Методы снижения галлюцинаций в больших языковых моделях

Oganyan Robert
High School of Economics
Nizhny Novgorod
ogrobertino2@gmail.com

Abstract

На фоне значительных успехов языковых моделей, таких как GPT-3, выявляется проблема галлюцинаций, когда модель генерирует содержание, содержащее фактически недостоверный ответ. Эта статья представляет исследование методов снижения галлюцинаций в контексте больших языковых моделей. Мы рассматривем три существущих подхода к снижению галлюцинаций у языковых моделей и проводим эксперименты на различных типах запросов. Полученные результаты открывают новые перспективы для улучшения надежности и точности крупных языковых моделей в различных приложениях.

Keywords галлюцинации · большие языковые модели

1 Вступление

С внушительными достижениями в области языкового моделирования, представленными современными глубокими нейронными сетями, такими как GPT-3, встает вопрос о доверии и надежности сгенерированных ими текстов. Одной из основных проблем, выделяющихся в данном контексте, является явление галлюцинаций.Галлюцинации в больших языковых моделях – ситуации, когда модель генерирует бессмыслицу или фактически недостоверный ответ. Можно выделить 2 типа галлюцинаций: внутренние (intrinsic) и внешние(extrinsic).

1. Внутренние. Сгенерированный вывод, который противоречит исходному содержимому. Например, “The first Ebola vaccine was approved in 2021” противоречит исходному содержанию “The first vaccine for Ebola was approved by the FDA in 2019.”

2. Внешние. Сгенерированные выходные данные, которые невозможно проверить по исходному содержимому (т. е. выходные данные, которые не могут быть ни подтверждены, ни опровергнуты источником). Например, сгенерированный ответ «Китай уже начал клинические испытания вакцины против COVID-19» не упоминается в обучающих данных. То есть мы не можем найти ни доказательства сгенерированного вывода в источниках, ни утверждать, что он неверен. Примечательно, что внешняя галлюцинация не всегда ошибочна, поскольку может возникнуть на основе фактически правильной внешней информации. Такая фактическая галлюцинация может быть полезной, поскольку она напоминает о дополнительных фоновых знаниях и повышает информативность сгенерированного текста. Однако в большей части литературы к внешним галлюцинациям по-прежнему относятся с осторожностью, поскольку непроверяемый аспект этой дополнительной информации увеличивает риск с точки зрения фактической безопасности.

В данной статье мы рассмотрим три способа снижения внутренних галлюцинаций больших языковых моделей.

2 Данные

В качестве данных используется небольшой датасет[1], состоящий из 19 запросов перевода английского текста на русский и 20 запросов объяснения смысла термина. Пример промптов:

Q: Translate the next paragraphs into Russian. Be as precise in terminology as possible. Everyone has the right to life, liberty and security of person.

A (ground-truth):Каждый человек имеет право на жизнь, на свободу и на личную неприкосновенность.

Q:Can you explain what is ology?

A: Ology is the study of a particular subject or field of knowledge. It usually involves research and analysis to gain a deeper understanding of that subject.

3 Способы снижения галлюцинаций

3.1 Изменение температуры

В области обработки естественного языка (NLP), температура используется в контексте генерации текста. Это такой гиперпараметр, регулирующий степень разнообразия и случайности в генерируемом тексте. Вероятность сгенерировать i-ый токен на текущем шаге вычисляется следующим образом:

$$p'_i = \frac{\exp\left(\frac{p_i}{T}\right)}{\sum_j \exp\left(\frac{p_j}{T}\right)}$$

Температура принимает значения от 0 до бесконечности. Разные значения температуры приводят к разным результатам:

1. Низкая температура (близкая к 0):
   - Модель становится более уверенной в своих предсказаниях.
   - Вероятности слов смещаются в сторону наиболее вероятных токенов, делая генерацию менее разнообразной и более детерминированной.
   - Менее вероятные варианты слова подавляются, что приводит к более жесткой генерации текста.
2. Высокая температура (близкая к бесконечности)
   - Распределение вероятностей становится более равномерным.
   - Вероятности всех слов уравниваются, что делает генерацию более разнообразной и менее предсказуемой.
   - Модель становится менее уверенной в выборе наиболее вероятных слов, и вероятности менее вероятных вариантов увеличиваются.

