Министерство науки и высшего образования Российской Федерации

САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО

Физико-механический институт
Высшая школа прикладной математики и вычислительной физики

КУРСОВАЯ РАБОТА НА ТЕМУ

«Система сравнения мультимодальных языковых моделей»

по дисциплине
«Автоматизация научных исследований»

Выполнил

студент гр. 5040102/50201: Мартынов А. Д.

Преподаватель: Новиков Ф.А., Пестряков Д.Д.

Санкт-Петербург

2026

Ключевые слова: мультимодальные языковые модели; мультимодальные LLM; vision-language модели; оценка качества; сравнение моделей; бенчмарки; датасеты; метрики качества; протокол оценки; воспроизводимость; системные инструкции; промптинг; параметры генерации; мультимодальное понимание; мультимодальное рассуждение; VQA; визуальный диалог; суммаризация текста; генерация текста; семантическая близость; ROUGE; BERTScore; human evaluation; LLM-as-a-judge; мультикритериальная оценка; ранжирование; агрегирование критериев; стоимость инференса; задержка; ограничения контекста; мониторинг качества; отчетность; трассируемость; MLOps; логирование экспериментов; артефакты; регрессии качества.

Аннотация: В работе рассматривается общий подход к построению системы сравнения мультимодальных языковых моделей, предназначенной для упорядоченного и воспроизводимого сопоставления моделей в прикладных сценариях. Обсуждаются типичные причины, по которым сравнение мультимодальных моделей затруднено: неоднородность задач и входных данных, множественность метрик качества, влияние промптов и параметров генерации, а также практические ограничения по стоимости и вычислительным ресурсам. Предлагается концептуальная структура системы сравнения как набора модулей (подготовка задач, запуск моделей, расчет метрик, агрегация критериев, формирование отчета) и описывается, какие артефакты целесообразно фиксировать для обеспечения воспроизводимости. В качестве иллюстративных сценариев приводятся VQA и суммаризация текста, а архитектура системы задается в виде UML-диаграмм.

Keywords: multimodal language models; multimodal LLMs; vision-language models; evaluation; benchmarking; datasets; metrics; evaluation protocol; reproducibility; prompting; decoding parameters; multimodal reasoning; visual question answering; summarization; multi-criteria assessment; ranking; inference cost; latency; reporting; traceability; MLOps; experiment tracking.

Abstract: This work outlines a general approach to building a comparison system for multimodal language models, aimed at structured and reproducible assessment for applied use cases. We discuss common reasons why comparing multimodal models is challenging: heterogeneity of tasks and inputs, the need for multiple quality metrics, sensitivity to prompts and generation parameters, and practical constraints such as inference cost and compute availability. We propose a conceptual architecture of the comparison system as a set of modules (task preparation, model execution, metric computation, criteria aggregation, reporting) and describe which artifacts should be stored to support reproducibility. VQA and text summarization are used as illustrative scenarios, and the system architecture is presented via UML diagrams.

Содержание

Введение

Мультимодальные языковые модели (MLLM) стали одним из наиболее заметных направлений развития современных систем искусственного интеллекта. В прикладных задачах они выступают как универсальные «интерфейсы» для работы со смешанными данными: изображениями, текстом и производными представлениями (таблицы, скриншоты документов, инфографика). При этом практическая ценность модели почти всегда определяется не тем, что она «в целом сильная», а тем, насколько хорошо она подходит под конкретный сценарий: входные данные, формат ответа, требования к качеству, допустимая стоимость и ограничения инфраструктуры.

Несмотря на наличие бенчмарков и метрик, сравнение мультимодальных моделей остается задачей, в которой легко получить неоднозначный или непереносимый вывод. Причины этого лежат одновременно в методической и инженерной плоскости. С методической стороны, разные наборы задач и метрик измеряют разные аспекты поведения моделей, а «универсальной» метрики, которая одинаково хорошо описывает качество в разных сценариях, фактически нет. С инженерной стороны, результаты заметно зависят от условий запуска: промптов, параметров генерации, ограничений контекста, способа обработки изображений и даже версий библиотек.

