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 *«Фундаментальные науки»*

Кафедра *«Математическое моделирование»*

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

к научно-исследовательской работе

на тему:

ОБНОВЛЯЕМЫЙ БЕНЧМАРК ДЛЯ ОЦЕНКИ ОБЩИХ ЗНАНИЙ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Студент группы ФН12-31М _____ *Р.Н. Новиков*
(подпись, дата)

Руководитель НИР _____ *И.А. Ташков*
(подпись, дата)

Москва — 2025

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 ФН-12

_____ А.П. Крищенко

«____» _____ 20___ г.

ЗАДАНИЕ
на выполнение научно-исследовательской работы

по теме

«Обновляемый бенчмарк для оценки общих знаний больших языковых моделей»

Студент группы ФН12-31М

Новиков Руслан Николаевич

(Фамилия, Имя, Отчество)

Направленность НИР учебная

(учебная, исследовательская, практическая, производственная и др.)

Источник тематики кафедра

График выполнения работы: 25% к __ нед., 50% к __ нед., 75% к __ нед., 100% к __ нед.

Задание _____

Оформление научно-исследовательской работы:

Расчетно-пояснительная записка на ____ листах формата А4.

Перечень графического (иллюстративного) материала _____

Дата выдачи задания «____» _____ 20___ г.

Руководитель НИР _____ *И.А. Ташков*

(подпись, дата)

Студент _____ *Р.Н. Новиков*

(подпись, дата)

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

Содержание

1. Введение

Прогресс в области искусственного интеллекта последних лет во многом определяется успехами в разработке больших языковых моделей (Large Language Models, LLM). Способность этих систем к пониманию, генерации и рассуждению на естественном языке открывает новые перспективы для науки, образования и индустрии. Однако вместе с ростом возможностей моделей возникает фундаментальный вопрос: как достоверно и объективно измерить их реальную интеллектуальную эффективность. Традиционные методы оценки, основанные на фиксированных наборах данных, сталкиваются с непреодолимым противоречием: модели развиваются быстрее, чем успевают создаваться адекватные инструменты для их тестирования.

Основная проблема современных подходов к оценке заключается в трёх ключевых ограничениях.

- Первое ограничение – это статичность и насыщаемость большинства существующих наборов данных для тестирования.
- Второе – отсутствие механизмов для проверки работы модели со свежей, актуальной информацией.
- Третье – сложность масштабирования и адаптации методов извлечения структурированных данных из неоднородных и меняющихся источников.

Эти вызовы делают необходимым создание не просто нового набора вопросов, а целой методологии и инфраструктуры для их автоматизированного, гибкого и масштабируемого пополнения.

Целью данной научно-исследовательской работы является разработка и апробация прототипа обновляемого бенчмарка для оценки общих знаний LLM. Предметом исследования выступают методы построения динамических оценочных систем с использованием самих языковых моделей для автоматизации ключевых процессов. Ключевой задачей является создание конвейера, способного извлекать вопросы и ответы из сложноструктурированных, но регулярно обновляемых источников.

В качестве основного источника данных для бенчмарка была выбрана популярная российская телевикторина “Своя игра”. Её текстовые расшифровки, публикуемые на специализированном форуме, представляют собой качественный, верифицированный и постоянно пополняемый массив вопросов. Данный источник позволяет проверять широкий спектр общих знаний и эрудицию на русском языке. Для решения задачи извлечения структурированных данных из вариативных текстовых описаний был применён инновационный подход – использование LLM в качестве универсального парсера. В работе использовалось API модели DeepSeek для преобразования неструктурированного текста в формализованные пары “вопрос–ответ”. Этот подход преодолевает хрупкость регулярных выражений и вычислительную затратность обучения специализированных моделей. Он обеспечивает устойчивость к изменениям формата исходных данных и высокую степень масштабируемости для подключения новых источников в будущем.

Таким образом, разрабатываемый бенчмарк представляет собой саморазвивающуюся оценочную систему. Его архитектура основана на автоматизированном конвейере, который обеспечивает постоянное пополнение базы новыми, актуальными

вопросами. Это позволяет проводить оценку LLM на данных, которые гарантированно не входили в их предобучение, и проверять способность работать со знаниями о постоянно меняющемся мире.

2. Обзор существующих бенчмарков

Современные подходы к оценке больших языковых моделей можно разделить на две основные категории: статические и динамические бенчмарки. В данном разделе рассматриваются ключевые представители каждого типа, их методология, сильные стороны и ограничения.

MMLU (Massive Multitask Language Understanding)

MMLU [4] является одним из наиболее распространённых статических бенчмарков. Он состоит из 15 908 вопросов с множественным выбором, охватывающих 57 академических дисциплин – от математики и права до питания и религии. Формат тестирования предполагает выбор одного правильного ответа из четырёх предложенных, что позволяет автоматически вычислять точность. Бенчмарк был создан в 2020 году и быстро стал стандартом для сравнения моделей благодаря своей широте и сложности.

Однако MMLU страдает от фундаментальных проблем статичности: вопросы фиксированы и со временем попадают в тренировочные данные новых LLM, что приводит к «загрязнению» (contamination) и искусственному завышению результатов. Несмотря на это, MMLU продолжает широко упоминаться в исследованиях и сравнительных таблицах. Это связано с его исторической ролью как первого масштабного многозадачного теста, а также с тем, что он предоставляет простую и сопоставимую метрику, удобную для отслеживания общего прогресса в отрасли.

Для преодоления ограничений MMLU был предложен его улучшенный вариант – **MMLU-Pro**. Он расширяет набор ответов с четырёх до десяти вариантов, устраняет тривиальные и шумные вопросы из оригинального набора и делает акцент на сложных рассуждениях, а не только на фактологических знаниях. Это приводит к значительному падению точности моделей (на 16–33

Таблица 1: Примеры вопросов из различных предметных областей MMLU

Question 1 (Abstract Algebra):
Find all c in $\mathbb{Z}_3$ such that $\mathbb{Z}_3[x]/(x^2 + c)$ is a field.

(A) 0

(B) 1

(C) 2

(D) 3

Question 2 (International Law):
Would a reservation to the definition of torture in the **International Covenant on Civil and Political Rights (ICCPR)** be acceptable in contemporary practice?

(A) This is an acceptable reservation if the reserving country's legislation employs a different definition.

(B) This is an unacceptable reservation because it contravenes the object and purpose of the ICCPR.

(C) This is an unacceptable reservation because the definition of torture in the ICCPR is consistent with customary international law.

(D) This is an acceptable reservation because under general international law States have the right to enter reservations to treaties.

Question 3 (Professional Medicine):
A 33-year-old man undergoes a radical thyroidectomy for thyroid cancer. During the operation, moderate hemorrhaging requires ligation of several vessels in the left side of the neck. Postoperatively, serum studies show a calcium concentration of 7.5 mg/dL, albumin concentration of 4 g/dL, and parathyroid hormone concentration of 200 pg/mL. Damage to which of the following vessels caused the findings in this patient?

(A) Branch of the costocervical trunk.

(B) Branch of the external carotid artery.

(C) Branch of the thyrocervical trunk.

(D) Tributary of the internal jugular vein.

Humanity's Last Exam (HLE)

Humanity's Last Exam [6] – это статический бенчмарк, созданный как ответ на стремительный прогресс LLM. Его цель – предложить вопросы, находящиеся на переднем крае человеческих знаний. Набор состоит из более чем 2500 экспертных вопросов, охватывающих математику, физику, биологию, гуманитарные науки и другие области. Около 24

Ключевая особенность HLE – строгий процесс контроля качества. Вопросы создаются и проверяются экспертами с учёными степенями (PhD), а за лучшие вопросы авторы получают денежные призы. Такой подход обеспечивает высокую сложность и надёжность вопросов, но делает бенчмарк крайне затратным и медленным в обновлении. Кроме того, в 2025 году независимое исследование выявило потенциальные

ошибки в части ответов, что подчёркивает трудности поддержания безупречного качества в таких масштабных экспертных проектах. HLE служит «высокой планкой» для самых передовых моделей, но его статичность и медленный цикл обновления не позволяют ему стать инструментом для оперативного сравнения.

