LLM Basics

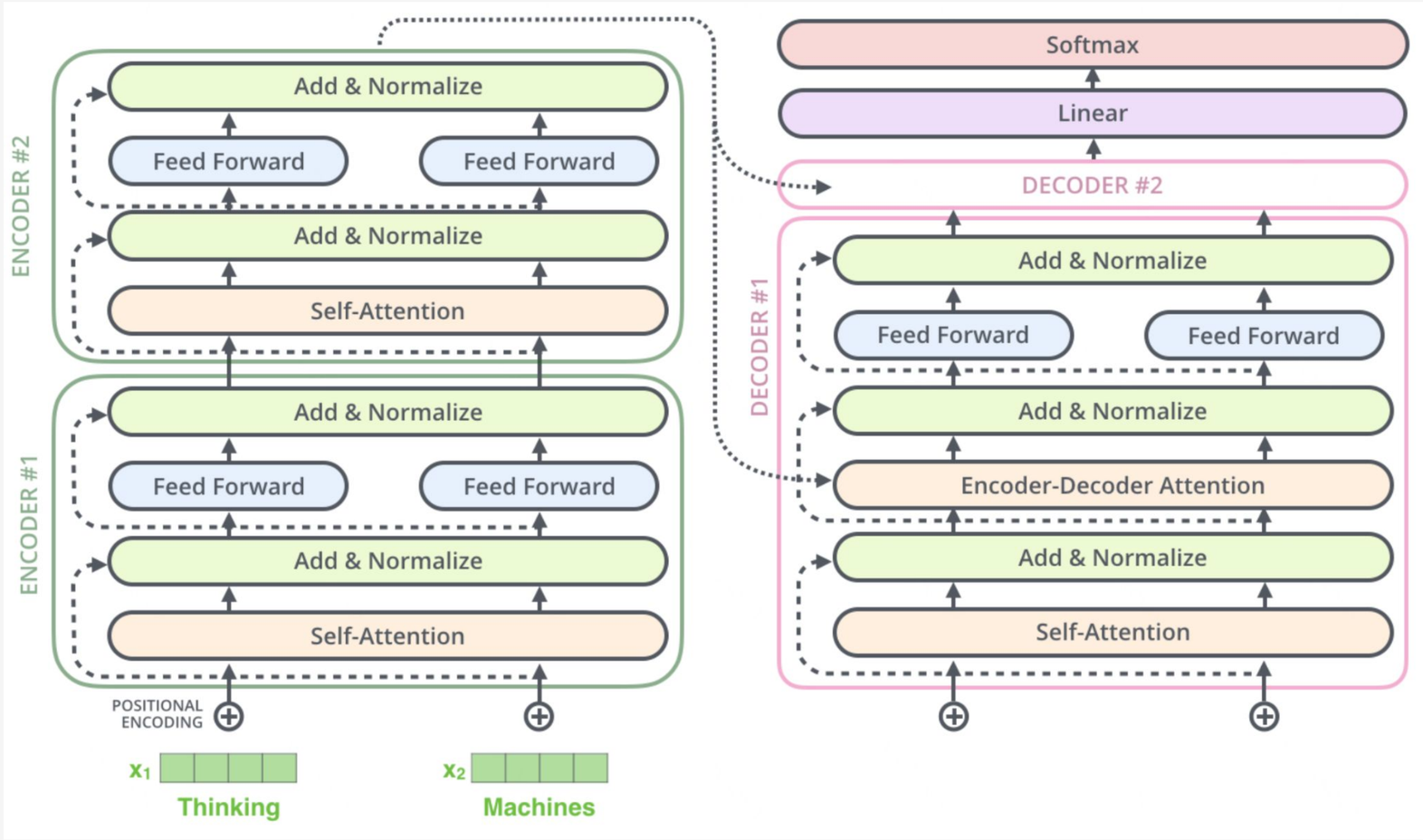
Natural Language Processing

Ника Зыкова, 2025/11/06

Transformer целиком

Итоговый Трансформер состоит из последовательности преобразований с помощью механизма внимания.

