Министерство науки и высшего образования Российской Федерации
Московский физико-технический институт
(национальный исследовательский университет)
Физтех-школа прикладной математики и информатики
Магистерская программа
«Методы и технологии искусственного интеллекта»

Олисов Валерий Александрович

White-box и Black-box подходы для оценки галлюцинаций языковых моделей

Обзор исследований

Научный руководитель:
Коновалов Василий Павлович

Москва

2024 г.

Содержание

1 Введение **3**

2 Цели и задачи работы **3**

2.1 Особенности архитектуры модели и обучения 4

2.2 Этап предварительного обучения 4

2.3 Этап дообучения 4

3 Обзор методов оценки галлюцинаций LLM **6**

3.1 Постановка задачи 6

3.2 White-box методы 7

3.2.1 Методы на основе энтропии 7

3.2.2 Методы на основе ансамблей 8

3.2.3 Метод самопроверки модели для оценки неопределенности . . . 9

3.3 Black-box методы 10

3.3.1 Методы, использующие матрицу схожести 10

3.3.2 Оценка достоверности на основе дополнения знаний 11

3.3.3 Методы обнаружения галлюцинаций на основе сэмплирования . 13

4 Заключение **15**

1 Введение

Современные большие языковые модели (LLM) достигли отличных результатов в решении множества задач, включая генерацию текста, диалоговые системы и автоматический перевод. Они открывают новые горизонты для приложений в самых разных областях, от науки до индустрии. Однако, несмотря на их достижения, остаётся серьёзная проблема: склонность моделей к "галлюцинациям то есть генерации недостоверного, несогласованного или вымышленного контента.

В контексте этой работы под галлюцинацией понимаются фактические ошибки, когда выход модели не основывается ни на предоставленном контексте, ни на знаниях о мире. Галлюцинации можно разделить на два основных типа:

- In-context галлюцинации — ошибки, связанные с несоответствием исходному контексту.
- Extrinsic галлюцинации — ошибки, возникающие из-за отсутствия привязки к данным предварительного обучения или знаниям о мире.

Обзор направлен на изучение методов оценки фактической достоверности (Extrinstic) генераций LLM. Рассматриваются как классические подходы к оценке неопределённости (Uncertainty Estimation), так и современные методы, включающие использование Retrieval-Augmented Generation (RAG) и крупных LLM для проверки качества генераций.

2 Цели и задачи работы

Галлюцинации в LLM представляют собой генерацию контента, который не соответствует знаниям о реальном мире, или заданному контексту. Такие явления могут включать вымышленные утверждения, логические несоответствия или некорректные выводы. Галлюцинации неизбежны в работе современных моделей, поскольку они опираются на вероятностные подходы для генерации текста и обучаются на больших, часто неоднородных корпусах данных.

Цель данной работы — изучение природы галлюцинаций в LLM и разработка методов для их детекции. Несмотря на значительные успехи в генерации текстов, модели продолжают генерировать информацию с фактическими ошибками. Эти ошибки могут быть обусловлены как архитектурными особенностями моделей, так и этапами

их обучения. Учитывая, что полностью устранить галлюцинации невозможно, задача их обнаружения становится одной из ключевых в области разработки надежных языковых моделей.

Рассмотрим некоторые причины, обуславливающие неизбежность возникновения галлюцинаций LLM.

2.1 Особенности архитектуры модели и обучения

LLM используют вероятностные подходы для генерации текста, где каждое следующее слово выбирается на основе распределения вероятностей. Это может приводить к:

- Ошибкам обобщения, особенно в случаях, когда модель не имеет достаточного контекста или сталкивается с вопросами за пределами своих знаний.
- Неспособности корректно учитывать семантические связи между токенами, что усложняет создание точных и последовательных ответов.

Цель модели формализуется как максимизация условной вероятности следующего токена, что означает стремление модели предсказывать наиболее вероятное продолжение текста на основе предыдущих токенов. Однако такой подход никак не учитывает риск появления галлюцинаций.

2.2 Этап предварительного обучения

LLM обучаются на огромных корпусах данных, представляющих знания о языке и мире, включая информацию из открытых источников в Интернете. Однако такие данные часто теряют актуальность, могут содержать ошибки, быть неполными или противоречивыми. Эти недостатки приводят к галлюцинациям, поскольку LLM в ходе предобучения стремятся максимизировать правдоподобие своих генераций без учета их точности. Более того, объём данных настолько велик, что невозможно полноценно проверять их на достоверность и актуальность, что увеличивает риск генерации недостоверной информации.