Таблица 2: Пример вопроса из Humanity's Last Exam

Question:
Hummingbirds within Apodiformes uniquely have a bilaterally paired oval bone, a sesamoid embedded in the caudolateral portion of the expanded, cruciate aponeurosis of insertion of m. depressor caudae. How many paired tendons are supported by this sesamoid bone? Answer with a number.

Chatbot Arena

Chatbot Arena [2] – это краудсорсинговый бенчмарк для оценки языковых моделей, основанный на прямых сравнениях. Пользователи заходят на веб-платформу и вводят любой интересующий их запрос. Система анонимно выбирает две разные языковые модели и генерирует их ответы на этот запрос. Пользователю показываются два ответа без указания, какая модель их сгенерировала. Задача пользователя – сравнить ответы и выбрать лучший, либо отметить ничью. Каждое такое голосование называется «баттлом». На основе миллионов подобных парных сравнений вычисляется рейтинг для каждой модели. Для расчёта рейтинга используется система Эло, аналогичная шахматной. Чем чаще модель выигрывает в сравнениях, тем выше её позиция на публичном лидерборде. Таким образом, лидерборд отражает коллективные человеческие предпочтения, а не автоматические метрики.

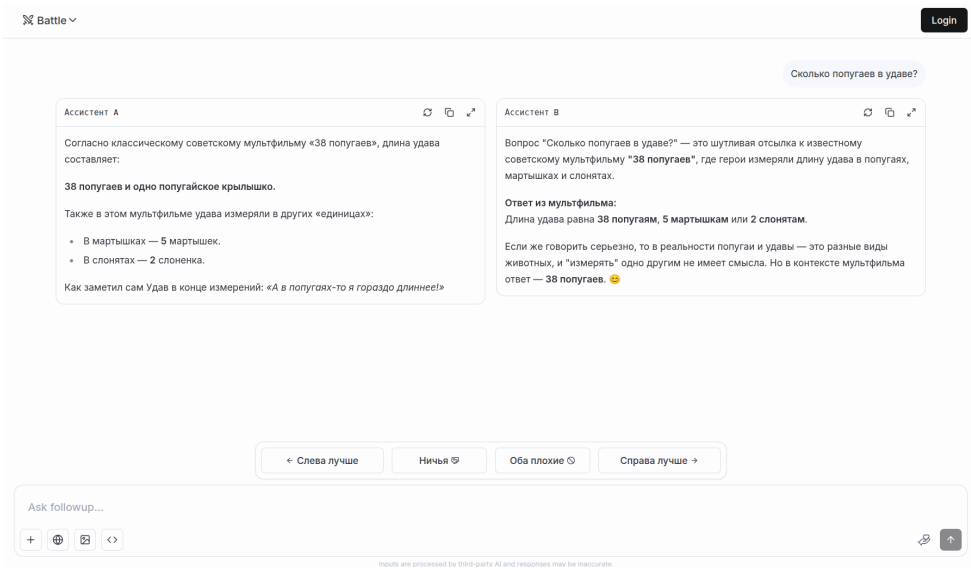

Рис. 1: Интерфейс оценивания в Chatbot Arena

Плюсы Chatbot Arena

Ключевым преимуществом Chatbot Arena является динамическое обновление лидерборда. В отличие от статических тестов, рейтинг здесь «живой» и меняется в реальном времени по мере поступления новых голосов. Это позволяет мгновенно отражать выход новых версий моделей или обновлений. Система устойчива к целевой оптимизации или «переобучению» под себя. Разработчики не могут намеренно «подстроить» модель под конкретные вопросы бенчмарка, так как набор запросов непредсказуем. Модели в каждом баттле выбираются анонимно, поэтому невозможно подготовиться к соревнованию с конкретным оппонентом. Критерий оценки – человеческое предпочтение – сложно имитировать с помощью автоматических методов. Сам массив запросов крайне разнообразен и создаётся реальными пользователями, что делает его практически невоспроизводимым для тренировки. Это обеспечивает высокую объективность оценки способности модели общаться в реальных условиях.

Минусы Chatbot Arena

Главным минусом является предвзятость пользовательской аудитории, формирующей лидерборд. Активнее всего голосуют технические специалисты, разработчики и энтузиасты ИИ. Это приводит к тематическому смещению в запросах и оценках. В бенчмарке непропорционально много вопросов по программированию, машинному обучению и технологиям. В то же время сильно недопредставлены бытовые, творческие или узкопрофессиональные темы (например, медицина или право). Стилевые предпочтения этой аудитории также влияют на результат: часто ценится излишняя детализация и технический жаргон, что не всегда оптимально для среднего пользователя. Существует и асимметрия в данных: популярные коммерческие модели участвуют в большем числе баттлов, что позволяет точнее оценить их рейтинг. Модели с открытыми весами часто оказываются в менее выгодном положении из-за меньшего количества сравнений. Наконец, сохраняется теоретический риск манипуляций с голосованием, хотя команда разработчиков активно внедряет защитные меры, такие как CAPTCHA и системы обнаружения аномалий.

LiveBench

LiveBench [7] – это открытый автоматизированный бенчмарк для оценки больших языковых моделей, созданный для решения проблем устаревания и «загрязнения» тестовых данных. Его ключевая идея – использование часто обновляемых вопросов, основанных на недавно опубликованных данных. Вопросы генерируются из актуальных источников: свежих научных статей на arXiv, новостей, датасетов, описаний фильмов и задач с математических олимпиад. Каждый вопрос имеет объективный, проверяемый эталонный ответ (ground truth), что позволяет оценивать модели автоматически, без привлечения людей или других ИИ-судей. Бенчмарк охватывает шесть основных категорий: математика, программирование, логические рассуждения, понимание языка, следование инструкциям и анализ данных. Внутри каждой категории есть несколько специализированных задач, всего более 18 задач. Для минимизации риска, что модели уже встречали тестовые данные во время обучения, набор вопросов регулярно обновляется, а публикация самых свежих вопросов задерживается. Оценка модели по LiveBench даёт единый процентный балл, а также

детальную разбивку по категориям, что позволяет понять её сильные и слабые стороны.

Model	Organization	Global Average	Reasoning Average	Coding Average	Agentic Coding Average	Mathematics Average	Data Analysis Average	Language Average	IF Average
Claude 4.5 Opus Thinking High Effort	Anthropic	75.61	80.09	79.65	63.33	90.39	71.98	81.26	62.55
GPT-5.1 Codex Max High	OpenAI	74.34	83.65	80.68	53.33	83.22	72.63	76.48	70.38
GPT-5.2 High	OpenAI	74.07	83.21	76.07	51.67	93.17	72.79	79.81	61.77
GPT-5.2 Codex	OpenAI	73.53	77.71	83.62	51.67	88.77	72.80	73.68	66.45
Gemini 3 Pro Preview High	Google	73.46	77.42	74.60	55.00	81.84	74.91	84.62	65.85

Рис. 2: Вершина лидерборда LiveBench на момент 08.01.2026

Плюсы LiveBench

Главным преимуществом LiveBench является его динамическая и устойчивая к загрязнению природа. Регулярное обновление вопросов из актуальных источников делает бенчмарк «живым» и значительно усложняет целевое переобучение моделей под него. Поскольку модели не могли заранее видеть эти конкретные вопросы в своих тренировочных данных, их результаты точнее отражают реальные способности к рассуждению и решению задач, а не навык запоминания. Объективная проверка по эталонным ответам исключает субъективные предубеждения, которые присущи оценке с помощью LLM-судей или краудсорсинга. Например, известно, что LLM-судьи часто предпочитают более многословные ответы или ответы, сгенерированные моделями, схожими с ними самими. LiveBench избегает этой ловушки, обеспечивая полностью автоматическую и беспристрастную оценку. Широта охвата – от решения математических задач до анализа табличных данных – делает бенчмарк комплексным инструментом для оценки общих возможностей модели. Сложность задач подобрана так, что даже лучшие модели на момент создания показывали accuracy ниже 70%, что позволяет эффективно различать современные мощные модели.

