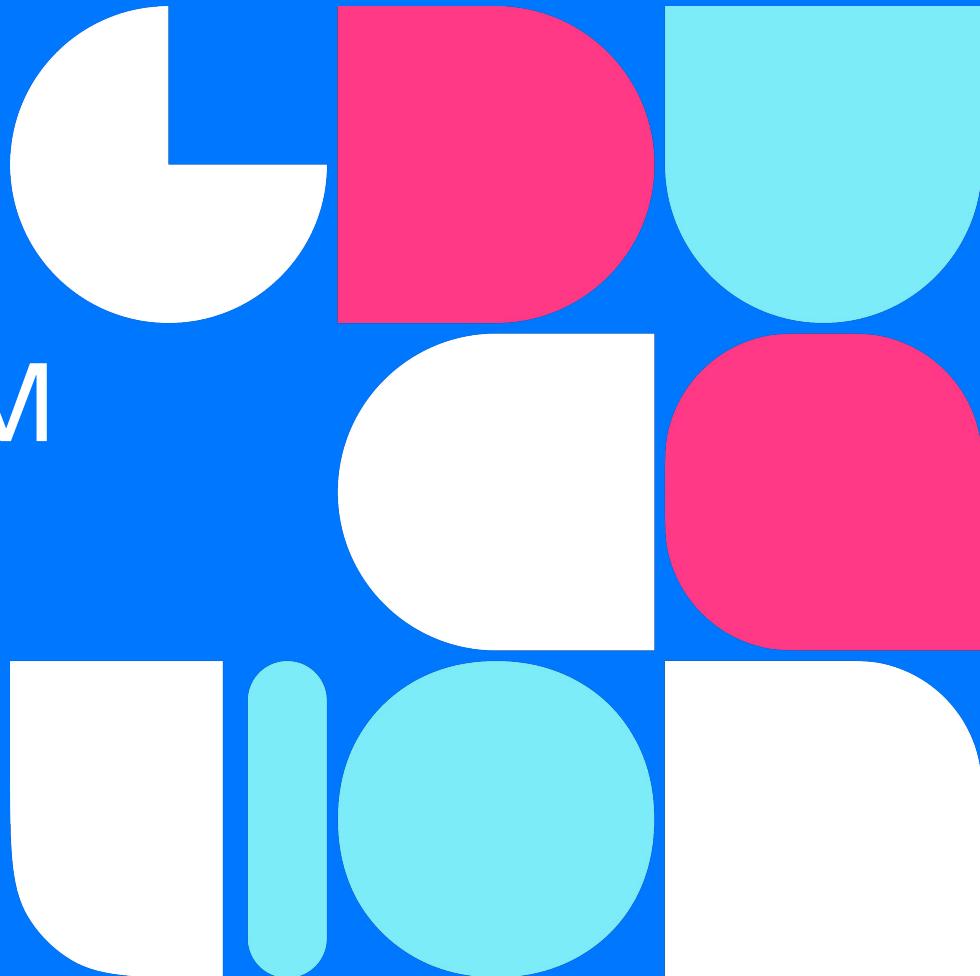




Данные для VLM

Курс “Мультимодальные модели”



Игорь Рябков

Разработчик мультимодальных VLM
Группа мультимодальных языковых моделей



Цели занятия:

- Познакомиться с особенностями данных для мультимодальных моделей
- Не потеряться в обилии современных датасетов
- Научиться корректно оценивать современные модели

О чём поговорим?

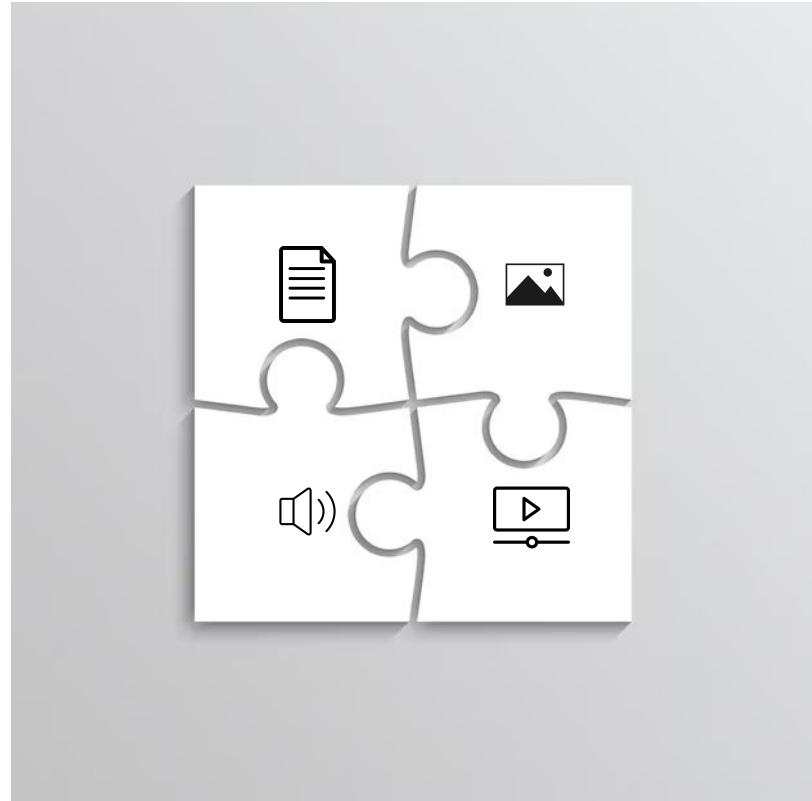
- Типы мультимодальных данных
- Обзор ключевых датасетов
- Бенчмарки для VLM/LMM
- Метрики и процедуры оценки мультимодальных моделей
- Сбор и препроцессинг мультимодальных данных
- Синтетические данные, преимущества и недостатки

Типы мультимодальных данных

Основные модальности:

- Текст
- Картинки
- Видео
- Аудио
- Таблицы
- Информация с сенсоров
- *3D модели

Мультимодальные данные - комбинации модальностей



Обзор ключевых мультимодальных датасетов



Модальности: Картинки + Текст. Датасет: COCO

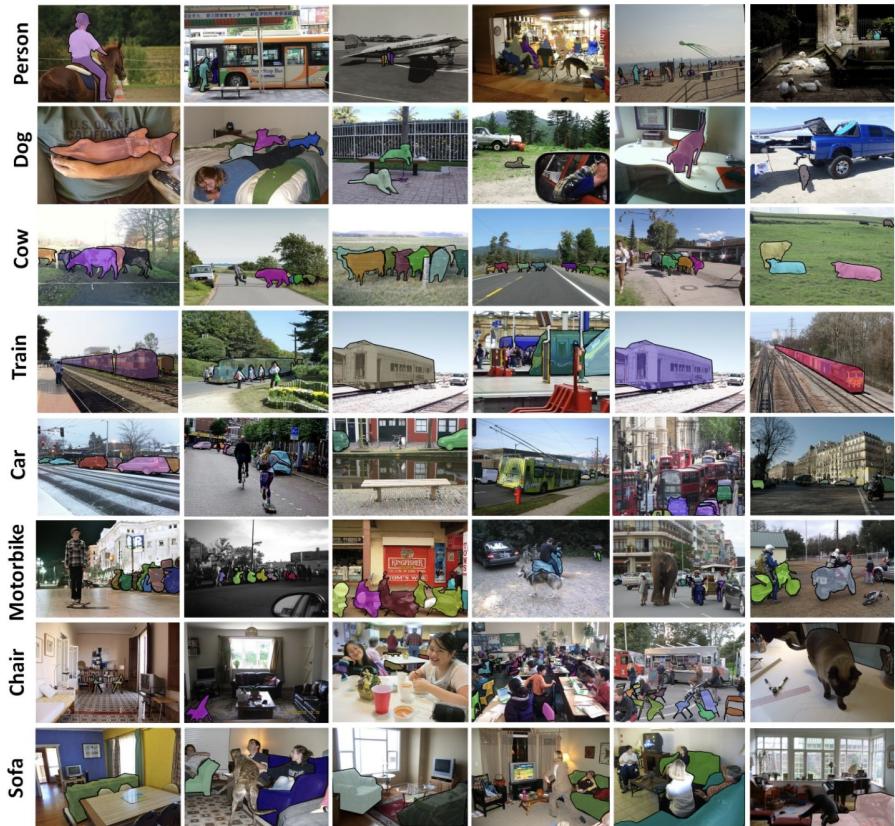
Размер: 330т

Задачи (основные):

