УДК 004.8

ПРОБЛЕМА ГАЛЛЮЦИНАЦИЙ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ И МЕТОДЫ ПОВЫШЕНИЯ ДОСТОВЕРНОСТИ ГЕНЕРИРУЕМЫХ ОТВЕТОВ

Красова А.Д., [1]

студент,

Поволжский государственный университет телекоммуникаций и информатики,

Самара, Россия

Аннотация

В статье рассматривается проблема недостоверной генерации ответов больших языковых моделей - галлюцинации. Цель статьи – рассмотреть причины возникновения галлюцинаций в больших языковых моделях и методы улучшения качества генерируемых ответов. Рассмотрены методы Retrieval-Augmented generation (RAG), проверка фактов, Chain-of-Verification, промпт-инжиниринг. В результате исследования сделан вывод: наиболее эффективно комплексное использование нескольких методов. Дальнейшее исследование и развитие технологий проверки достоверности информации является одним из ключевых направлений модернизации систем искусственного интеллекта.

Ключевые слова: большие языковые модели, искусственный интеллект, генеративный искусственный интеллект, галлюцинации, генерация, дополненная поиском (RAG), проверка фактов, цепочка верификации, промпт-инжиниринг.

[1] Научный руководитель: **Куваева Е.Н.**, ст. преподаватель, Поволжский государственный университет телекоммуникаций и информатики, Самара, Россия.
Kuvaeva E.N., Senior Lecturer, Povolzhskiy State University of Telecommunications and Informatics, Samara, Russia.

THE PROBLEM OF HALLUCINATIONS IN LARGE LANGUAGE MODELS AND METHODS FOR IMPROVING THE FACTUAL ACCURACY OF GENERATED RESPONSES

Krasova A. D.

Student,

Povolzhskiy State University of Telecommunications and Informatics,

Russia, Samara

Abstract

The article examines the problem of inaccurate response generation by large language models, known as hallucinations. The purpose of the article is to examine the causes of hallucinations in large language models and methods for improving the quality of generated responses. The methods of Retrieval-Augmented Generation (RAG), fact-checking, Chain-of-Verification, and prompt engineering are considered. As a result of the study, it was concluded that the combined use of several methods is the most effective approach. Further research and the development of technologies for verifying the reliability of information are among the key directions for the modernization of artificial intelligence systems.

Keywords: large language models, artificial intelligence, generative artificial intelligence, hallucinations, Retrieval-Augmented Generation (RAG), fact-checking, Chain-of-Verification, prompt engineering.

В последние годы большие языковые модели (LLM) все чаще стали использоваться не только для повседневных запросов, но и в учебной и профессиональной деятельности. С помощью языковых моделей пользователи могут быстро получить ответ на интересующий вопрос, не тратя время на самостоятельный поиск информации в нескольких источниках. Однако высокая скорость генерации не гарантирует достоверность ответов. Зачастую LLM может допускать фактические ошибки, хотя ответ может выглядеть правдоподобно.

Иногда предоставленная генерация контента оказывается выдуманной, не имеющей отношения к исходным данным или не соответствующей им. Такие генерации принято называть галлюцинациями LLM. [1]

Основными причинами возникновения галлюцинаций в больших языковых моделях являются особенности обучения и работы этих моделей. При формировании ответа LLM не определяет достоверность каждого утверждения напрямую. Основой генерации является предсказание следующего токена с учетом уже имеющегося контекста. В такой последовательности она не учитывает истинность предоставляемой информации, а ориентируется на вероятность появления определенного слова на основе данных, использованных в обучении. Из этого следует еще одна причина галлюцинаций: обучающие данные могут содержать недостоверную, неполную или устаревшую информацию. Усугубляет эту проблему отсутствие у модели доступа к актуальным внешним источникам. Кроме того, при недостатке контекста в пользовательском запросе модель может неправильно интерпретировать поставленную задачу. При этом современные LLM стремятся предоставить правдоподобный ответ даже при недостатке данных, что приводит к «выдумыванию» информации. [2]

Рассмотрим основные методы улучшения качества генерируемых ответов:

1. Retrieval-Augmented Generation (RAG).

Retrieval-Augmented Generation (RAG) – подход, при котором модель способна получать информацию из заранее подготовленных внешних источников при генерации ответа. Обычная языковая модель полагается только на то, что она выучила во время обучения. RAG же работает иначе: система ищет нужную информацию в базе документов и формирует ответ с учетом найденных материалов. Модель не просто пытается вспомнить всё по памяти, а работает с конкретными фактами.

Принцип работы RAG включает два основных этапа. Сначала система сканирует внешнюю базу знаний — это могут быть статьи, корпоративные

инструкции, научные работы или любые другие документы. Найденные фрагменты передаются модели в качестве контекста, и уже на их основе строится итоговый ответ.

За счет того, что модель может опираться на проверенные источники, использование RAG снижает вероятность галлюцинаций и значительно упрощает работу с узкоспециализированной информацией. Также этот метод позволяет добавлять новые источники данных вместо переобучения всей модели, что снижает затраты [3, 283]. Однако у метода есть и свои недостатки. Качество ответа сильно зависит от качества поиска. Не всегда возможно извлечь наиболее релевантную информацию из базы знаний, что приводит к неточности ответов. Также внедрение RAG усложняет систему: нужно настраивать инфраструктуру, следить за качеством поиска и выделять дополнительные вычислительные мощности [3, 285]

2. Проверка фактов (Fact Checking).

Проверка фактов (Fact Checking) – метод, направленный на обнаружение ошибок и неточностей в уже готовом ответе. Суть заключается в сопоставлении информации, предоставленной моделью, с внешними источниками и проверке ее достоверности.

