Сокращение галлюцинаций в LLM с помощью модицифированных запросов

Научно-исследовательский семинар "Методы интеллектуального анализа данных и компьютерной лингвистики"

Выполнил: студент группы 22МАГИАД Рухович Игорь Владимирович
Руководитель: профессор, д. т. н., Крылов Владимир Владимирович
Факультет информатики, математики и компьютерных наук
НИУ ВШЭ
Нижний Новгород, 603000

сентябрь-декабрь 2023 г.

Abstract

По мере развития больших языковых моделей, перед исследователями возникают всё новые вызовы, и одной из важных проблем являются галлюцинации в LLM. В то время как появляются различные методы смягчения для решения проблемы галлюцинаций, также крайне важно исследовать их основные причины. В данной работе будет рассмотрено, как лингвистические факторы в промптах модели, такие как читаемость, формальность и конкретность, влияют на возникновение галлюцинаций. Экспериментальные результаты показывают, что промпты, характеризующиеся меньшей формальностью и абстрактностью, склонны вызывать меньше галлюцинаций. Однако результаты, касающиеся читабельности, остаются неоднозначными и демонстрируют смешанный характер.

Keywords LLM · Hallucination · GPT

1 Введение

Недавнее быстрое развитие больших языковых моделей привлекло много внимания. Для LLM находят всё больше возможных применений и начинают использовать в совершенно разных областях. Тем не менее, есть одно препятствие, которое мешает с полной уверенностью повсеместно внедрять LLM. Это так называемые “галлюцинации” - ситуации, когда языковая модель генерирует ответ с фактическими неточностями и несоответствиями.

“Prompt engineering” может сыграть ключевую роль в смягчении галлюцинаций в генеративных моделях искусственного интеллекта. Предоставляя ясные и конкретные промпты, пользователи могут направлять модель искусственного интеллекта на создание контента, соответствующего их заданному контексту или требованиям. Это может снизить вероятность того, что модель выдаст галлюцинации или неточную информацию. Промпты могут включать контекстные подсказки, которые помогут модели понять контекст запроса. Этот дополнительный контекст может направлять модель к созданию ответов, более точно соответствующих контексту и менее подверженных галлюцинациям. Сложные подсказки могут использоваться для направления модели через серию шагов, обеспечивая логическую последовательность мышления и создание последовательных ответов.

Современные большие языковые модели обладают способностью обрабатывать длинные промпты в качестве входных данных. Тем не менее, исследования показывают, что эти модели показывают наилучшие результаты, когда ключевая информация находится в начале или конце ввода. Производительность значительно снижается, когда моделям нужно получить доступ к важной информации в середине контекста. Более того, по мере увеличения длины промпта даже модели, специально разработанные для обработки более длинных вводов, испытывают существенное ухудшение производительности.

$$206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

Рис. 1: Индекс удобочитаемости Флеша

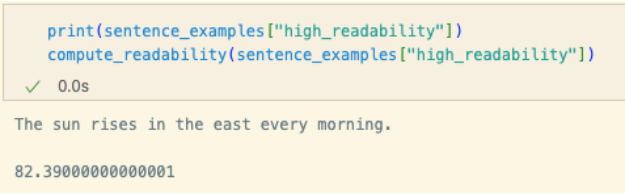

Рис. 2: Пример предложения с высоким индексом удобочитаемости

2 Измерение характеристик запроса

Лингвистические свойства относятся к различным характеристикам и атрибутам языка и его компонентов. Эти свойства помогают определить и понять язык. Некоторые фундаментальные лингвистические свойства включают: синтаксические, семантические, прагматические и лексические. Исходя из этого, мы более детально рассмотрим три интересных лингвистических характеристики.

2.1 Удобочитаемость

Удобочитаемость количественно определяет легкость, с которой текст может быть понят. Несколько факторов, включая сложность текста, знакомость слов, разборчивость и типографику, в совокупности способствуют его удобочитаемости.

Показатель легкости чтения Флеша (FRES) (Flesh, 1948) (Рис. 1) - это показатель удобочитаемости текста. Он был разработан для оценки того, насколько легко или сложно читать и понимать фрагмент текста. Оценка рассчитывается на основе двух факторов: (а) Длины предложения и (б) сложности слова.

Как показано на (Рис. 2, 3), в первом предложении язык простой, предложение совсем легко понять, что приводит к высокому индексу удобочитаемости. Напротив, второе предложение содержит сложную лексику и длинные фразы, что затрудняет его понимание, а значит приводит к более низкому индексу удобочитаемости Флеша. Здесь и далее примеры будут приведены на английском языке, поскольку использованные в работе метрики больше подходят для английского языка.