Поэтому в практическом контексте часто требуется не столько «доказать превосходство» одной модели над другой, сколько создать прозрачную и повторяемую процедуру сравнения: чтобы получить объяснимую картину различий, быстро пересчитать результаты при обновлении модели и иметь возможность формировать единый отчет для принятия решения. В работе описывается именно такой общий подход: концепция системы сравнения, структура этапов и типовые артефакты, которые полезно фиксировать.

Постановка задачи

Объектом исследования являются мультимодальные языковые модели и процессы их прикладной оценки. Предметом исследования выступают организационные и методические подходы к сравнению моделей по совокупности критериев качества и эксплуатационных требований.

Цель работы — описать концепцию системы сравнения мультимодальных языковых моделей, которая позволяет:

- задавать единый протокол тестирования;
- сравнивать модели по нескольким критериям (качество, стоимость, ограничения);
- сохранять артефакты и формировать отчет таким образом, чтобы сравнение можно было повторить и обновить.

Для достижения цели формулируются задачи, но в общем виде, поскольку в исходной постановке известна только тема:

- определить типовые сценарии сравнения MLLM и их особенности;
- описать набор классов метрик и критериев, которые обычно используют для мультимодальных задач;
- предложить структуру этапов сравнения (подготовка данных, запуск, метрики, агрегирование, отчет);
- описать возможную архитектуру системы сравнения и взаимосвязь модулей;
- указать ограничения и факторы, из-за которых результаты сравнения могут быть неустойчивыми.

Общий обзор мультимодальных языковых моделей

Под мультимодальной языковой моделью в прикладном смысле обычно понимают систему, которая принимает на вход текст и изображения (иногда дополнительные модальности) и генерирует текстовый ответ, ориентируясь на инструкцию пользователя. Архитектурно такие системы часто включают языковую модель, визуальный энкодер и механизм «сшивки» представлений (проекция, адаптеры, совместное внимание и т. п.). В практических описаниях различия между моделями удобно рассматривать через призму того, как они решают три взаимосвязанные задачи: восприятие изображения, связывание визуальных элементов с текстом и генерация ответа.

С точки зрения сравнения, мультимодальные модели можно условно рассматривать как «черный ящик», который зависит от:

- входных данных и их предварительной обработки (масштабирование, качество изображения, формат);
- инструкций (системный промпт, пользовательский запрос, примеры);
- параметров генерации (температура, максимальная длина, ограничения словаря);
- ограничений контекста и внутренней политики отказов.

Даже при одинаковой теме сравнения, конкретное поведение модели может существенно меняться при смене шаблона запроса или при добавлении небольшого контекстного уточнения. Поэтому сравнение MLLM на практике неизбежно связано с фиксацией протокола запуска.

Подходы к сравнению и оценке: что обычно учитывают

В прикладных сценариях сравнение MLLM редко сводится к одной оси «лучше–хуже». Чаще рассматривают набор аспектов, часть из которых относится к качеству ответа, а часть — к применимости модели в системе.

С точки зрения качества, в задачах, где модель генерирует текст, могут использоваться метрики лексического совпадения (например, ROUGE-подобные), метрики семантической близости (например, BERTScore), а также экспертная оценка человеком. В некоторых протоколах используется подход «LLM-as-a-judge», когда отдельная модель оценивает качество ответа по заданным критериям; такой подход может быть удобен для унификации, но требует осторожности, поскольку оценщик сам может иметь систематические смещения.

С точки зрения эксплуатации, в сравнение нередко включают стоимость, задержку, требования к памяти и вычислениям, стабильность, доступность, ограничения по данным, а также удобство интеграции. Даже если «качество» выглядит предпочтительным, модель может оказаться непрактичной, если она слишком дорогая, слишком медленная или требует инфраструктуры, которой нет в целевом контуре.

