|  |  |
| --- | --- |
| Gerb-BMSTU_01 | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

|  |  |
| --- | --- |
| ФАКУЛЬТЕТ | Робототехника и комплексная автоматизация (РК) |
| КАФЕДРА | Системы автоматизированного проектирования (РК6) |

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

***НА ТЕМУ:***

***«Исследование характеристик различных больших языковых моделей»***

|  |  |  |
| --- | --- | --- |
| Студент РК6-84Б | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | **Больных А.С.** |
|  | (Подпись, дата) | И.О. Фамилия |
| Руководитель | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | **Витюков Ф.А.** |
|  | (Подпись, дата) | И.О. Фамилия |

*2024 г.*

**Министерство науки и высшего образования Российской Федерации**

**Федеральное государственное бюджетное образовательное учреждение**

**высшего образования**

**«Московский государственный технический университет имени Н.Э. Баумана**

**(национальный исследовательский университет)»**

**(МГТУ им. Н.Э. Баумана)**

УТВЕРЖДАЮ

Заведующий кафедрой РК6

А.П. Карпенко

«\_\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_ 2024 г.

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

по теме: исследование характеристик различных больших языковых моделей \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Студент группы \_\_РК6-84Б\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Больных Андрей Сергеевич\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.) \_учебная\_\_\_

Источник тематики (кафедра, предприятие, НИР) \_кафедра \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

График выполнения НИР: 25% к 5 нед., 50% к 11 нед., 75% к 14 нед., 100% к 16 нед.

Техническое задание:\_Провести продуктовое исследование существующих больших языковых моделей (LLM). Проанализировать технические параметры выбранных моделей. Исследовать адекватность, качество текстовых ответов LLM моделей. Выяснить производительности разных типов моделей. \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

***Оформление научно-исследовательской работы:***

Расчетно-пояснительная записка на 17 листах формата А4.

Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.):

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дата выдачи задания «11» февраля 2024 г.

|  |  |  |
| --- | --- | --- |
| **Руководитель НИР** | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | **Витюков Ф.А.** |
|  | (Подпись, дата) | И.О. Фамилия |
| **Студент** | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | **Больных А.С.** |
|  | (Подпись, дата) | И.О. Фамилия |

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре

СОДЕРЖАНИЕ

[ВВЕДЕНИЕ 4](#_Toc165659142)

[ОСНОВНЫЕ ПОНЯТИЯ И ОПРЕДЕЛЕНИЯ 5](#_Toc165659143)

[1. Большие языковые модели 6](#_Toc165659144)

[2. Сравнительный анализ LLM 8](#_Toc165659145)

[3. Характеристики LLM 9](#_Toc165659146)

[4. Система оценки LLM (Benchmark) 11](#_Toc165659147)

[ЗАКЛЮЧЕНИЕ 14](#_Toc165659148)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 15](#_Toc165659149)

ВВЕДЕНИЕ

Цели НИРС:

1. Провести продуктовое исследование существующих больших языковых моделей (LLM).
2. Проанализировать технические параметры выбранных моделей.
3. Исследовать адекватность, качество текстовых ответов LLM моделей.

Искусственный интеллект продолжает добиваться огромных успехов в различных отраслях, и одним из ключевых событий последних лет является рост больших языковых моделей (LLM).

За последние десять лет разработки больших языковых моделей качество их ответов значительно улучшилось, помимо простого продолжения текста они развили возможность нахождения логических связей в предоставленных запросах и выдаваемых ими ответах, а также анализа и структуризации предоставленных данных и формировании выводов из них.

Повышение качества работы LLM также привело к значительному возрастанию затрат на содержание аппаратного обеспечения, поддерживающего их работу, в полной мере компенсировать которые не могут даже инновационные нейронные процессоры (NPU), поэтому разработчики вынуждены применять горизонтальное масштабирование уже существующих языковых моделей для повышения качества работы своих систем без увеличения расходов на аппаратное обеспечение. [1]

В данной работе рассмотрены различные LLM и их характеристики, а также произведена оценка их производительности.

ОСНОВНЫЕ ПОНЯТИЯ И ОПРЕДЕЛЕНИЯ

**CPU (Central Processing Unit)** – центральный процессор.

**GPU (Graphics Processing Unit)** – графический процессор (видеокарта).

**FPS (Frames per second)** – количество кадров в секунду.

**LLM** **(Large Language Model)** ­– большая языковая модель.

**Token (токен)** **–** это специальные маркеры, которые используются для обозначения определенных элементов в тексте, например, начала и конца предложения, начала и конца абзаца, маркеров времени и т.д. [3]