- Detection
- Segmentation
- Image Captioning

Особенности:

- Разнообразные изображения
- По каждому изображению - много разнообразной разметки
- Стандарт для задач выше



Модальности: Картинки + Текст. Датасет: Flickr 8k/30k



A man with **pierced ears** is wearing **glasses** and **an orange hat**.
A man with **glasses** is wearing a **beer can crotched hat**.
A man with **gauges** and **glasses** is wearing a **Blitz hat**.
A man in **an orange hat** staring at **something**.
A man wears **an orange hat** and **glasses**.

Размер: 8k/30k

Задачи (основные):

- Image Captioning,
- Object Detection,
- Image Retrieval,
- Visual Relationship Detection

Особенность:

- По каждому изображению - набор из пяти разнообразных описаний.
- Изображения с Flickr (очень качественные) - одновременно и преимущество, и недостаток

Модальности: Картинки + Текст. Датасет: LAION-5B

Размер: 5 млрд

Задачи (основные):

- Image Captioning
- Object Detection
- Image Retrieval
- Multimodal Understanding
- Text-to-Image Generation

Особенности:

- Огромный размер и разнообразие данных
- По каждому изображению - много разнообразной разметки и подписей
- Много шума, риск сомнительного содержания

English



BMW-M2-M-
Performance-Dekor-
Long-Beach-Blue-05

Becoming More Than a
Good Bible Study Girl:
Living the Faith after Bible
Class Is Over [...]

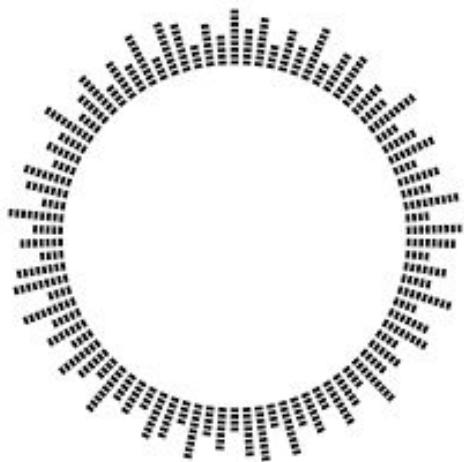
Multilingual



Episcopia Ortodoxa
a Maramuresului si
Satmarului are un
nou Arhiereu vicar

DON QUIJOTE
DE LA MANCHA
(SELECCIÓN DE
TEXTOS)

Модальности: Аудио + Текст. Датасет: LibriSpeech



Размер: 1,000 часов

Задачи (основные):

- Automatic Speech Recognition (ASR),
- Speaker Identification,
- Speech Retrieval

Особенность:

- Данные собраны из аудиокниг с LibriVox, начитаны разными дикторами (английский язык)

Модальности: Видео + Текст. Датасет: WebVid

Размер: 10.5 млн видеоклипов

Задачи (основные):

- Video-Text Retrieval
- Video Captioning
- Video-Language Pretraining

Особенность:

- Автоматически собран из веб-источников
- К каждому видео — краткая текстовая аннотация/описание
- Большой охват сюжетов



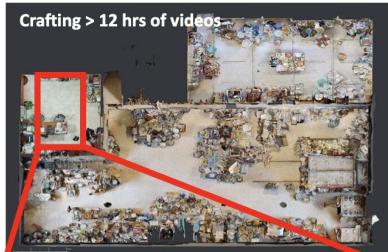
“Lonely beautiful woman sitting on the tent looking outside. wind on the hair and camping on the beach near the colors of water and shore. freedom and alternative tiny house for traveler lady drinking”



“Female cop talking on walkie-talkie, responding emergency call, crime prevention”

“Get anchor for departure safari dive boat scuba diving maldives”

Модальности: Текст + Аудио + Видео. Датасет: Ego4D



Размер: 3 670+ часов видео, 3 500+ видео, 900 участников

Основные задачи:

- Отслеживание и поиск объектов
- Распознавание действий
- Антиципация событий
- Суммаризация взаимодействий

Особенности:

- Эгоцентрические видеозаписи (вид от первого лица)
- Аннотации рук, объектов, субтитры
- Многообразие сценариев и культур
- Реалистичные, неинсцинированные повседневные ситуации

ShareGPT4V | ShareGPT4Video

Размер: 1.2 млн

Тип задач: Open Ended

Задача: Image Captioning

Язык: Английский

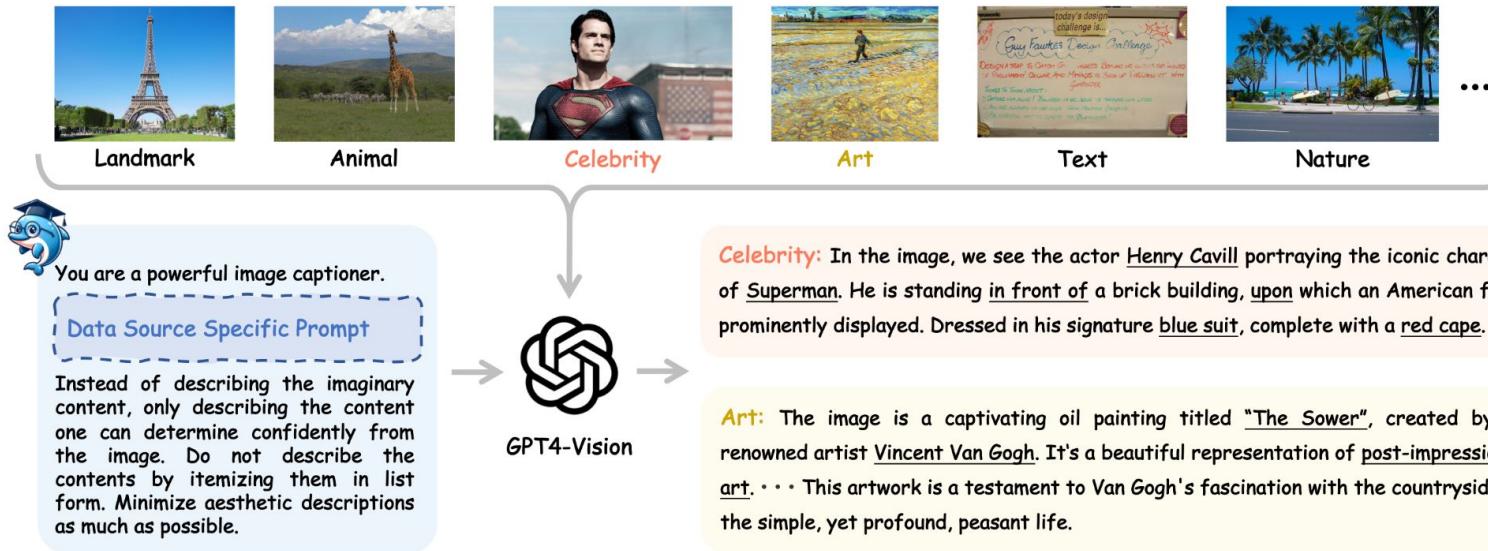
Особенности:

- Подробные, насыщенные описания
- Является полностью синтетическим
- Хорошо себя показывает в SFT
- Источники изображений включают: COCO, LAION, CC, SBU, Flickr



ShareGPT4V | ShareGPT4Video (Процесс сбора)

(a) ShareGPT4V Data Collection



(b) Pretrain Captions Generation



ShareGPT4V | ShareGPT4Video (Примеры)



COCO: Young children standing on a platform waiting for a train to arrive.