Влияние температуры на распределение вероятностей токенов можно увидеть на Рис.1.
Другими словами, регулирование температуры меняет баланс таких свойств языковой модели, как разнообразие и осмысленность.

- Разнообразие (Diversity) - способность модели генерировать различные, непохожие фрагменты текста.
- Осмысленность (Coherence) - способность модели генерировать текст, который логически связан и имеет смысл

Использование температуры позволяет регулировать этот баланс, формируя характер генерации текста от более предсказуемого и логичного до более разнообразного и креативного. Мы исследуем влияние температуры в этой статье.

3.2 Factual-nucleus sampling

Существуют различные способы сэмплирования токенов в процессе генерации текста. В противовес жадной стратегии, где модель просто выбирает наиболее вероятный следующий токен, и стандартному

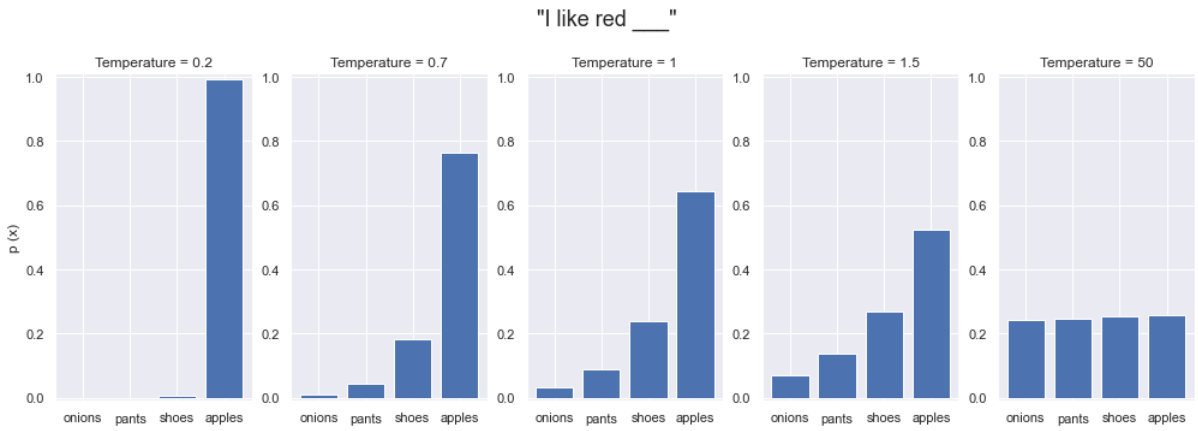

Рис. 1: Влияние температуры на распределение вероятностей токенов

случайному сэмплированию, которое выбирает токен с разной вероятностью из всего словаря, существует подход, известный как "Nucleus sampling"(или top-p sampling). Принцип работы nucleus sampling заключается в определении динамического набора вероятных кандидатов для следующего слова на основе кумулятивных вероятностей. В качестве кандидатов используются токены, представляющее собой наименьший набор слов, сумма вероятностей которых превышает заранее заданный порог (гиперпараметр).
Авторы статьи [2] предлагают новый алгоритм сэмплирования на основе nucleus sampling, который достигает лучшего баланса между качеством генерации и фактичностью, чем существующие алгоритмы декодирования. Авторы утверждают, что "случайность"при сэмплировании оказывает более вредное воздействие на фактичность, когда она используется для создания заключительной части предложения, чем в начале предложения. Поскольку в начале предложения нет предшествующего текста, для языковой модели безопасно генерировать что угодно, лишь бы это было грамматически правильным и соответствующим контексту. Однако по мере продвижения в генерации предпосылка становится более определенной, и меньше выборов слов могут сделать предложение фактическим. Поэтому они представляют алгоритм сэмплирования factual-nucleus sampling, который динамически адаптирует "ядерную"(nucleus) вероятность p_t в течение генерации каждого предложения:

$$p_t = \max\{\omega, p \times \lambda^{t-1}\}$$

Объяснение гиперпараметров:

