МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ (НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)

Институт №8 «Компьютерные науки и прикладная математика» Кафедра 806 «Вычислительная математика и программирование»

Проект

по курсу «Генеративные модели и Prompt Engineering»

Генерация квестов, диалогов и описание локаций для текстовой игры с использованием RAG.

Выполнил: *Стрыгин Д.Д.* Группа: *M8O-211M-23*

Преподаватель: Ахмед С.Х.

Условие

1. **Цель работы**: разработать систему, генерирующую уникальные сценарии, диалоги и квесты для текстовых видеоигр, используя LLM.

Программное и аппаратное обеспечение

Характеристики графического процессора Compute capability: 8.9

Name: NVIDIA GeForce RTX 4060

Laptop GPU

Total Global Memory: 8188 MiB Shared memory per block: 49152 Registers per block: 65536 Warp size: 32

Max threads per block: (1024, 1024, 64)

Max block: (2147483647, 65535, 65535)

Total constant memory: 65536 Multiprocessors count: 24

Процессор: 13th Gen Intel(R) Core(TM) i7-13700H

Оперативная память: 7806 Мб Накопитель: Virtual Disk 1T OS: Ubuntu 22.04.3 LTS (WSL 2) IDE: Visual Studio Code 1.97.2

compiler: Cuda compilation tools, release 12.6, V12.6.77

Метод решения

Для решения поставленной задачи используется API для LLM Gigachat-2, также создается система RAG, заполняемая данными из фэнтези книг: The Hobbit, The world war, The king Arthur. Содержимой книг обрабатывается, вычленяя предполагаемые диалоги, описания квестов и локаций, в дальнейшим полученные данные индексируются для быстрого поиска и подкладываются в запрос к LLM.

Описание программы

project.ipynb

Функции	Описание
def retrieve(query, method="hyde")	Принимает на вход запрос к LLM – query, а также method для генерации. Ищет подходящие фрагменты в индексе и подкладывает их в запрос.
<pre>def gigachat_request(access_token, message, model_name="GigaChat-2", max_tokens=512, repetition_pen- alty=1.0, stream=False, url="https://gigachat.devices.sber- bank.ru/api/v1/chat/completions")</pre>	Принимает токен доступа к апи, сообщение, название модели, максимальное кол-во токенов, штраф ща повторяемость, ответы в режиме stream, а также url. Отправляет rest запрос к апи Gigachat.

Результаты

Запрос	Результат	
Generate a location in the style of a futuristic war with aliens.	Hypothetical Location: The Battle of Novaya Zemlya Location Description: Name: Novaya Zemlya Region: Arctic Circle, Siberia Environment: A desolate, icy archipelago, where winds howl relentlessly and snow drifts across frozen plains. The landscape is rugged, with jagged mountains, glaciers, and vast stretches of ice-covered tundra. Key Features: 1. Fortress of Novaya Zemlya - A massive, heavily fortified military base built deep beneath the ice. 2. Alien Outposts - Scattered across the icy plains, equipped with cloaking technology and advanced weaponry. 3. Tundra Battlefields - Frozen plains where intense battles rage between human soldiers and alien forces.	
Create a dialogue in the style of medieval conversations.	Scene: The Great Hall of the Royal Castle. King Valdemar the Just: "Hear ye, hear ye, all good people of my realm! It has been many years since we gathered here to discuss the matters of the land. Today, we have important business to attend to. Sir Gawain, tell us what news you have brought from the outer marches?" Sir	

Запрос	Результат
Generate a quest in the style of a fantasy adventure.	The Quest: The Dragon's Hoard of Remembrance Introduction: In the distant kingdom of Valdoria, a legendary tale has been whispered for centuries: a dragon named Malador, the "Guardian of the Golden Coast," has been slain by an unknown hero. His treasure was buried deep beneath the waves, guarded by a powerful enchantment. Quest Objectives: 1. Discover the location of the dragon's lair. 2. Retrieve the Key of the Deep, a magical artifact that opens the underwater catacombs. 3. Confront the ghost of Malador and lift the curse of the treasure.

Выводы

RAG - мощный инструмент для регулирования работы LLM, не требующий переобучения модели, способный масштабироваться за счет расширения базы знаний, и способный поддерживать актуальность ответов модели пользователю.