Adults and children watching a train slowly leave.

A family near a railroad track watching the train pass.

People waiting on a platform as a train pulls up.

A train station with a green chain on the tracks and children waiting for it to go by.

LLaVA: At a train station, a group of people, including both young children and adults, are standing on a platform waiting for a train to arrive. The train is already present on the tracks, partially visible on the right side of the image. Some of the people watch the train closely, while others seem to be patiently anticipating its departure.

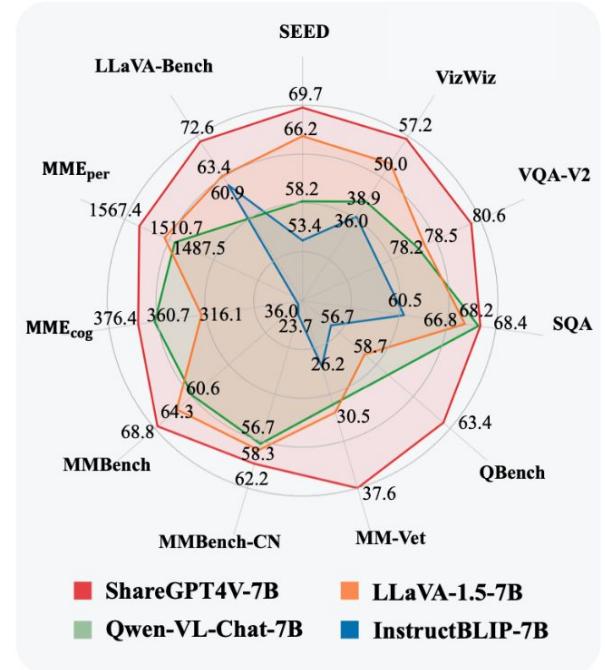
There is a total of eight individuals waiting for the train, with one child in the middle of the platform and the others scattered around. A backpack can be found on the far left side of the platform, suggesting that someone may have set it down while waiting.

ShareGPT4V: The image captures a moment at a train station. A green train is on the tracks, moving away from the platform labeled "Platform 2". The train's motion is observed by people standing on the platform, their attention drawn to the departing vehicle.

A red signal stands on the tracks, its vibrant color contrasting with the green of the train. Signs are prominently displayed around the platform. One warns "Beware of Trains", a cautionary reminder of the powerful machines that frequent these tracks. Another sign advises "Please Use The Footbridge To Cross The Line. If You Require Assistance Please Ask The Station Staff", guiding passengers to safely navigate the station.

The sky above is a clear blue, providing a serene backdrop to the bustling activity of the station. In the distance, trees can be seen, adding a touch of nature to this man-made setting. The image is a snapshot of everyday life at a train station, capturing both its routine operations and its inherent charm.

(a) Comparison of Captions' Quality



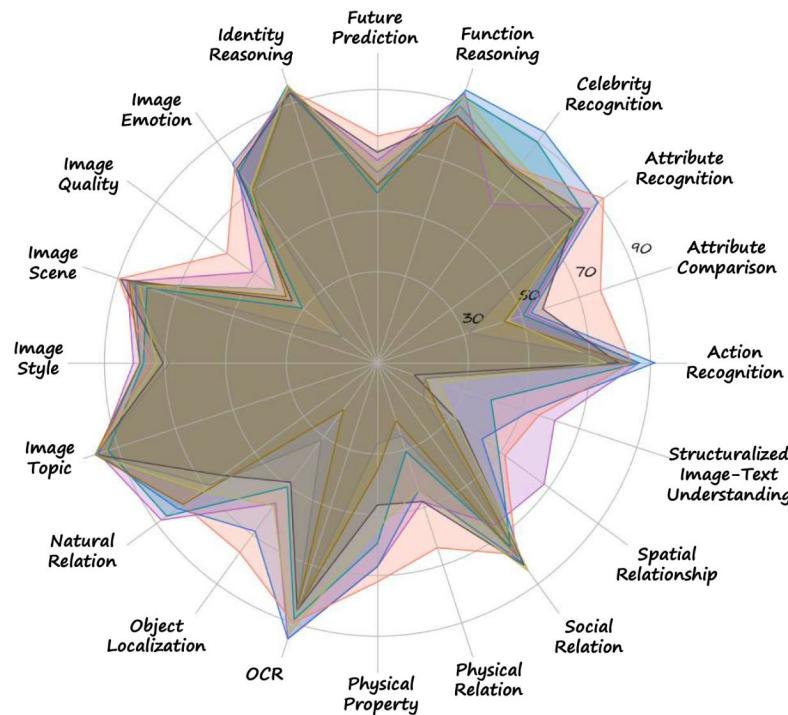
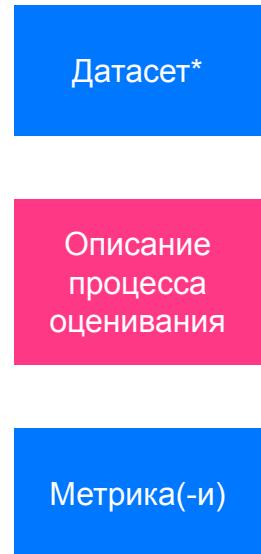
(b) Comparison of Performance

Бенчмарки для VLM/LMMs



Что такое бенчмарк?

Ответ: Метод оценивания VLM



GPT4-V
Gemini-Pro-V
Qwen-VL-Max
InternLM-XComposer2
LLaVA-v1.5-13B
CogVLM-Chat-17B
Yi-VL-34B
MiniCPM-V

MMBench

Размер: 3T

Тип задач: Multiple Choice

Язык: Английский

Особенности:

- Охватывает множество задач на Reasoning и Perception
- Вводит метод CircularEval

Упоминание в статьях: LLaVA-OneVision



The original VL problem:

Q: How many apples are there in the image?