- λ-decay: Учитывая, что набор для сэмплирования top-p выбирается как набор подслов, суммарная вероятность которых превышает значение p, мы постепенно уменьшаем значение p с помощью коэффициентом затухания λ на каждом шаге генерации для уменьшения "случайности"во времени.
- p-reset: Вероятность ядра p может быстро уменьшиться до небольшого значения после длительной генерации. Поэтому мы сбрасываем значение p к значению по умолчанию в начале каждого нового предложения в генерации (начало нового предложения определяется проверкой, был ли предыдущий шаг завершен точкой). Это снижает избыточные затраты на разнообразие для длительных генераций.
- ω-bound Если применять только λ-затухание, значение p может стать слишком маленьким и эквивалентным жадному декодированию, что может навредить разнообразию. Для преодоления этого мы вводим нижний предел ω-bound, чтобы ограничить, насколько сильно может уменьшаться значение p.

В результате экспериментов авторы заключают, что оптимальные значения гиперпараметров: $\lambda = 0.9, p = 0.9, \omega = 0.3$

3.3 Detecting and Mitigating Hallucinations of LLMs by Validating Low-Confidence Generation

Авторы статьи[3] предлагают следующий алгоритм действий для обнаружения и устранения галлюцинаций:

1. IDENTIFY KEY CONCEPTS. Для начала выявляют кандидатов на потенциальную галлюцинацию, то есть «важных понятий» из сгенерированного предложения. Проверять весь ответ на галлюцинацию неправильно, так как текст содержит множество аспектов, которые нельзя проверить сразу. Поэтому надо из сгенерированного ответа выделить несколько концепций. То есть, концепция (важное понятие) — это, по сути, фрагмент текста, состоящий из одного или нескольких токенов. Находить эти важные понятия можно разными способами:
   - Использование других специально обученных моделей для обнаружения концепций. Например, Keyphrase Boundary Infilling with Replacement[4] – трансформер на основе roBERT, умеющий выделять важные понятия.
   - Инструктирование модели: поскольку современные LLM прекрасно работают в широком диапазоне задач, в этом методе мы напрямую инструктируем модель идентифицировать важные понятия из сгенерированного предложение.

   Авторе статьи провели исследование и выяснили, что инструктирование модели работает лучше.
2. CALCULATE MODEL'S UNCERTAINTY. Далее среди обнаруженных концепций надо выбрать те, в которых модель меньше всего уверена. Каждое слово модель генерировала с какой-то вероятностью (напомню, что мы кормим в модель контекст и модель выдает нам распределение вероятностей слов), на основе вероятностей слов из концепции можем посчитать 3 метрики:
   - Среднее вероятностей
   - Нормированное произведение вероятностей
   - Минимальная вероятность

   Авторы утверждают, что метрика минимума подходит лучше всего. Однако, если нет доступа к распределению логитов, рассматривать можно все концепции.
3. CREATE VALIDATION QUESTION. Теперь на основе концепций надо сгенерировать проверяющие вопросы. Можно использовать либо другие question generation models, либо делать промпт к нашей же модели. Можно генерировать как ДА/НЕТ вопрос, так и специальные вопросы (w/h questions). Авторы статьи утверждают, что лучше просто генерить ДА/НЕТ, так как на них ответить проще.
4. FIND RELEVANT KNOWLEDGE. Следующим шагом нам нужно найти подходящую информацию в интернете для данного вопроса и создать промпт.
5. ANSWER VALIDATION QUESTION REPAIR HALLUCINATED SENTENCE. С помощью нового запроса просим модель с учетом полученных знаний ответить на вопрос еще раз.

Для рассматриваемых нами данных мы немного изменили алгоритм.
Для задачи объяснения смысла мы создаем новый запрос вида
Q: *Соответствующие знания*. Using provided evidence above, *первоначальный запрос*

В качестве соответствующих знаний были взяты:

- summary у Википедии[5]
- информация о термине из freeDictionary[6]

Пример такого нового запроса, в котором сочетается информация из Википедии и freeDictionary:

Q:Here are some ground truth information about billon:
Billon may refer to:
Billon (alloy), a metal alloy containing mostly copper or bronze with small quantity of silverBillon () is an alloy of a precious metal (most commonly silver, but also gold) with a majority base metal content (such as copper). It is used chiefly for making coins, medals, and token coins. The word comes from the French bille, which means "log". Any of several alloys of precious metals and base metals that are used (where legal) to make coins, medals etc. Using provided evidence above, Can you explain what is billon?