2.2 Формальность

Формальность языка относится к степени изысканности, благопристойности или вежливости, передаваемой выбором слов, структурой предложения и общим тоном общения. Это способ указать уровень этикета, уважения или профессионализма в данном контексте. Формула расчета формальности представлена на (Рис. 4)

В примере, приведенном ниже (Рис. 5, 6), оба предложения передают идентичное сообщение, однако в первом из них значительно больше формальности. Такие стилистические различия часто оказывают более существенное влияние на понимание предложения читателем, чем само буквальное значение.

2.3 Конкретность

Конкретность оценивает степень, в которой слово представляет осязаемую или воспринимаемую концепцию. Согласно теории, предполагается, что конкретные слова легче обрабатывать по сравнению с абстрактными

Рис. 3: Пример предложения с низким индексом удобочитаемости

$$\begin{aligned} F = (&\text{noun frequency} + \text{adjective freq.} \\ &+ \text{preposition freq.} + \text{article freq.} \\ &- \text{pronoun freq.} - \text{verb freq.} \\ &- \text{adverb freq.} - \text{interjection freq.} + 100)/2 \end{aligned}$$

Рис. 4: Формула расчета индекса формальности. Частоты измерены в процентах

Рис. 5: Пример предложения с высоким индексом формальности

словами. Степень конкретности, связанная с каждым словом, выражается с помощью 5- балльной шкалы оценок, которая варьируется от абстрактной до конкретной.

Слово получает более высокую оценку, если относится к чему-то, что физически существует в реальности, то есть человек может непосредственно ощутить это через органы чувств (обоняние, вкус, осязание, слух, зрение) и действия. Абстрактные слова получают более низкую оценку и относятся к чему-то, что недоступно непосредственно вашим чувствам или действиям. Их значения зависят от языка и обычно разъясняются с помощью других слов, поскольку нет простого метода прямой демонстрации.

Оценивать конкретность каждого слова предлагается с помощью экспертно размеченного набора слов и словосочетаний на английском языке. Конкретность предложения предлагается рассчитывать как среднюю конкретность слов в нём. Аналогично, конкретность текста равна средней конкретности его предложений.

На (Рис. 7, 8) представлены примеры более конкретного и более абстрактного предложений.

3 Данные

Данные для экспериментов были собраны и предоставлены коллегой-студентом Дмитрием Трониным. Было получено 3 датасета, по 20 запросов в каждом. Имаются запросы 3-х типов: запросы на объяснение смысла слова, запросы с логическими цепочками и запросы на перевод текста с английского на русский язык. Запросы были поданы 3 языковым моделям: GigaChat (Сбер), ChatGPT (OpenAI), Claude (Anthropic) и размечены по наличию или отсутствию галлюцинаций.

В итоге был получен набор данных (Рис. 9) из 59 записей. Для каждой записи рассчитаны оценки удобочитаемости, формальности и конкретности.

4 Результаты

Полные результаты и другие данные эксперимента можно найти на

`https://github.com/RukhovichIV/HSE-NIS-IDA-Methods`

На рисунках (Рис. 10, 11, 12) для моделей GigaChat, ChatGPT (3.5T), Claude соответственно предствалены попарные распределения рассчитанных метрик, а также распределения значений самих метрик. Синим цветом отмечены точки и кривые, в которых в ответе модели отсутствуют галлюцинации, а оранжевым - сгаллюцинированные ответы.

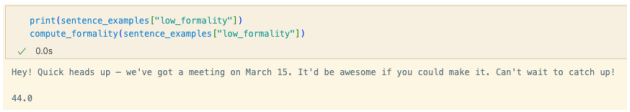

Рис. 6: Пример предложения с более низким индексом формальности

```
print(sentence_examples["high_concreteness"])
compute_concreteness(word_concreteness_df, sentence_examples["high_concreteness"])
```

✓ 0.0s

```
I can smell the flowers and feel the warmth of the sun as I walk through the colorful garden.

2.779375
```

Рис. 7: Пример предложения с высоким индексом конкретности

```
print(sentence_examples["low_concreteness"])
compute_concreteness(word_concreteness_df, sentence_examples["low_concreteness"])
```

✓ 0.0s

```
The beauty and calmness of nature create a sense of harmony and connection.

2.1261538461538465
```

