

LLM-инженер

online course ✨

модуль #2

лекция: современные LLM

Современные LLM – часть 1

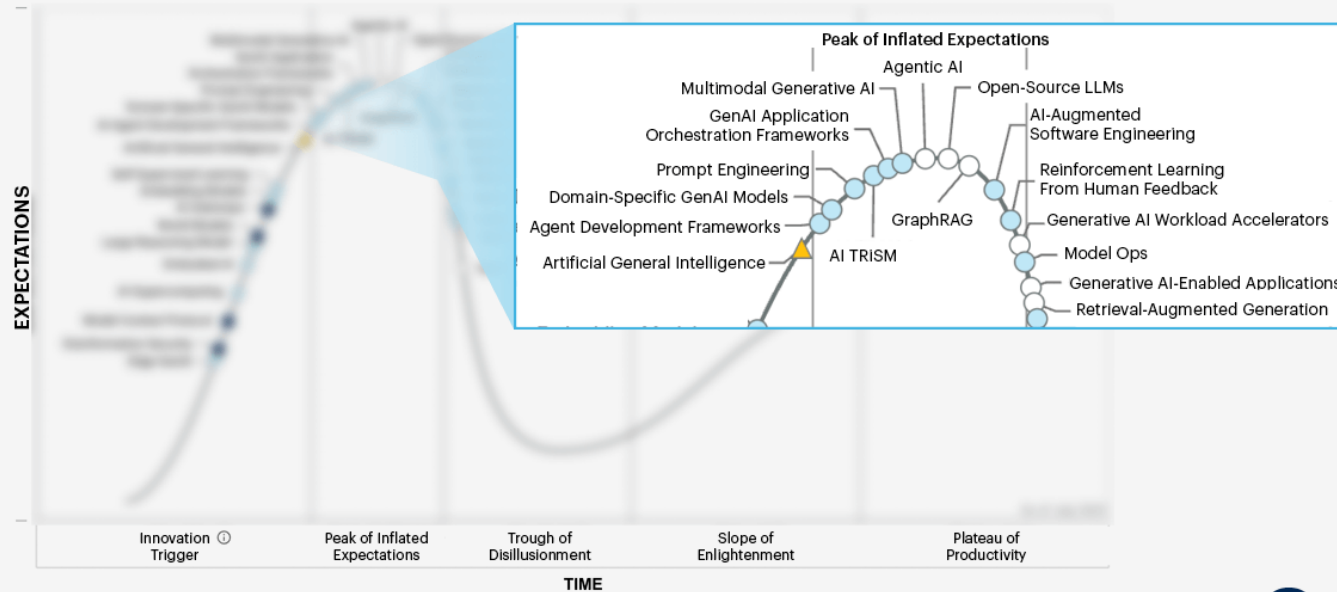
Некоторые тренды

- Рост частоты релизов - крупные игроки индустрии выпускают новые модели или значительные обновления в среднем каждые 2-4 недели
- Технологические прорывы – рост длины контекста, оптимизация обучения, ...
- Экспоненциальный рост требований к ресурсам
- Инфраструктурные вызовы
- Рост хайпа

AI Hype Cycle

Hype Cycle for Generative AI, 2025

Plateau will be reached: ☐ < 2 yrs. ☐ 2-5 yrs. ☒ 5-10 yrs. ☐ >10 yrs.



Source: Gartner
© 2025 Gartner, Inc. and/or its affiliates. All rights reserved. CTMKT_3881100

Gartner®

Closed-source VS open-weight

Closed-source vs. open-weight models

Llama 3 405B from Meta closes the gap between closed-source and open-weight models.

MMLU (5-shot)













Closed-source LLMs

Comprehensive Comparison

Feature	Claude 3	Gemini	GPT-4
Developer	Anthropic	Google	OpenAI
Model Sizes	Haiku, Sonnet, Opus	-	-
Training Data	Unknown	Unknown	Trained on online data up to Sept 2021
Benchmarks	Leads in MMLU, GPQA, GSM8K and others	Lags behind Claude 3	Trails Claude 3 Opus but GPT-4 Turbo leads
Coding	Excels with detailed explanations	Good	Good, but lacks details
Math Reasoning	Strongest with Opus model	Weaker than Claude 3	Weaker than Claude 3
Vision	On par, faster response times	Good, but ethical limits	On par with Claude 3
General Knowledge	Detailed scientific explanations	More general explanations	Detailed like Claude 3
Prompt Following	Opus leads, up to 10 logical outputs	Weaker than Claude 3	Up to 9 logical outputs
Multilingual	Unknown	Specialty, over 100 languages	Good, 25+ languages
Input Modalities	Text, images	Text, images, audio	Text, images, audio, video