1. Большие языковые модели

LLM – это мощные модели искусственного интеллекта, которые используют возможности обработки естественного языка для эффективного понимания, создания и управления человеческим языком. Эти модели обучаются на огромных объемах текстовых данных, что позволяет им преуспеть в различных языковых задачах, таких как контекстно-зависимое понимание естественного языка, генерация текста, перевод и обобщение. [9]

В настоящее время доступно несколько LLM, и новые разрабатываются по мере роста спроса на возможности обработки естественного языка. Некоторые популярные примеры включают в себя: [7]

* GPT-4 от OpenAI
* BERT от Google
* RoBERTa от Facebook
* Microsoft Turing-NLG
* LLaMA (Large Language Model Meta AI)
* Mistral (Mistral AI)
* Bloom от группы экспертов Hugging Face

Все эти модели работают с помощью текстовых запросов. Принцип работы состоит в следующем:

1. на вход модели подается контекст (любой текст);
2. модель анализирует весь контекст и генерирует всего один токен (токен - слово, часть слова или символ);
3. модель добавляет токен к контексту и передает новый контекст в пункт 1, до тех пор, пока не будет получен ответ.

Языковая модель имеет ограничение на суммарное количество токенов, которое мы можем передать на вход и получить на выходе. В основном, это зависит от размера модели. Токен это слово или часть слова, символ и тд. 1 токен не всегда равен 1 слову. Для английского языка 1000 токенов в среднем равны 750 словам. Для русского языка 1000 токенов это всего около 375 слов. [5]

1. Сравнительный анализ LLM

Можно определить следующие ключевые аспекты сравнительного анализа:

1. Размер обучающих данных: Оценка влияния объема обучающих данных на качество моделей. Анализ важности разнообразия данных для более широкого обзора и лучшего обобщения моделей.

2. Адаптация к контексту: Исследование способности моделей адаптироваться к контексту, учитывать предыдущие запросы пользователя, что имеет большое значение в сфере диалоговых систем.

3. Интерпретируемость: Анализ того, насколько языковые модели являются интерпретируемыми, и их способность предоставлять понятные решения. Это критически важно для установления доверия со стороны конечных пользователей.

4. Обработка многозначности: Оценка эффективности моделей в разрешении многозначных запросов, двусмысленных и сложных сценариев. Рассмотрение способности моделей к правильной интерпретации разнообразных вопросов.

5. Вычислительные требования и доступность: Исследование вычислительных требований каждой модели, выявление критических факторов, которые могут влиять на практическое применение в различных проектах. [2]

1. Характеристики LLM

Существуют различные параметры LLM моделей (примечание: названия параметров могут отличаться в зависимости от разработчика LLM):

1. **Temperature** (Температура). Чем ниже температура, тем более детерминированными являются результаты в том смысле, что всегда выбирается наиболее вероятный следующий токен. Повышение температуры может привести к большей случайности, что будет способствовать более разнообразным и творческим результатам. Таким образом, увеличивается вес других возможных токенов. С точки зрения применения можно использовать более низкое значение температуры для таких задач, как контроль качества на основе фактов, чтобы стимулировать более фактические и краткие ответы. Для создания стихотворений или других творческих задач может быть полезно увеличить значение температуры.
2. **Top P** — метод выборки с температурой, называемый выборкой ядра, при котором можно контролировать степень детерминированности модели. Если нужны точные и фактические ответы, данный параметры должен быть на низком уровне. Если требуются более разнообразные ответы, значение следует увеличить. Если в модели есть параметр Top P, это означает, что для ответов учитываются только токены, составляющие вероятностную массу top\_p, поэтому низкое значение top\_p выбирает наиболее уверенные ответы. Это означает, что высокое значение top\_p позволит модели просмотреть больше возможных слов, в том числе менее вероятных, что приведет к более разнообразным результатам. Общая рекомендация – изменить температуру или Top P, но не оба сразу.
3. **Max Length (**максимальная длина). Данные параметр управляет количеством токенов, генерируемых моделью, регулируя максимальную длину. Указание максимальной длины помогает предотвратить длинные или ненужные ответы и контролировать затраты.
4. **Стоп-последовательности**. Это строка, которая останавливает модель от генерации токенов. Указание последовательности остановки — это еще один способ контролировать длину и структуру ответа модели. Например, можно указать модели генерировать списки, содержащие не более 10 элементов, добавив «11» в качестве стоп-последовательности.
5. **Frequency Penalty** (штраф за частоту). Штраф за частоту накладывает штраф на следующий токен, пропорциональный тому, сколько раз этот токен уже появлялся в ответе и подсказке. Чем выше штраф за частоту, тем меньше вероятность того, что слово появится снова. Этот параметр уменьшает повторение слов в ответе модели, назначая более высокий штраф за лексемы, которые кажутся более выраженными.
6. **Presence Penalty (**штраф за присутствие). Штраф за присутствие также применяет штраф к повторяющимся токенам, но, в отличие от штрафа за частоту, штраф одинаков для всех повторяющихся токенов. Токен, который появляется дважды, и токен, который появляется 10 раз, наказываются одинаково. Эта настройка не позволяет модели слишком часто повторять фразы в ответе. Если требуется, чтобы модель генерировала разнообразный или креативный текст, можно использовать более высокий штраф за присутствие. Чтобы модель оставалась детерминированной, используется более низкий штраф за присутствие. [8]
7. Система оценки LLM (Benchmark)

В качестве моделей для сравнения выбраны LLaMa и Gemma. Версии моделей следующие: LLaMa-2-7b и Gemma-7b.