Минусы LiveBench

Основным и наиболее критичным для русскоязычного контекста минусом является языковая ограниченность бенчмарка. LiveBench создан и существует исключительно на английском языке. Все вопросы, инструкции и эталонные ответы представлены на английском. Бенчмарк напрямую не оценивает и не учитывает способности модели понимать, генерировать или рассуждать на русском языке или любом другом языке, кроме английского. Это делает его результаты малорелевантными для оценки применимости модели в русскоязычной среде, будь то бизнес-задачи, поддержка клиентов или создание контента. Несмотря на заявленную сложность, некоторые эксперты отмечают, что по мере роста популярности любого бенчмарка, включая LiveBench, возникает риск его «взлома» – целевой оптимизации моделей для достижения высоких результатов на конкретных типах задач бенчмарка, что может не коррелировать с общим качеством. Кроме того, несмотря на широкий охват категорий, бенчмарк всё же может иметь перекос в сторону формальных,

структурированных задач (математика, код, головоломки) в ущерб оценке креативности, социального интеллекта или сложного многоступенчатого диалога. Наконец, хотя автоматическая оценка объективна, она может быть менее гибкой для задач, где возможны несколько правильных или частично правильных ответов, которые сложно свести к одному эталону.

Бенчмарки специального назначения

Также можно выделить узкоспециализированные бенчмарки, которые проверяют отдельные способности модели.

- Например, игровой бенчмарк LLMArena [1] тестирует модель в 6 разных игровых средах: покер, аукционы, командные игры и др. Этот бенчмарк позволяет выяснить, отдельные аспекты моделей: пространственное-мышление, способность оценивать риск, математические способности, способность работать в команде и др.
- Бенчмарк PingPong [3], проверяющий, насколько LLM способна поддерживать ролевое общение в диалоге, насколько правдоподобно это получается у модели.
- Бенчмарк RuQualBench [5] оценивает, насколько модель знает грамматику русского языка. Но он тестирует исключительно грамматику, а не общие знания.

Поскольку в рамках данной работы стоит задача разработки бенчмарка для оценивания общих знаний, то не будем подробно вдаваться в детали данных бенчмарков, а лишь обойдёмся их упоминанием. Полезно понимать, что общие знания - не единственный аспект, который можно измерять у LLM.

3. Методология работы бенчмарка

Данный раздел описывает методологию построения обновляемого бенчмарка для оценки больших языковых моделей. Ключевой принцип заключается в создании не статичного набора данных, а динамического конвейера, способного автоматически извлекать, верифицировать и актуализировать вопросы для тестирования.

Выбор источника данных

В основе бенчмарка лежит идея использования публичных, регулярно обновляемых источников вопросов. В качестве первичного и концептуального источника была выбрана популярная российская телевизионная викторина **«Своя игра»**.

Обоснование выбора:

- **Высокое качество контента:** Вопросы викторины составлены профессиональными редакторами и тестируются на живых игроках, что гарантирует их содержательность, корректность и адекватный уровень сложности.
- **Проверяемость:** Для каждого вопроса существует точный, верифицированный ответ, что критически важно для объективной автоматической оценки моделей.
- **Разнообразие тем:** Вопросы охватывают широкий спектр областей знаний: история, наука, искусство, литература, география, спорт и другие, что позволяет оценивать *общие знания* и эрудицию модели.
- **Регулярность и актуальность:** Новые выпуски игры публикуются на официальном форуме в текстовом формате с высокой частотой, обеспечивая постоянный приток **свежих данных**, которые гарантированно отсутствуют в обучающих наборах моделей, выпущенных ранее.
- **Структурированность:** Текстовые отчёты на форуме сохраняют базовую структуру (темы, стоимость вопросов, ответы), что упрощает их последующий анализ.

Масштабируемость подхода: Выбранная методология не ограничивается одним источником. Архитектура конвейера спроектирована с учётом возможности интеграции дополнительных потоков данных:

- Другие викторины и интеллектуальные игры (как русскоязычные, так и мультиязычные).
- Новостные ленты и дайджесты, позволяющие формулировать вопросы на основе текущих событий.
- Специализированные форумы и энциклопедии.

Таким образом, бенчмарк эволюционирует из фиксированного набора в **самообновляемую систему** агрегации оценочных данных из множества источников.

4. Техническая реализация

Выбор и обоснование технологий

LLM как универсальный парсер

Извлечение структурированных вопросно-ответных пар из неоднородных текстовых отчётов представляет собой нетривиальную задачу.

- **Проблематика традиционных подходов:**
  - **Регулярные выражения (RegEx):** Из-за постоянных вариаций в формате текстовых отчётов (разметка, порядок блоков, оформление) создание и поддержка набора правил является *хрупким и трудномасштабируемым* решением. Любое изменение в источнике требует модификации кода.
  - **Специализированные ML-модели (NER, Information Extraction):** Обучение собственной модели для распознавания сущностей и отношений требует размеченного датасета, значительных вычислительных ресурсов для обучения и не гарантирует обобщения на новые форматы данных без дообучения.
- **Предлагаемое решение:** Использование API большой языковой модели (в данном прототипе – **DeepSeek**) в качестве ядра парсинга. LLM получает на вход необработанный текст и текстовую инструкцию (prompt), детально описывающую ожидаемый выход – структурированный JSON-объект.
- **Преимущества подхода:**
  - **Устойчивость к изменениям:** При изменении формата исходных данных корректируется лишь текстовая инструкция для LLM, а не логика парсера.
  - **Контекстное понимание:** Модель способна корректно интерпретировать сокращения, косвенные формулировки, извлекать ответы из повествовательных предложений.
  - **Нулевое обучение (Zero-shot/Few-shot):** Для работы не требуется предварительное обучение на размеченных данных конкретного источника.
  - **Лёгкость масштабирования:** Для подключения нового источника (новостная лента, другая викторина) достаточно создать для него новый промпт.

Пример промпта, используемого для парсинга “Своей игры”:

```
Извлеки все пары вопрос-ответ из текста ниже. Верни ТОЛЬКО валидный JSON-массив.
Не добавляй пояснений, комментариев, маркеров кода (типа ‘‘‘json).
Каждый элемент массива - объект с двумя полями: "question" (строка) и "answer" (стр
Выведи пары в формате JSONL: один JSON-объект на строку.

Пример:
{"question": "Какая столица в России?", "answer": "Москва"}
```

```
{"question": "Сколько будет 2 + 2?", "answer": "4"}
{"question": "Сколько округов в Париже?", "answer": "20"}

Текст: {text}
```

Хранение данных

Для хранения данных была выбрана иерархическая схема, начиная с простого решения для прототипа с перспективой роста.

- **Текущая СУБД: SQLite**
  - **Причина выбора:** Простота развёртывания (один файл), отсутствие необходимости в отдельном сервере и управлении пользователями. Идеально подходит для этапа прототипирования и локальной разработки.
  - **Недостатки:** Ограниченная производительность при высокой конкурентной нагрузке, отсутствие продвинутых механизмов параллелизма, что делает её неподходящей для production-среды с множеством одновременных запросов.
- **Планируемая СУБД: PostgreSQL**
  - **Причина миграции:** Потребность в надёжной системе управления данными для будущего публичного API и веб-интерфейса. PostgreSQL обеспечит стабильность, производительность, поддержку сложных запросов, полнотекстового поиска и удобное масштабирование.