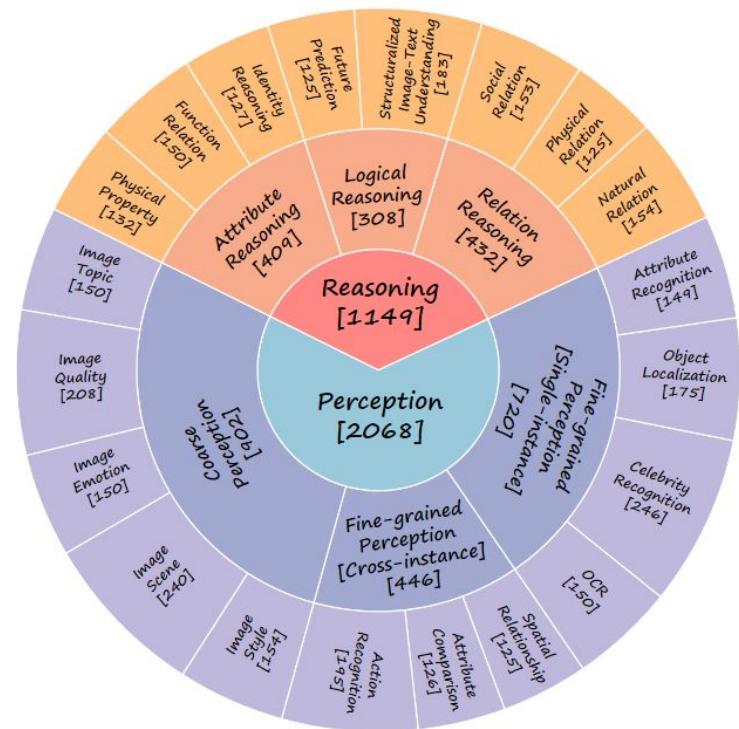
A. 4; B. 3; C. 2; D. 1

GT: A

4 Passes in Circular Evaluation (choices with circular shift):

1. Q: How many apples are there in the image? Choices: A. 4; B. 3; C. 2; D. 1. VLM prediction: A. GT: A ✓
2. Q: How many apples are there in the image? Choices: A. 3; B. 2; C. 1; D. 4. VLM prediction: D. GT: D ✓
3. Q: How many apples are there in the image? Choices: A. 2; B. 1; C. 4; D. 3. VLM prediction: B. GT: C X
4. Q: How many apples are there in the image? Choices: A. 1; B. 4; C. 3; D. 2. VLM prediction: B. GT: B ✓

VLM failed at pass 3. Thus wrong.



MMBench (примеры)



Q. Which image is the second brightest?

- A. upper-left
- B. upper-right
- C. lower-left
- D. lower-right

Answer: C

$A_{max}=61.3\%$

(a). Image Quality



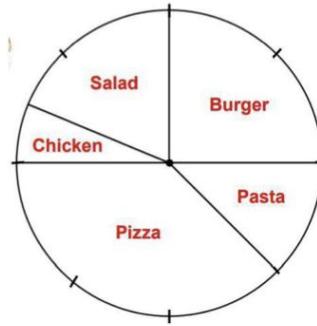
Q. What is the positional relationship between the two shapes in the picture?

- A. The two shapes are positioned apart or separated from each other.
- B. The two shapes are tangentially positioned or externally tangent to each other.
- C. The two shapes intersect with each other.
- D. One shape is contained within the other or there is an inner shape enclosed by an outer shape.

Answer: C

$A_{max}=68.0\%$

(c). Spatial Relationship



Q. The graph shows the meals purchased in a restaurant in one day. What is the least popular meal?

- A. Salad
- B. Burger
- C. Chicken
- D. Pasta

Answer: C

$A_{max}=61.5\%$

(b). Structralized Image-Text Understanding



Q. From the perspective of the driver of the blue truck, in what position is the person riding a bike relative to the blue truck?

- A. Left front
- B. Right front
- C. Right rear
- D. Left rear

$A_{max}=64.0\%$

Answer: A

(d). Physical Relation Reasoning

МММУ

Размер: 11.5т

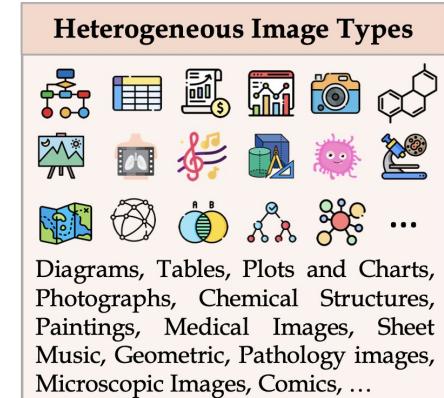
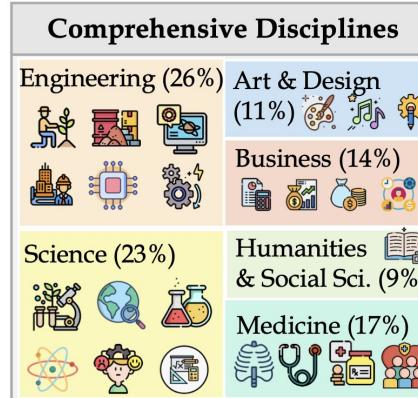
Тип задач: Multiple Choice

Язык: Английский

Особенности:

- Собирали студенты из различных областей
- Оценивает: Reasoning, Perception и Knowledge

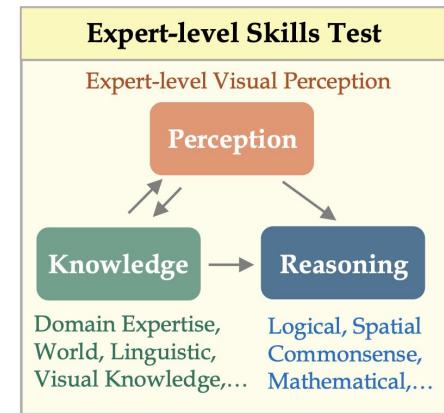
Упоминание в статьях: LLaVA-OneVision, Qwen3, InternVL3



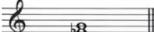
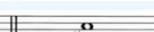
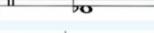
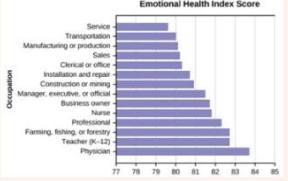
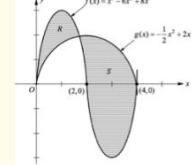
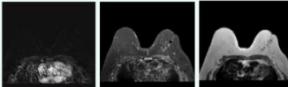
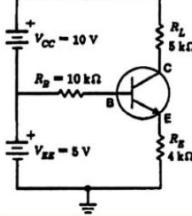
Interleaved Text and Images

Question: You are shown subtraction *<image 1>*, T2 weighted *<image 2>* and T1 weighted axial *<image 3>* from a screening breast MRI. What is the etiology of the finding in the left breast?

<image 1> *<image 2>* *<image 3>*



МММУ (примеры)

Art & Design	Business	Science
<p>Question: Among the following harmonic intervals, which one is constructed incorrectly?</p> <p>Options:</p> <ul style="list-style-type: none"> (A) Major third  (B) Diminished fifth  (C) Minor seventh  (D) Diminished sixth  	<p>Question: ...The graph shown is compiled from data collected by Gallup </p> <p>Options:</p> <ul style="list-style-type: none"> (A) 0 (B) 0.2142 (C) 0.3571 (D) 0.5 	<p>Question:  The region bounded by the graph as shown above. Choose an integral expression that can be used to find the area of R.</p> <p>Options:</p> <ul style="list-style-type: none"> (A) $\int_0^{1.5} [f(x) - g(x)] dx$ (B) $\int_0^{1.5} [g(x) - f(x)] dx$ (C) $\int_0^2 [f(x) - g(x)] dx$ (D) $\int_0^2 [g(x) - x(x)] dx$
<p>Subject: Music; Subfield: Music; Image Type: Sheet Music; Difficulty: Medium</p>	<p>Subject: Marketing; Subfield: Market Research; Image Type: Plots and Charts; Difficulty: Medium</p>	<p>Subject: Math; Subfield: Calculus; Image Type: Mathematical Notations; Difficulty: Easy</p>
Health & Medicine	Humanities & Social Science	Tech & Engineering
<p>Question: You are shown subtraction , T2 weighted  and T1 weighted axial  from a screening breast MRI. What is the etiology of the finding in the left breast?</p> <p>Options:</p> <ul style="list-style-type: none"> (A) Susceptibility artifact (B) Hematoma (C) Fat necrosis (D) Silicone granuloma 	<p>Question: In the political cartoon, the United States is seen as fulfilling which of the following roles? </p> <p>Option:</p> <ul style="list-style-type: none"> (A) Oppressor (B) Imperialist (C) Savior (D) Isolationist 	<p>Question: Find the VCE for the circuit shown in .</p> <p>Answer: <u>3.75</u></p> <p>Explanation: ...IE = [(VEE) / (RE)] = [(5 V) / (4 k-ohm)] = 1.25 mA; VCE = VCC - IERL = 10 V - (1.25 mA) 5 k-ohm; VCE = 10 V - 6.25 V = 3.75 V</p>
<p>Subject: Clinical Medicine; Subfield: Clinical Radiology; Image Type: Body Scans: MRI, CT.; Difficulty: Hard</p>	<p>Subject: History; Subfield: Modern History; Image Type: Comics and Cartoons; Difficulty: Easy</p>	<p>Subject: Electronics; Subfield: Analog electronics; Image Type: Diagrams; Difficulty: Hard</p>

