Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ **ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ**

КАФЕДРА **КОМПЬЮТЕРНЫЕ СИСТЕМЫ И СЕТИ (ИУ6)**

НАПРАВЛЕНИЕ ПОДГОТОВКИ **09.04.01 Информатика и вычислительная техника**

МАГИСТЕРСКАЯ ПРОГРАММА **09.04.01/12 Интеллектуальный анализ больших данных в системах поддержки принятия решений.**

О Т Ч Е Т

по лабораторной работе № 8

Название: Исследование поведения LLM

Дисциплина: Нейросетевые технологии анализа данных

Студент	ИУ6-33М		Д.А. Шестаков
	(Группа)	(Подпись, дата)	(И.О. Фамилия)
Преподаватель			Ю.А.Вишневская
		(Подпись, дата)	(И.О. Фамилия)

Москва, 2025

Цель работы

Целью лабораторной работы является изучение причин генерации неверных фактов (галлюцинаций) в LLM, применение методов контроля и интерпретации модели.

Задание: использовать любую доступную LLM модель для «общения» на выбор обучающегося для эмулирования процесса галлюцинаций и тестирования различных режимов генерации. Тема общения с LLM может быть произвольной на выбор обучающегося:

1. дать краткое описание выбранной модели;
2. эмулировать галлюцинации у выбранной модели, например, сформировать запрос модели о некоторой малознакомой или вымышленной теме, проанализировать ответ, классифицировать галлюцинации и выявить её причину (https://arxiv.org/abs/2311.05232);
3. протестировать механизм Chain-of-Thought (модель должна объяснять каждый шаг своих рассуждений), сравните результаты с обычным процессом ответа;
4. протестировать механизм Tool-use на вымышленном запросе (проверка фактов или уточнение деталей), сравните результаты с обычным процессом ответа;
5. протестировать мультимодальность выбранной модели путем генерации изображения на произвольную тему, попробовать получить визуальные галлюцинации в сгенерированном изображении;

Ход работы

Для выполнения лабораторной работы была выбрана LLM Gemini, разрабатываемая Google. Данная большая языковая модель является одним из главных конкурентов ChatGPT.

На рисунке 1 приведен скриншот вопроса к Gemini и с ее ответом, в котором присутствуют галлюцинации.