Архитектура базы данных

Ниже представлена упрощённая схема основных таблиц базы данных и их взаимосвязей.

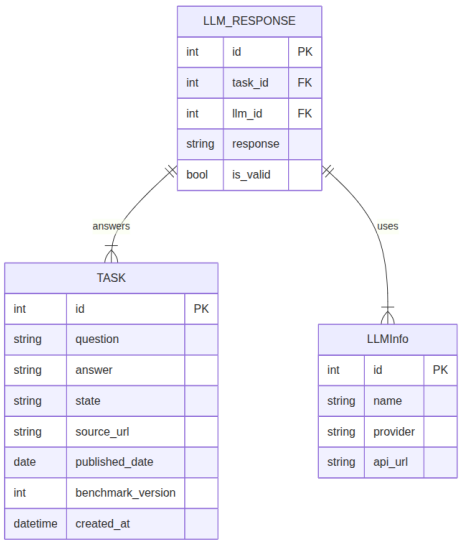

Рис. 3: Схема взаимосвязей таблиц в базе данных бенчмарка

Описание таблиц:

- `tasks` – Верифицированные задачи, готовые к использованию. Статус каждой задачи меняется в процессе эксплуатации бенчмарка: `'queue'` → `'benchmark'` → `'archive'`. На каждой итерации обновления бенчмарака используется поле `benchmark_version` - для версионирования бенчмарка и отслежавания истории измерений.
- `llm_infos` – Каталог тестируемых языковых моделей (GPT, Claude, Llama и др.).
- `llm_responses` – Ответы моделей на вопросы. Каждая запись связана с задачей и с моделью. Также в таблице хранится вердикт LLM-as-a-Judge - верен ли ответ модели.

Управление через командную строку (CLI)

На текущем этапе управление бенчмарком осуществляется через консольное приложение. Это позволяет гибко управлять процессом и интегрировать его в скрипты.

Команда `collect_tasks [url]`

Выполняет сбор и первоначальную обработку данных с указанного URL.
Алгоритм работы:

1. Загрузка HTML-страницы по указанному URL.
2. Извлечение основного текстового контента (игнорирование шапки, навигации, рекламы).
3. Отправка очищенного текста в LLM (DeepSeek API) с промптом для парсинга.
4. Получение и валидация JSON-ответа от LLM.
5. Сохранение извлечённых задач в таблицу `tasks` со статусом `'queue'`.

Получение структурированных данных от LLM выполняется в формате JSONL. Этот формат представляет из себя последовательность JSON-обектов, каждый на новой строке, без каких-либо лишних разделителей. Это позволяет считывать задачи в реальном времени и отслеживать прогресс. Также этот формат более устойчив к ошибкам форматирования. Объекты парсятся и сохраняются независимо друг от друга. Поэтому, если в одном из объектов возникнет синтаксическая ошибка, программа сможет безошибочно обработать остальные объекты - они будут добавлены в базу данных. В классическом же формате список объектов воспринимался бы как неделимый элемент. И ошибка в одном из элементов приводила бы к ошибке парсинга всего текста.

Команда `update_benchmark`

Ключевая команда, выполняющая ротацию задач в активном бенчмарке и запускающая тестирование.

Алгоритм работы:

1. **Проверка очереди:** Определяется количество задач со статусом `'queue'` в `tasks`. Если их меньше порогового значения (например, $N = 100$), процесс прерывается с соответствующим сообщением.

2. **Архивация текущей версии:** Задачи, входящие в текущую активную версию бенчмарка, помечаются как архивные.

3. **Создание новой версии:** В `benchmark_versions` создаётся новая запись с увеличенным номером версии и текущей датой. Она становится активной.

4. **Наполнение новой версии:** Первые N задач из очереди (`'queue'`) переносятся в бенчмарк (`'benchmark'`).

5. **Автоматический запуск тестирования:**

   (a) Для каждой модели из `llm_models` (помеченной для автотестирования) и для каждой новой задачи в созданной версии:

   - Вопрос отправляется в соответствующее API модели.
   - Ответ сохраняется в `model_responses`.
   - **Оценка ответов (LLM-as-a-Judge):** Для каждой тройки (`вопрос`, `эталонный_ответ`, `ответ_модели`) запускается процесс судейства. Отдельная LLM (на данный момент - DeepSeek) выступает в роли судьи. Она получает инструкцию сравнить два ответа на предмет семантической эквивалентности и вынести вердикт (`'OK'`, `'FAIL'`). Вердикт сохраняется в `llm_responses.is_valid`.

Планы по развитию

- **Автоматизация сбора:** Реализация фонового демона (daemon), который будет по расписанию (например, раз в сутки) проверять источники на наличие новых данных и запускать `collect_tasks` автоматически. Требует переноса системы на хостинг с постоянно работающим сервером.

- **Веб-интерфейс:** Разработка графического интерфейса пользователя (GUI) для:

  – Удобного просмотра лидербордов и истории версий.

  – Визуализации результатов (графики, сравнение моделей).

  – Управления источниками и моделями.

  – Ручной верификации спарсенных задач.

- **Расширение источников:** Интеграция парсеров для дополнительных викторин и новостных агрегаторов.

5. Эксперименты

Автоматизированное тестирование DeepSeek через API

Для автоматизированного тестирования была использована модель DeepSeek, доступная по API. Схема оценивания: "1 запрос - 1 ответ чтобы фокус внимания LLM был сконцентрирован только на одном вопросе. Это приоритетный способ оценивания, в отличие от пакетной оценки (несколько вопросов сразу). Тестирование проводилось на последовательных версиях бенчмарка (v1-v4). Результаты, полученные автоматически с использованием текущей версии системы LLM-as-a-Judge, представлены в Таблице 3.

Таблица 3: Результаты автоматизированного тестирования DeepSeek на разных версиях бенчмарка

Версия бенчмарка	**Модель**	**Accuracy**	**OK**	**FAIL**
1	DeepSeek	0.27	27	73
2	DeepSeek	0.42	42	58
3	DeepSeek	0.36	36	64
4	DeepSeek	0.34	34	66

Анализ результатов: Как видно из таблицы, метрика accuracy, рассчитанная автоматическим судьёй, варьируется в диапазоне от 0.27 до 0.42. Значительный разброс показателей между версиями бенчмарка указывает на возможную нестабильность либо самого судьи, либо на существенные различия в сложности вопросов между срезами данных. Среднее значение accuracy ≈ 0.35 является невысоким, что требует дополнительного исследования. Качественный анализ ошибочных ответов показал, что **LLM-as-a-Judge в текущей конфигурации демонстрирует излишнюю строгость**, засчитывая как некорректные ответы, которые являются семантически верными, но не совпадают с эталоном дословно (например, синонимичные формулировки или развернутые объяснения вместо краткого факта).

Ручное тестирование моделей через веб-интерфейс

Для получения предварительной сравнительной картины различных моделей был проведён эксперимент с использованием веб-интерфейсов. В данном случае использовался **альтернативный метод оценки**: ответы нескольких моделей на одну и ту же подборку вопросов были получены **единым пакетом** через веб-интерфейс. Затем были поданы **единым пакетом** в LLM-судью (DeepSeek) через её веб-интерфейс. Промпт судьи содержал инструкцию сравнить все ответы с эталоном сразу в рамках таблицы. Результаты данной пакетной оценки представлены в Таблице 4.

