billon" (англ. низкопробное золото или серебро) сгенерировали объяснение слова "billion" (англ. миллиард). Это может быть связанно с внутренней обработкой опечаток.

Согласно исследованию GigaChat имеет значительно меньшие знания об окружающем мире на английском языке, чем ChatGPT-3.5 и Claude 2.

Наблюдения на основе исследования на русском языке:

1. Claude 2 иногда отказывается отвечать на обращение пользователя (запрос №4, 7).

Согласно исследованию уровень знаний об окружающем мире на русском языке у GigaChat и Claude 2 примерно одинаков.

При сравнении случаев английского и русского языка видим, что число галлюцинаций Claude 2 на русском языке растет.

Абстрактное мышление

В качестве тестового набора данных для английского языка использовался открытый датасет `ruletaker-depth-3ext` из статьи [1].

В качестве тестового набора данных для русского языка использовались те же запросы, переведенные на русский язык. При переводе особое внимание было уделено сохранению однозначности признаков в разных фактов, в случае когда английское слово имеет несколько смыслов (напр. "red" можно перевести как красный и рыжий).

Запрос:

- `Given some premises, conduct reasoning to answer whether the given query is true, false or unknown.`
  `Premises: <некоторые факты>`
  `Query: <вопрос>`
- `Учитывая некоторые предпосылки, проведите рассуждения, чтобы ответить, является ли данный запрос истинным, ложным или неизвестным.`
  `Предпосылки: <некоторые факты>`
  `Запрос: <вопрос>`

Результаты представлены в таблице 3.

	Английский язык		
	GigaChat	ChatGPT-3.5	Claude 2
Верно	8	12	18
Неверно	12	8	2
	Русский язык		
	GigaChat	ChatGPT-3.5	Claude 2
Верно	8		15
Неверно	12		5

Таблица 3: Результаты исследования способности к абстрактному мышлению

Наблюдения на основе исследования на английском языке:

1. GigaChat иногда ошибается в тривиальных случаях, когда ответ выводится из 1 факта (запросы №2, 5, 7, 15)
2. ChatGPT-3.5 и Claude 2 лучше следуют запросу, так как явно генерируют цепочку размышлений.

3. ChatGPT-3.5 и Claude 2 чаще генерирует неверный ответ с ростом глубины цепочки размышлений (количестве использованных фактов, для корректного логического вывода).

Согласно исследованию GigaChat имеет самую низкую способность к абстрактному мышлению на английском языке, далее по качеству стоит ChatGPT-3.5, и наиболее качественно перевод генерирует Claude 2.

Наблюдения на основе исследования на русском языке:

1. GigaChat иногда ошибается в тривиальных случаях, когда ответ выводится из 1 факта (запросы №1, 3, 7).
2. Обе модели точно следуют запросу и генерируют цепочку размышлений.

Согласно исследованию GigaChat имеет более низкую способность к абстрактному мышлению на русском языке, далее по качеству стоит ChatGPT-3.5, и наиболее качественно перевод генерирует Claude 2.

При сравнении случаев английского и русского языка видим, что число галлюцинаций Claude 2 на русском языке растет.

5 Выводы экспериментального исследования

По результатам экспериментального исследования на всех уровнях мышления соблюдается следующий порядок (от большего числа галлюцинаций к большему):

1. GigaChat
2. ChatGPT-3.5
3. Claude 2

Интересно заметить, что использование запросов на русском языке снижает качество работы Claude 2, но не повышает качество GigaChat.

6 Пути развития работы

В качестве развития данной работы мы видим следующие направления:

- увеличение числа запросов к моделям
- более подробное исследование способности к абстрактному мышлению с учетом разной глубины логической цепочки и различных способах логического вывода
- инжиниринг запросов к LLM для увеличения качества

Список литературы

[1] Clark, P., Tafjord, O., and Richardson, K. Transformers as soft reasoners over language, 2020.

[2] Davies, M., and Gardner, D. A frequency dictionary of contemporary American English: Word sketches, collocates and thematic lists. Routledge, 2013.

[3] Du, L., Wang, Y., Xing, X., Ya, Y., Li, X., Jiang, X., and Fang, X. Quantifying and attributing the hallucination of large language models via association analysis, 2023.

[4] UNH. Universal declaration of human rights, 1961.

[5] Сбер GigaChat. `https://developers.sber.ru/portal/products/gigachat`. Просмотрено: 10.11.2023.

[6] OpenAI ChatGPT-3.5. `https://chat.openai.com/`. Просмотрено: 10.11.2023.

[7] Claude 2 Anthropic. `https://www.anthropic.com/index/claude-2`. Просмотрено: 10.11.2023.

[8] Ляшевская О. Н., Шаров С. А. Частотный словарь современного русского языка на материалах Национального корпуса русского языка. Общество с ограниченной ответственностью"Издательский центр"Азбуковник 2009.

[9] ООН. Всеобщая декларация прав человека, 1961.