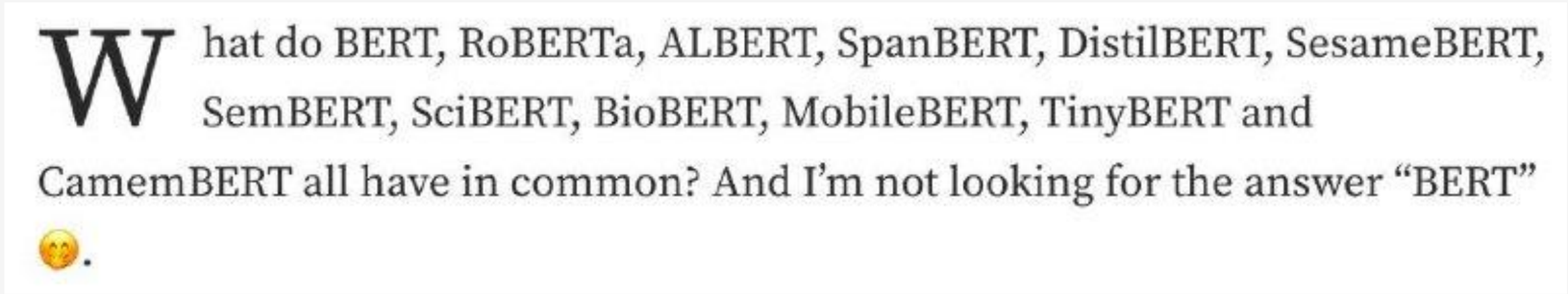
И еще нескольких нюансов.



В чем разница?

В разных моделях используют разные части архитектуры трансформера.

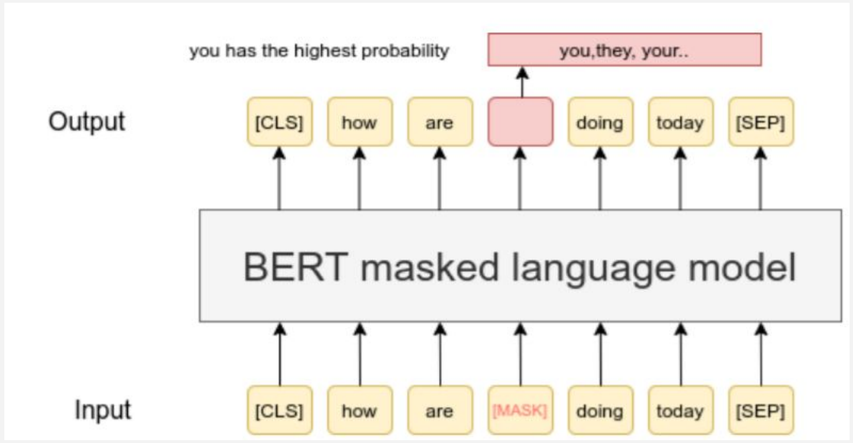
Encoder	Decoder	Encoder-Decoder
BERT	GPT1, 2, 3, 4	T5
RoBERTa	ChatGPT	Bart
ALBERT	PaLM	Pegasus
ELECTRA	LLaMA	ProphetNet



Encoder VS Decoder

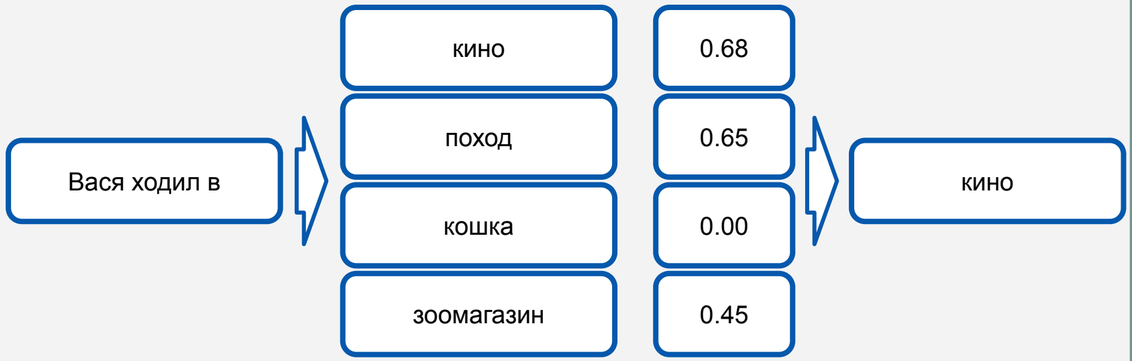
Encoder

- Слой эмбеддингов + позиционные эмбеддинги + несколько слоев энкодера;
- При векторизации конкретного слова смотрит на все слова в последовательности;
- Задачи Masked Language Modeling + Next Sentence Prediction.