Анализ: Пакетная оценка показала ожидаемую оценку качества среди моделей, при этом абсолютные значения метрик (от 0.365 до 0.717) оказались значительно выше, чем в эксперименте с индивидуальными API-запросами. Это позволяет сделать два важных вывода:

Таблица 4: Сравнительные результаты ручного тестирования моделей через веб-интерфейс (оценка эксперта)

Ранг	**Модель**	**Accuracy**	**OK**	**FAIL**
1	DeepSeek	0.717	167	66
2	Qwen	0.614	143	90
3	YandexGPT	0.575	129	95
4	GigaChat	0.442	103	130
5	Perplexity	0.365	85	148

1. **Консистентность ранжирования:** LLM-судья успешно выполняет сравнительную функцию, выстраивая модели в порядке, соответствующем их ожидаемой мощности (Модель A $>$ B $>$ C $>$ D $>$ E).

2. **Влияние метода оценки:** Существенное расхождение в абсолютных значениях accuracy (пакетная оценка дала результаты в ≈ 2 раза выше) указывает на критическую зависимость итоговой метрики от **методологии взаимодействия с судьёй**. Различия в промптах (объём инструкций, наличие few-shot примеров, формат ввода данных) и режиме обработки (пакетный vs. поштучный) напрямую влияют на строгость суждения.

План корректировок и дальнейших работ:

1. **Стандартизация и калибровка LLM-as-a-Judge:** Для перехода к полностью автоматизированному API-тестированию необходимо разработать и зафиксировать эталонный промпт и протокол взаимодействия, которые обеспечат баланс между строгостью и смысловой адекватностью оценок. За основу будет взят промпт, показавший более реалистичные результаты при пакетной оценке. Однако будет использован именно изолированный подход к тестированию ("1 вопрос - 1 ответ") для более прозрачного оценивания LLM.

2. **Переход на массовое API-тестирование:** После калибровки судьи система будет доработана для проведения массовых тестов всех целевых моделей исключительно через их официальные API. Это обеспечит воспроизводимость, объективность сравнения и возможность регулярного обновления лидерборда.

Таким образом, проведённые эксперименты в большей степени выявляют и количественно определяют методологическую неопределённость в системе автоматической оценки, нежели измеряют абсолютные способности моделей. Устранение этой неопределённости составляет основную инженерную задачу следующего этапа работы.

6. Заключение

В ходе данной научно-исследовательской работы была изучена проблема объективной оценки больших языковых моделей. Была исследована ограниченность традиционных статических бенчмарков, таких как MMLU. Основное внимание уделялось современным динамическим подходам, включая LiveBench и Chatbot Arena. На основе проведённого анализа была сформулирована цель создания нового инструмента. Целью стала разработка динамического, обновляемого бенчмарка, устойчивого к загрязнению данных.

Была предложена и детально спроектирована архитектура такого бенчмарка. Ключевым принципом стал автоматизированный сбор вопросов из публичных, постоянно обновляемых источников. В качестве демонстрационного источника была выбрана викторина «Своя игра». Для преобразования неструктурированных текстов в данные был применён инновационный метод. Этот метод использует саму большую языковую модель в качестве универсального и гибкого парсера. Была разработана модульная система, включающая этапы сбора, структурирования, хранения и оценки. Для хранения данных была создана реляционная база данных на SQLite с перспективой перехода на PostgreSQL. Управление системой реализовано через консольное приложение с базовым набором команд.

В рамках экспериментальной части был успешно развёрнут и протестирован конвейер сбора данных. Парсинг с использованием LLM показал высокую точность. Были получены первые оценки моделей в двух режимах: через API и веб-интерфейс. Результаты выявили существенное влияние методологии оценки на итоговые метрики. Было установлено, что подход LLM-as-a-Judge требует тщательной калибровки для обеспечения консистентности. Несмотря на это, бенчмарк продемонстрировал способность дифференцировать модели по уровню знаний.

По итогам работы намечены конкретные направления для дальнейшего развития.

- Первоочередной задачей является стандартизация и улучшение процедуры автоматического судейства. Необходимо интегрировать API ключевых коммерческих и открытых моделей для полной автоматизации тестирования.
- Планируется расширить спектр источников данных, добавив новостные ленты и другие викторины.
- Для удобства пользователей будет разработан веб-интерфейс с визуализацией результатов и лидербордами.

Долгосрочной целью является создание общедоступной, самообновляемой платформы для объективного оценивания LLM. Таким образом, работа закладывает основу для инструмента, способного эволюционировать вместе с развитием языковых моделей.

Список литературы

1. Junzhe Chen, Xuming Hu, Shuodi Liu, Shiyu Huang, Wei-Wei Tu, Zhaofeng He, and Lijie Wen. Llmarena: Assessing capabilities of large language models in dynamic multi-agent environments, 2024.

2. Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024.

3. Ilya Gusev. Pingpong: A benchmark for role-playing language models with user emulation and multi-model evaluation, 2025.

4. Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.

5. kristaller. Ruqualbench: A benchmark for evaluating the quality of the russian language in llm responses.

6. Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Dmitry Dodonov, Tung Nguyen, Jaeho Lee, Daron Anderson, Mikhail Doroshenko, Alun Cennyth Stokes, Mobeen Mahmood, Oleksandr Pokutnyi, Oleg Iskra, Jessica P. Wang, John-Clark Levin, Mstyslav Kazakov, Fiona Feng, Steven Y. Feng, Haoran Zhao, Michael Yu, Varun Gangal, Chelsea Zou, Zihan Wang, Serguei Popov, Robert Gerbicz, Geoff Galgon, Johannes Schmitt, Will Yeadon, Yongki Lee, Scott Sauers, Alvaro Sanchez, Fabian Giska, Marc Roth, Søren Riis, Saiteja Utpala, Noah Burns, Gashaw M. Goshu, Mohinder Maheshbhai Naiya, Chidozie Agu, Zachary Giboney, Antrell Cheatom, Francesco Fournier-Facio, Sarah-Jane Crowson, Lennart Finke, Zerui Cheng, Jennifer Zampese, Ryan G. Hoerr, Mark Nandor, Hyunwoo Park, Tim Gehrunger, Jiaqi Cai, Ben McCarty, Alexis C Garretson, Edwin Taylor, Damien Sileo, Qiuyu Ren, Usman Qazi, Lianghui Li, Jungbae Nam, John B. Wydallis, Pavel Arkhipov, Jack Wei Lun Shi, Aras Bacho, Chris G. Willcocks, Hangrui Cao, Sumeet Motwani, Emily de Oliveira Santos, Johannes Veith, Edward Vendrow, Doru Cojoc, Kengo Zenitani, Joshua Robinson, Longke Tang, Yuqi Li, Joshua Vendrow, Natanael Wildner Fraga, Vladyslav Kuchkin, Andrey Pupasov Maksimov, Pierre Marion, Denis Efremov, Jayson Lynch, Kaiqu Liang, Aleksandar Mikov, Andrew Gritsevskiy, Julien Guillod, Gözdenur Demir, Dakotah Martinez, Ben Pageler, Kevin Zhou, Saeed Soori, Ori Press, Henry Tang, Paolo Rissone, Sean R. Green, Lina Brüssel, Moon Twayana, Aymeric Dieuleveut, Joseph Marvin Imperial, Ameya Prabhu, Jinzhou Yang, Nick Crispino, Arun Rao, Dimitri Zvonkine, Gabriel Loiseau, Mikhail Kalinin, Marco Lukas, Ciprian Manolescu, Nate Stambaugh, Subrata Mishra, Tad Hogg, Carlo Bosio, Brian P Coppola, Julian Salazar, Jaehyeok Jin, Rafael Sayous, Stefan Ivanov, Philippe Schwaller, Shaipranesh Senthilkuma, Andres M Bran, Andres Algaba, Kelsey Van den Houte, Lynn Van Der Sypt, Brecht Verbeken, David Noever, Alexei Kopylov, Benjamin Myklebust, Bikun Li, Lisa Schut, Evgenii Zheltonozhskii, Qiaochu Yuan, Derek Lim, Richard Stanley, Tong Yang, John Maar, Julian Wykowski,

