УДК 001.811:004.89

Мустафин Альхас Амирович,
доцент кафедры общеобразовательных дисциплин,
ФГБОУ ВО «Ангарский государственный технический университет»,
e-mail: alhas355@mail.ru

ФАНТОМНЫЕ ЦИТИРОВАНИЯ: БИБЛИОГРАФИЧЕСКИЕ КОНФАБУЛЯЦИИ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ КАК ПОКАЗАТЕЛЬ ДОСТОВЕРНОСТИ НАУЧНОЙ КОММУНИКАЦИИ

Mustafin A.A.

PHANTOM CITATIONS: BIBLIOGRAPHICAL CONFABULATIONS OF LARGE LANGUAGE MODELS AS AN INDICATOR OF THE RELIABILITY OF SCIENTIFIC COMMUNICATION

Аннотация. В статье рассматривается системное явление генерации большими языковыми моделями (LLM) несуществующих, но правдоподобных библиографических ссылок – фантомных цитирований, или библиографических конфабуляций. На основе анализа литературных данных и характерных примеров предложена типология таких конфабуляций, обсуждаются их причины, укоренённые в архитектуре LLM. Показано, что частота и структура фантомных цитирований могут служить индикатором достоверности научной коммуникации при использовании генеративных моделей. Вводится понятие индекса библиографической конфабуляции как концептуального инструмента для оценки надёжности автоматически генерируемой библиографии. Анализируются риски «заражения» научного оборота несуществующими работами. Делается вывод, что конфабуляции являются не случайным дефектом, а диагностическим признаком, требующим системного учёта данного феномена в академической практике.

Ключевые слова: фантомные цитирования, большие языковые модели (LLM), библиографические конфабуляции, искусственный интеллект в науке.

Abstract. This article examines the systemic phenomenon of large language models (LLMs) generating nonexistent but plausible bibliographic references—phantom citations, or bibliographic confabulations. Based on an analysis of literature and representative examples, a typology of such confabulations is proposed and their causes, rooted in the LLM architecture, are discussed. It is shown that the frequency and structure of phantom citations can serve as an indicator of the credibility of scholarly communication using generative models. The concept of a bibliographic confabulation index is introduced as a conceptual tool for assessing the reliability of automatically generated bibliographies. The risks of "infection" of scholarly discourse with nonexistent works are analyzed. It is concluded that confabulations are not a random defect, but a diagnostic feature requiring systematic consideration of this phenomenon in academic practice.

Keywords: phantom citations, large language models (LLM), bibliographic confabulations, artificial intelligence in science.

В последние два года большие языковые модели (LLM), такие как GPT-4, Gemini, Llama, стали неотъемлемым инструментом академической работы. Их используют для поиска идей, редактирования текстов, составления обзоров литературы и библиографических списков. Казалось бы, машина, обученная на миллионах научных статей, должна свободно ориентироваться в миллионах ссылок. Однако практика обнаруживает обратное: LLM систематически порождают ссылки, которые оформлены как вполне реальные, но в действительности не

существуют. Это явление получило название – конфабуляция [лат. confābulārī – болтать, рассказывать]: продуцирование ложных, но связных и правдоподобных сообщений. LLM, без всякого намерения обмануть, действуют именно таким образом, достраивая вероятностные цепочки токенов, чтобы ссылка выглядела максимально реальной. В результате возникают фантомные цитирования (научные работы, которые в действительности не существуют), вводящие в заблуждение рецензентов, редакторов и читателей, если автор по невнимательности включит их в свою статью. Но есть и другая сторона: сам факт существования подобных фантомов, их частота и структура могут служить индикатором достоверности научного общения, опосредованного искусственным интеллектом [1]. Цель работы – проанализировать феномен библиографических конфабуляций LLM, предложить их типологию, обсудить причины и наметить принципы использования этого феномена в качестве определителя достоверности научной коммуникации.