2.3 Этап дообучения

Дообучение с использованием новых данных, например, в рамках этапа Supervisef Fine-Tuning или Alignment, может усиливать проблему галлюцинаций. Ис-

следования Gekhman et al. 2024 [1] показывают, что:

1. Модели усваивают новые знания медленнее, если они противоречат уже существующим.

2. Обучение на неизвестных данных увеличивает вероятность генерации галлюцинаций.

Особенно важно учитывать, что лучший результат достигается, когда модели обучаются на большинстве “известных” данных, избегая глубокого погружения в неизвестные области, которые увеличивают риск ошибок. Важно также понимать, что RLHF может непреднамеренно усиливать уверенность модели в её ошибочных ответах, если эти ответы воспринимаются как удовлетворяющие намерения пользователя.

На графике 1 показано, как в процессе усвоения моделью новых знаний в ходе fine-tuning меняется точность её ответов на валидационной выборке.

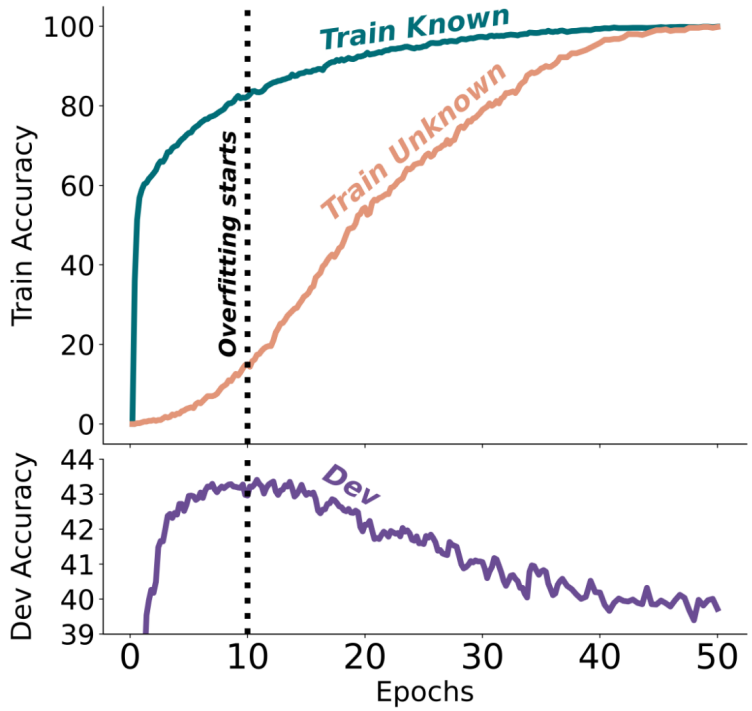

Рис. 1: Точность ответов модели в зависимости от продолжительности дообучения при использовании в дообучающей выборке 50% известных и 50% неизвестных примеров.

Таким образом, данная работа направлена на изучение природы галлюцинаций, разработку методов их детекции и оценку существующих подходов, чтобы увеличить качество текстов, генерируемых LLM. В последующих главах будут рассмотрены:

- White-box методы Uncertainty Estimation, применяемые к моделям текстовой генерации для детекции галлюцинаций.
- Black-box подходы, включая техники RAG и использование крупных LLM для проверки точности генераций.

3 Обзор методов оценки галлюцинаций LLM

Методы оценки неопределенности в LLM играют ключевую роль в анализе надежности их предсказаний. В данном разделе будут рассмотрены существующие подходы, их особенности, преимущества и ограничения.

3.1 Постановка задачи

В рамках оценки неопределенности формализуется следующая задача: пусть имеется входная последовательность $X = [x_1, x_2, \ldots, x_n]$, а языковая модель f с параметрами w генерирует выходную последовательность $Y = [y_1, y_2, \ldots, y_m]$ с использованием авторегрессионного декодирования:

$$y_j = f([X, y_1, y_2, \ldots, y_{j-1}] \mid w), \tag{3.1}$$

где y_j — j-й сгенерированный токен. Метод оценки неопределенности g предназначен для вычисления значения u, которое отражает степень неопределенности предсказания:

$$u = g(f(\cdot), X, Y, P), \tag{3.2}$$

где P — распределение вероятностей или top-k вероятности для каждого токена.