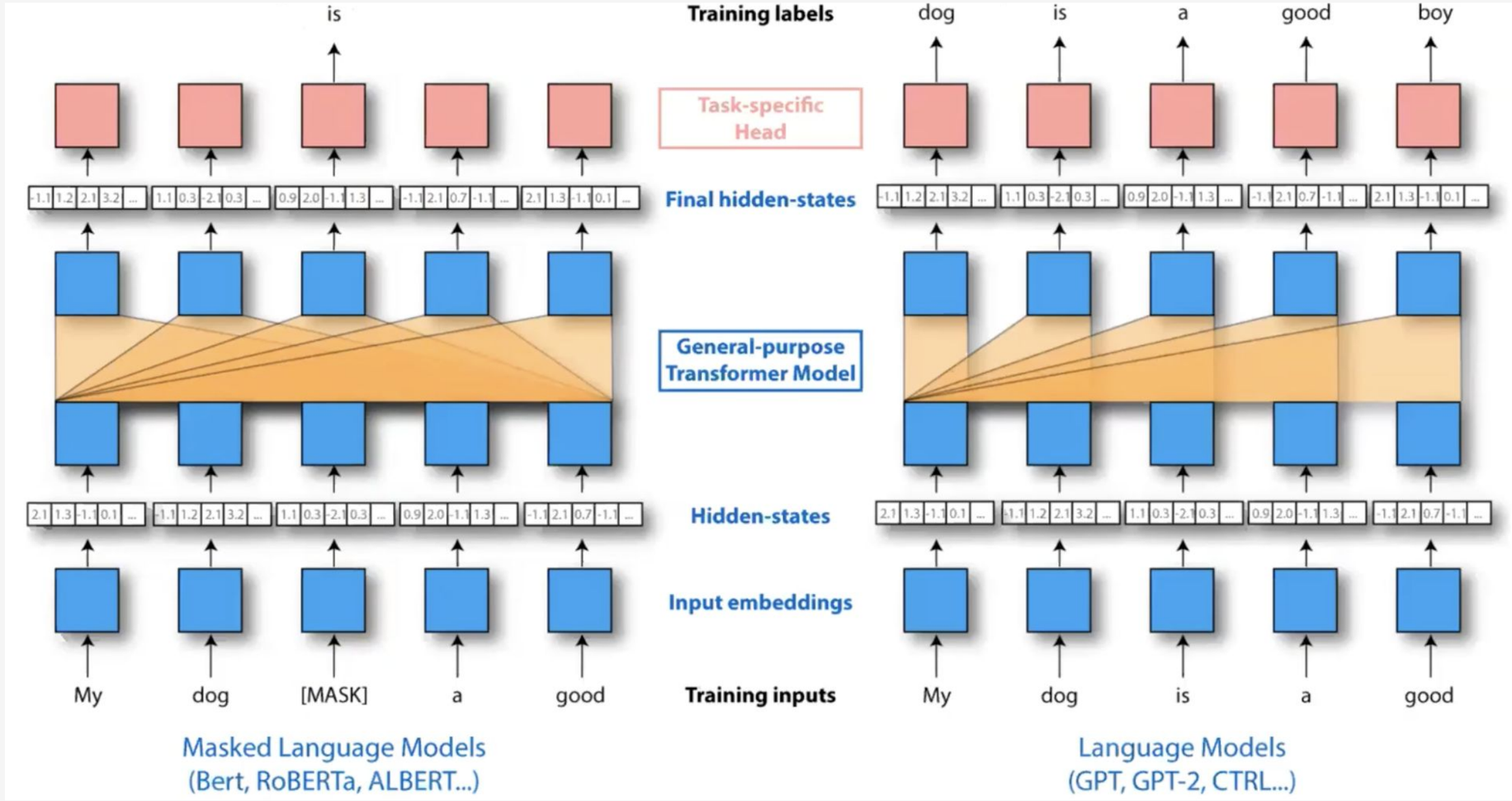


Decoder

- Слой эмбеддингов + позиционные эмбеддинги + несколько слоев декодера;
- При векторизации смотрит только на уже сгенерированные слова;
- **Не содержит cross-attention** ;
- Задача Next Token Prediction.