Text 2 days ago

Rank (UB) ↑	Model ↓	Score ↓	Votes ↓
1	 gemini-2.5-pro	1456	30 752
1	 gpt-5-high	1456	8 532
1	 claude-opus-4-1-20250805-thi...	1451	5 812
2	 o3-2025-04-16	1446	36 660
2	 claude-opus-4-1-20250805	1441	8 407
3	 chatgpt-4o-latest-20250326	1442	33 809
3	 gpt-4.5-preview-2025-02-27	1438	15 271
8	 grok-4-0709	1426	15 580
8	 gpt-5-chat	1422	5 715
8	 qwen3-235b-a22b-instruct-2507	1421	8 642
View all			

Rank* (UB) ▲	Model ▲	Arena Elo ▲	95% CI ▲	Votes ▲	Organization ▲	License ▼	Knowledge Cutoff ▲
1	Google: Gemini Pro 2.5 Preview 03-25	1152	+24/-24	662	Google	Proprietary	March 2025
2	gpt-4.1-2025-04-14	1100	+21/-21	765	OpenAI	Proprietary	June 2024
2	claude-3.7-sonnet-20250219:thinking	1095	+18/-20	934	Anthropic	Proprietary	February 2025
2	o4-mini-2025-04-16	1091	+22/-19	768	OpenAI	Proprietary	In training
2	Google: Gemini Pro 1.5	1089	+16/-15	1429	Google	Proprietary	Unknown

Pricing

GPT-5

The best model for coding and agentic tasks across industries

Price

Input:

\$1.250 / 1M tokens

Cached input:

\$0.125 / 1M tokens

Output:

\$10.000 / 1M tokens

GPT-5 mini

A faster, cheaper version of GPT-5 for well-defined tasks

Price

Input:

\$0.250 / 1M tokens

Cached input:

\$0.025 / 1M tokens

Output:

\$2.000 / 1M tokens

GPT-5 nano

The fastest, cheapest version of GPT-5
—great for summarization and classification tasks

Price

Input:

\$0.050 / 1M tokens

Cached input:

\$0.005 / 1M tokens

Output:

\$0.400 / 1M tokens

- 5 RPS – 432000 запроса в день
- Пусть каждый запрос ~1000 токенов, длина ответа – 500 токенов
- Для GPT-5 – $432000 * (1000 * 1.25\$ + 500 * 10\$) / 1000000 = 2700\$$

Closed-source LLMs

- Плюсы:
 - Быстрое прототипирование
 - Дешевая проверка гипотез
 - Лучшее* качество
- Минусы:
 - Нельзя* дообучать модели
 - Отсылаем данные на сторонний сервер
 - Юридически подтвержденные проблемы с ПД
 - При больших RPS и на долгой дистанции – дорого
 - Alignment и модерация не всегда удобны

Open-weight LLMs - Qwen

	Qwen3-30B-A3B MoE	QwQ-32B	Qwen3-4B Dense	Qwen2.5-72B-Instruct	Gemma3-27B-IT	DeepSeek-V3	GPT-4o 2024-11-20
ArenaHard	91.0	89.5	76.6	81.2	86.8	85.5	85.3
AIME'24	80.4	79.5	73.8	18.9	32.6	39.2	11.1
AIME'25	70.9	69.5	65.6	15.0	24.0	28.8	7.6
LiveCodeBench v5, 2024.10-2025.02	62.6	62.7	54.2	30.7	26.9	33.1	32.7
CodeForces Elo Rating	1974	1982	1671	859	1063	1134	864
GPQA	65.8	65.6	55.9	49.0	42.4	59.1	46.0
LiveBench 2024-11-25	74.3	72.0	63.6	51.4	49.2	60.5	52.2
BFCL v3	69.1	66.4	65.9	63.4	59.1	57.6	72.5
Multif 8 Languages	72.2	68.3	66.3	65.3	69.8	55.6	65.6

1. AIME 24/25: We sample 64 times for each query and report the average of the accuracy. AIME'25 consists of Part I and Part II, with a total of 30 questions.

2. Aider: We didn't activate the think mode of Qwen3 to balance efficiency and effectiveness.

3. BFCL: The Qwen3 models are evaluated using the FC format, while the baseline models are assessed using the highest scores obtained from either the FC or prompt formats.