Формально вероятность генерации последовательности y для авторегрессионных моделей задаётся как:

$$P(y \mid x, \theta) = \prod_{l=1}^{L} P(y_l \mid y_{<l}, x, \theta), \tag{3.3}$$

где $y_{<l} = \{y_1, \ldots, y_{l-1}\}$, а θ обозначает параметры модели.

Дальнейший текст посвящён детальному описанию каждого из подходов, их применению в современных LLM и оценке их эффективности.

Методы для оценки неопределенности в LLM можно разделить на 2 принципиальные группы методов:

1. **White-box методы** предполагают доступ к внутренним структурам модели, таким как логиты, промежуточные представления слоёв или параметры модели. Эти методы позволяют проводить глубокий анализ модели, но зачастую требуют значительных вычислительных ресурсов.

2. **Black-box методы** работают только с текстами, сгенерированными моделью, и могут быть легко интегрированы с API сторонних сервисов, например, OpenAI. Такие методы ограничены в доступе к информации о модели, но обладают большей практической применимостью благодаря низким требованиям к доступу.

Эти подходы различаются также по вычислительным затратам. Например, некоторые методы накладывают существенные ограничения из-за необходимости повторных запусков модели или использования больших объёмов памяти, что делает их менее подходящими для реального применения.

3.2 White-box методы

В данном разделе описываются white-box методы оценки галлюцинаций LLM. Эти методы предполагают доступ к внутренним структурам модели, включая вероятности токенов $P(y_l \mid y_{<l}, x, \theta)$ и последовательностей $P(y \mid x, \theta)$, логитам, а также промежуточным представлениям слоёв. Рассмотрим основные категории таких методов.

3.2.1 Методы на основе энтропии

Суть методов, ииспользующих энтропию, заключается в использовании этими методами вероятностей предсказанных токенов или последовательностей. Энтропия может быть вычислена как на уровне токенов, так и на уровне последовательностей. Энтропия определяется следующим образом:

$$H(Y) = -\sum_{y \in Y} P(y) \log P(y),$$

где $P(y)$ — вероятность сгенерированного токена или последовательности.

Преимуществами таких методов являются их низкая вычислительная сложность и простота реализации. Однако качество таких методов зачастую уступает

более сложным подходам, поэтому они чаще всего используются в качестве baseline-моделей. В последние годы были предложены улучшения методов на основе энтропии, такие как семантическая энтропия [2].

Семантическая энтропия, представленная Kuhn et al. (2023), является методом измерения неопределённости в LLM с учётом феномена семантической эквивалентности — ситуации, когда разные предложения передают одно и то же значение. Традиционные меры неопределённости часто упускают этот аспект, что приводит к неточным оценкам. Семантическая энтропия решает эту проблему, группируя семантически эквивалентные выходы и вычисляя энтропию по этим кластерам, что обеспечивает более точное представление неопределённости модели относительно значений, а не конкретных последовательностей слов.

Опишем расчет семантической энтропии, согласно данной работе. Пусть дан входной контекст x и множество сгенерированных последовательностей $\{s^{(1)}, s^{(2)}, \ldots, s^{(M)}\}$. Эти последовательности разделяются на кластеры $\{c_1, c_2, \ldots, c_K\}$, где каждый кластер c_k представляет уникальное значение. Семантическая энтропия H_{semantic} вычисляется как:

$$H_{\text{semantic}} = -\sum_{k=1}^{K} P(c_k \mid x) \log P(c_k \mid x), \tag{3.4}$$

где $P(c_k \mid x)$ - агрегированная вероятность всех последовательностей в кластере c_k:

$$P(c_k \mid x) = \sum_{s \in c_k} P(s \mid x). \tag{3.5}$$

3.2.2 Методы на основе ансамблей

Основная идея метода заключается в построении ансамбля из M независимых моделей, каждая из которых обучается на одном и том же наборе данных $D = \{(x_i, y_i)\}_{i=1}^{N}$, но с различными инициализациями параметров и (опционально) различными подвыборками данных. Пусть $f_\theta(x)$ — предсказание модели с параметрами θ для входного объекта x. Тогда ансамбль моделей представляет собой набор функций $\{f_{\theta_m}(x)\}_{m=1}^{M}$, где θ_m — параметры m-й модели.