Gemma выделяется своими расширенными архитектурными функциями, включая механизмы многозапросного внимания, встраивания RoPE, активации GeGLU и стратегическое размещение нормализаторов. Эти инновации позволяют Gemma преуспеть в обработке сложных задач в контекстах увеличенной длины. Обученная на колоссальном наборе данных, содержащем до 6 триллионов токенов, Gemma фокусируется на текстах на английском языке, оттачивая свои модели, настроенные на инструкции, посредством сочетания контролируемой точной настройки и обучения с подкреплением на основе отзывов людей.

LLaMa, разработанная Meta, использует архитектуру на основе трансформатора, оптимизированную для масштабируемости и эффективности. Модель предназначена для адаптации к широкому кругу задач с минимальной тонкой настройкой благодаря обучению на разнообразном наборе данных. В моделях LLaMa приоритет отдается универсальности и простоте интеграции в различные приложения. [4]

Оценочные показатели LLM, такие как правильность ответов, семантическое сходство — это показатели, которые оценивают результаты LLM на основе заранее отобранных критериев. Например, если приложение с LLM предназначено для обобщения страниц новостных статей, вам понадобится показатель оценки LLM, который оценивается на основе информативности и противоречий в тексте, собранным LLM моделью.

Результаты LLM, как известно, трудно оценить. Существует множество устоявшихся методов расчета показателей: некоторые используют нейронные сети, в том числе модели внедрения и LLM (Model-Based), тогда как другие полностью основаны на статистическом (лингвистическом и грамматическом) анализе (Statistical) (см рис. 1). [11]

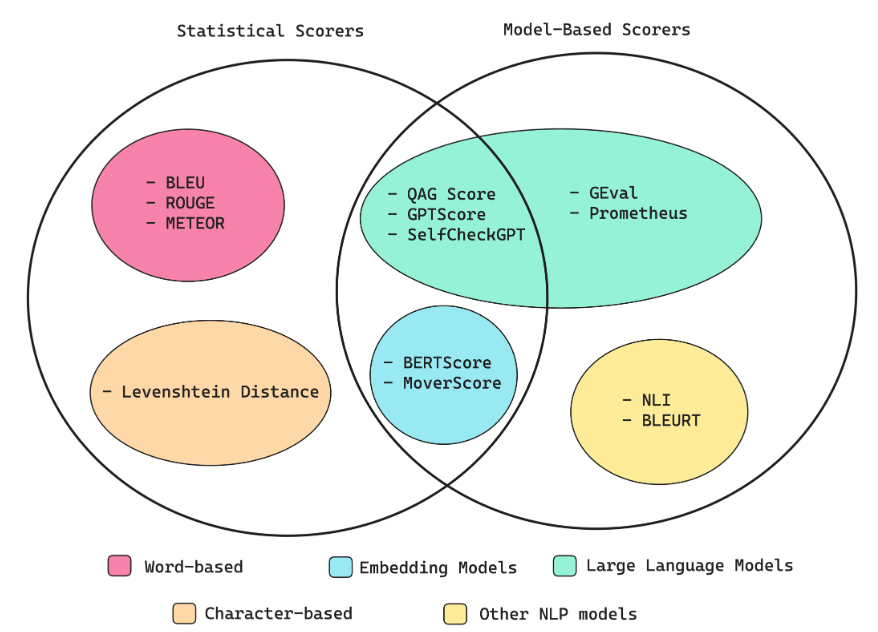


Рисунок 1. Виды метрик (оценочные показатели).

Выбранная система оценки от HuggingFace [10] оценивает модели по 6 ключевым критериям, используя модель Eleuther AI Language Model Evaluation Harness, — унифицированную среду для тестирования генеративных языковых моделей на большом количестве различных задач оценки.

Виды задач (столбцы таблицы 1):

AI2 Reasoning Challenge (ARC) — набор вопросов по естественным наукам для начальной школы.

HellaSwag — тест на логические выводы, который прост для людей (~95%), но сложен для моделей.

MMLU — тест для измерения точности многозадачности текстовой модели. Тест охватывает 57 задач, включая элементарную математику, информатику, право и многое другое.

