Санкт-Петербургский государственный университет
Кафедра системного программирования

Система для автоматизированной проверки больших языковых моделей на основе заданий государственных экзаменов

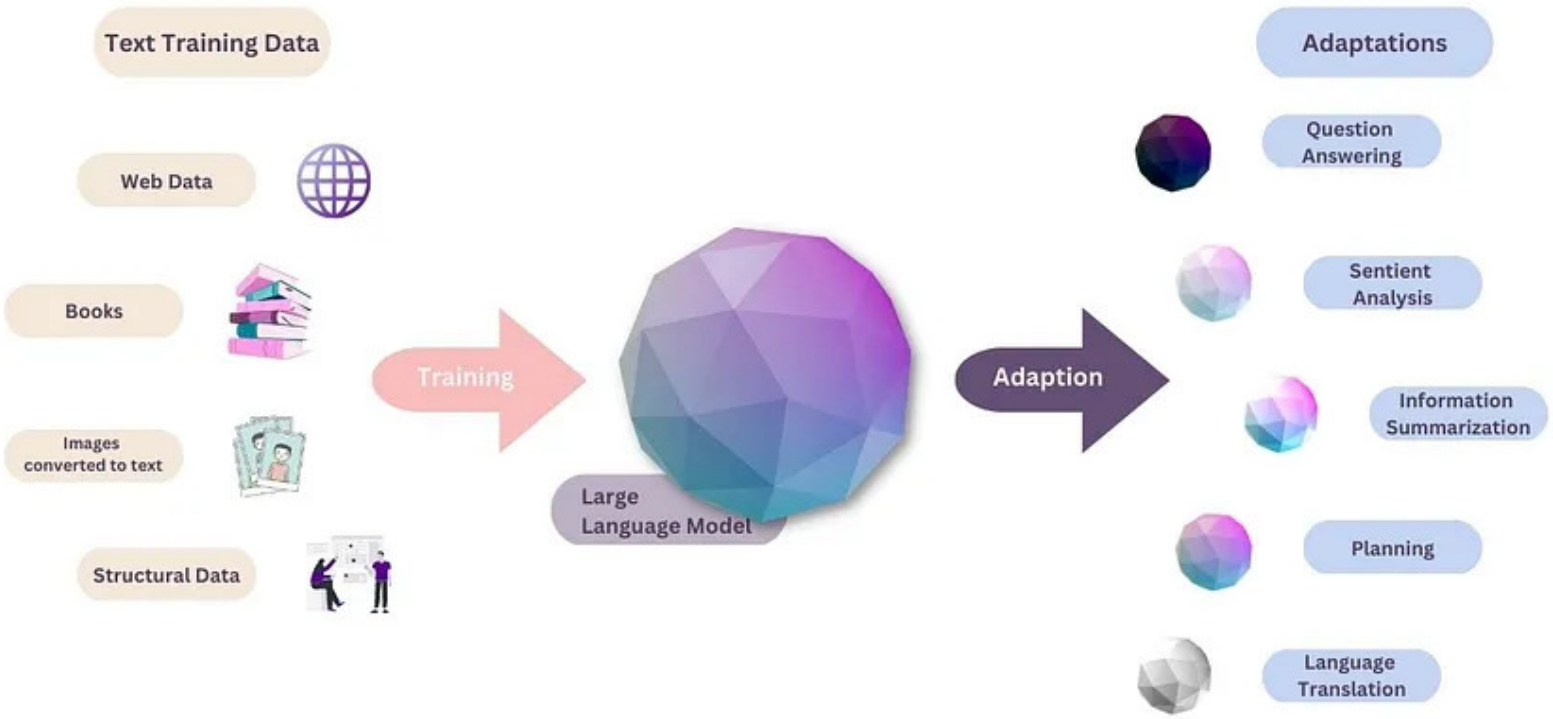
Михаил Владимирович Выродов, 371 группа

Научный руководитель: Е. С. Спирин, ассистент кафедры СП

Санкт-Петербург
2024

Введение

- В современном мире большие языковые модели (БЯМ) развиваются с огромной скоростью. Они уже успели принести много пользы людям;



Введение

- MMLU [1] — пример набора данных для валидации языковых моделей (ЯМ) на английском языке;

As of 2017, how many of the world's 1-year-old children today have been vaccinated against some disease?
(A) 80%
(B) 60%
(C) 40%
(D) 20%

Рис. 1: Пример вопроса из MMLU по теме "Global Facts"

- MT-Bench [2] — валидационный набор данных, состоящий из заданий на генерацию связного текста;
- Задания из Единого и Основного Государственных Экзаменов могут послужить хорошими данными для валидации ЯМ;

