Московский государственный технический университет им. Н.Э. Баумана

# Разработка метода интеграции больших языковых моделей (LLM) средствами REST API для управления виртуальными агентами в Unreal Engine 5

Докладчик: Больных А. С., РК6-84Б

Руководитель: Витюков Ф. А.

# Цель и задачи

Цель:

Разработать симуляцию, моделирующую взаимодействие LLM-модели и виртуальных агентов, которые имеют различные внутренние параметры.

Управляемые агенты должны взаимодействовать друг с другом и с окружающей средой, стараясь удовлетворить свои потребности.

Задачи:
1. Создание сцены с окружающим миром;
2. Создание персонажей с различными характеристиками;
3. Интеграция больших языковых моделей в движок через REST API;
4. Синхронизация действий виртуальных агентов;
5. Оценка эффективности LLM.

# Актуальность

- LLM является мощным и универсальным инструментом, позволяющим сократить код
- На рынке Epic Game MarketPlace представлено мало решений (особенно бесплатных) для интеграции различных LLM в движок Unreal Engine 5
- В настоящий момент решения на базе LLM занимаются в основном генерацией текста, реплик, картинок, анимированных аватаров (NVIDIA ACE)

# Компьютерный (виртуальный) агент

Определение

**Компьютерный агент - это автономная сущность, которая действует в окружающей среде и наблюдает за ней.**

Агент ставит цели и принимает решения исходя из состояния внутренних и внешних параметров.

Об интеллектуальности агента можно говорить, если он взаимодействует с окружающей средой примерно так же, как действовал бы человек.

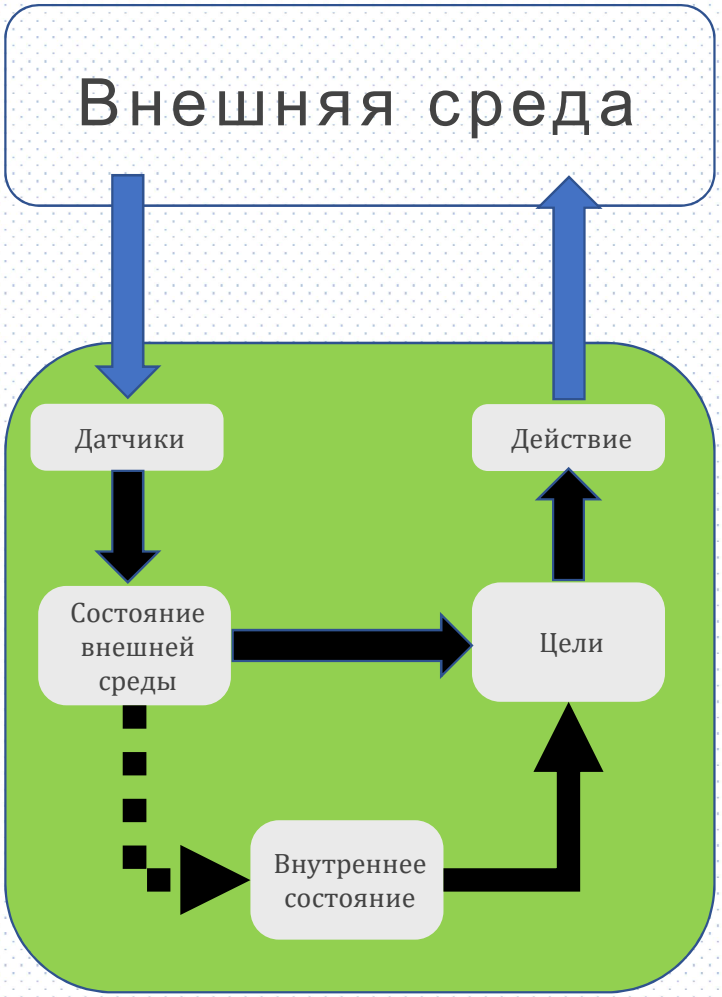



Схема агента с простым поведением

# Большая языковая модель

Определение

**Большая языковая модель (large language model, LLM) – это нейросеть с огромным числом весовых коэффициентов (параметров), обученная на большом количестве текста.**

Считается, что языковая модель является большой, если содержит больше одного миллиарда параметров.

Gemini

OpenAI GPT-4

Llama 3

BERT

a BigScience initiative BLOOM

NVIDIA Nemotron-4 340B

# Принцип работы LLM

## Токен

Это самая маленькая единица текста, например слово или знак препинания.

Для английского языка 1000 токенов в среднем равны 750 словам.