Для входного объекта x предсказания ансамбля объединяются с использованием усреднения вероятностных распределений:

$$P(y \mid x, D) = \frac{1}{M} \sum_{m=1}^{M} P(y \mid x, \theta_m), \tag{3.6}$$

где $P(y \mid x, \theta_m)$ — вероятностное распределение, выдаваемое m-й моделью ансамбля. Это усреднение позволяет учесть эпистемическую неопределенность, связанную с вариативностью предсказаний различных моделей.

Представителем этого класса методов является метод Deep Ensembles, описанный в работе Lakshminarayanan et al. (2017) [3], основанный на построении и использовании нескольких независимых моделей. Этот метод обеспечивает возможность учета эпистемической неопределенности (неопределенности, обусловленной недостатками модели) и демонстрирует простоту реализации в сочетании с высокой точностью предсказаний.

Каждая модель в ансамбле обучается независимо с использованием метода максимального правдоподобия. Пусть $p(y \mid f_\theta(x))$ — распределение целевой переменной y, параметризованное предсказанием $f_\theta(x)$. Тогда функция потерь для одной модели имеет вид:

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \log p(y_i \mid f_\theta(x_i)). \tag{3.7}$$

Для регрессионных задач, где целевая переменная распределена нормально $y \sim \mathcal{N}(\mu(x), \sigma^2(x))$, функция потерь принимает вид:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{(y_i - \mu_\theta(x_i))^2}{2\sigma_\theta^2(x_i)} + \log \sigma_\theta(x_i) \right]. \tag{3.8}$$

Deep Ensembles демонстрируют конкурентоспособные результаты по сравнению с более сложными методами оценки неопределенности, однако обучение и инференс для нескольких моделей требуют значительных ресурсов.

3.2.3 Метод самопроверки модели для оценки неопределенности

Метод самопроверки модели, рассмотренный в Kadavath et al.(2022) [4] используется для оценки неопределенности, измеряя вероятность правильности сгенерированного ответа при помощи самой модели.

Идея метода заключается в том, что модель сначала генерирует ответ на заданный запрос, а затем выполняет дополнительный шаг инференса, в рамках которого оценивает вероятность $P(\text{True})$ корректности своего ответа. Это позволяет модели прямо оценивать уровень уверенности в своих предсказаниях, основываясь на собственных вероятностных предположениях.

Рассмотрим процесс более детально: модель предсказывает последовательность токенов $y = \{y_1, y_2, \ldots, y_L\}$ на основе входного запроса x. Вероятность генерации последовательности y выражается как:

$$P(y \mid x, \theta) = \prod_{l=1}^{L} P(y_l \mid y_{<l}, x, \theta), \tag{3.9}$$

где $y_{<l} = \{y_1, y_2, \ldots, y_{l-1}\}$ — предыдущие токены, а θ — параметры модели.

После генерации ответа y модель оценивает вероятность его истинности. Для этого она принимает на вход пару (x, y) и вычисляет вероятность $P(\text{True} \mid y, x, \theta)$. Эта вероятность может быть получена различными способами, включая использование дополнительного классификатора или повторного инференса с модифицированным запросом.

Метод самопроверки модели предоставляет эффективный инструмент для оценки неопределенности в её предсказаниях. Несмотря на дополнительные вычислительные затраты, этот подход способствует повышению достоверности и интерпретируемости результатов, что особенно важно в критически значимых приложениях.

3.3 Black-box методы

В данном разделе описываются black-box методы оценки неопределенности. Эти методы предполагают, что доступ к архитектуре модели, её скрытым состояниям или логитам отсутствует. Вместо этого, для анализа используются только сгенерированные ответы модели. Рассмотрим основные подходы к оценке неопределенности в рамках black-box методов.

3.3.1 Методы, использующие матрицу схожести

Такие методы основываются на семантическом разбросе сгенерированных ответов LLM. Например, такая идея рассматривается в работе Lin et al. (2023) [5] в рамках оценки галлюцинаций LLM в условиях ограниченного доступа к внутренним параметрам и архитектуре модели. Авторы различают понятия «неопределенность» и «уверенность»: первая характеризует разброс возможных предсказаний для заданного ввода, а вторая — степень уверенности в конкретном сгенерированном ответе.