MME

Размер: 14 подзадач, 2-3т

Тип задач: Yes/No

Язык: Английский

Особенности:

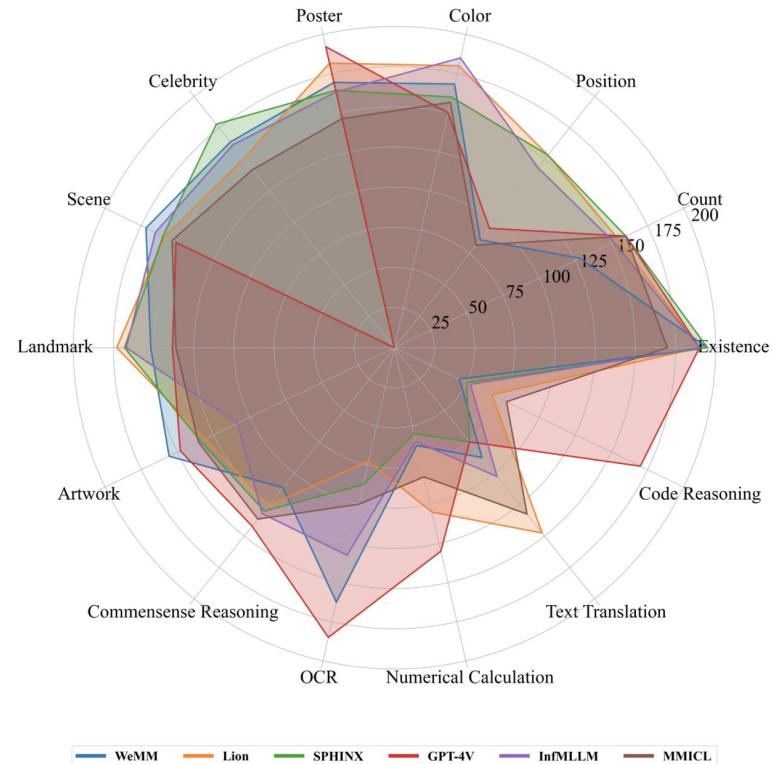
- Содержит 2 метрики Acc, Acc+
- Для каждого изображения - 2 вопроса
- Данные собирались вручную

Упоминание в статьях: Qwen-omni, LLaVA-OneVision

Acc - доля правильных ответов

Acc+ - доля правильных пар ответов (на оба вопроса)

Итоговый score = "Acc" + "Acc"



MME (Примеры)

Perception (Coarse-Grained Tasks)

Existence

[Y] Is there a **elephant** in this image?
 [N] Is there a **hair drier** in this image?

[Y] Is there a **refrigerator** in this image?
 [N] Is there a **donut** in this image?

Count

[Y] Is there a total of **two** person appear in the image?
 [N] Is there only **one** person appear in the image?

[Y] Are there **two** pieces of pizza in this image?
 [N] Is there only **one** piece of pizza in this image?

Position

[Y] Is the motorcycle on the **right** side of the bus?
 [N] Is the motorcycle on the **left** side of the bus.

[Y] Is the baby on the **right** of the dog in the image?
 [N] Is the baby on the **left** of the dog in the image?

Color

[Y] Is there a **red** coat in the image?
 [N] Is there a **yellow** coat in the image?

[Y] Is there a **red** couch in the image?
 [N] Is there a **black** couch in the image?

Perception (OCR Task)

OCR

[Y] Is the phone number in the picture "**0131 555 6363**"?
 [N] Is the phone number in the picture "**0137 556 6363**"?

[Y] Is the word in the logo "**high time coffee shop**"?
 [N] Is the word in the logo "**high tite coffee shop**"?

Commonsense Reasoning

Numerical Calculation

[Y] Should I **stop** when I'm about to **cross** the street?
 [N] When I see the sign in the picture, can I **cross** the street?

[Y] Is there **one** real cat in this picture?
 [N] Is there **two** real cats in this picture?

[Y] Is the answer to the arithmetic question in the image **65**?
 [N] Is the answer to the arithmetic question in the image **56**?

[Y] Should the value of "a" in the picture equal **3**?
 [N] Should the value of "a" in the picture equal **2**?

Perception (Fine-Grained Tasks)

Poster

[Y] Is this movie directed by **francis ford coppola**?
 [N] Is this movie directed by **franklin j. schaffner**?

[Y] Is this movie titled **twilight (2008)**?
 [N] Is this movie titled the **horse whisperer (1998)**?

Celebrity

[Y] Is the actor inside the red box called **Audrey Hepburn**?
 [N] Is the actor inside the red box called **Chris April**?

[Y] Is the actor inside the red box named **Jim Carrey**?
 [N] Is the actor inside the red box named **Jari Kinnunen**?

Scene

[Y] Does this image describe a place of **moat water**?
 [N] Does this image describe a place of **marsh**?

[Y] Is this picture captured in a place of **galley**?
 [N] Is this picture captured in a place of **physics laboratory**?

Landmark

[Y] Is this an image of **Beijing Guozijian**?
 [N] Is this an image of **Klinikkirche (Pfafferode)**?

[Y] Is this a picture of **Church of Saint Giles in Prague**?
 [N] Is this a picture of **Pfarrkirche St. Martin an der Raab**?

Artwork

[Y] Does this artwork belong to the type of **still-life**?
 [N] Does this artwork belong to the type of **mythological**?

[Y] Is this artwork displayed in **musee du louvre**?
 [N] Is this artwork displayed in **galleria nazionale d'arte moderna e contemporanea**?

Cognition (Reasoning Tasks)

Text Translation

老味道

[Y] Appropriate to translate into English '**classic taste**'?
 [N] Appropriate to translate into English '**strawberry flavor**'?

共同努力

[Y] Appropriate to translate into English '**work hard together**'?
 [N] Appropriate to translate into English '**be filled with intrigue**'?

Code Reasoning

[Y] Python code. Is the output of the code '**Hello**'?

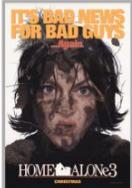
```
a = 10
if a < 20:
    print("Hello")
else:
    print("World")
```

 [N] Python code. Is the output of the code '**World**'?

```
b = "12345"
a = b.find('1')
print(b)
```

[Y] Python code. Is the output of the code '**0**'?
 [N] Python code. Is the output of the code '**I**'?

MME (Примеры)



[Y] Is this movie originated from the country or region of usa? Please answer yes or no.