Martí Oller, Anmol Sahu, Cesare Giulio Ardito, Yuzheng Hu, Ariel Ghislain Kemogne Kamdoum, Alvin Jin, Tobias Garcia Vilchis, Yuexuan Zu, Martin Lackner, James Koppel, Gongbo Sun, Daniil S. Antonenko, Steffi Chern, Bingchen Zhao, Pierrot Arsene, Joseph M Cavanagh, Daofeng Li, Jiawei Shen, Donato Crisostomi, Wenjin Zhang, Ali Dehghan, Sergey Ivanov, David Perrella, Nurdin Kaparov, Allen Zang, Ilia Sucholutsky, Arina Kharlamova, Daniil Orel, Vladislav Poritski, Shalev Ben-David, Zachary Berger, Parker Whitfill, Michael Foster, Daniel Munro, Linh Ho, Shankar Sivarajan, Dan Bar Hava, Aleksey Kuchkin, David Holmes, Alexandra Rodriguez-Romero, Frank Sommerhage, Anji Zhang, Richard Moat, Keith Schneider, Zakayo Kazibwe, Don Clarke, Dae Hyun Kim, Felipe Meneguitti Dias, Sara Fish, Veit Elser, Tobias Kreiman, Victor Efren Guadarrama Vilchis, Immo Klose, Ujjwala Anantheswaran, Adam Zweiger, Kaivalya Rawal, Jeffery Li, Jeremy Nguyen, Nicolas Daans, Haline Heidinger, Maksim Radionov, Václav Rozhoň, Vincent Ginis, Christian Stump, Niv Cohen, Rafał Poświata, Josef Tkadlec, Alan Goldfarb, Chenguang Wang, Piotr Padlewski, Stanislaw Barzowski, Kyle Montgomery, Ryan Stendall, Jamie Tucker-Foltz, Jack Stade, T. Ryan Rogers, Tom Goertzen, Declan Grabb, Abhishek Shukla, Alan Givré, John Arnold Ambay, Archan Sen, Muhammad Fayez Aziz, Mark H Inlow, Hao He, Ling Zhang, Younesse Kaddar, Ivar Ängquist, Yanxu Chen, Harrison K Wang, Kalyan Ramakrishnan, Elliott Thornley, Antonio Terpin, Hailey Schoelkopf, Eric Zheng, Avishy Carmi, Ethan D. L. Brown, Kelin Zhu, Max Bartolo, Richard Wheeler, Martin Stehberger, Peter Bradshaw, JP Heimonen, Kaustubh Sridhar, Ido Akov, Jennifer Sandlin, Yury Makarychev, Joanna Tam, Hieu Hoang, David M. Cunningham, Vladimir Goryachev, Demosthenes Patramanis, Michael Krause, Andrew Redenti, David Aldous, Jesyin Lai, Shannon Coleman, Jiangnan Xu, Sangwon Lee, Ilias Magoulas, Sandy Zhao, Ning Tang, Michael K. Cohen, Orr Paradise, Jan Hendrik Kirchner, Maksym Ovchynnikov, Jason O. Matos, Adithya Shenoy, Michael Wang, Yuzhou Nie, Anna Sztyber-Betley, Paolo Faraboschi, Robin Riblet, Jonathan Crozier, Shiv Halasyamani, Shreyas Verma, Prashant Joshi, Eli Meril, Ziqiao Ma, Jérémy Andréoletti, Raghav Singhal, Jacob Platnick, Volodymyr Nevirkovets, Luke Basler, Alexander Ivanov, Seri Khoury, Nils Gustafsson, Marco Piccardo, Hamid Mostaghimi, Qijia Chen, Virendra Singh, Tran Quoc Khánh, Paul Rosu, Hannah Szlyk, Zachary Brown, Himanshu Narayan, Aline Menezes, Jonathan Roberts, William Alley, Kunyang Sun, Arkil Patel, Max Lamparth, Anka Reuel, Linwei Xin, Hanmeng Xu, Jacob Loader, Freddie Martin, Zixuan Wang, Andrea Achilleos, Thomas Preu, Tomek Korbak, Ida Bosio, Fereshteh Kazemi, Ziye Chen, Biró Bálint, Eve J. Y. Lo, Jiaqi Wang, Maria Inês S. Nunes, Jeremiah Milbauer, M Saiful Bari, Zihao Wang, Behzad Ansarinejad, Yewen Sun, Stephane Durand, Hossam Elgnainy, Guillaume Douville, Daniel Tordera, George Balabanian, Hew Wolff, Lynna Kvistad, Hsiaoyun Milliron, Ahmad Sakor, Murat Eron, Andrew Favre D. O., Shailesh Shah, Xiaoxiang Zhou, Firuz Kamalov, Sherwin Abdoli, Tim Santens, Shaul Barkan, Allison Tee, Robin Zhang, Alessandro Tomasiello, G. Bruno De Luca, Shi-Zhuo Looi, Vinh-Kha Le, Noam Kolt, Jiayi Pan, Emma Rodman, Jacob Drori, Carl J Fossum, Niklas Muennighoff, Milind Jagota, Ronak Pradeep, Honglu Fan, Jonathan Eicher, Michael Chen, Kushal Thaman, William Merrill, Moritz Firsching, Carter Harris, Stefan Ciobâcă, Jason Gross, Rohan Pandey, Ilya Gusev, Adam Jones, Shashank Agnihotri, Pavel Zhelnov, Mohammadreza Mofayezi, Alexander Piperski, David K. Zhang, Kostiantyn Dobarskyi, Roman Leventov, Ignat Soroko, Joshua

