Опишите пожалуйста, как боролись с галюцинированием модели, соблюдая фактологическую связность, как происходила работа в асинхронном режиме и какие трудности были при этом.

Ответы:

Для борьбы **с галлюцинациями модели, соблюдая фактологическую связанность,** в рамках разработки чат-бота в стиле персонажа Рагнара («Викинги») были предприняты следующие шаги:

* **Тщательный подбор и очистка данных:** основой для обучения модели послужил качественно подготовленный и тщательно очищенный датасет, состоящий из реплик Рагнара на основании скрипта по всем 8 сезонам сериала «Викинги». Была проведена работа по удалению аномальных и нерелевантных примеров, что снижает вероятность генерации несвязного или некорректного контента.
* **Дообучение на специализированном датасете:** модель GPT-2 была дополнительно обучена на этом специально подготовленном датасете, что позволило ей лучше улавливать контекст и стилистику речи персонажа, а также сократить количество фактических ошибок и галлюцинаций.
* **Ограничение длины генерации:** установление пределов на максимальную длину сгенерированного текста помогает контролировать качество вывода, поскольку чем длиннее текст, тем выше вероятность появления некорректной информации.
* Была проведена работа с параметрами, такими как **top\_k, top\_p, temperature**. Эти параметры помогают предотвращать галлюцинации, так как влияют на случайность ответов.

**Работа в асинхронном режиме:**

Разработка чат-бота включала асинхронную обработку запросов пользователей для обеспечения более быстрого и стабильного взаимодействия. Это было реализовано через **асинхронное API:** использование асинхронных вызовов API Telegram для обработки входящих сообщений без блокирования основного потока выполнения. Это позволяло боту одновременно обрабатывать несколько запросов.

**Трудности**: несмотря на асинхронную обработку, задержки в ответах все еще могли возникать из-за времени, необходимого на генерацию текста, особенно при высокой нагрузке на сервер. Кроме этого, в какой-то момент обрывались фразы (помогло увеличение максимальной длины ответа). Также потребовалась настройка GPT-2, так как бот далеко не сразу стал отвечать, что он викинг. Однако в конечном счете, это получилось.