- Mixture of Experts (MoE)
- Hybrid Reasoning System
- Grouped Query Attention (GQA)
- RoPE (Rotary Positional Encoding)
- Лицензия: Apache 2.0 для большинства моделей

[Qwen3: Think Deeper, Act Faster](#)

Open-weight LLMs - Mistral/Mixtral

- Sparse Mixture of Experts - Mixtral
- Sliding Window Attention (SWA) и Grouped Query Attention (GQA)
- Лицензия: Apache 2.0

Модель	Количество параметров	Контекстное окно	Архитектура	Цена (контекст/генерация)	Мультимодальность	Особенности
Mistral Large v2407	123B	128K токенов	Transformer	1.00P/2.40P	✗	Многоязычность, 80+ языков программирования
Mistral Small	22B	128K токенов	Transformer	0.20P/0.60P	✗	Open source, оптимизирована для простых задач
Ministral 8B	8B	128K токенов	Transformer	0.10P/0.10P	✗	Высокая производительность для своего размера
Ministral 3B	3B	128K токенов	Transformer	0.04P/0.04P	✗	Самая компактная и бюджетная модель
Mixtral 8x22B	176B (8x22B, 39B активных)	32K токенов	Sparse MoE Transformer	1.20P/1.80P	✗	Многоязычность, сильна в математике
Mixtral 8x7B	56B (8x7B, 12.9B активных)	32K токенов	Sparse MoE Transformer	0.60P/0.90P	✗	Производительность на уровне GPT-3.5
Codestral Mamba	7.3B	256K токенов	Mamba2	0.15P/0.15P	✗	Open source, специализация на коде, линейное время вывода
Pixtral 12B	12.4B (12B декодер + 400M энкодер)	128K токенов	12B Multimodal Decoder + 400M Vision Encoder	0.12P/0.12P	✓	Мультимодальная, работа с изображениями любого размера

Open-weight LLMs – Llama 4



- Mixture of Experts (MoE)
- iRoPE и межслойное внимание
- Лицензия: Llama 3.1 Community License - разрешает использование выходных данных моделей для улучшения других моделей

[The Llama 3 Herd of Models](#)

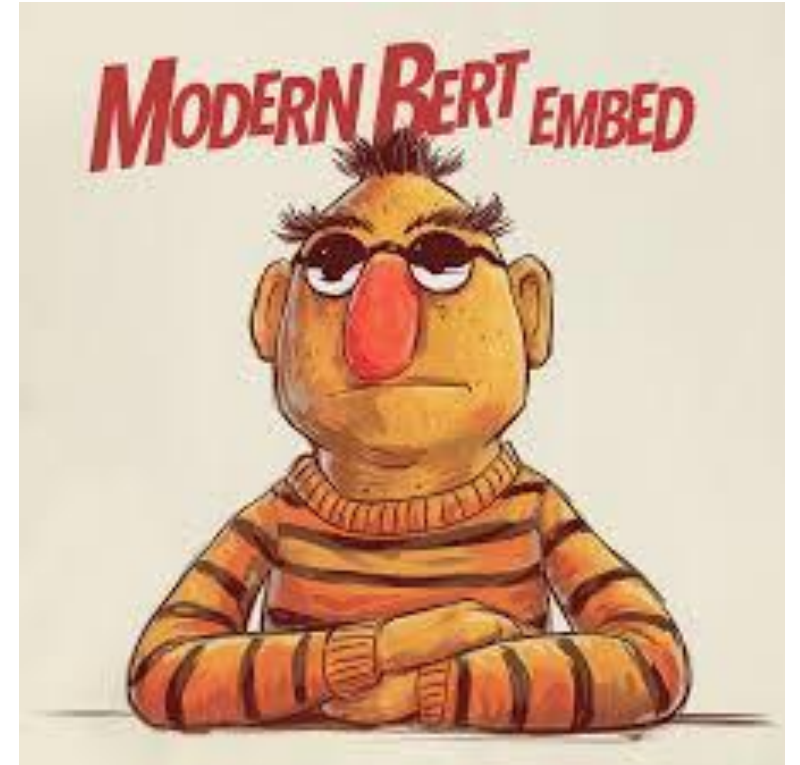
[The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation](#)

Русскоязычные модели

- [GigaChat](#): флагманская закрытая экосистема Сбера
- [YandexGPT](#): экосистема Яндекса & [YandexGPT-Lite](#)
- [T-Lite & T-Pro](#) от Т-банка
- [Saiga](#)
- [Vikhr](#)

Не декодерами едиными... ModernBert

- Контекст – 8к токенов
- Rotary Positional Embeddings – RoPE
- flash attention и unpadding для оптимизации
- Sliding window attention
- 22 и 28 слоев, для base- и large- версий