Encoder VS Decoder

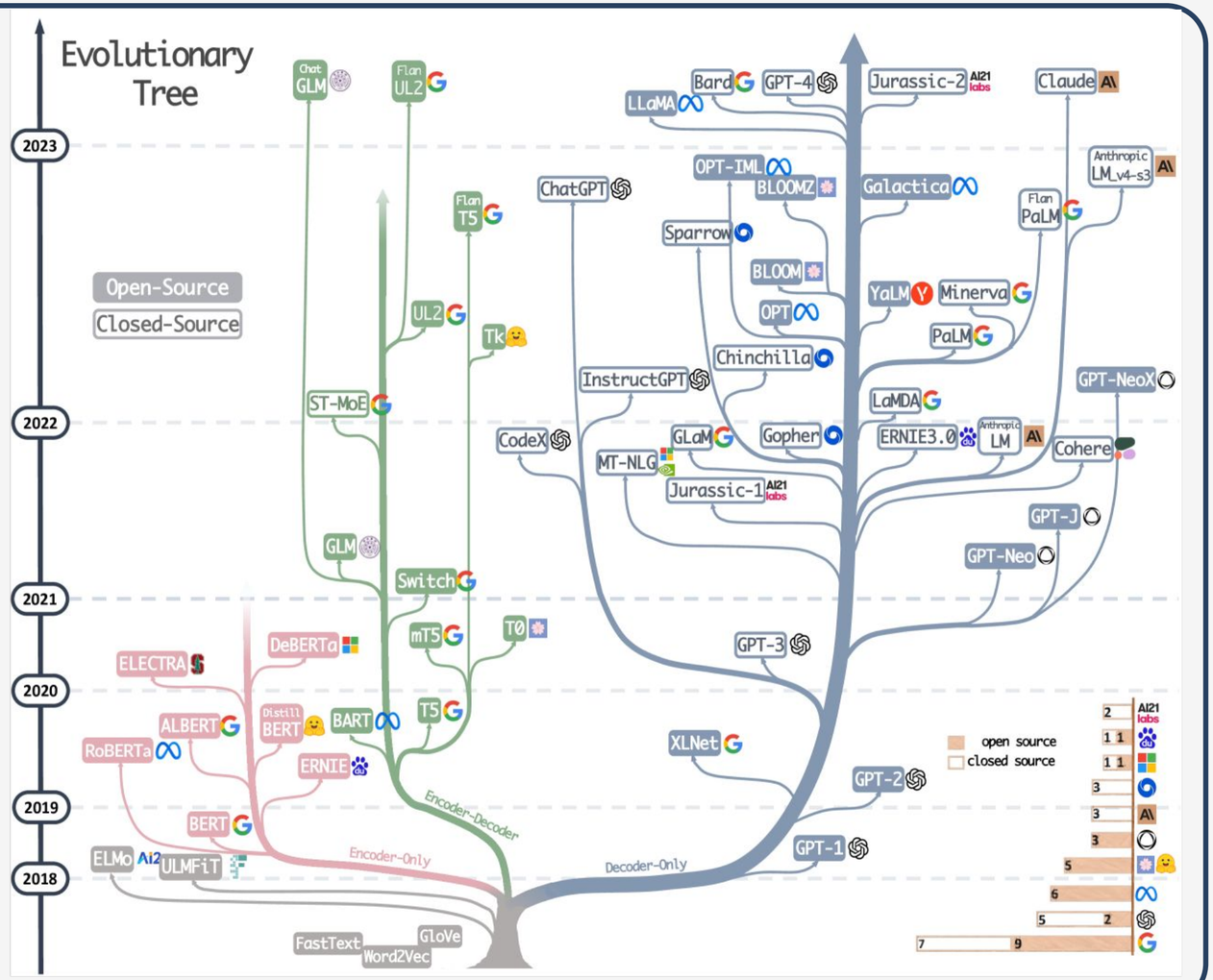


https://www.youtube.com/watch?v=t86G11tfVNw

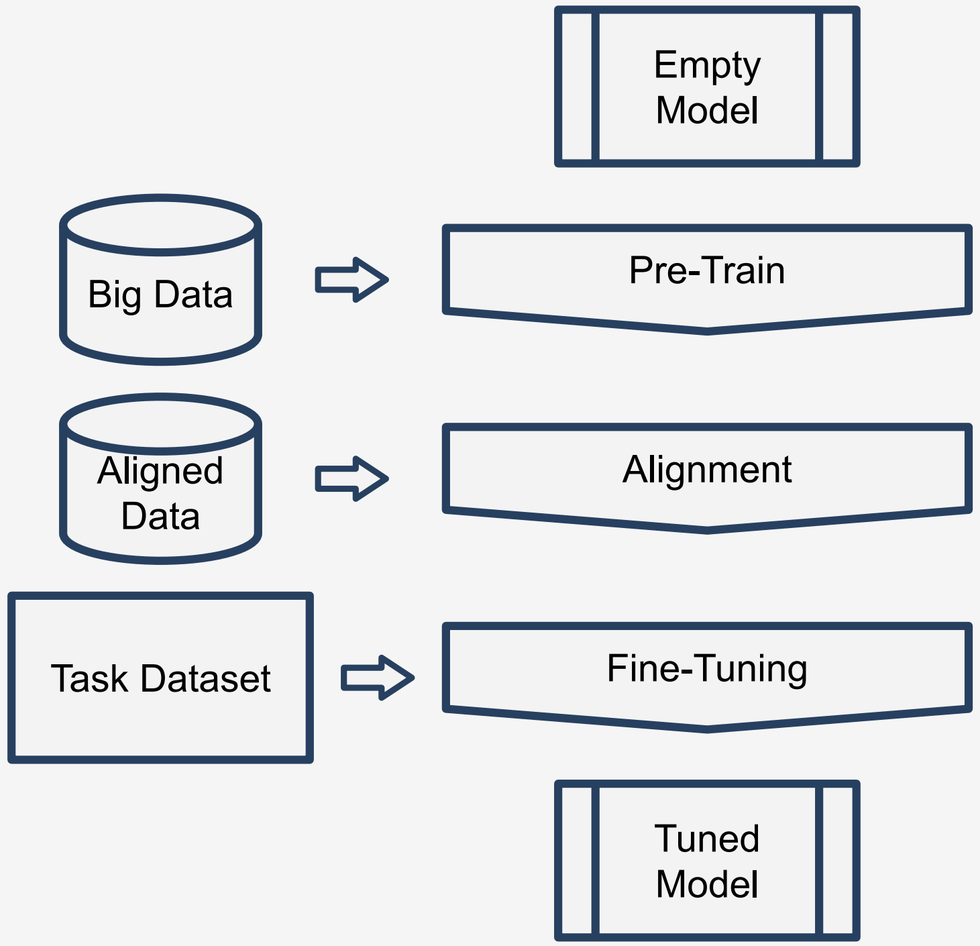
Large Language Model

У понятия большой языковой модели (как и у просто языковой модели) нет конкретного определения. Однако приятно считать, что к ним относятся:

- Decoder-only модели, хотя первые GPT так называют редко (сюда относятся почти все LLM);
- Большие генеративные модели, вышедшие после 2022 года (здесь самая известная Mamba, которая не Трансформер).