Проверка может происходить как автоматически, так и вручную. В первом случае автоматическая проверка может выполняться с помощью специальных алгоритмов, выделяющих факты из ответа и сравнивающих их с информацией из баз знаний, поисковых систем или прочих структурированных источников данных. Если обнаруживается несовпадение, система помечает токен как ненадежный [4]. Особенно важно применение этого метода в задачах, где ошибка или неточность может привести к серьезным последствиям, например, в технической документации, медицине и тд. В некоторых системах, если первичный ответ содержит неточности, модель получает дополнительные контекст и повторно генерирует исправленный вариант [5].

Несмотря на очевидную полезность, проверка фактов не является идеальным решением. Процесс декомпозиции, верефикации и переработки ответа увеличивает количество используемых токенов [5] и время отклика системы. Также эффективность метода зависит от доступности внешних источников, качества алгоритмов извлечения фактов и корректности интерпретации результатов.

3. Chain-of-Verification.

Chain-of-Verification (CoVe) – метод, основанный на поэтапной самопроверке ответа большой языковой модели. Сначала модель формирует предварительный, базовый ответ на пользовательский вопрос. Затем, опираясь на запрос и первоначальный ответ, она составляет набор контрольных вопросов для проверки. После этого модель последовательно отвечает на эти вопросы, чтобы в конечном итоге подготовить улучшенный пересмотренный ответ.

Преимущество данного подхода заключается в том, что модель не просто генерирует текст, но и дополнительно анализирует собственные утверждения и проверяет их согласованность. Отдельные проверочные вопросы обычно позволяют получить более точные сведения, чем ответы, данные в первоначальной, неподтвержденной форме [6]. Однако эффективность метода зависит от способности самой модели планировать проверку и честно обнаруживать собственные ошибки. Также дополнительный этап проверки увеличивает вычислительные затраты и время получения ответа.

4. Промпт-инжиниринг.

Помимо выше перечисленных методов не стоит забывать и о грамотном составлении промпта. Правильно составленный запрос задает модели нужные ограничение и направляет ее к более осторожному стилю ответа. Например, в запросе можно указать, что модель должна указать источник информации, не выдумывать факты, а при отсутствии или неточности данных прямо отвечать, что она не владеет нужной информацией. Кроме того, можно указать, какую роль должна играть модель, в каком формате она должна сгенерировать ответ и задать

правила для генерации [7]. Такие инструкции могут уменьшить количество галлюцинаций, пусть и не устраняют их полностью.

Для наглядного сравнения основных подходов повышения достоверности ответов LLM была составлена таблица методов и их ключевых характеристик. В таблице представлены выше описанные методы, их сильные стороны и ограничения.

Таблица 1 – Сравнение методов повышения достоверности ответов LLM

Метод	**Преимущества**	**Недостатки**
RAG	Высокая точность, актуальность	Требует наличия внешней базы знаний
Проверка фактов	Уменьшает количество ошибок	Увеличивает время ответа
Chain-of-Verification	Снижает число галлюцинаций	Повышает вычислительные затраты
Промпт-инжиниринг	Простота использования	Зависит от качества запроса и пользователя

Анализ данных, представленных в таблице, позволяет сделать вывод, что ни один из рассмотренных методов не является универсальным. В разных ситуациях разные методы будут показывать себя лучше остальных. Однако наиболее эффективным решением можно считать комплексное использование нескольких методов, дополняющих друг друга и компенсирующих ограничения каждого из них.

В настоящее время технологии больших языковых моделей еще далеки от идеала, однако плотно вошли в повседневную жизнь большинства пользователей. А потому одним из ключевых направлений модернизации систем искусственного интеллекта являются исследование и развитие технологий проверки достоверности информации. Совершенствование методов контроля,

верификации и интеграции внешних знаний будет способствовать созданию более надежных, безопасных и практически применимых интеллектуальных систем.

Библиографический список

1. Школьников А. Руководство для начинающих по галлюцинациям в больших языковых моделях (перевод) [Электронный ресурс] // Habr.com. – 2024. – URL: https://habr.com/ru/articles/826146/ (дата обращения: 25.07.2026)
2. Сухарева А. Почему нейросети галлюцинируют и как решить эту проблему? [Электронный ресурс] // t-j.ru. – 2024 – URL: https://t-j.ru/ai-hallucinations/ (дата обращения: 25.07.2026)
3. Науменко А.О. Технология RAG (Retrieval-Augmented Generation) как инновационный подход в LLM / Науменко А.О. // Вестник науки. – 2025. – №8. – С.280-289.
4. Fadeeva E. Fact-Checking the Output of Large Language Models via Token-Level Uncertainty Quantification [Электронный ресурс] // 2024. – URL: https://arxiv.org/pdf/2403.04696 (дата обращения: 25.07.2026)
5. Juraj Vladika. Improving Reliability and Explainability of Medical Question Answering through Atomic Fact Checking in Retrieval-Augmented LLMs [Электронный ресурс] // 2025. – URL: https://www.alphaxiv.org/abs/2505.24830 (дата обращения: 25.07.2026)
6. Shehzaad Dhuliawala. Chain-of-Verification Reduces Hallucination in Large Language Models [Электронный ресурс] // 2023. – URL: https://arxiv.org/pdf/2309.11495 (дата обращения: 25.07.2026)
7. LLM-галлюцинации: Как заставить нейросеть говорить правду (или хотя бы не врать так очевидно) [Электронный ресурс] // AiManual. – 2026. – URL: https://ai-manual.ru/article/llm-gallyutsinatsii-kak-zastavit-nejroset-govorit-pravdu-ili-hotya-byi-ne-vrat-tak-ochevidno/ (дата обращения: 25.07.2026)