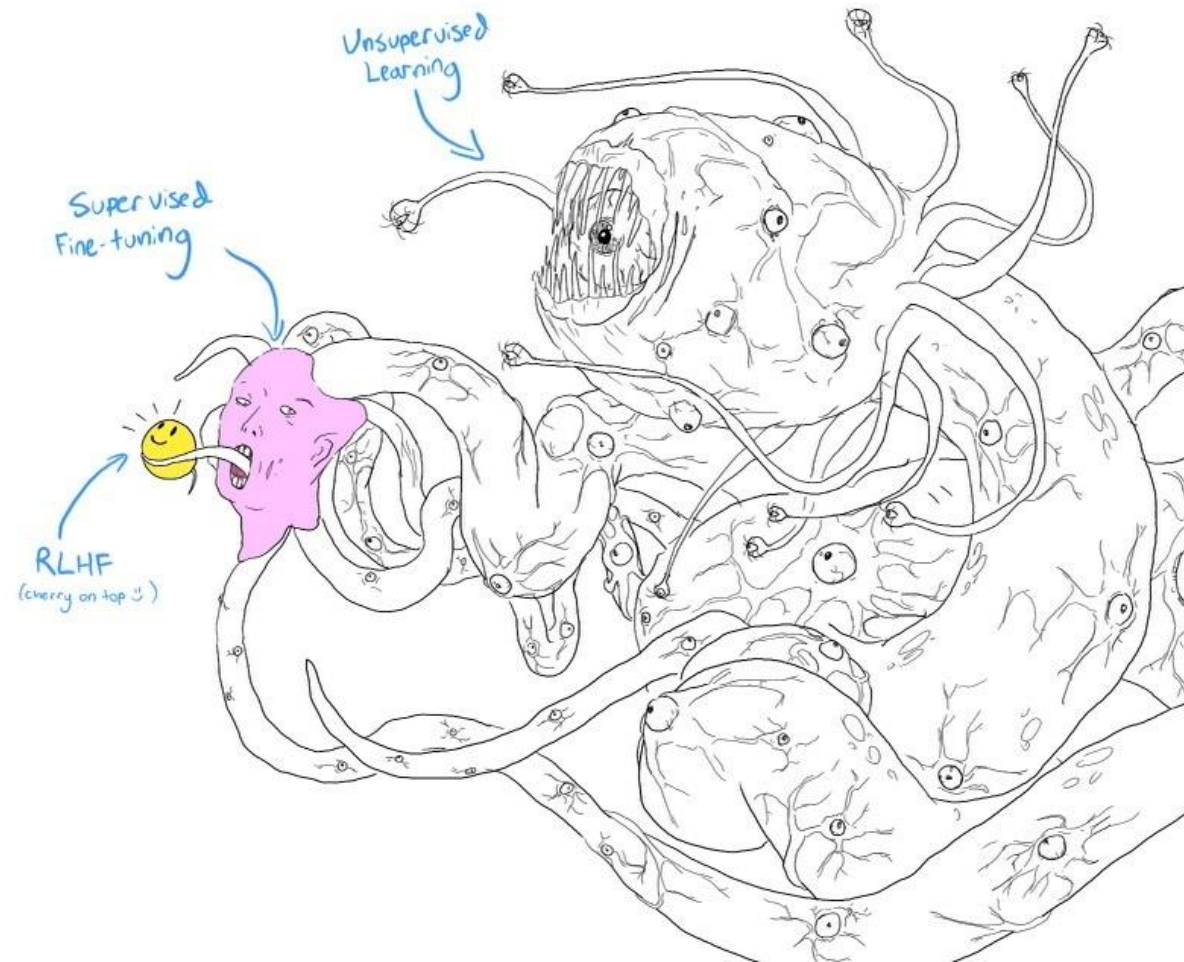


FRIDA



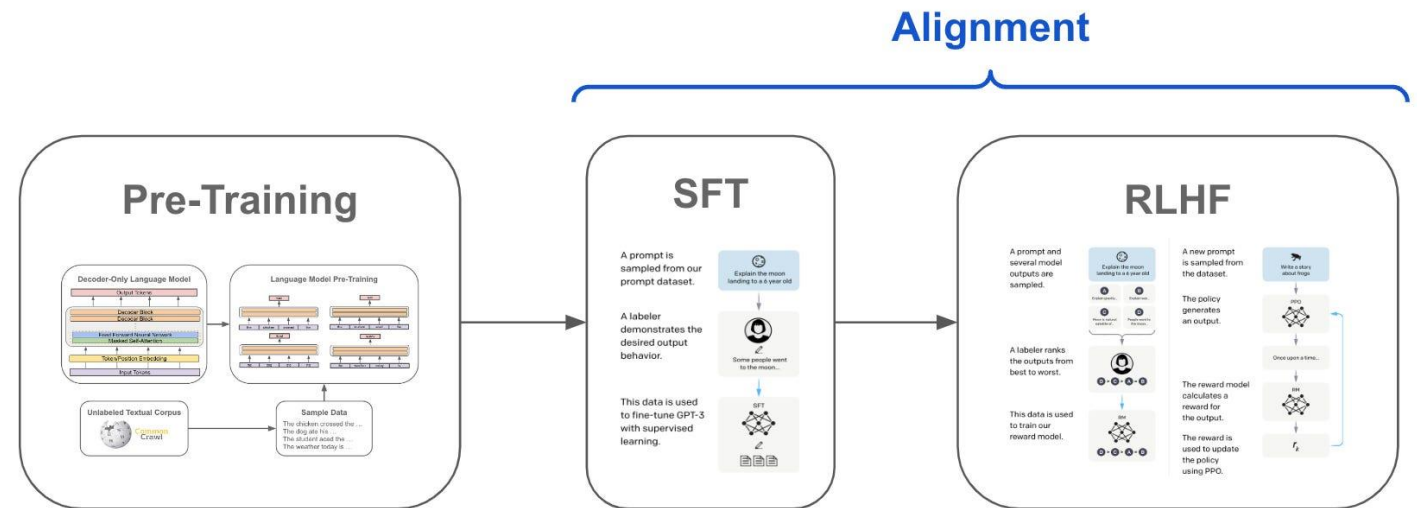
- Энкодерная часть FRED-T5-1.7B