1. Для заданного входного запроса x модель генерирует K различных ответов $\{y_1, y_2, \ldots, y_K\}$.

2. Строится матрица сходства $S \in \mathbb{R}^{K \times K}$, где каждый элемент $S_{ij} = s(y_i, y_j)$ представляет собой меру семантического сходства между ответами y_i и y_j. В качестве меры сходства может использоваться, например, оценка, основанная на естественном языковом выводе (*Natural Language Inference*, NLI).

3. На основе матрицы сходства S вычисляется средняя семантическая дисперсия, которая служит индикатором неопределенности модели относительно данного запроса.

Высокая семантическая дисперсия указывает на значительную вариативность в сгенерированных ответах, что свидетельствует о повышенной неопределенности модели. Этот метод не требует доступа к внутренним параметрам модели, таким как логиты или скрытые состояния, и может быть применен в условиях ограниченного доступа к модели.

Экспериментальные результаты, представленные в работе, показывают, что предложенная мера семантической дисперсии является надежным предиктором качества ответов, сгенерированных LLM. Этот подход предоставляет практикам эффективный инструмент для управления неопределенностью при использовании больших языковых моделей в различных приложениях.

3.3.2 Оценка достоверности на основе дополнения знаний

В работе Lee et al. (2022) [6] представлен подход для количественной оценки галлюцинаций языковых моделей с использованием эталонного набора данных **FactualityPrompt**. Этот набор данных содержит как фактические, так и нефактические запросы и использует Wikipedia в качестве базы знаний для проверки фактической корректности (*factuality grounding*). База знаний включает документы из датасета FEVER, проверенные как истинные, а также предложения, отобранные на основе сходства с помощью *tf-idf* или sentence embeddings.

Для количественного анализа ошибок модели используются две основные метрики:

- **Ошибки именованных сущностей (Hallucination NE errors)**. С помощью предварительно обученной модели для обнаружения именованных сущностей вычисляется доля сущностей, отсутствующих в эталонном документе. Данная

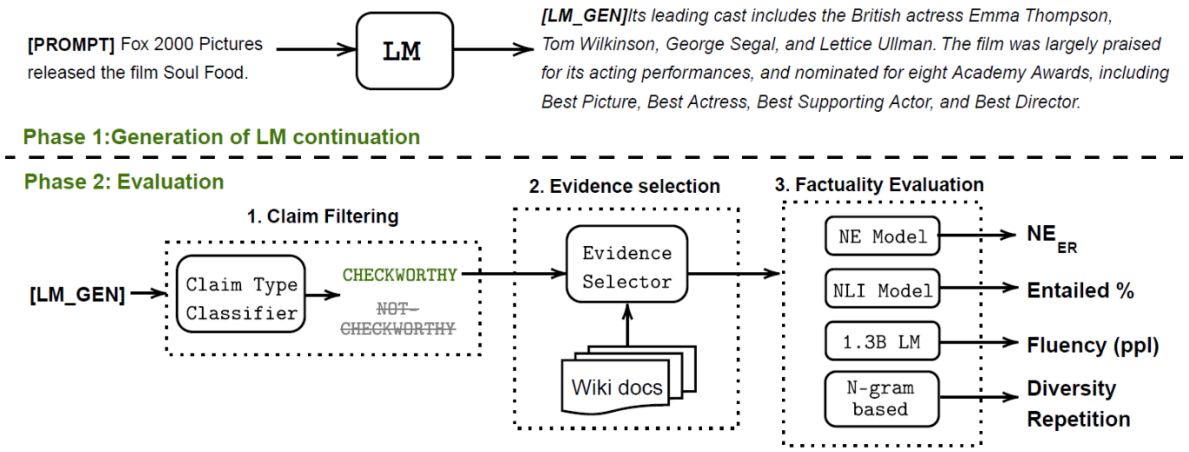

Рис. 2: Схема для оценки достоверности, предложенная авторами.