Чтобы понять, почему LLM систематически генерируют несуществующие библиографические записи, необходимо разобраться в их устройстве. LLM не являются базами данных и не обращаются к внешним хранилищам знаний (базам данных, поисковым системам). Это вероятностные языковые модели, которые работают следующим образом. Получив последовательность фрагментов текста, называемых токенами [англ. token – знак, символ], они предсказывают следующий фрагмент с наибольшей вероятностью. Их обучение происходит на корпусах текстов, включающих миллионы научных статей. Модель не запоминает конкретные ссылки целиком, усваивая только статистические закономерности, то есть как выглядят типичные названия научных работ, какие фамилии авторов и ключевые слова встречаются в библиографических описаниях, какие журналы характерны для определённых научных областей. Когда мы просим модель составить список литературы по какой-то теме, она не «вспоминает» реальные статьи, а конструирует наиболее вероятную последовательность символов, имитирующую библиографическую запись. Конфабуляция – это не сбой, а побочный эффект генеративного подхода. Модель оптимизирует, то есть «настраивает» правдоподобие, а не стремится к истинности. И чем разнообразнее формулировки названий в конкретной научной дисциплине, тем легче модели «выдумать» новое название, не противоречащее статистическим закономерностям обучающих данных.

Анализ примеров, описанных в блогах, препринтах и экспертных дискуссиях, позволяет выделить несколько типов библиографических конфабуляций. Первый тип – полностью синтетическая ссылка. Автор, название, журнал, год, DOI – всё вымышлено, но подобрано правдоподобно. Например*:* «Смирнов, А. В. Квантовая герменевтика и постнеклассическая рациональность // Вестник цифровых гуманитарных наук, 2003. – 8(4), – С. 112–125.» Ни автора, ни журнала, ни статьи такой не существует. Второй тип – реальный автор с выдуманной

работой. Модель берёт известного учёного и приписывает ему несуществующую статью. Имя автора вызывает доверие, хотя учёный никогда не писал ничего подобного. Третий тип – гибридный. Модель комбинирует фрагменты разных реальных статей – например, заимствует начало названия из одной статьи, а окончание – из другой, или приписывает реальному названию год и журнал, взятые из другой реальной работы. Каждый элемент по отдельности существует, однако в такой комбинации их нет. Верификация требует проверки всех библиографических полей (автора, названия, журнала, года выпуска, тома, номера, страниц, DOI). Четвёртый тип – искажённые метаданные. Статья реальна, но неверно указан год, том, страницы или DOI. Формально это не полный «фантом», однако ссылка бесполезна, поскольку по ней невозможно найти документ. Каждый тип опасен по-своему. Синтетическую ссылку ещё можно заподозрить, тогда как гибридные ссылки и ссылки с реальным автором выглядят почти достоверно.

Традиционные наукометрические показатели (индекс Хирша, импакт-фактор) оценивают продуктивность, но не достоверность. В ситуации с LLM требуются индикаторы достоверности автоматически генерируемой библиографии. Например, мы даём модели задание подготовить список источников. Потом проверяем долю фантомных ссылок. Если эта доля высокая, значит достоверность коммуникации будет низкая. Эта идея лежит в основе *индекса* библиографической конфабуляции – концептуального показателя для конкретной модели, и чем он ниже, тем надёжнее библиографические возможности модели. Сегодня ни одна LLM не может гарантировать низкий уровень конфабуляций. Считается, что в гуманитарных науках, где названия более разнообразны, модели конфабулируют чаще, чем в естественнонаучных дисциплинах со стандартизированными формулировками заголовков научных статей. Это объясняется тем, что модели чутко реагируют на предсказуемость языка. И чем свободнее и разнообразнее формулировки названий в конкретной дисциплине, тем выше риск конфабуляции.

Фантомные цитирования порождают несколько проблем.

Эффект «заражения». Включённая в статью фантомная ссылка становится частью информационной среды. Другие исследователи начинают цитировать её, не проверяя оригинал, в связи с чем возникает замкнутый круг. Несуществующая ссылка начинает многократно цитироваться в научной литературе [1].

Уязвимость рецензирования. Рецензент не обязан проверять каждую ссылку в рукописи, поскольку это не входит в его обязанности. Однако он может проверить их, если сомневается в какой-то отдельной, например, критически важной для выводов. Таким образом фантомы, вписанные в текст, могут остаться незамеченными.