- Контрастивное предобучение и fine-tuning

Как устроены LLM на самом деле?

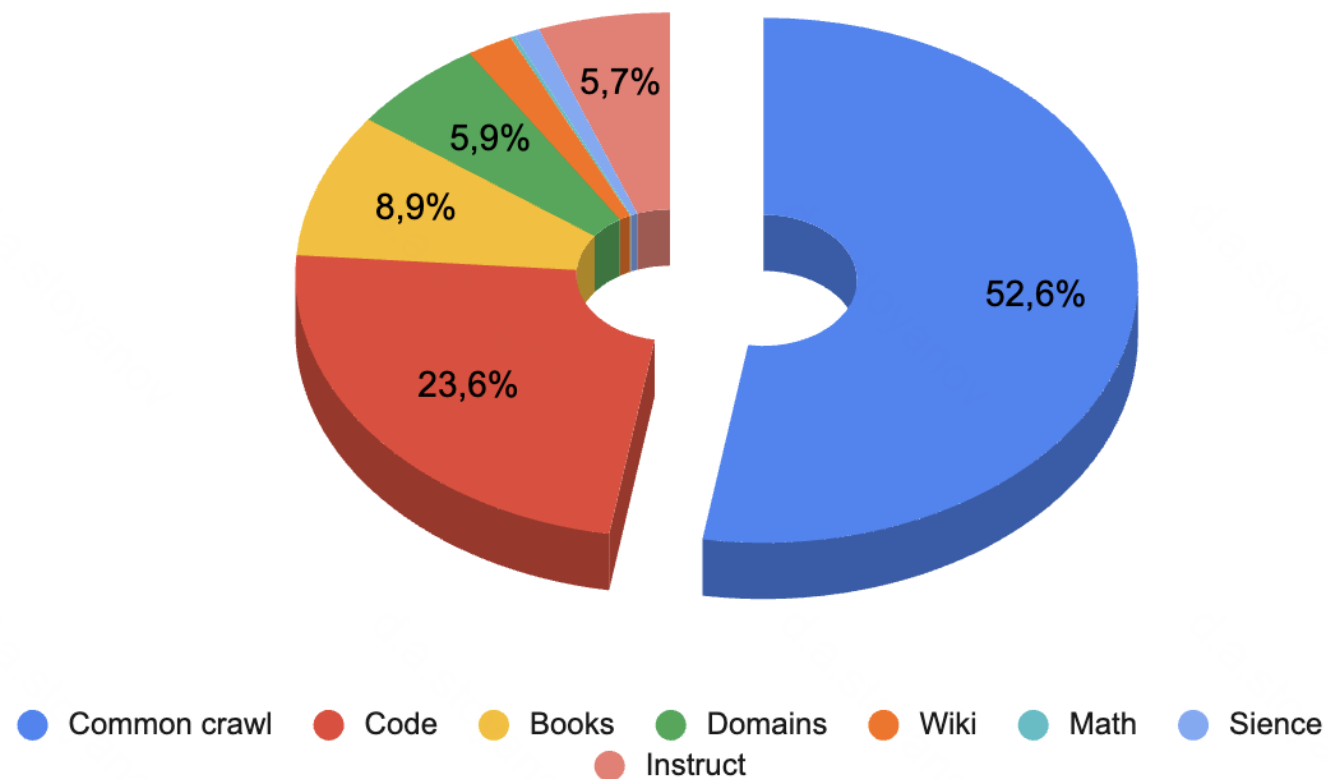


Как обучаются LLM?