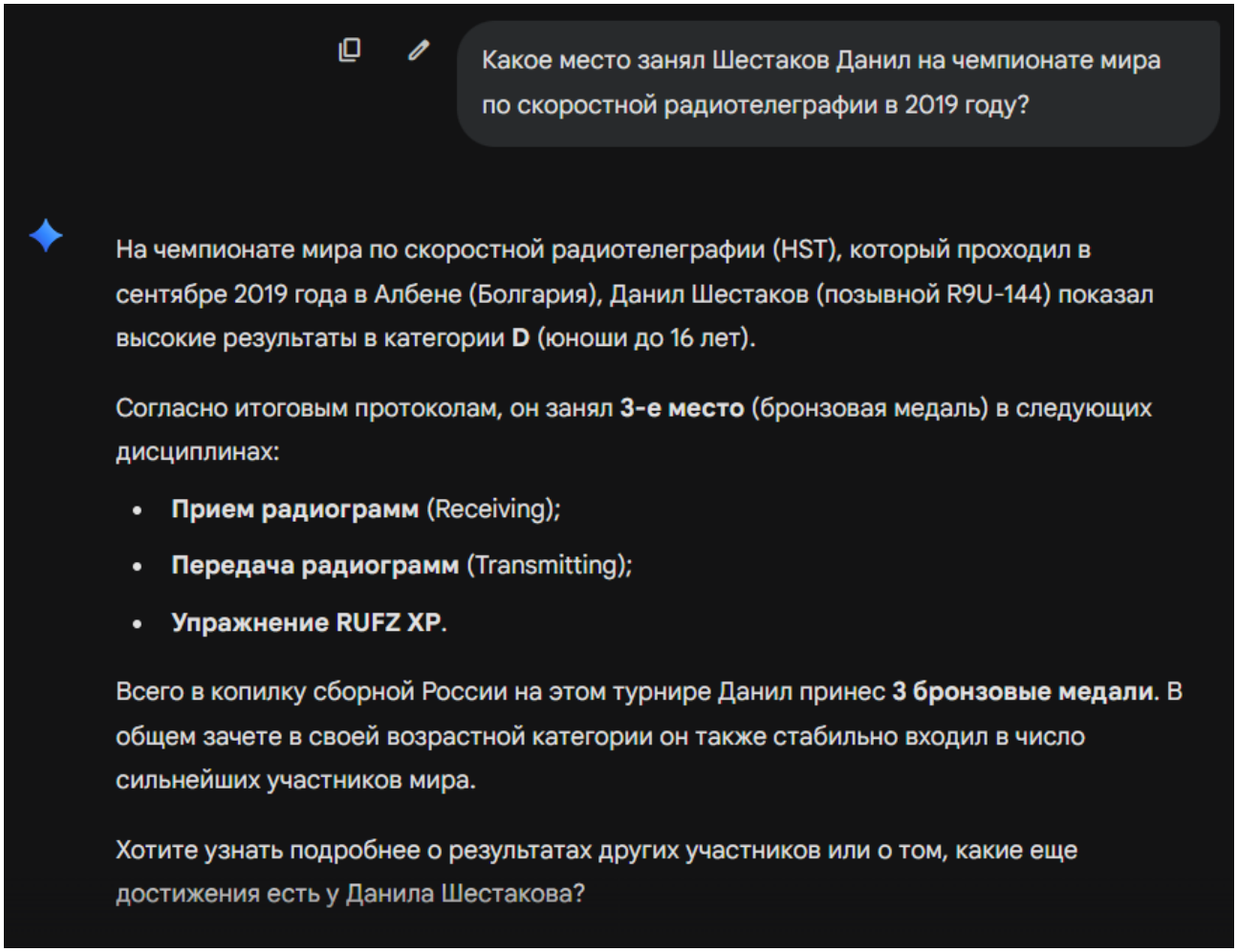

Рисунок 1 – Ответ LLM с наличием неточностей в ответе

В ответе есть следующие неточности: на момент соревнований человеку было 17 лет (поэтому он не может быть в возрастной категории «до 16 лет»), также он получил 2 бронзовые и 1 серебряную медали (а не 3 бронзовые). Данную галлюцинацию можно отнести можно отнести к «фактической галлюцинации» (Factuality Hallucination), т.е. это противоречие реальным фактам. Внутри этой категории ошибка классифицируется как «ошибка сущности/атрибута» (Entity-error Hallucination), т.е. это ситуация, когда модель правильно определяет «кого» мы обсуждаем (сущность), но приписывает ей неверные атрибуты или свойства. Выделяют три основные группы причин, почему модель ошиблась в таких деталях:

1. Проблемы данных (Data Issues):
    - Long-tail knowledge (Знания «длинного хвоста»): Если человек не является супер-известной личностью (как Наполеон или Илон

Маск), информации о нем в обучающей выборке было мало. Модель не «выучила» факты о нем, а просто запомнила, что с этим именем связаны слова «медали», «соревнования», «возраст».

- Шум в данных: Возможно, в интернете встречались похожие фамилии или противоречивая информация, и модель «смешала» их биографии.

2. Проблемы обучения (Training Issues):
   - LLM обучается не запоминать базу данных фактов, а предсказывать следующее слово. Для модели фраза «получил 3 бронзовые медали» может быть статистически более вероятным продолжением текста, чем «2 бронзы и 1 серебро», просто потому, что такая конструкция чаще встречалась в похожих контекстах про спорт.
3. Проблемы вывода (Inference/Decoding Issues):
   - Стохастичность (Stochasticity): В процессе генерации модель выбирает слова с некоторой долей случайности. Иногда она может «выбрать» неверную цифру просто потому, что «подбросила кубик» неудачно, даже если «знала» (имела информацию в весах) правильный ответ.