[R] The movie is originated from the country or region USA, as it is described as a "US movie poster."

[N] Is this movie originated from the country or region of spain? Please answer yes or no.

[R] This movie is originated from the country Spain. It is described as a " Spanish movie " or " Spanish film."



[Y] Is this photo taken in a place of greenhouse indoor? Please answer yes or no.

[R] The photo is taken inside a greenhouse, as the man is holding a tray of plants and flowers, which are typically grown and nurtured in such environments

[N] Is this photo taken in a place of aqueduct? Please answer yes or no.

[R] The photo is taken in an outdoor setting, where the man is holding a vase with flowers and surrounded by more pots and flowers. There is no visible aqueduct in the image



[Y] Are there two bananas in the image? Please answer yes or no.

[R] Yes, there are two bananas in the image.

[N] Are there three bananas in the image? Please answer yes or no.

[R] No, there are not three bananas in the image. There is **only one banana** on the table.

$$15 \times 15 =$$

[Y] Is the answer to the arithmetic question in the image 225? Please answer yes or no.

[R] No, the answer to the arithmetic question in the image is not 225. The question asks for the **sum of 11 and 15**, which is 26.

[N] Is the answer to the arithmetic question in the image 1515? Please answer yes or no.

[R] No, the answer to the arithmetic question in the image is not 1515. The question asks for the **sum of 15 and 15**, which is 30.



[Y] Is this photo taken in a place of bazaar outdoor? Please answer yes or no.

[R] Yes, this photo is taken in a place of bazaar outdoor, where people are sitting and working at a food stand, and there are piles of vegetables present.

[N] Is this photo taken in a place of office? Please answer yes or no.

[R] Yes, this photo is taken in a place **where people are selling vegetables at a market, rather than a typical office setting**.

$$17 \times 20 =$$

[Y] Is the answer to the arithmetic question in the image 340? Please answer yes or no.

[R] No, the answer to the arithmetic question in the image is not 340. The question asks for the result of multiplying 17 by 20, which is 340. **However, the answer provided in the image is 34.**

[N] Is the answer to the arithmetic question in the image 17? Please answer yes or no.

[R] Yes, the answer to the arithmetic question in the image is 17.



[Y] Is there a sports ball in this image? Please answer yes or no.

[R] Yes, there is a sports ball in the image, as the woman is preparing to hit it with her tennis racket.

[N] Is there a horse in this image? Please answer yes or no.

[R] Yes, there is a horse in the image. **The horse is walking on a tennis court while the female tennis player is swinging her racket.**



[Y] Is there a motorcycle in this image? Please answer yes or no.

[R] Yes, there is a motorcycle in the image. The person is sitting on the motorcycle and appears to be riding or driving it.

[N] Is there a airplane in this image? Please answer yes or no.

[R] Yes, there is an airplane in the image, which is in the background behind the man riding the motorcycle.

Ограничения современных бенчмарков

- Ограничено время жизни
- Мультимодальные бенчи трудно, долго и дорого собирать
- *Недостаток русскоязычных бенчмарков*

MARKER: Русскоязычный бенчмарк

Размер: 3.5т

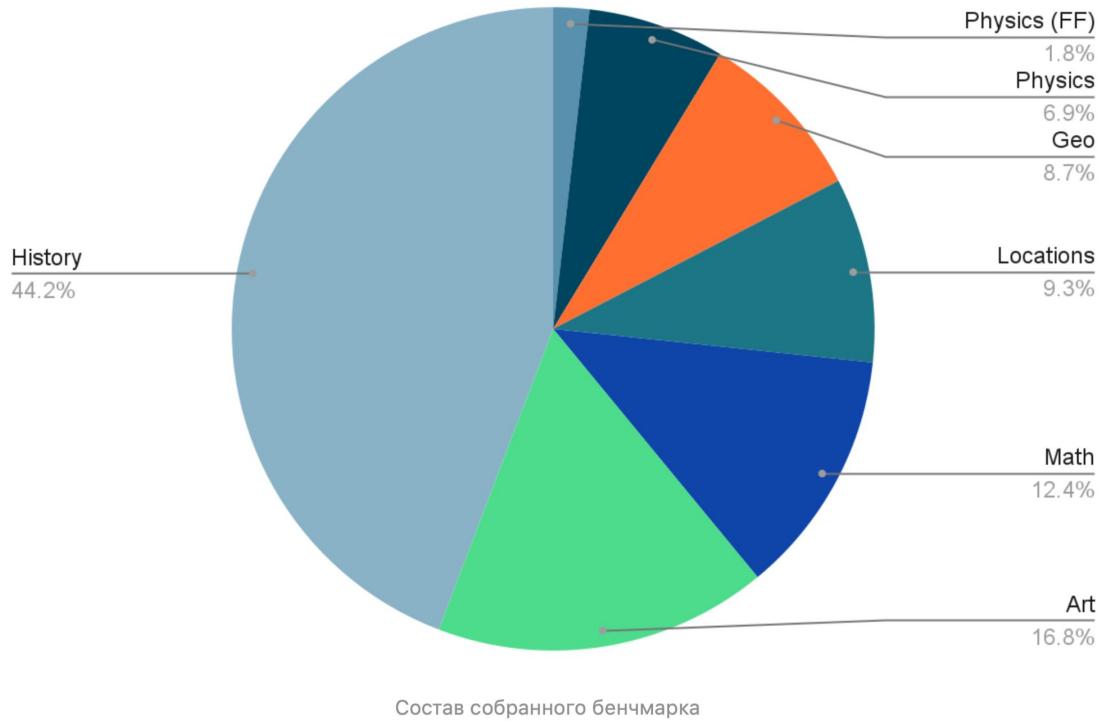
Тип задач: Multiple choice / open - ended

Язык: Русский

Особенности:

- Опирается в основном на задачи школьной программы

Метрики: Accuracy | Circular Eval



MARKER: Русскоязычный бенчмарк (Примеры)



Задача: На какой век ссылается данное изображение?

Варианты ответа: 'XVII в.',
'XIII в.', 'XIV в.', 'XI в.'

Ответ: XVI в.

Задача: Какой из предложенных вариантов, лучше всего описывает данное изображение?

Варианты ответа: '...', 'Лжедмитрий II. Портретная фантазия с гравюры 1698 г.', 'Портрет А. С. Матвеева. С гравюры XVIII в.', '...'
Ответ: Портрет А. С. Матвеева. С гравюры XVIII в.

Примеры задач из блока «История»

Вопросы?



Метрики и процедуры оценки Мультимодальных моделей



Классические метрики

BLEU, ROUGE, METEOR — сходство с референсным текстом, качество генерации описаний и ответов

CIDEr, SPICE — семантическая близость в captioning (для изображений и видео)

Perplexity — Мера неопределённости языковой модели при генерации текста; чем ниже, тем предсказания более вероятны.

F1, Accuracy — для задач классификации, ответа на вопросы

Метрики: Cer | Wer

Модальности: Аудио + Текст.

Используются символы для CER:

- **S** - кол-во замен слов
- **D** - кол-во удаленных символов
- **I** - кол-во новых символов
- **N** - кол-во слов всего

WER - тоже самое, но для слов

Чем меньше **WER/CER** тем лучше

Пример:

the cat sat -> the dog sat

[S = 1]; WER = 1/3

$$\text{CER} = \frac{S + D + I}{N}$$

$$\text{WER} = \frac{S + D + I}{N}$$

How CLIPScore works

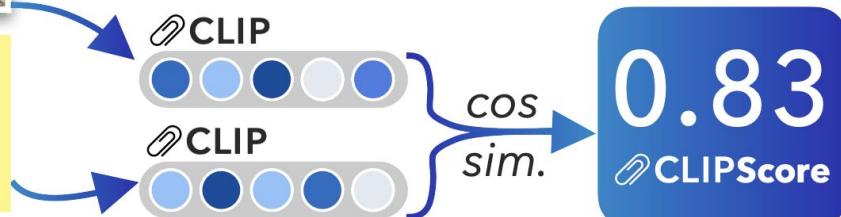


CANDIDATE

Two dogs run towards each other on a marshy area.