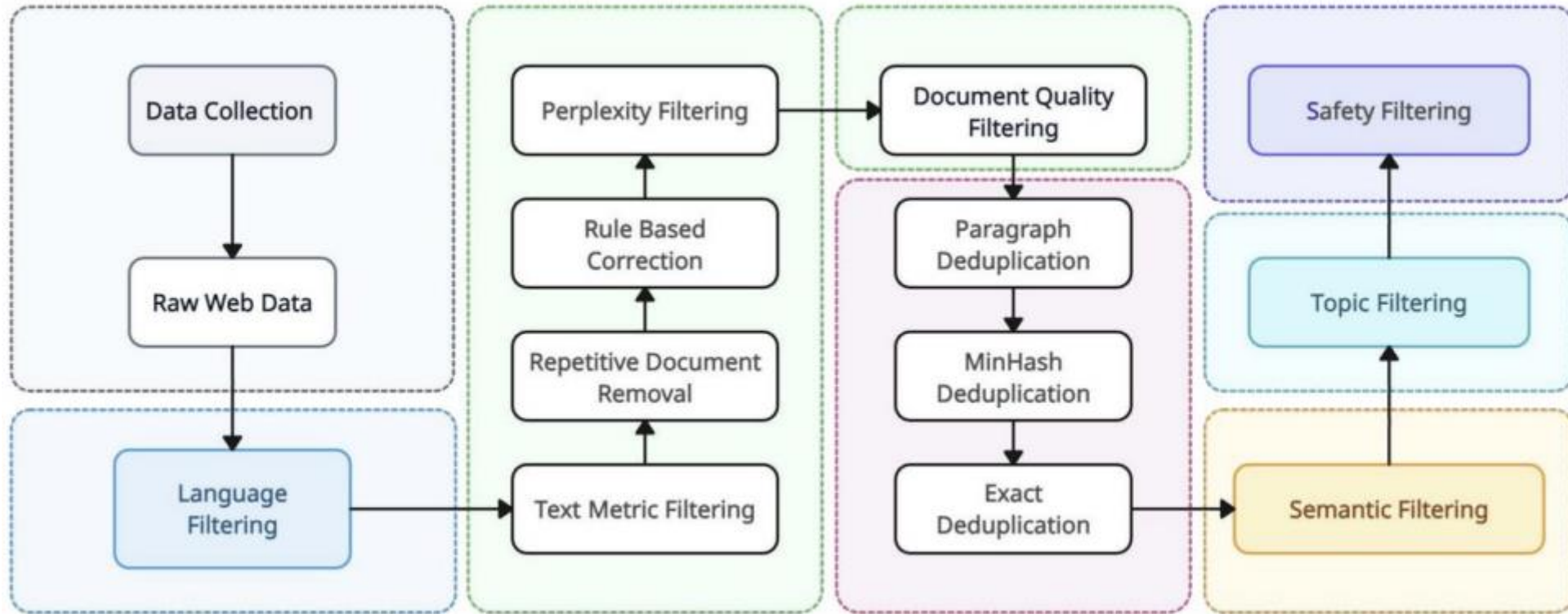
- Pretraining
 - Выучивание «знания о мире»
- Supervised Fine-Tuning
 - Адаптация модели под конкретные задачи
 - Instruction-tuning – обучение способности следовать инструкциям
- Preference Tuning
 - Следование человеческим предпочтениям