Процесс обучения



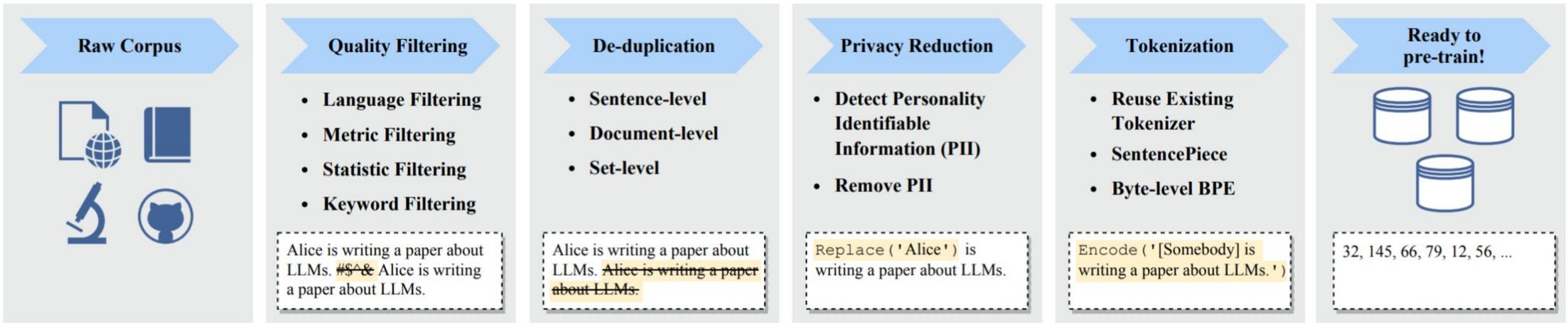
Обычно процесс разделяют на три больших этапа:

- Pre-train: на данном этапе происходит основное обучение на задачу NTP. Длинный и вычислительно дорогой процесс;
- Alignment: здесь модель учится соблюдать определенный набор правил, “выравнивается” под некоторое распределение (этика, домен и т.д.)
- Fine-tuning: дообучение модели на конкретную задачу (суммаризация, консультация клиента и т.д.)

Pre-train

На самом деле претрейн тоже разделяется на части:

- Pre-train: триллионы токенов, из подготовки данных чаще всего только дедупликация и анонимизация (не всегда);
- Continual Pre-Training (CPT): миллиарды токенов, учим модель конкретным навыкам (рассуждения, использование тулов и т.д.)
- Domain-Adaptive Pre-training (DAP): часто чуть больше CPT, адаптируем модель к домену (банковский, медицинский и т.д.)



Параметры обучения

Обучение модели обычно существенно зависит от трех параметров:

- Количества итераций обучения: редко делают больше 1 эпохи;
- Размера датасета: чаще всего ограничивается сверху доступными ресурсами;
- Размера самой модели: ограничивается ресурсами + кол-вом данных (чем больше параметров, тем больше надо данных.

И, конечно, архитектуры, но тут чаще всего выбор между Трансформером и Трансформером.

Alignment

На данный момент есть целая активно развивающаяся область на стыке AI, информационной безопасности и этики, которая занимается вопросами элаймента.

Безопасность и этика

Учим модель правильно реагировать на запросы:

- Никакой симметричной реакции на негатив;
- Нельзя разглашать персональные данные.

Реакция на инструкции

Современная llm должна уметь следовать произвольным наборам инструкций:

- “Напиши код для задачи”;
- “Расскажи рецепт тирамису”.

Выбросы и смещения

Модель не должна галлюцинировать и выходить из строя на странных примерах:

- Мир меняется;
- Пользователи не всегда пишут что-то осмысленное.

Fine-Tuning

Существует множество задач, на которые можно дообучать модель. Причем современные LLM часто дообучают на задачи, которые мы привыкли считать задачами для encoder-only моделей (классификация, оценки).