TruthfulQA — тест для измерения склонности модели воспроизводить ложь, часто встречающуюся в Интернете.

Winogrande — сложный тест, проверяющий модели на здравые рассуждения.

GSM8k — разнообразные математические задачи для начальной школы, позволяющие оценить способность модели решать многоэтапные математические задачи. [10]

Для всех этих оценок более высокий балл означает лучший результат. HuggingFace выбрала эти тесты, поскольку они проверяют различные рассуждения и общие знания в самых разных областях.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Модель | Среднее | ARC | HellaSwag | MMLU | TruthfulQA | Winogrande | GSM8K |
| Gemma | 64,29 | 61,09 | 82,47 | 66,03 | 44,91 | 78,45 | 52,77 |
| LLaMa2 | 50,97 | 53,07 | 78,59 | 46,87 | 38,76 | 74,03 | 14,48 |

Таблица 1. Параметры Gemma и LLaMa2.

Из таблицы 1 видно, что Gemma превосходит LLaMa2; особенно в тесте GSM8K. Стоит отметить, что Gemma более «тяжелая модель». Её работа может быть нестабильна на слабых машинах, т.к. она требует больше вычислительных мощностей. Однако, при работе с «облачными» версиями LLM необходимость в мощных вычислительных машинах отпадает – требуется только механизм общения с моделью, например, REST API.

ЗАКЛЮЧЕНИЕ

В работе описан принцип функционирования LLM моделей. Рассмотрены популярные модели LLM, алгоритм их анализа, а также общие характеристики. Дано определение «оценочные показатели LLM», описаны их виды и различия. Проанализированы качественные показатели и производительность выбранных моделей: Gemma и LLaMa2.

Gemma продемонстрировала превосходные результаты в академических тестах, особенно в областях, требующих глубокого понимания, таких как математика, естественные науки, программирование и задачи на рассуждение. Модель заметно превосходит модель LLaMa2, демонстрируя свое мастерство в решении сложных сценариев решения проблем.

Хотя LLaMa2 также демонстрируют высокую производительность при выполнении различных задач, передовая архитектура Gemma и целенаправленный подход к обучению дают ей преимущество в специализированных областях.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Системы LLM-агентов для решения задач ретроспективного анализа / А. Б. Смольников, С. А. Черномуров, О. Д. Евстратова, Е. М. Заботкина // Наукосфера. – 2024. – № 2-2. – С. 215-219. – EDN CTHGQD. Дата обращения: 05.03.2024.
2. Иванов, К. Н. Русскоязычные LLM: сравнительный анализ и перспективы применения в проектах / К. Н. Иванов, С. В. Кушуков // Актуальные проблемы информатики, радиотехники и связи : Материалы XXXI Российской научно-технической конференции, Самара, 01–02 февраля 2024 года. – Самара: Поволжский государственный университет телекоммуникаций и информатики, 2024. – С. 240-241. – EDN VSSFUY. Дата обращения: 05.03.2024.
3. Использование токенов в LLM // URL: https://dzen.ru/a/ZBKqdO5TiGGPDjJN. Дата обращения: 05.03.2024.
4. Gemma vs Llama vs Mistral: A Comparative Analysis with a Coding Twist// URL: <https://medium.com/@kagglepro.llc/gemma-vs-llama-vs-mistral-a-comparative-analysis-with-a-coding-twist-8eb4d849e4d5>. Дата обращения: 05.03.2024.
5. ChatGPT: влияем на галлюцинации или как потопаешь, так и полопаешь // URL: [https://habr.com/ru/articles/727458/](https://habr.com/ru/articles/727458/%20). Дата обращения: 05.03.2024.
6. Generative Agents: Interactive Simulacra of Human Behavior // URL: <https://arxiv.org/abs/2304.03442>. Дата обращения: 05.03.2024.
7. Большие языковые модели: полное руководство в 2024 году // URL: <https://research.aimultiple.com/large-language-models/>. Дата обращения: 05.03.2024.
8. Настройки LLM // URL: <https://www.promptingguide.ai/ru/introduction/settings>. Дата обращения: 05.03.2024.
9. Using Large-Language Models (LLM) in Game Development // URL: <https://gamedevacademy.org/using-large-language-models-llm-in-game-development-tutorial-list/#How_Can_Large_Language_Models_Be_Used_for_Games>. Дата обращения 05.03.2024.
10. Open LLM Leaderboard // URL: <https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard>. Дата обращения 05.03.2024.
11. LLM Evaluation Metrics: Everything You Need for LLM Evaluation // URL <https://www.confident-ai.com/blog/llm-evaluation-metrics-everything-you-need-for-llm-evaluation#what-are-llm-evaluation-metrics->. Дата обращения 05.03.2024.