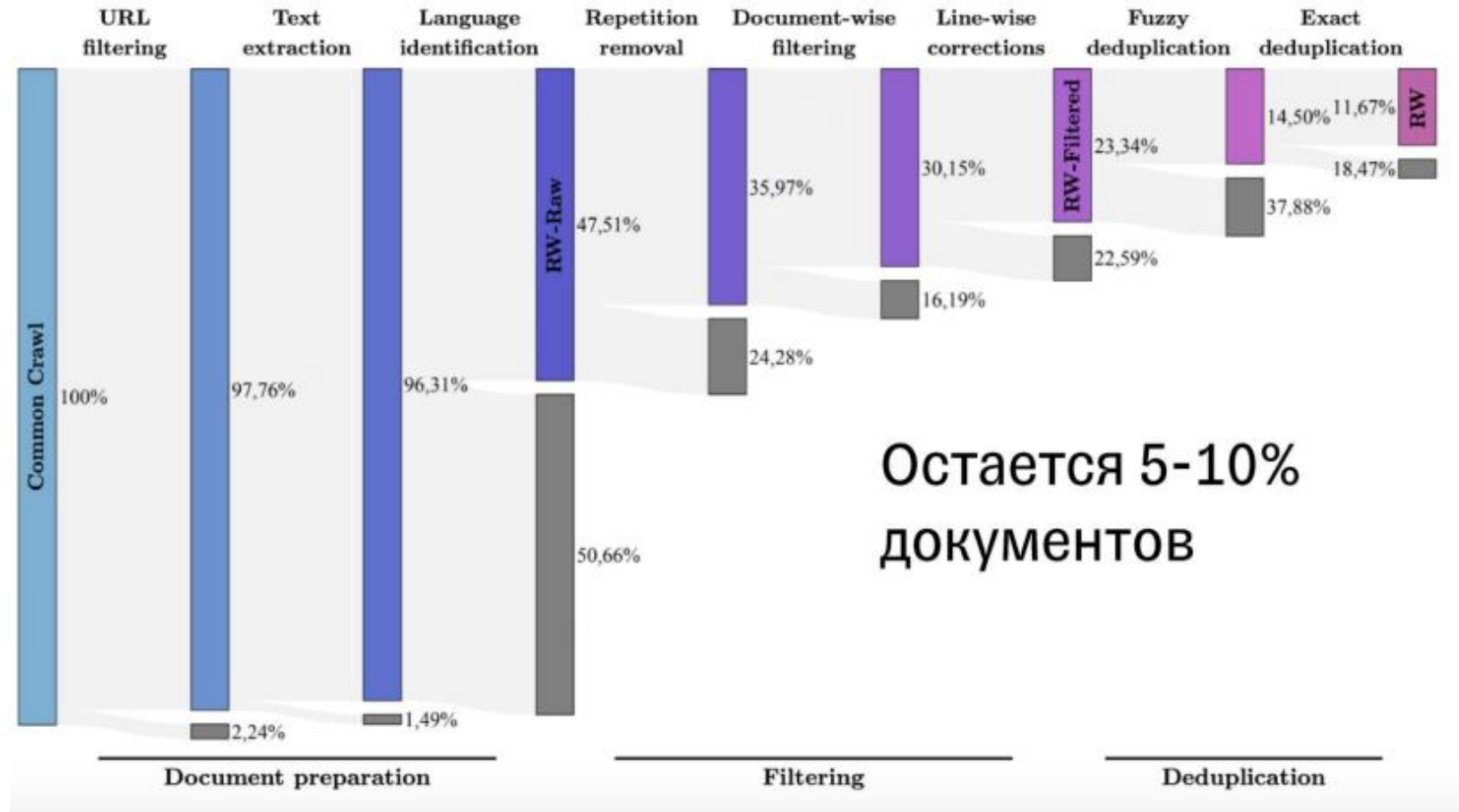
Pretraining - данные



Pretraining - предобработка



Pretraining - предобработка



Pretraining – некоторые техники

- Настройка пропорций языков в корпусе; перевзвешивание недопредставленных языков
- Data Mixture: комбинирование различных типов данных (web, книги, код, диалоги, SFT). Часто используют взвешивание по этапам обучения.
- Curriculum learning:
 - Начальный этап — объёмные, но простые тексты.
 - Переход к узкоспециальным и сложным (код, математика).
 - Финальная фаза — данные SFT и preference data

Бенчмарки

- <https://lmarena.ru/>
- <https://mera.a-ai.ru/ru/text>
- <https://lmarena.ai/>
- <https://huggingface.co/spaces/mteb/leaderboard>
- <https://habr.com/ru/companies/sberdevices/articles/831150/> - ruMTEB
- <https://huggingface.co/spaces/Samoed/Encodechka>

Бенчмарки – статические VS динамические

Динамические - Arena-стиль

- Турниры моделей, где две версии «сражаются» в ответах на одни и те же запросы, а победителя выбирает либо автоматический скорер, либо люди.

Статические – например, MT-Bench (Multi-Turn Benchmark)

- 80 многоэтапных вопросов, оценивает поток ведения беседы и способность следовать инструкциям, автоматическая оценка на основе согласия с «золотыми» ответами.