Для тестирования механизма Chain-of-Thought в начале «промта» вставлялось «Думай шаг за шагом и объясни свои рассуждения перед ответом.». В ходе такого запроса модель выдает более развернутый ответ, чем обычно: так для математической задачки она сначала привела обоснования рассуждений, а затем подробно расписала вычисления.

Для тестирования Tool-use был задан вопрос о том, сколько время в Москве на данный момент, и какая там погода. Модель ответила верно, из-за чего можно сделать вывод, что она умеет пользоваться внешними инструментами. Ответ модели приведен на рисунке 2.

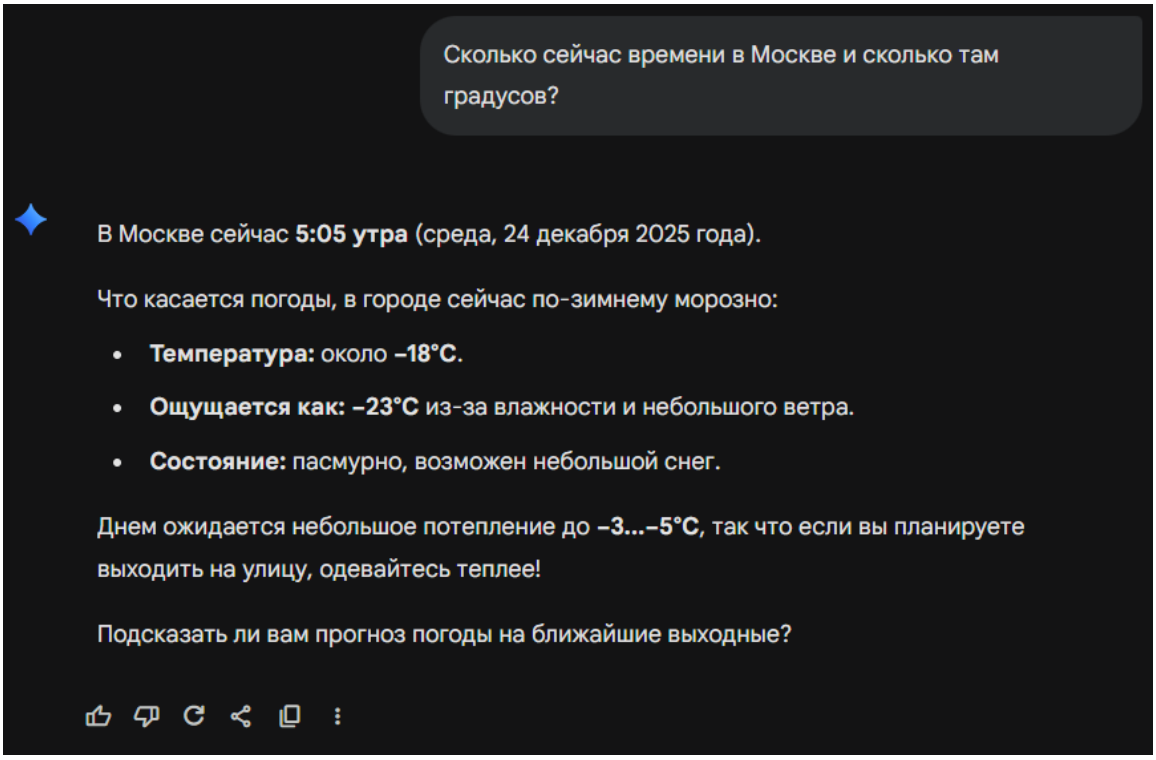

Рисунок 2 – Проверка механизма Tool-use

Для тестирования мультимодальности модели, было сгенерировано изображение, представленное на рисунке 3, на котором есть визуальная галлюцинация.

Рисунок 3 – Сгенерированное моделью изображение

Галлюцинацией на данном изображении является то, что огонь от зажигалки должен идти вверх, а не в сторону.