Падение доверия. Систематическое обнаружение фантомных ссылок в научных текстах, созданных при участии LLM, подрывает доверие ко всей

научной коммуникации. Возникает вполне закономерный вопрос: не являются ли выдумкой и другие цитирования в таких текстах?

Этические и издательские дилеммы. Фантомные цитирования порождают вопрос об ответственности за несуществующие ссылки, то есть, кто в ответе – автор, издательство или разработчик модели? Ведущие издательства требуют от авторов декларировать использование ИИ, но при этом не требуют от них проверять ссылки. Из этой несогласованности возникает другой вопрос: должна ли проверка библиографии стать обязательной частью рецензирования или издательского процесса при декларированном использовании LLM?

Практические рекомендации для тех, кто использует LLM:

Во-первых, никогда не следует доверять сгенерированным библиографическим ссылкам без проверки. Каждую следует проверять через Crossref, Google Scholar, PubMed – специальные платформы, используемые в научной среде для поиска, индексации и отслеживания научных публикаций;

Во-вторых, некоторые современные LLM поддерживают функцию поиска в интернете RAG (Retrieval Augmented Generation – генерация с дополненным поиском). Этот режим позволяет модели сверять свои ответы с актуальными веб-источниками. Использование такого подхода резко снижает количество фантомных ссылок, однако полагаться только на него без последующей проверки всё равно не следует;

В-третьих, в декларации следует обязательно указывать, что библиографический список частично сгенерирован LLM, и подтверждать факт проверки всех ссылок (например, отдельной подписью автора).

Выводы:

1. Библиографические конфабуляции LLM – это не случайный сбой, а закономерное следствие вероятностной архитектуры моделей. LLM не хранят базы знаний и не проверяют факты; они генерируют правдоподобные имитации библиографических записей на основе статистических закономерностей.

2. Предложена типология фантомных ссылок, включающая четыре типа. 1. Полностью синтетическая ссылка: всё (авторы, название, журнал, год) выдумано модель.; 2. Ссылка содержит реального автора, но конкретные название, журнал и год вымышлены. 3. Гибридная (комбинированная) ссылка с комбинацией элементов библиографической записи (автор, год публикации, название статьи, название журнала, том, выпуск, номера страниц) из разных реальных статей; 4. Искажённая реальная ссылка – реальная статья существует, но модель воспроизвела её метаданные с ошибкой (неверный год, том, страницы или незначительное искажение названия).

3. Частота конфабуляций зависит от дисциплинарных особенностей языка: в гуманитарных науках с более разнообразными и вариативными формулировками названий, модели конфабулируют чаще, чем в естественнонаучных дисциплинах со стандартизированными заголовками статей.

4. Фантомные цитирования создают серьёзные риски для научной коммуникации: эффект «заражения» научного оборота, уязвимость системы рецензирования и падение доверия к публикациям, подготовленным с участием LLM.

5. Введён концептуальный индекс библиографической конфабуляции как возможный инструмент для оценки надёжности автоматически генерируемой библиографии. Частота и структура фантомов могут служить индикатором достоверности научного общения, опосредованного искусственным интеллектом.

6. Практические рекомендации для исследователей включают обязательную проверку сгенерированных ссылок через базы данных, использование режимов поиска (RAG), декларирование применения LLM и критическое отношение к вызывающим сомнения в своей достоверности библиографическим спискам.

ЛИТЕРАТУРА

1. **Морозова С.А.** Библиографические конфабуляции генеративных моделей ИИ как индикатор достоверности научной коммуникации : доклад на конференции «Science Analytics 2026» (Москва, ИНТЦ МГУ «Воробьёвы горы», 2–3 апреля 2026 г.) / С.А. Морозова ; РГПУ им. А.И. Герцена. – Санкт-Петербург, 2026.

2. **Надаф, М.** Галлюцинаторные цитаты загрязняют научную литературу. Что можно сделать? / М. Надаф, Э. Куилл ; пер. с англ. // Nature. – 2026. – URL: https://www.nature.com/articles/d41586-026-00969-z (дата обращения: 25.04.2026).