Рис. 8: Пример предложения с более низким индексом конкретности (абстрактный)

	prompt	hal_gigachat	hal_chatgpt	hal_claude	readability	formality	concreteness
0	Can you explain what is ology?	False	False	False	78.872857	28.571429	2.844000
1	Can you explain what is playset?	False	False	False	90.958571	14.285714	3.153333
2	Can you explain what is footwell?	True	False	False	90.958571	14.285714	2.844000
3	Can you explain what is fathering?	False	False	False	78.872857	14.285714	2.844000
4	Can you explain what is conjunctiva?	False	False	False	66.787143	14.285714	2.844000
5	Can you explain what is seediness?	True	False	False	78.872857	28.571429	2.710000
6	Can you explain what is followership?	False	False	False	66.787143	14.285714	2.844000
7	Can you explain what is restenosis?	False	False	False	66.787143	28.571429	2.844000
8	Can you explain what is lectionary?	False	False	False	66.787143	28.571429	2.844000
9	Can you explain what is soundwave?	False	False	False	90.958571	14.285714	2.844000

Рис. 9: Сэмпл итогового набора данных, использованного в эксперименте

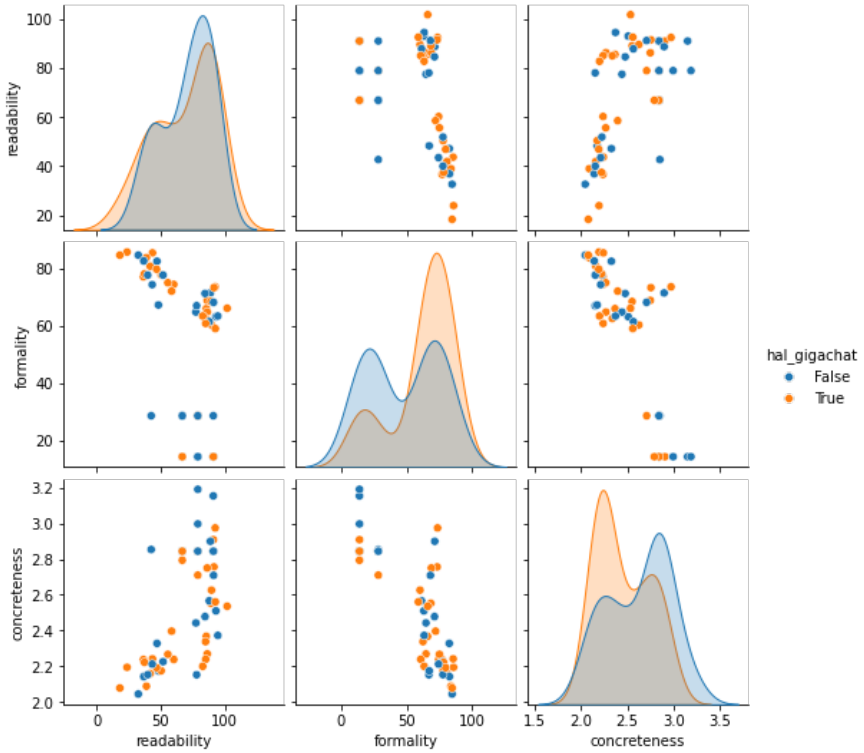

Рис. 10: Попарные распределения метрик с отображением наличия галлюцинаций. Модель GigaChat (Сбер)

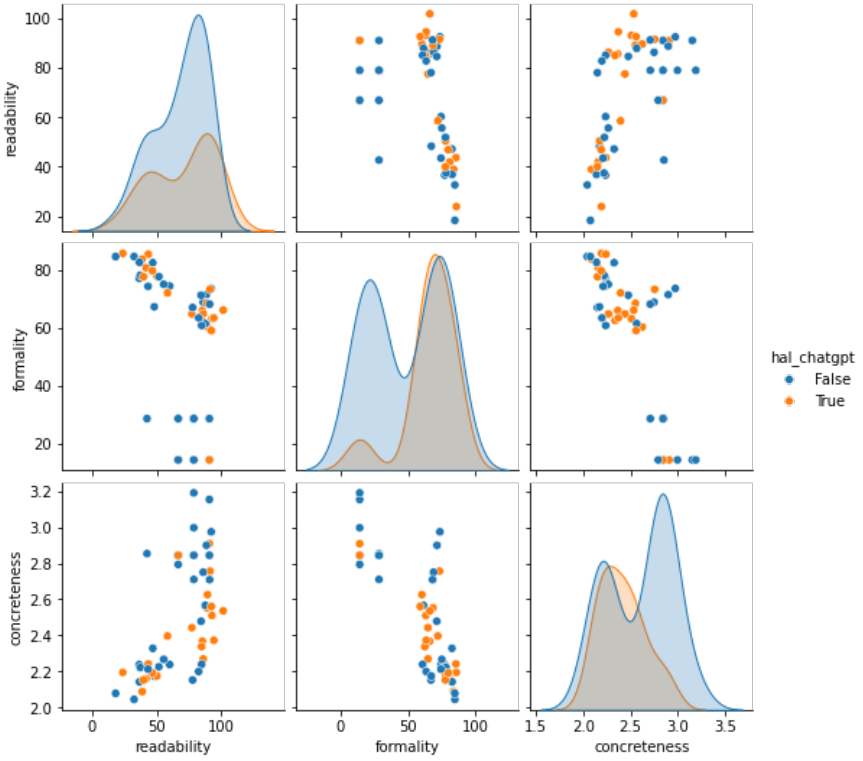

Рис. 11: Попарные распределения метрик с отображением наличия галлюцинаций. Модель ChatGPT (OpenAI)