Для русского языка 1000 токенов – это около 375 слов.

1. Модель получает на вход текстовый запрос, который разбивается на «токены».
2. Модель анализирует информацию и подбирает ещё один токен.
3. Полученный текст снова подаётся на вход модели.

Так получается «разумное продолжение» на основе изначального запроса.

Для пользователя это выглядит как ответ, который имеет смысл.

# Механизм взаимодействия с LLM

**Representational State Transfer (REST) –** архитектурные рекомендации по взаимодействию компонентов распределённого приложения в сети.

**Application protocol interface (API) –** описание способов взаимодействия одной компьютерной программы с другими.

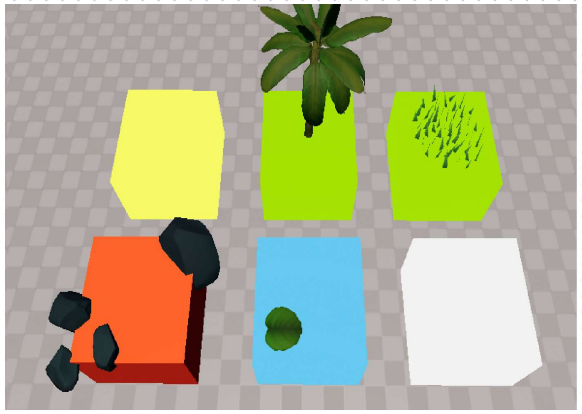
**VaRest –** это open-source плагин к движку Unreal Engine для обеспечения REST коммуникаций между клиентом и сервером.

За запуск локальных LLM отвечает кроссплатформенное приложение Ollama

# Создание симуляции

# Разработка поля для симуляции

Инстансинг геометрии (дублирование геометрии) – подход, позволяющий отрисовывать множество копий одного и того же 3d-объекта за один проход

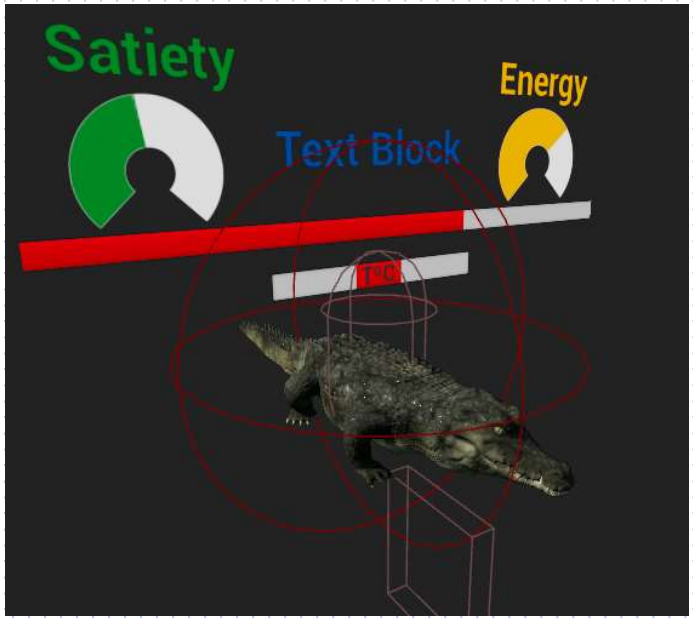


Все различные биомы (Пустыня, лес, равнина, гнездо крокодила, вода, снежная область)