Наконец, отдельной группой факторов является воспроизводимость: возможность повторить сравнение позже, а также возможность понять, почему модели «поменялись местами». Если протокол сравнения не фиксирует промпты, параметры генерации и версию данных, то даже аккуратные метрики теряют смысл, потому что сравниваются разные условия.

Концепция системы сравнения: этапы, модули и артефакты

Система сравнения MLLM может быть описана как конвейер, где каждый шаг выполняет ограниченную функцию и производит артефакты, пригодные для аудита и повторного использования.

На первом этапе задается конфигурация сравнения: список моделей, список задач и датасетов, набор метрик, а также параметры запуска (шаблоны промптов, параметры генерации). Важно, что конфигурация — это не «внутренний файл», а часть результата работы: именно она позволяет повторить сравнение или объяснить различия.

На втором этапе выполняется прогон моделей по выбранным примерам. В прикладной постановке результаты целесообразно сохранять в виде сырых ответов модели, а не только в виде агрегированных чисел. Даже если в текущей версии работы не предполагается получение «итогового победителя», сохранение ответов позволяет в будущем пересчитать метрики, заменить метрику или провести экспертный аудит.

На третьем этапе рассчитываются метрики и формируются табличные представления результатов. На практике часто используют не одну метрику, а несколько, и отдельно фиксируют направление «чем больше — тем лучше» или «чем меньше — тем лучше», чтобы избежать путаницы.

На четвертом этапе выполняется агрегирование критериев и формирование интерпретируемого результата сравнения. Здесь возможны разные стратегии: от простого описания «портрета» модели до построения ранжирования и групп

сопоставимых моделей. В рамках данной работы достаточно подчеркнуть, что агрегирование — это отдельное место, где легко получить спорные выводы, поэтому оно должно быть описано и зафиксировано в протоколе.

На пятом этапе формируется отчет: текстовое описание, таблицы, диаграммы, ссылки на артефакты и ограничения. Такой отчет может рассматриваться как «снимок» состояния сравнения, пригодный для повторной проверки и обновления.

Для воспроизводимости полезно заранее определить минимальный набор сохраняемых артефактов:

- конфигурация эксперимента (модели, задачи, датасеты, метрики, параметры);
- шаблоны промптов и примеры входов;
- сырые ответы моделей и журналы выполнения;
- таблицы метрик и вспомогательные промежуточные данные;
- итоговый отчет (включая ограничения и допущения).

Иллюстративные сценарии: VQA и суммаризация

Чтобы описание системы сравнения оставалось «по теме», удобно опираться на типовые сценарии, которые часто встречаются в мультимодальной практике.

Визуально-текстовое вопросно-ответное взаимодействие (VQA) используется как пример сценария, где модель должна интерпретировать изображение и связать его с текстовым вопросом. В таких задачах важно учитывать, что ответ может быть коротким, но при этом требует точного визуального восприятия. В качестве примера набора данных нередко упоминают VQA v2, однако конкретный выбор датасета и формат оценивания зависят от сценария: вопросно-ответные пары, ограничения ответа, наличие нескольких допустимых ответов и т. п.

Суммаризация текста рассматривается как пример сценария, где модель генерирует более длинный структурированный ответ, и качество зависит не только от совпадения формулировок, но и от адекватного покрытия смысла. В качестве примера датасета часто приводят CNN/DailyMail, однако на практике выбор набора данных определяется жанром текста, длиной входа, требованиями к стилю резюме и допустимой степенью перефразирования.

Важно подчеркнуть, что в рамках данной работы эти сценарии рассматриваются как иллюстративные: они помогают описать, какие виды входов и выходов встречаются, какие метрики обычно обсуждают и почему сравнение оказывается многокритериальным.

Ограничения и обсуждение

Даже хорошо организованная система сравнения не устраняет фундаментальные ограничения, характерные для мультимодальных моделей. Во-первых, метрики качества лишь частично отражают полезность ответа: «правильный» ответ может быть сформулирован по-разному, а некоторые ошибки могут быть критичны для одного сценария и малозаметны для другого. Во-вторых, чувствительность к промптам делает сравнение зависимым от выбранного протокола: одна и та же модель может выглядеть по-разному при разных инструкциях и примерах.