Encoder + Classification Head

- Natural Language Inference
- Sentiment Analysis
- Spam Detection
- Hate Speech Recognition
- Topic classification

Decoder | Encoder-Decoder

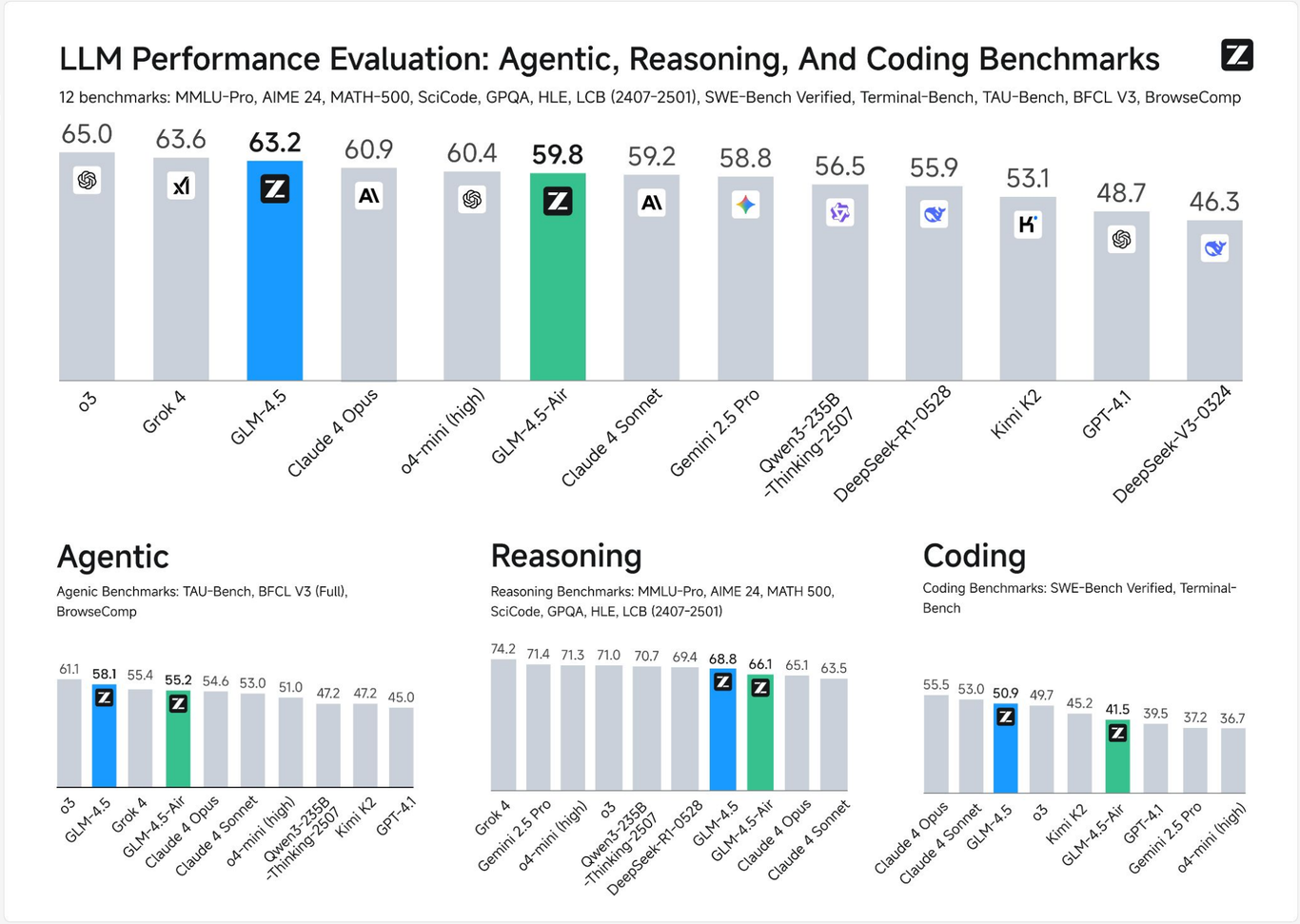
- Summarization
- Question Answering
- Machine Translation
- Style Transfer
- Text Generation in general

Оценка качества

Обычно общие способности LLM тестируются с помощью бенчмарков: датасетов, проверяющих один конкретный навык.

Особенность претрейна здесь в том, что проверка сложных взаимодействий (следование инструкциям, ведение диалога) пока невозможна.