метрика выявляет фактические ошибки модели, связанные с некорректным использованием именованных сущностей. Формально:

$$\text{NE Error} = \frac{\text{Количество несоответствующих именованных сущностей}}{\text{Общее количество именованных сущностей в ответе}}.$$

- **Коэффициент следования (Entailment ratio)**. Для оценки степени релевантности сгенерированных моделью предложений используются RoBERTa, дообученная на MNLI, и механизм привязки знаний к предложениям Wikipedia. Коэффициент следования определяется как доля предложений, признанных соответствующими знаниям из базы:

$$\text{Entailment Ratio} = \frac{\text{Количество предложений с entailment}}{\text{Общее количество предложений}}.$$

Высокое значение NE ошибок и низкий коэффициент следования указывают на значительное количество галлюцинаций модели.

В ходе экспериментов в этой работе авторы сделали следующие выводы:

- Крупные языковые модели (например, GPT-3) показывают лучшие результаты на FactualityPrompt благодаря лучшему обобщению знаний и генерации более контекстно осмысленных ответов.

- Качество датасетов для обучения модели напрямую влияет на её способность избегать галлюцинаций, особенно в задачах с высокими требованиями к фактической корректности.

Для дальнейшего развития метода FactualityPrompt были предложены несколько подходов:

- **FActScore (Min et al., 2023)**: Длинный текст модели разбивается на атомарные факты, каждый из которых проверяется на соответствие базе знаний (например, Wikipedia). Итоговая метрика рассчитывается как средняя точность для всех проверенных фактов:

$$\text{FActScore} = \frac{\sum_{i=1}^{N} \text{Precision}(f_i)}{N},$$

где f_i — атомарный факт, а N — общее количество фактов.

- **SAFE (Wei et al., 2024)**: Использует пошаговый процесс запросов в поисковых системах для проверки фактов в длинных текстах. Итоговый показатель — $F1@K$, который учитывает точность (*precision*) и полноту (*recall*) предоставленных фактов:

$$F1@K = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}},$$

где K — количество проверяемых фактов.

Эти методы продемонстрировали высокую корреляцию с аннотациями людей, что подтверждает их пригодность для оценки галлюцинаций языковых моделей.

Методы, основанные на расширении базы знаний, позволяют с высокой точностью определять достоверность фактов в генерации модели, однако обычно имеют высокую вычислительную сложность.

3.3.3 Методы обнаружения галлюцинаций на основе сэмплирования

В работе Manakul et al. (2023) [7] был предложен эффективный метод для оценки галлюцинаций LLM, не требующий внешних ресурсов и применимый к любой LLM в режиме black-box. Данный метод направлен на обнаружение ошибок фактической достоверности в сгенерированных текстах и основан на проверке согласованности ответов модели при многократном сэмплировании.

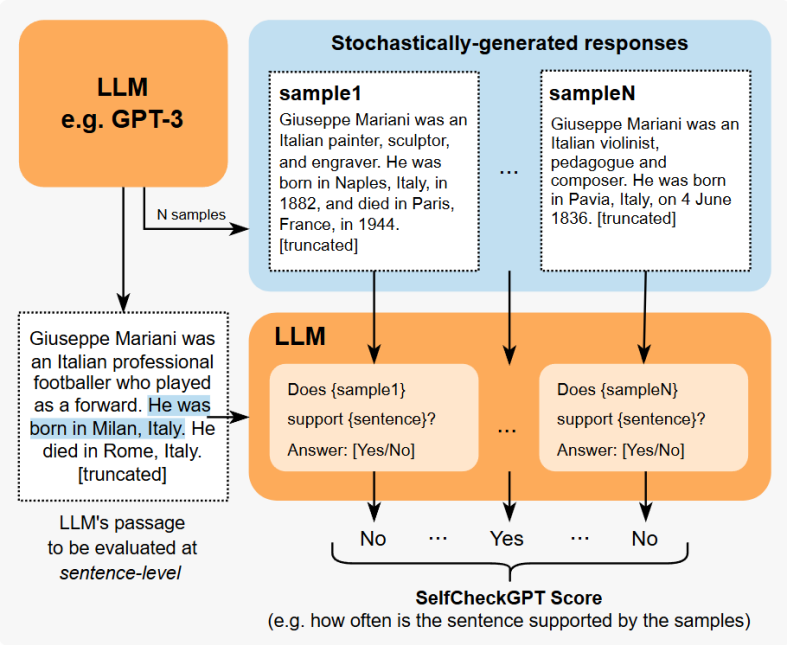

Рис. 3: SelfCheckGPT с оценкой Prompting

SelfCheckGPT основывается на проверке согласованности между различными ответами, сгенерированными моделью в условиях «черного ящика». В отличие от подходов «серого ящика», которые требуют доступа к логитам модели, данный метод использует исключительно результаты генерации. Более того, SelfCheckGPT не зависит от внешних баз знаний, что делает его применимым для языковых моделей с ограниченным доступом.