Duersch, Vage Taamazyan, Andrew Ho, Wenjie Ma, William Held, Ruicheng Xian, Armel Randy Zebaze, Mohanad Mohamed, Julian Noah Leser, Michelle X Yuan, Laila Yacar, Johannes Lengler, Katarzyna Olszewska, Claudio Di Fratta, Edson Oliveira, Joseph W. Jackson, Andy Zou, Muthu Chidambaram, Timothy Manik, Hector Haffenden, Dashiell Stander, Ali Dasouqi, Alexander Shen, Bita Golshani, David Stap, Egor Kretov, Mikalai Uzhou, Alina Borisovna Zhidkovskaya, Nick Winter, Miguel Orbegozo Rodriguez, Robert Lauff, Dustin Wehr, Colin Tang, Zaki Hossain, Shaun Phillips, Fortuna Samuele, Fredrik Ekström, Angela Hammon, Oam Patel, Faraz Farhidi, George Medley, Forough Mohammadzadeh, Madellene Peñaflor, Haile Kassahun, Alena Friedrich, Rayner Hernandez Perez, Daniel Pyda, Taom Sakal, Omkar Dhamane, Ali Khajegili Mirabadi, Eric Hallman, Kenchi Okutsu, Mike Battaglia, Mohammad Maghsoudimehrabani, Alon Amit, Dave Hulbert, Roberto Pereira, Simon Weber, Handoko, Anton Peristyy, Stephen Malina, Mustafa Mehkary, Rami Aly, Frank Reidegeld, Anna-Katharina Dick, Cary Friday, Mukhwinder Singh, Hassan Shapourian, Wanyoung Kim, Mariana Costa, Hubeyb Gurdogan, Harsh Kumar, Chiara Ceconello, Chao Zhuang, Haon Park, Micah Carroll, Andrew R. Tawfeek, Stefan Steinerberger, Daattavya Aggarwal, Michael Kirchhof, Linjie Dai, Evan Kim, Johan Ferret, Jainam Shah, Yuzhou Wang, Minghao Yan, Krzysztof Burdzy, Lixin Zhang, Antonio Franca, Diana T. Pham, Kang Yong Loh, Joshua Robinson, Abram Jackson, Paolo Giordano, Philipp Petersen, Adrian Cosma, Jesus Colino, Colin White, Jacob Votava, Vladimir Vinnikov, Ethan Delaney, Petr Spelda, Vit Stritecky, Syed M. Shahid, Jean-Christophe Mourrat, Lavr Vetoshkin, Koen Sponselee, Renas Bacho, Zheng-Xin Yong, Florencia de la Rosa, Nathan Cho, Xiuyu Li, Guillaume Malod, Orion Weller, Guglielmo Albani, Leon Lang, Julien Laurendeau, Dmitry Kazakov, Fatimah Adesanya, Julien Portier, Lawrence Hollom, Victor Souza, Yuchen Anna Zhou, Julien Degorre, Yiğit Yalın, Gbenga Daniel Obikoya, Rai, Filippo Bigi, M. C. Boscá, Oleg Shumar, Kaniuar Bacho, Gabriel Recchia, Mara Popescu, Nikita Shulga, Ngefor Mildred Tanwie, Thomas C. H. Lux, Ben Rank, Colin Ni, Matthew Brooks, Alesia Yakimchyk, Huanxu, Liu, Stefano Cavalleri, Olle Häggström, Emil Verkama, Joshua Newbould, Hans Gundlach, Leonor Brito-Santana, Brian Amaro, Vivek Vajipey, Rynaa Grover, Ting Wang, Yosi Kratish, Wen-Ding Li, Sivakanth Gopi, Andrea Caciolai, Christian Schroeder de Witt, Pablo Hernández-Cámara, Emanuele Rodolà, Jules Robins, Dominic Williamson, Vincent Cheng, Brad Raynor, Hao Qi, Ben Segev, Jingxuan Fan, Sarah Martinson, Erik Y. Wang, Kaylie Hausknecht, Michael P. Brenner, Mao Mao, Christoph Demian, Peyman Kassani, Xinyu Zhang, David Avagian, Eshawn Jessica Scipio, Alon Ragoler, Justin Tan, Blake Sims, Rebeka Plecnik, Aaron Kirtland, Omer Faruk Bodur, D. P. Shinde, Yan Carlos Leyva Labrador, Zahra Adoul, Mohamed Zekry, Ali Karakoc, Tania C. B. Santos, Samir Shamseldeen, Loukmane Karim, Anna Liakhovitskaia, Nate Resman, Nicholas Farina, Juan Carlos Gonzalez, Gabe Maayan, Earth Anderson, Rodrigo De Oliveira Pena, Elizabeth Kelley, Hodjat Mariji, Rasoul Pouriamanesh, Wentao Wu, Ross Finocchio, Ismail Alarab, Joshua Cole, Danyelle Ferreira, Bryan Johnson, Mohammad Safdari, Liangti Dai, Siriphan Arthornthurasuk, Isaac C. McAlister, Alejandro José Moyano, Alexey Pronin, Jing Fan, Angel Ramirez-Trinidad, Yana Malysheva, Daphiny Pottmaier, Omid Taheri, Stanley Stepanic, Samuel Perry, Luke Askew, Raúl Adrián Huerta Rodríguez, Ali M. R. Minissi, Ricardo Lorena, Krishnamurthy Iyer, Arshad Anil Fasiludeen, Ronald Clark, Josh Ducey, Matheus Piza, Maja Somrak, Eric Vergo, Juehang Qin, Benjámin

Borbás, Eric Chu, Jack Lindsey, Antoine Jallon, I. M. J. McInnis, Evan Chen, Avi Semler, Luk Gloor, Tej Shah, Marc Carauleanu, Pascal Lauer, Tran Đuc Huy, Hossein Shahrtash, Emilien Duc, Lukas Lewark, Assaf Brown, Samuel Albanie, Brian Weber, Warren S. Vaz, Pierre Clavier, Yiyang Fan, Gabriel Poesia Reis e Silva, Long, Lian, Marcus Abramovitch, Xi Jiang, Sandra Mendoza, Murat Islam, Juan Gonzalez, Vasilios Mavroudis, Justin Xu, Pawan Kumar, Laxman Prasad Goswami, Daniel Bugas, Nasser Heydari, Ferenc Jeanplong, Thorben Jansen, Antonella Pinto, Archimedes Apronti, Abdallah Galal, Ng Ze-An, Ankit Singh, Tong Jiang, Joan of Arc Xavier, Kanu Priya Agarwal, Mohammed Berkani, Gang Zhang, Zhehang Du, Benedito Alves de Oliveira Junior, Dmitry Malishev, Nicolas Remy, Taylor D. Hartman, Tim Tarver, Stephen Mensah, Gautier Abou Loume, Wiktor Morak, Farzad Habibi, Sarah Hoback, Will Cai, Javier Gimenez, Roselynn Grace Montecillo, Jakub Łucki, Russell Campbell, Asankhaya Sharma, Khalida Meer, Shreen Gul, Daniel Espinosa Gonzalez, Xavier Alapont, Alex Hoover, Gunjan Chhablani, Freddie Vargus, Arunim Agarwal, Yibo Jiang, Deepakkumar Patil, David Outevsky, Kevin Joseph Scaria, Rajat Maheshwari, Abdelkader Dendane, Priti Shukla, Ashley Cartwright, Sergei Bogdanov, Niels Mündler, Sören Möller, Luca Arnaboldi, Kunvar Thaman, Muhammad Rehan Siddiqi, Prajvi Saxena, Himanshu Gupta, Tony Fruhauff, Glen Sherman, Mátyás Vincze, Siranut Usawasutsakorn, Dylan Ler, Anil Radhakrishnan, Innocent Enyekwe, Sk Md Salauddin, Jiang Muzhen, Aleksandr Maksapetyan, Vivien Rossbach, Chris Harjadi, Mohsen Bahaloohoreh, Claire Sparrow, Jasdeep Sidhu, Sam Ali, Song Bian, John Lai, Eric Singer, Justine Leon Uro, Greg Bateman, Mohamed Sayed, Ahmed Menshawy, Darling Duclosel, Dario Bezzi, Yashaswini Jain, Ashley Aaron, Murat Tiryakioglu, Sheeshram Siddh, Keith Krenek, Imad Ali Shah, Jun Jin, Scott Creighton, Denis Peskoff, Zienab EL-Wasif, Ragavendran P V, Michael Richmond, Joseph McGowan, Tejal Patwardhan, Hao-Yu Sun, Ting Sun, Nikola Zubić, Samuele Sala, Stephen Ebert, Jean Kaddour, Manuel Schottdorf, Dianzhuo Wang, Gerol Petruzella, Alex Meiburg, Tilen Medved, Ali ElSheikh, S Ashwin Hebbar, Lorenzo Vaquero, Xianjun Yang, Jason Poulos, Vilém Zouhar, Sergey Bogdanik, Mingfang Zhang, Jorge Sanz-Ros, David Anugraha, Yinwei Dai, Anh N. Nhu, Xue Wang, Ali Anil Demircali, Zhibai Jia, Yuyin Zhou, Juncheng Wu, Mike He, Nitin Chandok, Aarush Sinha, Gaoxiang Luo, Long Le, Mickaël Noyé, Michał Perełkiewicz, Ioannis Pantidis, Tianbo Qi, Soham Sachin Purohit, Letitia Parcalabescu, Thai-Hoa Nguyen, Genta Indra Winata, Edoardo M. Ponti, Hanchen Li, Kaustubh Dhole, Jongee Park, Dario Abbondanza, Yuanli Wang, Anupam Nayak, Diogo M. Caetano, Antonio A. W. L. Wong, Maria del Rio-Chanona, Dániel Kondor, Pieter Francois, Ed Chalstrey, Jakob Zsambok, Dan Hoyer, Jenny Reddish, Jakob Hauser, Francisco-Javier Rodrigo-Ginés, Suchandra Datta, Maxwell Shepherd, Thom Kamphuis, Qizheng Zhang, Hyunjun Kim, Ruiji Sun, Jianzhu Yao, Franck Dernoncourt, Satyapriya Krishna, Sina Rismanchian, Bonan Pu, Francesco Pinto, Yingheng Wang, Kumar Shridhar, Kalon J. Overholt, Glib Briia, Hieu Nguyen, David, Soler Bartomeu, Tony CY Pang, Adam Wecker, Yifan Xiong, Fanfei Li, Lukas S. Huber, Joshua Jaeger, Romano De Maddalena, Xing Han Lù, Yuhui Zhang, Claas Beger, Patrick Tser Jern Kon, Sean Li, Vivek Sanker, Ming Yin, Yihao Liang, Xinlu Zhang, Ankit Agrawal, Li S. Yifei, Zechen Zhang, Mu Cai, Yasin Sonmez, Costin Cozianu, Changhao Li, Alex Slen, Shoubin Yu, Hyun Kyu Park, Gabriele Sarti, Marcin Briański, Alessandro Stolfo, Truong An Nguyen, Mike Zhang, Yotam Perlitz, Jose Hernandez-Orallo, Runjia Li, Amin Shabani, Felix Juefei-Xu,