REFERENCE CAPTIONS

- Two dogs are running toward each other across the sand.
- Two dogs run toward each other.
- Two dogs are running towards each other on a beach.



Human Evaluation | LLM-as-judge



Счастливый котик

Оцени описание
по шкале 1-3:
1: Не релевантно
2: Частично релевантно
3: Релевантно



3

Ты модель оценщик,
поставь оценку 1-3:
1: Не релевантно
2: Частично релевантно
3: Релевантно



Постановки задач для бенчмарков

Exact Match:

Точное совпадение с эталоном

- Только короткие ответы
- + Минимальная подготовка данных

Multiple Choice:

Выбор из вариантов ['A', 'B', ...]

- Сложнее составлять варианты
- + Чёткая автоматическая оценка

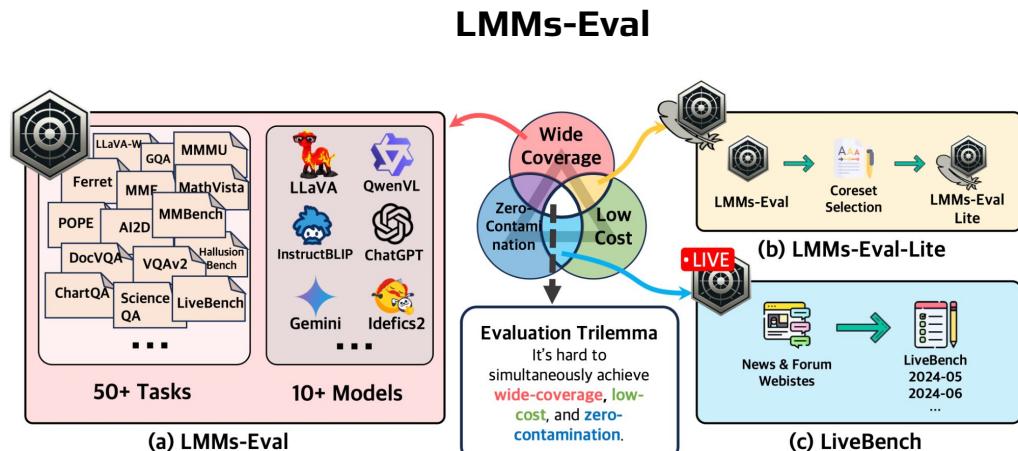
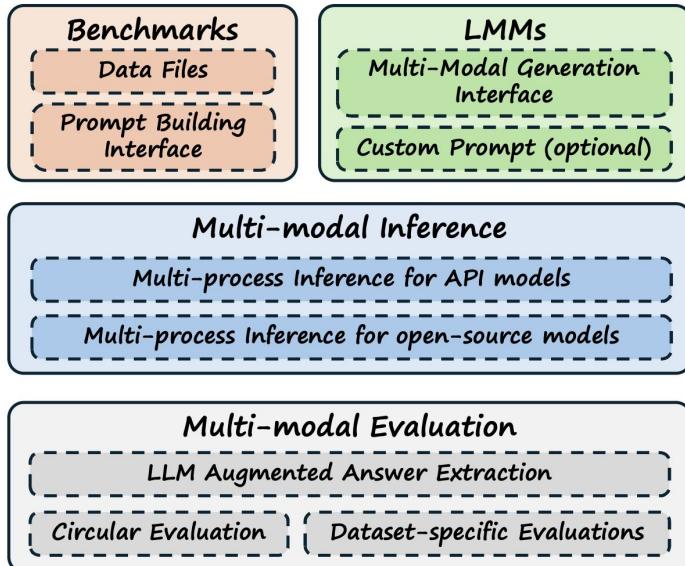
Open Ended:

Свободная форма ответа

- Сложно автоматизировать оценку
- + Гибко, подходит для сложных задач

Инструменты для оценивания LMMs

VLM-Eval-Kit



Сбор и препроцессинг мультимодальных данных



Сбор и препроцессинг мультимодальных данных

Веб-парсинг и скрейпинг

Beautiful Soup, Scrapy, WebAutoGUI

Краудсорсинг

Получение быстрой разметки, через специализированные сайты/платформы

Найм ассесоров

Найм и обучение специализированных людей для получения качественной разметки

Промышленные и корпоративные данные

Использование данных собираемых биг технами

Открытые датасеты

COCO, LAION-5B, Libri Speech, Common Voice, WebVid

Синтетические данные

Создание искусственных данных

Веб-парсинг и скрейпинг

Инструменты для сбора данных:

- BeautifulSoup - парсинг HTML
- Scrapy - парсинг HTML
- WebAutoGUI - Эмуляция поведения пользователя

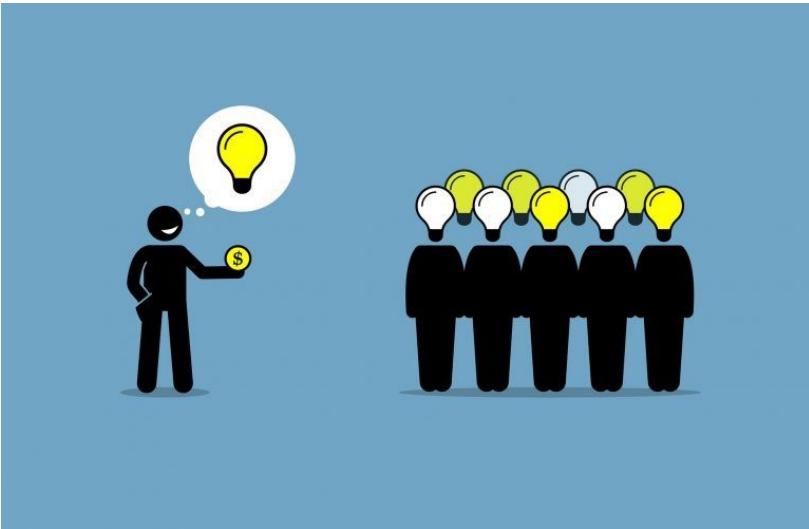
Плюсы:

- Эффективно и предсказуемо
- Дёшево

Уязвимости:

- “Анти бот системы”
- Личные данные
- Интеллектуальный труд

Краудсорсинг



Плюсы:

- Наличие специализированных сервисов
- Быстрый способ получения “человеческой” разметки
- Относительно дёшево

Уязвимости:

- Неквалифицированность и низкая мотивация разметчиков
- Риск низкого качества данных
- Требуется инструкция для разметки
- Требуется перекрёстная проверка, “золотые” задания, фильтрация
- Не подходит для сложных, экспертных или конфиденциальных задач

Найм асессоров



Плюсы:

- Проверенные, обученные и мотивированные специалисты
- Высокое качество разметки
- Возможность решать нетривиальные или экспертные задачи
- Можно давать тщательные инструкции, проводить обучение и ревью результатов

Уязвимости:

- Сильно дороже по сравнению с краудсорсингом
- Более низкая масштабируемость и скорость
- Нужно время на отбор, обучение и управление
- Меньше гибкости для экспериментальных или очень разнообразных задач