Для задачи перевода алгоритм выглядит следующим образом:

1. Написание исходного запроса
2. Выделение важных фраз из ответа модели
3. Поиск релевантных знаний в интернете и создание соответствующего промпта

Выделение концепций происходит так же, как и в первоначальной статье - с помощью инструктирования языковой модели. Соответствующие знания состоят из перевода ключевой фразы и контекстов, в которых она встречается. Эта информация извлекается с помощью сервиса Reverso[7]. Пример диалога:

Q: Translate the next paragraphs into Russian. Be as precise in terminology as possible. Everyone has the right to life, liberty and security of person
A: Каждый человек имеет право на жизнь, свободу и безопасность личности.
Q: Identify all the important keyphrases from the above sentence in english and return a comma separated list in quotes
A: right to life, liberty, security of person
Q: Here are some ground truth information about translations: right to life translates as право на жизнь; liberty translates as свобода; security of person translates as личную неприкосновенность. Based on provided evidence above, Translate the next paragraphs into Russian. Be as precise in terminology as possible. Everyone has the right to life, liberty and security of person
A: Каждый человек имеет право на жизнь, свободу и личную неприкосновенность

4 Эксперименты

4.1 Результаты

Исследования проводились для следующих моделей:

1. ChatGPT [WEB, API]
2. Claude 2 [WEB]
3. Gigachat [WEB]
4. Llama 2, 3B parameters [Collab]

Для задачи объяснения смысла:

Таблица 1: Верно/Не Верно для задачи объяснения смысла

Модель	Верно	Не верно
GigaChat	14	6
ChatGPT	19	1
Claude2	19	1
llamav2	6	14
GigaChat_stitch	14	6
ChatGPT_stitch	20	0
Claude_stitch	20	0
llamav2_3B_temp	4	16
llamav2_3B_factual	5	15
ChatGPT_temp	20	0
ChatGPT_factual	20	0

- Третий способ починил gpt и claude, однако gigachat оказался слишком слабой моделью. Ответ модели никак не менялся после написания запроса с релевантными знаниями.
- Первый и второй способ починили gpt.
- Первый и второй способы для llama некоторые галлюцинации устранили, но сгаллюционировали в других запросах

Что касается задачи перевода,

- Третий способ починил gpt и claude, однако gigachat оказался слишком слабой моделью. Ответ модели никак не менялся после написания запроса с релевантными знаниями.
- Первый способ починил один запрос и сгаллюцинировал в другом.
- Второй способ плохо работает с chatgpt-api. Скорее всего разработчики не подразумевали генерацию по одному токену. Тем более на русском языке, где единица измерения это буква (в отличие от английского, там - слово и буквы).

Таблица 2: Верно/Не Верно для задачи перевода

Модель	Верно	Не верно
GigaChat	7	12
ChatGPT	12	7
Claude2	17	2
GigaChat_stitch	7	12
ChatGPT_stitch	19	0
Claude_stitch	19	0
ChatGPT_temp	12	7
ChatGPT_factual	9	10

4.2 Примеры

Примеры есть в репозитории[8] в данных и в презентации. TODO: вынести сюда красивые примеры.

5 Заключение

В данной работе были рассмотрены 3 способа снижения галлюцинаций в больших языковых моделях. Результаты показали, что метод с поиском релевантных знаний чинит 100 процетов галлюцинаций. Изменение температуры и динамическое ядерное семплирование также могут убрать галлюцинации, но могут и создать новые. Надеюсь, что результаты данного исследования по снижению галлюцинаций в языковых моделях окажутся ценным вкладом в разработку более точных и надежных алгоритмов генерации текста.

6 Список литературы

1. https://github.com/TroninDV, Данные
2. Lee N. et al. Factuality enhanced language models for open-ended text generation //Advances in Neural Information Processing Systems. – 2022. – Т. 35. – С. 34586-34599.
3. Varshney N. et al. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation //arXiv preprint arXiv:2307.03987. – 2023.
4. https://huggingface.co/bloomberg/KBIR, KBIR
5. https://www.mediawiki.org/wiki/API:Main$_page$
6. https://dictionaryapi.dev/
7. https://www.reverso.net/text-translation
8. https://github.com/OganyanRV/LLM$_Hallucinations_Reducing, project repository$