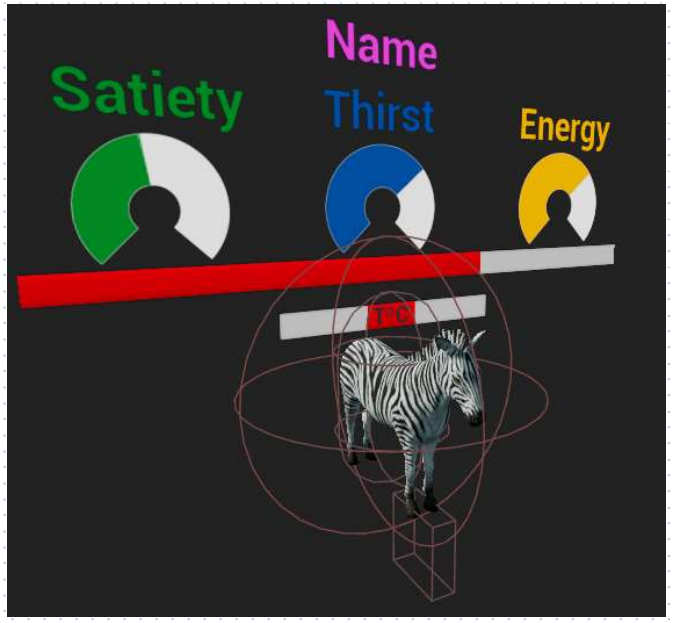
Пример поля в виде континента Африки

Разработка компьютерных агентов

Модели и компоненты



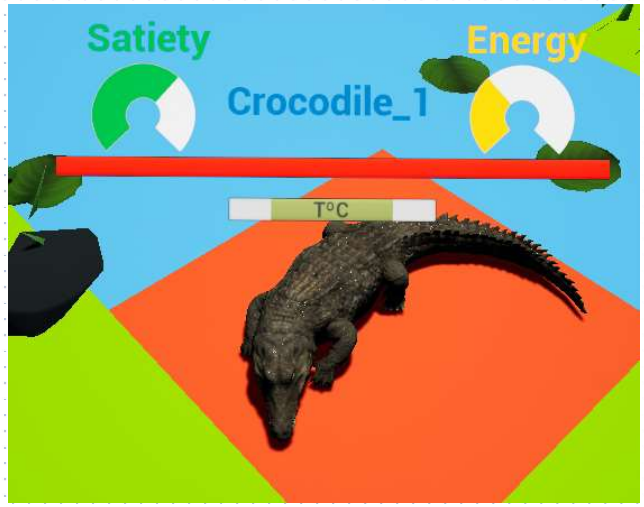

Виртуальный агент «Крокодил»




Виртуальный агент «Зебра»