Обзор существующих решений

- Open-source проект от Сбера MERA[1]:
  - Поддерживает валидацию на нескольких популярных наборах данных;
  - Есть валидационный набор данных USE[2], состоящий из заданий ЕГЭ по русскому языку;
  - Есть либерборд ЯМ, который позволяет сравнивать их между собой и выбирать наиболее эффективные;
- Минусы:
  - В проекте есть задания из ЕГЭ только по русскому языку;
  - Не поддерживает валидацию на заданиях, требующих генерацию связного текста;
  - MERA делает валидацию более удобной, но не автоматизирует её полностью;

[1] https://mera.a-ai.ru/ru
[2] https://mera.a-ai.ru/ru/tasks/15

Постановка задачи

Целью работы является создание системы, предназначенной для валидации больших ЯМ, которая будет основана на наборе данных, состоящем из заданий государственных экзаменов;
Задачи:

- Провести обзор предметной области, анализ существующих фреймворков для создания веб-приложений и существующих бенчмарков для валидации ЯМ;
- Собрать банк вопросов из государственных экзаменов, провести очистку и классифицировать задания по типам вопросов и способам взаимодействия с ЯМ;
- Создать систему, которая будет валидировать подаваемую на вход ЯМ на собранном наборе данных;
- Апробировать систему на популярных ЯМ, построить их качественное сравнение;

Сбор данных

- Всего с сайта sdamgia.ru было извлечено 6211 заданий по обществознанию и литературе. Результаты парсинга приведены в таблице. Архив с извлечёнными заданиями доступен по ссылке;
- Все извлечённые задания были разбиты на 3 типа, в зависимости от способа взаимодействия с ЯМ;

Тип экзамена	Предмет	Количество вопросов	Выбор ответа	Соответствие	Текстовый ответ
ЕГЭ	Обществознание	2936	1751	580	605
ЕГЭ	Литература	1174	108	66	1000
ОГЭ	Обществознание	1673	1130	126	417
ОГЭ	Литература	428	0	0	428