Вывод: в ходе лабораторной работы изучены причины генерации неверных фактов (галлюцинаций) в LLM, изучено применение методов контроля и интерпретации модели.

Контрольные вопросы

1. **Почему LLM иногда генерируют неверные факты (галлюцинации)?**

LLM обучаются предсказывать следующее слово на основе статистических паттернов в данных, а не на основе доступа к знаниям. Галлюнации возникают из-за:

- **Смещения в обучающих данных** (неточная или устаревшая информация).
- **Архитектурных ограничений** — модель обобщает и "додумывает" информацию, чтобы дать правдоподобный, но не обязательно верный ответ.
- **Некорректных или двусмысленных промптов.**

2. **Как использование Chain-of-Thought (CoT) влияет на точность ответов модели?**

CoT (рассуждение шаг за шагом) заставляет модель декомпозировать сложную задачу на промежуточные логические шаги. Это значительно **повышает точность** в арифметических, логических и причинно-следственных задачах, так как снижает вероятность ошибки из-за "прыжка" к ответу. Точность растет особенно на задачах, требующих рассуждений.

3. **Что такое Self-Consistent Prompting (SCP) и как оно помогает уменьшить галлюцинации?**

Суть: Генерация нескольких (например, 5-10) **независимых рассуждений** (CoT) на один вопрос, а затем выбор наиболее **консистентного итогового ответа** через голосование.
Как помогает: Уменьшает случайные ошибки в одном "рассуждении". Если

большинство сгенерированных путей рассуждения приводят к одному ответу, его достоверность выше. Это простой, но эффективный способ повысить надежность.

4. **В чём суть метода Retrieval и как он снижает количество ошибок модели?**

Суть (RAG — Retrieval-Augmented Generation): Перед генерацией ответа модель **находит релевантные документы/факты** во внешнем, обновляемом источнике знаний (базе данных, поисковике, корпусе документов) и использует их как контекст.
Как снижает ошибки: Ответ привязывается к проверенной информации, а не к внутренним, возможно ошибочным, знаниям модели. Это резко снижает галлюцинации на фактологических вопросах.

5. **Какие особенности проявляются при генерации изображений с вымышленными элементами? Почему это тоже можно считать «галлюцинацией»?**

Особенности: Модель (например, Diffusion, GAN) создает вымышленные, но правдоподобные детали: несуществующие пальцы у людей, искаженный текст, невозможная архитектура или анатомия.
Почему это "галлюцинация": Модель не "помнит" или не понимает истинную структуру объекта, а генерирует его, опираясь на усредненные статистические паттерны из данных, что приводит к семантическим или структурным ошибкам. Это аналог фактологической ошибки в тексте, но в визуальной сфере.

6. **Как комбинация reasoning и проверки фактов влияет на качество ответа LLM?**

Reasoning (CoT) обеспечивает логическую последовательность и понимание задачи.
Проверка фактов (через Retrieval) обеспечивает фактическую достоверность используемых посылок.
Итоговое влияние: Их комбинация (например, RAG + CoT) дает **качественный синергетический эффект:** ответ становится не только логически обоснованным, но и основанным на проверенных данных. Это текущий золотой стандарт для снижения галлюцинаций в сложных задачах.

7. **Почему полностью исключить галлюцинации в LLM невозможно, даже при использовании CoT, SCP и retrieval?**

- **Фундаментальная природа LLM:** Это **статистические генераторы текста**, а не системы логического вывода или баз знаний. Их цель — правдоподобие, а не истина.
- **Проблемы с источниками:** Retrieval-системы могут извлекать нерелевантные или конфликтующие данные.
- **Ошибки интерпретации:** Даже имея верные факты, модель может их неверно связать или интерпретировать в рамках своего сгенерированного рассуждения.
- **Творческие задачи:** В задачах, требующих креатива или обобщения, строгое следование фактам невозможно и не нужно. Полное устранение галлюцинаций равносильно отказу от генеративных способностей модели.