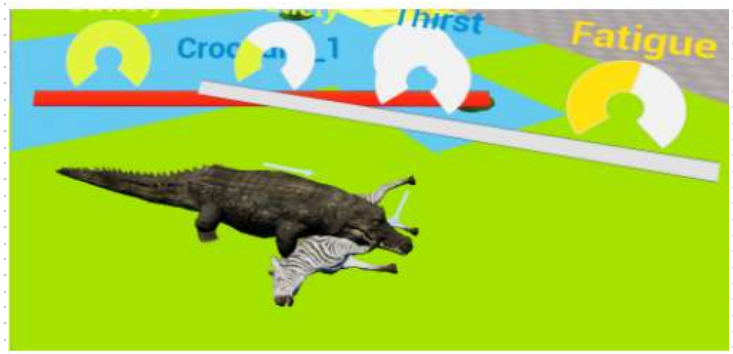
Разработка компьютерных агентов

Правила симуляции

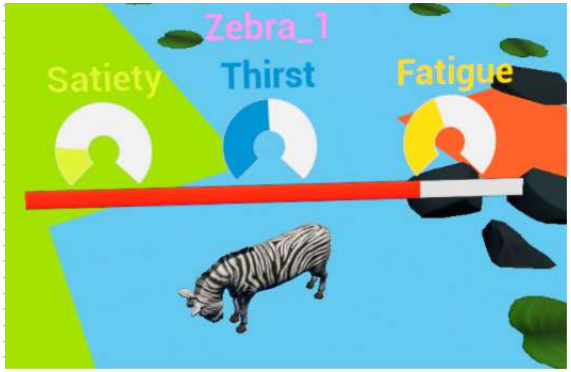



Крокодил отдыхает


Крокодил добывает ветку

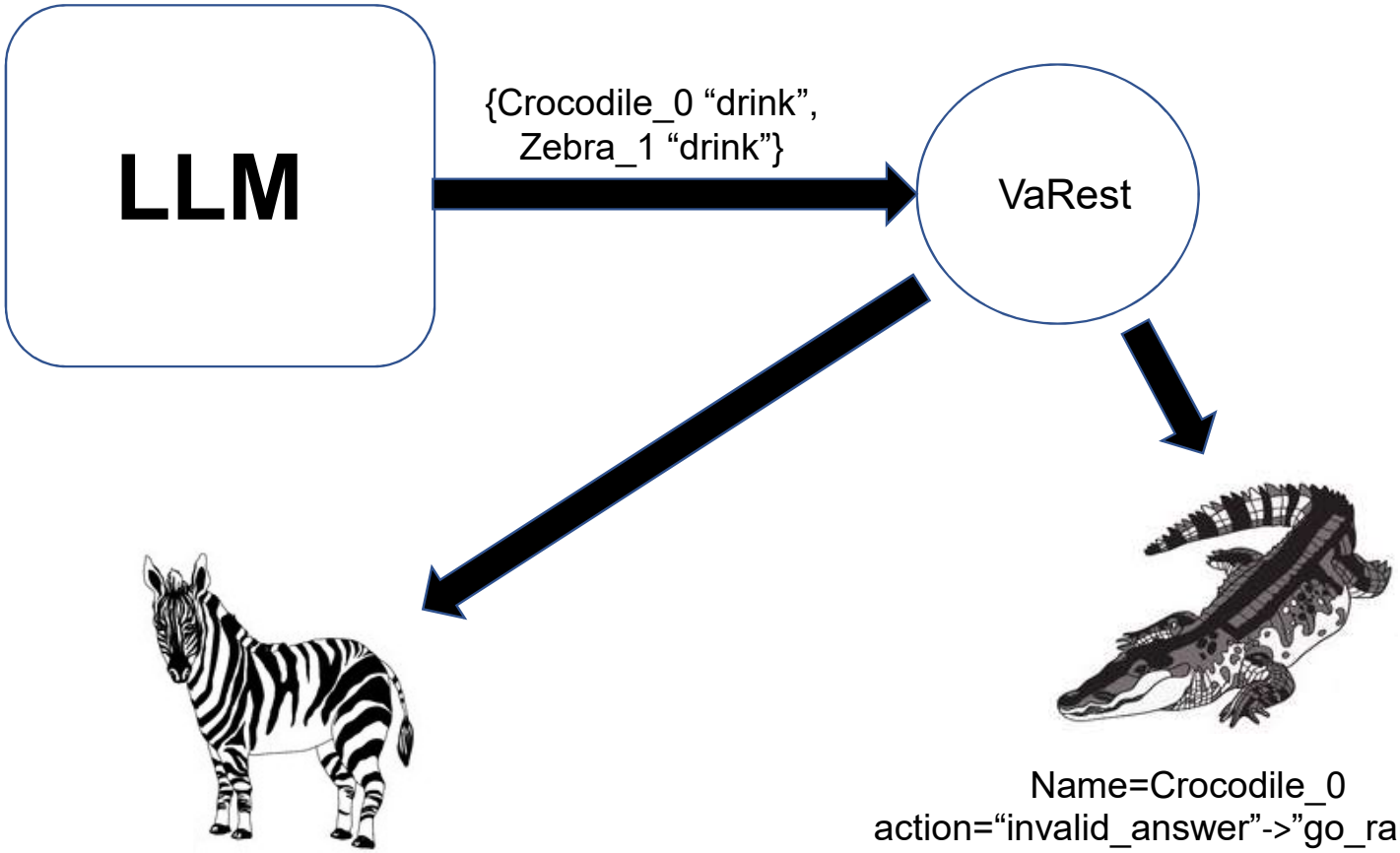



Крокодил успешно поохотился




Зебра утоляет жажду

# Обработка ответов




Name=Zebra_1
action=“drink” (thirst=45->100)
Hint: character called Zebra_1 restored thirst, try restore other parameters

Name=Crocodile_0
action=“invalid_answer”->”go_random”
Hint: you typed wrong action for character called Crocodile_0, so he moved to random location. Here are all possible moves [allowed_moves]

## Тестирование


Процессор Intel(R) Core(TM) i5-2400 CPU @ 3.30 GHz

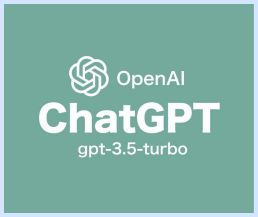
Оперативная память 16,0 ГБ

Видеокарта NVIDIA GeForce GTX 1050 Ti 4 ГБ

Тип системы 64-разрядная операционная система, процессор x64

Время ответа LLM

От 25 сек. до 1 мин. 30 сек.


Процессор Intel(R) Core(TM) i7-7700K @ 4.50 GHz

Оперативная память 16,0 ГБ

Видеокарта NVIDIA GeForce RTX 3060 Ti 8 ГБ

Тип системы 64-разрядная операционная система, процессор x64

Время ответа LLM

От 3 сек. до 5 сек.

# Заключение

- Создана трёхмерная сцена из ISMC кубов и других моделей;
- Спроектированы персонажи с различными характеристиками и поведением;
- Интегрирован механизм взаимодействия с LLM по REST API;
- Проведены тесты на разных по мощности ПК;
- Проанализировано влияние параметров моделей на результат их ответов;
- Определены преимущества, недостатки локальных и онлайн LLM;
- Выяснено, что для коррекции ответов LLM нужно эффективно использовать «промпт-инжиниринг».