-

Бенчмарки - MERA

Все сеты можно разбить условно на три класса:

- **Проблемные:** это тип задач, который описывает некоторую проблему, по которой есть однозначное решение; для решения задачи (чаще всего это некоторая классификационная задача) нужны знания о мире, логика, причинно-следственные связи, — все то, что для человека является некоторым набором базовых функций понимания языка и проявления способностей к качественному решению проблем.
- **Экзаменационные:** тип задач, для которых нужны специальные знания и подготовка, некоторая экспертность. Например, сдать ЕГЭ (USE датасет) или решить задачу на код (датасет HumanEval) сможет не любой человек. Это набор специальных знаний, поэтому для таких задач мы предоставляем некоторый условный HumanBenchmark, т. к. очевидно, что оценки экспертов в данных областях и оценки обычных людей будут различаться.
- **Этические:** диагностический сет, предназначенный для выявления байесов и стереотипов моделей. Т. к. тема может показаться субъективной, и нет устоявшихся на этот счет формальных правил относительно того, как мерить этику ИИ, 4 сета являются экспериментальными и не входят в общую оценку на лидерборде.

[MERA — инструктивный бенчмарк для оценки фундаментальных моделей](#)

Бенчмарки - ruMTEB

ruMTEB — это русскоязычное расширение англоязычного набора для оценки текстовых эмбеддингов MTEB (Massive Text Embedding Benchmark). Основные особенности ruMTEB:

- Основан на оригинальном MTEB: сохраняет ту же структуру
- Содержит 23 задачи, из которых 6 мультязычных наборов MTEB с русскими подмножествами и 17 полностью русскоязычных датасетов, созданных и проверенных научным сообществом.
- Задачи группируются по семи категориям:
 - Классификация (9 датасетов)
 - Кластеризация (3 датасета)
 - Мульти-label классификация (2 задачи)
 - Парная классификация (1 задача)
 - Reranking (2 задачи)
 - Retrieval (3 датасета)
 - Semantic Textual Similarity (3 датасета)

[mteb/leaderboard](https://mteb.github.io/leaderboard)

Бенчмарки - проблемы

- **Не существует стандартных метрик** оценки LLM для **систем LLM**, поскольку метрики зависят от точной архитектуры системы LLM
- **Не существует стандартного набора тестовых кейсов/набора данных** для оценки для конкретного варианта использования
- При переносе из синтетических бенчмарков на реальные пользовательские запросы **модели часто теряют в качестве**, особенно в малоформализованных сценариях
- Contamination - частичное искажение бенчмарк-наборов вследствие включения их фрагментов в тренировочные датасеты моделей

Как выбрать LLM под задачу?



Как выбрать LLM под задачу?

- Определение требований и ограничений
 - Тип задачи, домен, ожидания по качеству
 - Язык (моно/мульти)
 - Объем/количество запросов, скорость реагирования, допустимый latency
 - Чувствительность данных
 - Ограничения по ресурсам – CPU/GPU, облако / on-prem, бюджет, возможность адаптации
- Сравнение по бенчмаркам – общим или доменным (если есть)
- PoC, подготовка небольшого eval-сета
- MVP, сбор метрик

Немного о лицензиях

Apache License 2.0 (Apache-2.0)

- Позволяет коммерческое использование, модификацию, распространение и создание производных продуктов. Требует сохранения уведомлений об авторских правах и включения текста самой лицензии.

MIT License

- Свободное коммерческое использование, модификация, распространение и сублицензирование.

Llama License (Meta Llama 3 Community License)

- Специфическая лицензия для модели Llama 3 от Meta.
- Разрешено изменять и распространять с соблюдением лицензии, включая обязательную атрибуцию ("Built with Meta Llama 3").
- Ограничения: Для крупных сервисов (>700 млн пользователей в месяц) требуется дополнительная лицензия; запрещено использование в целях конкуренции; ограничено использование торговых марок.
- Коммерческое использование: Разрешено с учетом ограничений.

Чек-лист для лицензий

Коммерческое использование	Разрешено ли использовать лицензию для коммерческих продуктов или услуг?
Модификации	Можно ли создавать модифицированные версии и распространять их? Каким образом?
Перепродажа	Можно ли перепродавать продукт с лицензией? Есть ли ограничения?
Атрибуция	Обязательно ли указывать авторство, ссылаться на оригинальную лицензию и соблюдать другие нотификации?
Веса модели	Есть ли ограничения на использование, распространение или коммерческое применение обученных весов?