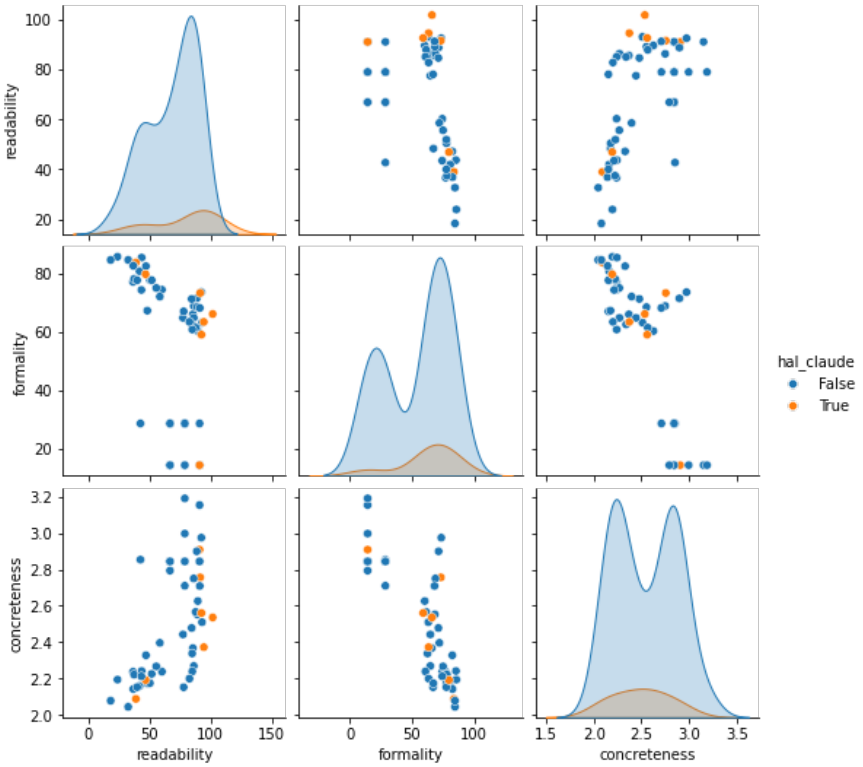

Рис. 12: Попарные распределения метрик с отображением наличия галлюцинаций. Модель Claude (Anthropic)

На первый взгляд, результаты экмперимента получились весьма неоднозначными, но если присмотреться к рисункам, можно заметить зависимость вероятности галлюцинаций от формальности и конкретности заданных запросов.

Модель Claude галлюцинирует меньше других, рассмотренных в эксперименте, и поэтому на её рисунке едва ли видны какие-либо зависимости. Но если взять GigaChat или ChatGPT, то видно, что **с увеличением "конкретности"** запроса вероятность получить несгаллюцинированный ответ резко повышается. Как и видно, что запросы, заданные **в менее формальной манере**, как ни странно, **меньше подвержены галлюцинациям** со стороны выбранных моделей. Что касается метрики удобочитаемости - в результате эксперимента не было выявлено какой-либо зависимости вероятности галлюцинаций от значения этой метрики.

Список литературы

[1] Vipula Rawte1, Prachi Priya, S.M Towhidul Islam Tonmoy, S M Mehedi Zaman, Amit Sheth, Amitava Das. Exploring the Relationship between LLM Hallucinations and Prompt Linguistic Nuances. *arXiv:2309.11064v1*, 2023.

[2] R. Flesch. *A new readability yardstick journal of applied psychology, 32: 221–233*, 1948.

[3] Francis Heylighen, Jean-Marc Dewaele. Formality of Language: definition, measurement and behavioral determinants. *Internal Report, Center "Leo Apostel Free University of Brussels*, 1999.

[4] Ellie Pavlick, Joel Tetreault. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics, 4:61–74*, 2016.

[5] Eduard Hovy. Generating natural language under pragmatic constraints. *Journal of Pragmatics, 11(6):689–719*, 1987.

[6] Allan Paivio. Dual coding theory, word abstractness, and emotion: a critical review of kousta et al, 2011.

[7] Marc Brysbaert, Amy Beth Warriner, Victor Kuperman. Concreteness ratings for 40 thousand generally known English word lemmas. *Psychonomic Society, Inc*, 2013.