В-третьих, эксплуатационные критерии (стоимость, задержка, ограничения по данным) часто меняются со временем и зависят от конкретного контура внедрения. Это означает, что сравнение следует рассматривать как обновляемый процесс, а не как одноразовый акт. В-четвёртых, переносимость результатов с одного набора данных на другой не гарантирована: даже при «похожих» задачах модели могут иметь разную чувствительность к домену изображений и стилю текста.

С практической точки зрения, полезно заранее фиксировать, что именно считается допустимым в сравнении: какие допущения сделаны, какие факторы не контролируются и почему. Такая «честная» рамка делает отчет более устойчивым и снижает риск неправильной интерпретации.

UML-диаграммы системы сравнения

Ниже приведены UML-диаграммы в формате PlantUML. Такой формат удобен тем, что диаграммы можно хранить как текстовые артефакты и версионировать вместе с отчетом.

Диаграмма вариантов использования

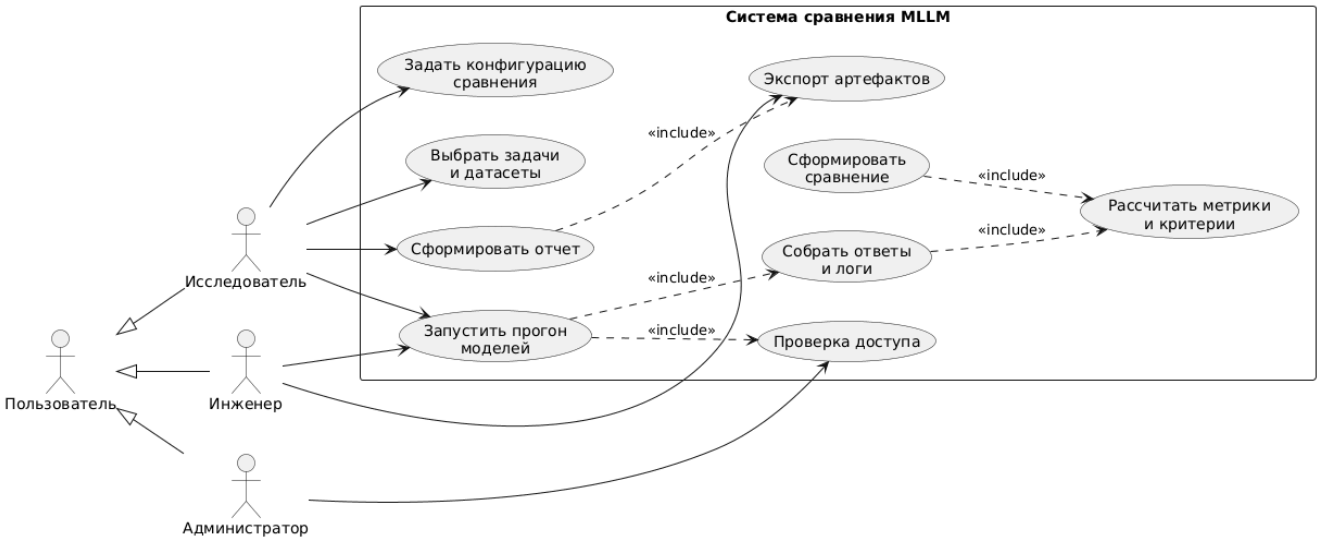
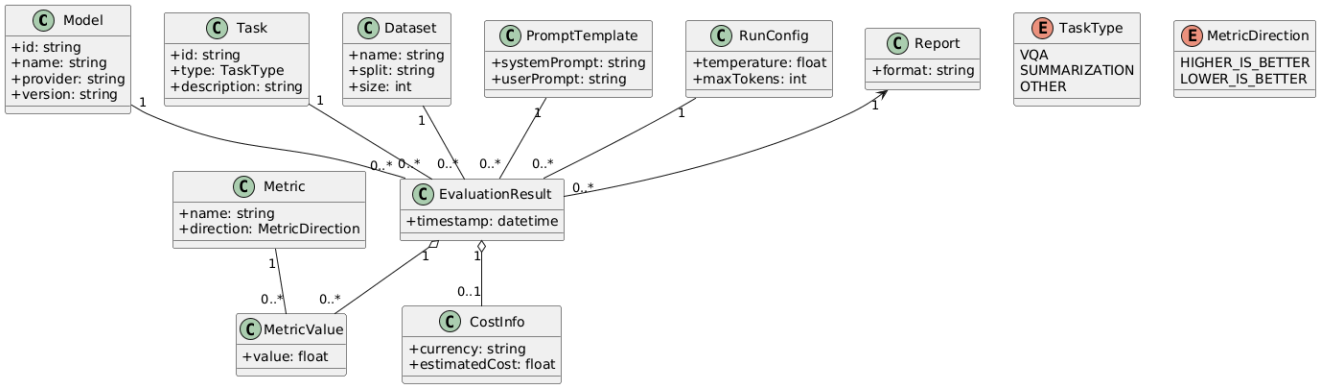


