Подходы общения с БЯМ для сокращения вызываемых галлюцинаций

Рафиков А.С.
студент ВШЭ
Нижний
Новгород
artemas.raf@gmail.com

Крылов В.В.
куратор ВШЭ
Нижний
Новгород
vladimir.v.krylov@gmail.com

Abstract

Большие языковые модели (БЯМ) используются во многих NLP задачах и важно, чтобы при выполнении этих задач, решения этой модели были корректными с большей вероятностью, так как это важно для всех областей. Однако человечество еще не готово доверять решениям таких моделей пока люди не проверят эти решения сами. В этой работе оценивается частота галлюцинаций в БЯМ, работа с контекстом и корректное указание источников в подтверждение своих утверждений на разные темы, например, шахматы, солнечная система, любовные отношения и другие. Также явно выделяются проблемы большинства моделей в соответствии ответа запросу, а также понимании минимально запрашиваемой информации. Эффективное общение с большими языковыми моделями крайне важно для использования их потенциала при минимизации ошибок, известных как "галлюцинации когда модель генерирует правдоподобную, но неточную или бессмысленную информацию. В этой статье исследуются различные подходы к снижению частоты галлюцинаций путем уточнения структуры запросов, внедрения механизмов обратной связи.

Keywords LLM · Hallucinations · GPT-4 · GPT-3.5 · Claude 2 · Zephyr · LLaMa-2

1 Введение

Большие языковые модели (БЯМ) или же Large Language Models (LLMs) представляют собой инструмент, огромный объем информации и работу с ним. Эти модели были обучены на огромных объемах текстовых данных и могут выполнять широкий спектр задач, связанных с языком, от банальных как ответы на вопросы, диалог, выжимка текста, перевод с разных языков и многое другое. В данной работе демонстрируется способность моделей подкреплять свои высказывания ссылаясь на какой-либо источник. А также умение работать с вспомогательной информацией. Контекст - это набор предложений, фактов, информации на одну или несколько тем, которые будут запрашиваться у модели. Большие языковые модели становятся все более сложными благодаря своей способности понимать и генерировать текст, похожий на человеческий. Несмотря на свои впечатляющие возможности, эти модели подвержены возникновению "галлюцинаций которые могут варьироваться от незначительных неточностей до полностью сфабрикованных утверждений. Это может быть проблематично в приложениях, где точность и надежность имеют первостепенное значение. Устранение таких случаев имеет решающее значение для пользователей, которые полагаются на информацию, предоставляемую БЯМ.

2 Понимание галлюцинаций

Прежде чем углубляться в тему способов уменьшения галлюцинаций, важно понять, что влекут за собой галлюцинации в БЯМ. Они часто являются результатом данных на которых обучались модели

или ее интерпретации. Поскольку БЯМ предсказывают следующее слово в последовательности, они иногда могут генерировать ответы, которые являются последовательными, но не связаны с реальностью или предполагаемым контекстом. В таких "проявлениях фантазии"чаще всего модель выдумывает несвязанные вещи от слова совсем. Однако существуют прецеденты из-за которых открывалось новое лекарство или обнаруживалась смертельная болезнь на ранних стадиях у ребенка. Поэтому стоит четко отметить галлюцинации это не всегда плохо, все зависит от их природы, так как это может быть идея выстроенная на логичных выводах или рамндомная несвязная последовательность слов или предложений.

3 Подходы к общению с БЯМ:

Чтобы снизить вероятность галлюцинаций при взаимодействии с БЯМ, можно использовать определенные способы общения.

3.1 Ясные и точные запросы:

Один из основных способов уменьшить количество галлюцинаций - задавать четкие, конкретные и хорошо структурированные вопросы. Двусмысленность может привести модель к "коллапсу мыслей". Пользователям следует избегать открытых вопросов, когда требуется точность, и предоставлять достаточный контекст для лучшего понимания у модели.

3.2 Поэтапное предоставление информации:

Основываясь на предыдущем пункте, постепенное предоставление информации может привести модель к генерированию ответов, основанных на постепенном раскрытии информации. Этот поэтапный подход может помочь модели не сбиться с пути и сохранить актуальность для рассматриваемой темы. Такое явление называется цепочка мыслей, модель поэтапно вникает в тему и с большей вероятностью даст правельный ответ без галлюцинаций.

3.3 Циклы обратной связи и обучение с подкреплением:

Включение механизма обратной связи, при котором результаты модели оцениваются, а исправления передаются обратно в систему, может помочь в обучении модели уменьшению галлюцинаций с течением времени. Обучение моделей распознавать и признавать неопределенность может предотвратить представление неверной информации как факта. Такие методы, как "штрафы за неопределенность"во время обучения, могут отбить у модели охоту делать утверждения, когда данные недостаточно подтверждающие.

3.4 Техническое вмешательство:

Помимо коммуникационных стратегий, также могут быть предприняты определенные технические вмешательства, чтобы свести к минимуму галлюцинации. Обеспечение разнообразного и сбалансированного набора обучающих данных может помочь моделям изучать широкий спектр тем и перспектив, тем самым уменьшая переобучение конкретным рассказам или шаблонам, которые могут привести к галлюцинациям. Язык динамичен, и модели, которые его имитируют, должны быть такими же. Регулярные обновления последней информацией могут поддерживать знания модели свежими и в большей степени соответствовать реальности. Разработка подсказок и шаблонов ответов для LLMS может упростить тип ответов, генерируемых моделью, гарантируя, что они основаны на входных данных и менее подвержены ошибочному изготовлению.

4 Эксперименты

Для подтверждения ранее сказанных рекомендаций, были проведены эксперименты на нескольких моделях:

Важные замечания, все модели, кроме Zephyr отвечали на запрос с контекстом в рамках одной сессии, Zephyr на второй запрос выводил информационную сводку о том, что ему не хватает информации и он не будет дальше с нами разговаривать.

Таблица 1: Сводные результаты

Context	Model	Wrong sources	Correct sources	Correct %	requested sources %
No	GPT-3.5	13	34	0.74	0.43
No	GPT-3.5 16K	0	35	1.00	0.33
No	Zephyr	0	35	1.00	0.33
No	LLaMa-2	6	91	0.94	0.92
No	Code-LLaMa	10	95	0.90	1.00
No	Claude 2	0	105	1.00	1.00
No	GPT-4	1	104	0.99	1.00
Yes	GPT-3.5	0	24	1.00	0.23
Yes	GPT-3.5 16K	0	56	1.00	0.53
Yes	Zephyr	6	115	0.97	1.15
Yes	LLaMa-2	0	45	1.00	0.43
Yes	Code-LLaMa	1	92	0.99	0.88
Yes	Claude 2	0	52	1.00	0.49
Yes	GPT-4	4	93	0.96	0.92

5 Вывод

Хотя большие языковые модели представляют собой значительный прогресс в области искусственного интеллекта, их предрасположенность к галлюцинациям остается препятствием. Используя точную тактику общения и внедряя надежные технические меры, частоту и воздействие таких галлюцинаций можно снизить. Непрерывная эволюция этих моделей в сочетании со стратегическим взаимодействием человека и искусственного интеллекта открывает путь к более надежной и точной коммуникации, управляемой искусственным интеллектом.

Список литературы