Из-за этого качество модели обычно проверяется либо через дообучение на датасет, либо через few-shot подход на отдельном наборе бенчмарков.



https://z.ai/blog/glm-4.5

Оценка качества

А еще для анонимной оценки от пользователей были созданы ресурсы с “ареной для моделей”:

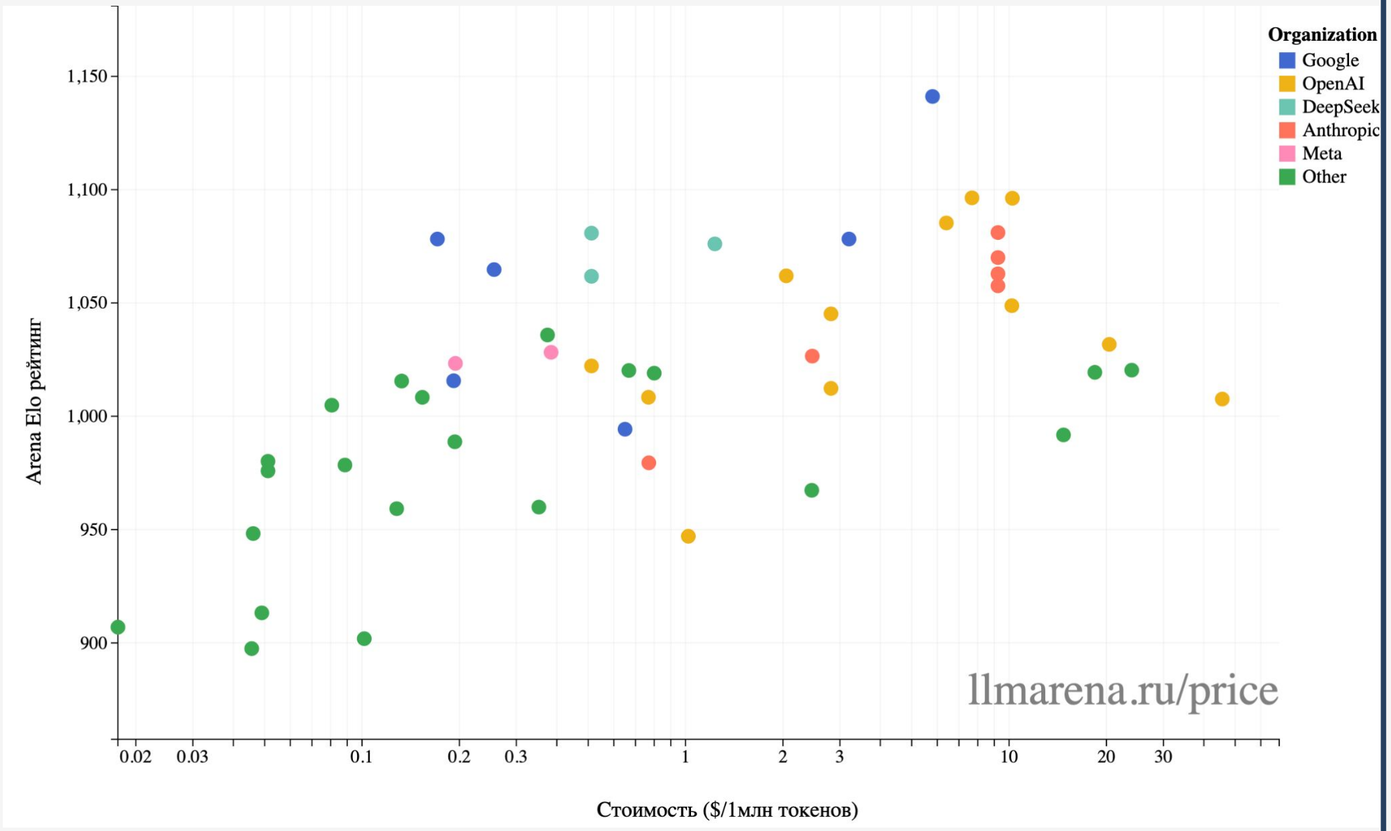
lmarena.ai (может не открыться)

llmarena.ru

Они достаточно медленные, если сравнивать с тем же сайтом openai, но зато вы можете попробовать все модели в одном месте.

А еще бывают демо на hugging face:

https://huggingface.co/spaces/Alibaba-NLP/Tongyi-DeepResearch



https://llmarena.ru/