Архитектура

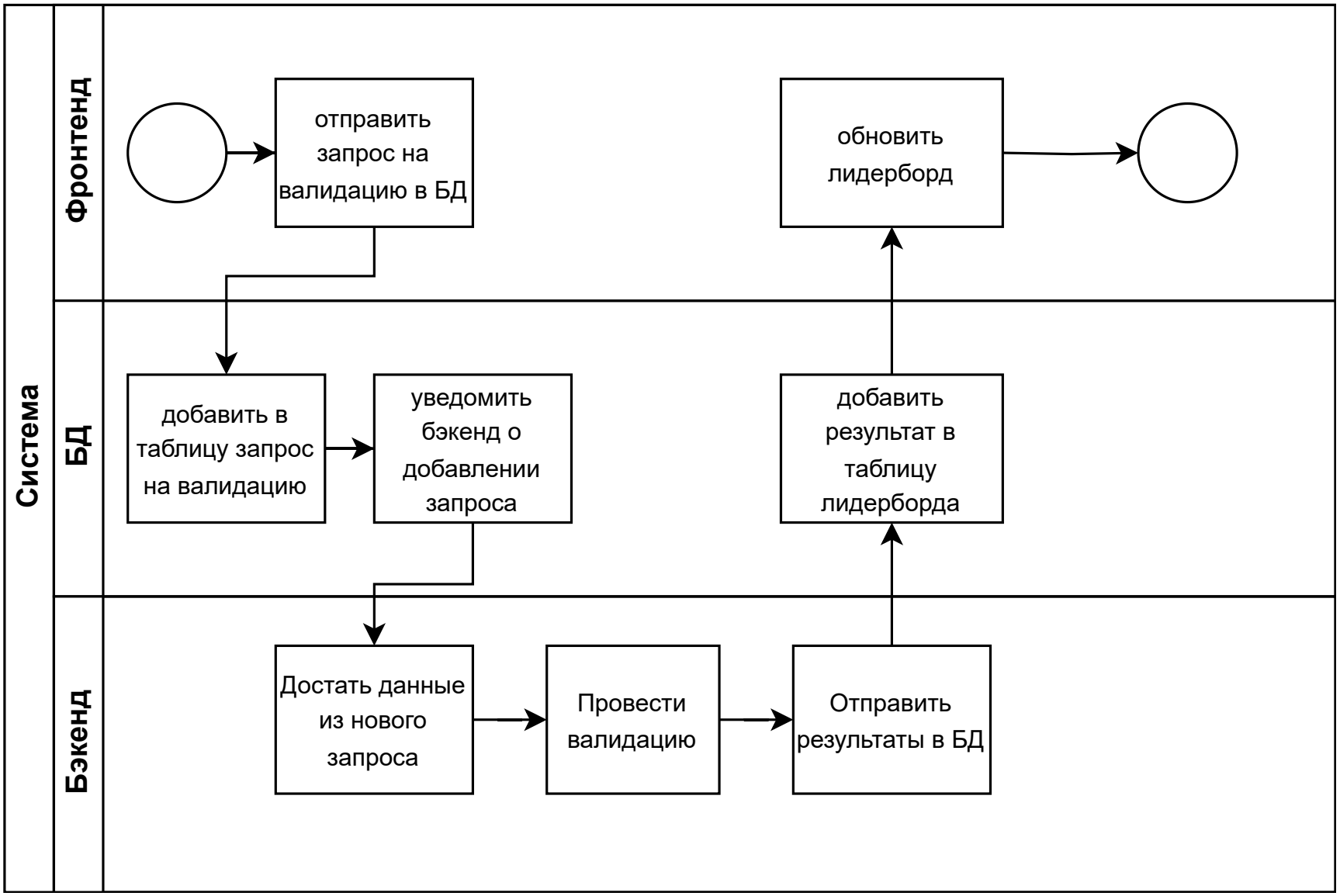


Рис. 2: Диаграмма активности системы

Стек технологий

- Парсинг сайтов с заданиями ЕГЭ/ОГЭ - scrapy;
- Frontend - HuggingFace space, gradio;
- База данных - PostgreSQL;
- Backend:
  - LM evaluation harness[3] ($\approx 5k$ звёзд);
  - FastChat[4] ($\approx 35k$ звёзд);

[3] https://github.com/EleutherAI/lm-evaluation-harness
[4] https://github.com/lm-sys/FastChat/

Веб-приложение

- Веб-приложение состоит из двух вкладок — лидерборд и отправка запроса на валидацию.

LLM Benchmark | Submit

model	single_choice	multiple_choice	word_gen
mistralai/Mistral-7B-Instruct-v0.2	0.77	0.23	0.1
meta-llama/Llama-2-7b-chat-hf	0.59	0.05	0.08
TinyLlama/TinyLlama-1.1B-Chat-v1.0	0.36	0.02	0.01

Рис. 3: Фрагмент лидерборда моего веб-приложения

Детали реализации

- Для поддержки микросервисности на сервере с GPU валидация моделей запускается в Docker контейнере;
- БД на удалённом сервере тоже запускается в Docker контейнере;
- Код проекта находится в репозитории[5] на GitHub;

[5] https://github.com/deepvk/goat/tree/leaderboard

Результаты

- В ходе обзора предметной области была рассмотрена система для облегчённой валидации ЯМ. Её преимущества и недостатки были проанализированы и в результате были сформированы требования, которым моя система для валидации должна удовлетворять;
- Был собран набор данных, состоящий из 6211 заданий государственных экзаменов. Все извлечённые задания были разбиты на 3 типа, в зависимости от способа взаимодействия с ЯМ;
- Было создано веб-приложение на HuggingFace для автоматической валидации ЯМ на собранном наборе данных, удовлетворяющее сформированным ранее требованиям;
- Несколько популярных open-source моделей были оценены с помощью созданной системы. Результаты валидации удовлетворили ожиданиям;

Метрики для валидации ЯМ I

- Для заданий, в которых надо выбрать один вариант ответа, использовалась метрика accuracy, которая считает процент заданий, в которых ЯМ выбрала правильный вариант ответа, от всех заданий в наборе данных. Данная метрика использовалась в бенчмарке MMLU [1];
- Для заданий, ответом на которые является число, слово или словосочетание, использовалась метрика Exact match, которая считает, что задание выполнено правильно, только тогда, когда ответ ЯМ полностью совпадает с правильным ответом на задание. Такая метрика используется в бенчмарке SQuAD [3];
- Для заданий, в которых надо выбрать несколько вариантов ответа, метрика Exact Match была модифицирована так, что порядок цифр, означающих варианты ответа в задании, не учитывался;

Метрики для валидации ЯМ II

- Для заданий, ответом на которые является связный текст из нескольких предложений, есть два варианта оценивания:
  - Найм ассесоров, которые будут оценивать ответ ЯМ по официальным критериям оценивания данных заданий;
  - Использование других БЯМ. Сначала выбирается достаточно хорошая БЯМ. После этого в неё загружается оцениваемый ответ на задание и официальные критерии оценивания этого задания, и выбранную БЯМ просят оценить ответ на задание по загруженным критериям;
- В проекте валидация на таких заданиях была реализована вторым способом через бенчмарк FastChat, который облегчает проверку ответов модели на такие задания с помощью GPT-4;

Список литературы

[1] Dan Hendrycks и др. “Measuring Massive Multitask Language Understanding”. В: *CoRR* abs/2009.03300 (2020). URL: https://arxiv.org/abs/2009.03300.

[2] Lianmin Zheng и др. “Judging LLM-as-a-judge with MT-Bench and Chatbot Arena”. В: *CoRR* abs/2306.05685 (2023). DOI: 10.48550/ARXIV.2306.05685. arXiv: 2306.05685. URL: https://doi.org/10.48550/arXiv.2306.05685.

[3] Pranav Rajpurkar и др. “SQuAD: 100, 000+ Questions for Machine Comprehension of Text”. В: *CoRR* abs/1606.05250 (2016). URL: http://arxiv.org/abs/1606.05250.