Промышленные и корпоративные данные

Особенности:

- Данные собираются автоматически благодаря созданным сервисам: Яндекс музыка, VK video
- Доступ только у крупных организаций
- Долгосрочное планирование
- Постоянное логирование действий пользователей
- Высокие требования к хранению и обработке
- Конфиденциальность и соблюдение законодательства

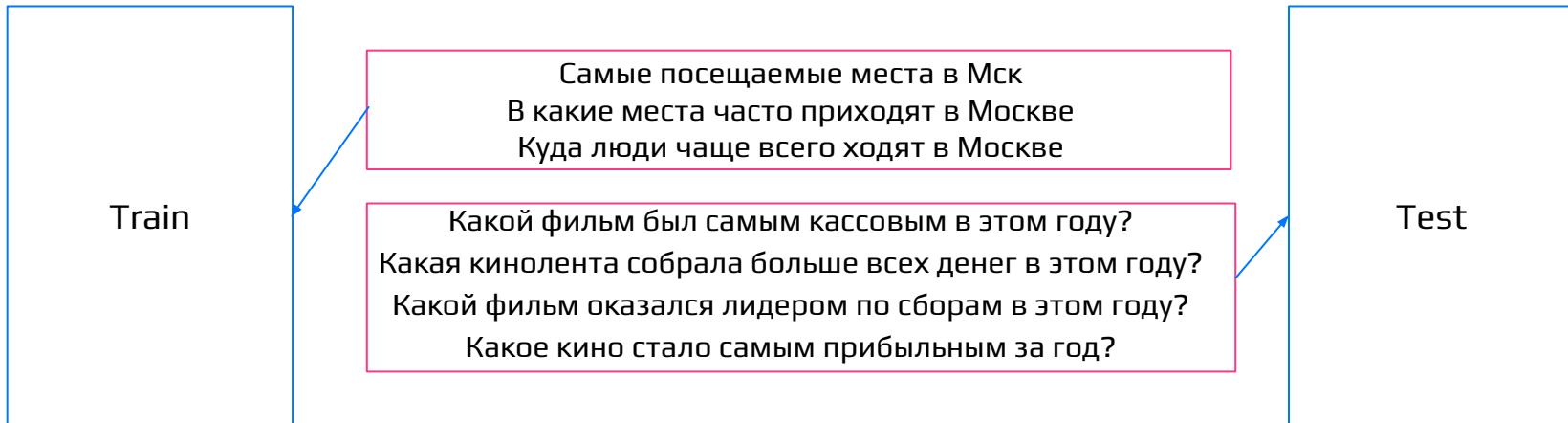
Дополнительно: Дедупликация

Train

Самые посещаемые места в Мск
В какие места часто приходят в Москве
Куда люди чаще всего ходят в Москве
Какой фильм был самым кассовым в этом году?
Какая кинолента собрала больше всех денег в этом году?
Какой фильм оказался лидером по сборам в этом году?
Какое кино стало самым прибыльным за год?

Test

Дополнительно: Дедупликация



Дополнительно: Проблема подбора дистракторов



Какой из предложенных вариантов лучше всего описывает данное изображение?

- 1) Пушкин А.С.
- 2) Бородинское сражение
- 3) Золотая монета IV века
- 4) XI-XIII века

Синтетические данные, преимущества и недостатки

• • • • •

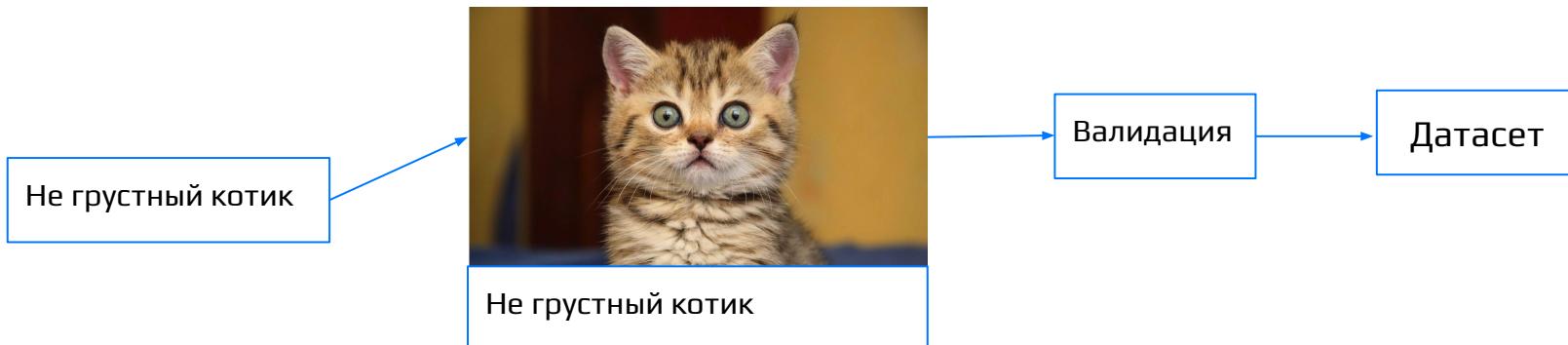
Как генерируются синтетические данные

Ключевые подходы к синтетической генерации данных для мультимодальных задач:

- Использование генеративных моделей (DALL-E, Stable Diffusion)
- Симуляторы и движки (Unreal, Unity)
- Аугментация и миксинг реальных данных
- Программные правила и шаблоны
- Комбинированные пайплайны (LLM + генератор изображений)

Workflow: LLM+GenAI для captioned images

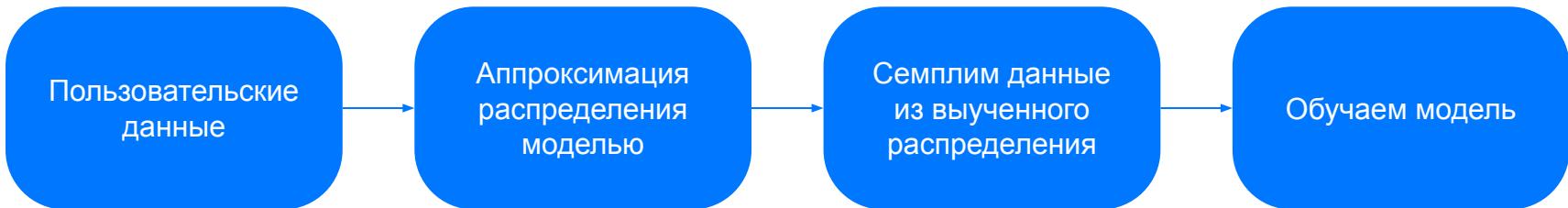
- Генерация текстовых данных с помощью LLM
- Генерация изображений (или другого модальности) по тексту
- Валидация и фильтрация
- (Опционально) Реверс-генерация
- Финальная сборка и подготовка



Плюсы синтетики

- Масштабируемость
- Контролируемость
- Балансировка классов
- Аугментация
- Защита приватности
- Ясная аннотация
- Тестирование крайних случаев (corner cases)

Подробнее про защиту приватности



Проблемы синтетики

- Domain gap
- Синтетические артефакты
- Риск переобучения на особенности генератора
- Ограниченностъ “жизненного” разнообразия
- Юридические и лицензионные вопросы
- Возможная низкая доверительность для финальной оценки
- “Непризнание” (low acceptance) в сообществе

Итоги

- Различные модальности и датасеты
- Современные бенчи и их проблемы
- Метрики и их особенности
- Особенности сбора и генерации данных



Вопросы?



Спасибо за
внимание!



education

Перерыв