Суть метода заключается в следующем:

1. Для каждого входного запроса x модель сэмплирует K различных ответов $\{y_1, y_2, \ldots, y_K\}$ с использованием стохастического процесса генерации (например, температурного сэмплирования).

2. Для каждого ответа y_i оценивается согласованность с остальными сэмплами $\{y_j\}_{j \neq i}$ с использованием различных метрик, таких как BERTScore, NLI, Prompting(задаются вопросы с ответом "да"/"нет"для оценки достоверности фактов)

3. На основе согласованности между сэмплами вычисляется интегральный показатель достоверности. Высокая степень согласованности указывает на меньшую вероятность галлюцинаций.

В экспериментах на синтетическом корпусе WikiBio, сгенерированном с использованием GPT-3, SelfCheckGPT показал высокую эффективность в выявлении ошибок достоверности. Лучшую производительность продемонстрировал подход, основанный на метрике prompting, при котором модель задает дополнительные уточняющие вопросы к своим ответам.

Независимость этого метода от внешних данных и простота реализации делают его универсальным решением для задач, требующих обнаружения галлюцинаций и повышения качества генерации.

4 Заключение

В данном обзоре рассмотрены причины возникновения галлюцинаций в LLM и методы их оценки. Галлюцинации обусловлены архитектурными особенностями моделей, этапами обучения и ограничениями данных. White-box и black-box подходы к оценке достоверности предоставляют различные инструменты, от глубокого анализа модели до практичных решений для реальных приложений.

Современные методы, такие как семантическая энтропия, ансамбли и RAG, позволяют снизить вероятность ошибок, но требуют дальнейших исследований для повышения их эффективности и снижения вычислительных затрат. Развитие надежных и интерпретируемых решений станет ключом к безопасному и широкому применению в критически важных задачах.

Список литературы

[1] Zorik Gekhman и др. “Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations?” В: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*. Под ред. Yaser Al-Onaizan, Mohit Bansal и Yun-Nung Chen. Association for Computational Linguistics, 2024, с. 7765—7784. URL: https://aclanthology.org/2024.emnlp-main.444.

[2] Lorenz Kuhn, Yarin Gal и Sebastian Farquhar. “Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation”. В: *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL: https://openreview.net/forum?id=VD-AYtP0dve.

[3] Balaji Lakshminarayanan, Alexander Pritzel и Charles Blundell. “Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles”. В: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Под ред. Isabelle Guyon и др. 2017, с. 6402—6413. URL: https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

[4] Saurav Kadavath и др. “Language Models (Mostly) Know What They Know”. В: *CoRR* abs/2207.05221 (2022). DOI: 10.48550/ARXIV.2207.05221. arXiv: 2207.05221. URL: https://doi.org/10.48550/arXiv.2207.05221.

[5] Zhen Lin, Shubhendu Trivedi и Jimeng Sun. “Generating with Confidence: Uncertainty Quantification for Black-box Large Language Models”. В: *Trans. Mach. Learn. Res.* 2024 (2024). URL: https://openreview.net/forum?id=DWkJCSxKU5.

[6] Nayeon Lee и др. “Factuality Enhanced Language Models for Open-Ended Text Generation”. В: *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*. Под ред. Sanmi Koyejo и др. 2022. URL: http://papers.nips.cc/paper%5C_files/paper/2022/hash/df438caa36714f69277daa92d608dd63-Abstract-Conference.html.

[7] Potsawee Manakul, Adian Liusie и Mark J. F. Gales. “SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models”. В: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*. Под ред. Houda Bouamor, Juan Pino и Kalika Bali. Association for Computational Linguistics, 2023, с. 9004—9017. DOI: 10.18653/V1/2023.EMNLP-MAIN.557. URL: https://doi.org/10.18653/v1/2023.emnlp-main.557.