

Лабораторная работа №8

Исследование поведения LLM

Цель работы – 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;
6. составить отчет о проделанной работе в соответствии с требованиями кафедры.

Требования к отчету. Отчет должен содержать постановку задачи, исходные данные, результаты решения задачи, необходимые иллюстративные материалы.

Требования к защите

Защита лабораторной работы происходит индивидуально. Система оценки – рейтинговая.

Критерии оценки:

- корректность выполненного исследования;
- адекватность полученных результатов;
- качество отчета;
- качество ответов на контрольные вопросы;
- срок выполнения работы.

Время выполнения работы – 4 академических часа.

Контрольные вопросы

1. Почему LLM иногда генерируют неверные факты (галлюцинации)?
2. Как использование Chain-of-Thought (CoT) влияет на точность ответов модели?
3. Что такое Self-Consistent Prompting (SCP) и как оно помогает уменьшить галлюцинации?
4. В чём суть метода Retrieval и как он снижает количество ошибок модели?
5. Какие особенности проявляются при генерации изображений с вымышленными элементами? Почему это тоже можно считать «галлюцинацией»?
6. Как комбинация reasoning и проверки фактов влияет на качество ответа LLM?
7. Почему полностью исключить галлюцинации в LLM невозможно, даже при использовании CoT, SCP и retrieval?