Shikhar Dhingra, Orr Zohar, My Chiffon Nguyen, Alexander Pondaven, Abdurrahim Yilmaz, Xuandong Zhao, Chuanyang Jin, Muyan Jiang, Stefan Todoran, Xinyao Han, Jules Kreuer, Brian Rabern, Anna Plassart, Martino Maggetti, Luther Yap, Robert Geirhos, Jonathon Kean, Dingsu Wang, Sina Mollaei, Chenkai Sun, Yifan Yin, Shiqi Wang, Rui Li, Yaowen Chang, Anjiang Wei, Alice Bizeul, Xiaohan Wang, Alexandre Oliveira Arrais, Kushin Mukherjee, Jorge Chamorro-Padial, Jiachen Liu, Xingyu Qu, Junyi Guan, Adam Bouyamourn, Shuyu Wu, Martyna Plomecka, Junda Chen, Mengze Tang, Jiaqi Deng, Shreyas Subramanian, Haocheng Xi, Haoxuan Chen, Weizhi Zhang, Yinuo Ren, Haoqin Tu, Sejong Kim, Yushun Chen, Sara Vera Marjanović, Junwoo Ha, Grzegorz Luczyna, Jeff J. Ma, Zewen Shen, Dawn Song, Cedegao E. Zhang, Zhun Wang, Gaël Gendron, Yunze Xiao, Leo Smucker, Erica Weng, Kwok Hao Lee, Zhe Ye, Stefano Ermon, Ignacio D. Lopez-Miguel, Theo Knights, Anthony Gitter, Namkyu Park, Boyi Wei, Hongzheng Chen, Kunal Pai, Ahmed Elkhanany, Han Lin, Philipp D. Siedler, Jichao Fang, Ritwik Mishra, Károly Zsolnai-Fehér, Xilin Jiang, Shadab Khan, Jun Yuan, Rishab Kumar Jain, Xi Lin, Mike Peterson, Zhe Wang, Aditya Malusare, Maosen Tang, Isha Gupta, Ivan Fosin, Timothy Kang, Barbara Dworakowska, Kazuki Matsumoto, Guangyao Zheng, Gerben Sewuster, Jorge Pretel Villanueva, Ivan Rannev, Igor Chernyavsky, Jiale Chen, Deepayan Banik, Ben Racz, Wenchao Dong, Jianxin Wang, Laila Bashmal, Duarte V. Gonçalves, Wei Hu, Kaushik Bar, Ondrej Bohdal, Atharv Singh Patlan, Shehzaad Dhuliawala, Caroline Geirhos, Julien Wist, Yuval Kansal, Bingsen Chen, Kutay Tire, Atak Talay Yücel, Brandon Christof, Veerupaksh Singla, Zijian Song, Sanxing Chen, Jiaxin Ge, Kaustubh Ponkshe, Isaac Park, Tianneng Shi, Martin Q. Ma, Joshua Mak, Sherwin Lai, Antoine Moulin, Zhuo Cheng, Zhanda Zhu, Ziyi Zhang, Vaidehi Patil, Ketan Jha, Qiutong Men, Jiaxuan Wu, Tianchi Zhang, Bruno Hebling Vieira, Alham Fikri Aji, Jae-Won Chung, Mohammed Mahfoud, Ha Thi Hoang, Marc Sperzel, Wei Hao, Kristof Meding, Sihan Xu, Vassilis Kostakos, Davide Manini, Yueying Liu, Christopher Toukmaji, Jay Paek, Eunmi Yu, Arif Engin Demircali, Zhiyi Sun, Ivan Dewerpe, Hongsen Qin, Roman Pflugfelder, James Bailey, Johnathan Morris, Ville Heilala, Sybille Rosset, Zishun Yu, Peter E. Chen, Woongyeong Yeo, Eeshaan Jain, Ryan Yang, Sreekar Chigurupati, Julia Chernyavsky, Sai Prajwal Reddy, Subhashini Venugopalan, Hunar Batra, Core Francisco Park, Hieu Tran, Guilherme Maximiano, Genghan Zhang, Yizhuo Liang, Hu Shiyu, Rongwu Xu, Rui Pan, Siddharth Suresh, Ziqi Liu, Samaksh Gulati, Songyang Zhang, Peter Turchin, Christopher W. Bartlett, Christopher R. Scotese, Phuong M. Cao, Ben Wu, Jacek Karwowski, Davide Scaramuzza, Aakaash Nattanmai, Gordon McKellips, Anish Cheraku, Asim Suhail, Ethan Luo, Marvin Deng, Jason Luo, Ashley Zhang, Kavin Jindel, Jay Paek, Kasper Halevy, Allen Baranov, Michael Liu, Advaith Avadhanam, David Zhang, Vincent Cheng, Brad Ma, Evan Fu, Liam Do, Joshua Lass, Hubert Yang, Surya Sunkari, Vishruth Bharath, Violet Ai, James Leung, Rishit Agrawal, Alan Zhou, Kevin Chen, Tejas Kalpathi, Ziqi Xu, Gavin Wang, Tyler Xiao, Erik Maung, Sam Lee, Ryan Yang, Roy Yue, Ben Zhao, Julia Yoon, Sunny Sun, Aryan Singh, Ethan Luo, Clark Peng, Tyler Osbey, Taozhi Wang, Daryl Echeazu, Hubert Yang, Timothy Wu, Spandan Patel, Vidhi Kulkarni, Vijaykaarti Sundarapandiyan, Ashley Zhang, Andrew Le, Zafir Nasim, Srikar Yalam, Ritesh Kasamsetty, Soham Samal, Hubert Yang, David Sun, Nihar Shah, Abhijeet Saha, Alex Zhang, Leon Nguyen, Laasya Nagumalli, Kaixin Wang, Alan Zhou, Aidan Wu, Jason Luo, Anwith Telluri, Summer Yue, Alexandr Wang, and Dan Hendrycks.

Humanity's last exam, 2025.

7. Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. Livebench: A challenging, contamination-limited llm benchmark, 2025.

Приложение

Исходный код бенчмарка доступен на GitHub:

- `https://github.com/ruslann19/llm_benchmark`