Диаграмма классов



Заключение

В работе представлен общий, описательный взгляд на задачу построения системы сравнения мультимодальных языковых моделей. Основная идея состоит в том, что сравнение MLLM целесообразно оформлять как воспроизводимую процедуру с фиксированным протоколом, набором критериев и отчетными артефактами. При известной только теме работы можно описать типовые этапы такого сравнения, типовые сценарии (VQA и суммаризация), а также предложить архитектурное разбиение системы на модули и представить его в виде UML-диаграмм.

Предложенное описание не претендует на получение «итогового победителя» и не фиксирует конкретные численные результаты. Вместо этого акцент сделан на структурировании процесса сравнения и на тех местах, где в реальных проектах чаще всего возникают неоднозначности: выбор задач, выбор метрик, чувствительность к промптам и влияние эксплуатационных ограничений.

Список литературы

1. Brans J.-P., Vincke Ph., Mareschal B. How to select and how to rank projects: The PROMETHEE method. _European Journal of Operational Research_. 1986. DOI: https://doi.org/10.1016/0377-2217(86)90044-5
2. Alayrac J.-B. et al. Flamingo: a Visual Language Model for Few-Shot Learning. arXiv: https://arxiv.org/abs/2204.14198
3. Li J. et al. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv: https://arxiv.org/abs/2301.12597
4. Liu H. et al. Visual Instruction Tuning (LLaVA). arXiv: `https://arxiv.org/abs/2304.08485`
5. Zhu D. et al. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. arXiv: https://arxiv.org/abs/2304.10592
6. Dai W. et al. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv: https://arxiv.org/abs/2305.06500
7. Liu Z. et al. MMBench: Is Your Multi-modal Model an All-around Player? arXiv: https://arxiv.org/abs/2307.06281
8. Li B. et al. SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension. arXiv: https://arxiv.org/abs/2307.16125
9. Bai J. et al. Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities. arXiv: https://arxiv.org/abs/2308.12966
10. Yue X. et al. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. arXiv: https://arxiv.org/abs/2311.16502
11. Lian D. et al. MLLM-Bench: Evaluating Multimodal LLMs with Per-sample Criteria. arXiv: https://arxiv.org/abs/2311.13951
12. M4U: Evaluating Multilingual Understanding and Reasoning for Large Multimodal Models. arXiv: https://arxiv.org/abs/2405.15638

13. Goyal Y. et al. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. arXiv: https://arxiv.org/abs/1612.00837
14. Hermann K. M. et al. Teaching Machines to Read and Comprehend. arXiv: https://arxiv.org/abs/1506.03340
15. Zhang T. et al. BERTScore: Evaluating Text Generation with BERT. arXiv: https://arxiv.org/abs/1904.09675