ChatGPT

2.1 Модели: Типы, Размеры и Производители

Классификация моделей и производители

Современный ландшафт LLM разделяется на две условные категории:

- **Большие языковые модели (LLM)** — это модели с десятками или сотнями миллиардов параметров. Они обеспечивают высокий уровень понимания и генерации, но требуют значительных ресурсов. Примером открытой коллекции является **Llama 3.1** от Meta, представленная в версиях **8B, 70B и 405B** параметров. Модель обучена на ~15 трлн токенов и поддерживает 128-тысячный контекст; она оптимизирована для многоязычных диалоговых задач и использует методики SFT+RLHF для выравнивания с человеческими предпочтениями [1]. Версия на 405B параметров требует серьёзной инфраструктуры (более 39 млн GPU-часов на H100), поэтому такие модели чаще применяются через облачные сервисы, а не on-prem.

- **Компактные модели (SLM / small LLM)** — весом до 10 млрд параметров. Благодаря архитектурным оптимизациям и обучению на качественных данных они демонстрируют впечатляющую производительность при скромных требованиях к «железу», что делает их подходящими для on-prem-развёртывания. Примером является **Phi-3.5 Mini**, опубликованная Microsoft: модель имеет **3,8 млрд параметров**, это плотный декодер-only Transformer с контекстом 128 тыс. токенов [2]. Он был обучен на 3,4 трлн токенов, включая тщательно отобранные публичные документы, синтетические «учебные» данные и код; поддерживает более 20 языков (в том числе русский) и был выпущен в августе 2024 года [3]. В модельной карточке отмечается, что с **3,8 млрд активных параметров** Phi-3.5 Mini на многоязычных задачах сопоставим или превосходит модели с гораздо большим количеством параметров [4].

Открытые семейства моделей

Llama 3.1 от Meta

- **Размеры**: 8B / 70B / 405B параметров; все версии используют оптимизированную трансформер-архитектуру с Grouped-Query Attention, что улучшает масштабируемость инференса [1].
- **Языки**: поддержка английского, немецкого, французского, итальянского, португальского, хинди, испанского и тайского; обучена на публичных данных с отсечкой по декабрю 2023 г. и предназначена для коммерческого и исследовательского применения [1].
- **Лицензия**: Llama 3.1 Community License — условно-коммерческая лицензия, разрешающая использование и дообучение моделей при соблюдении условий Meta [5].
- **Применение**: инструкция-тюн версии предназначены для чат-асистентов, в то время как базовые модели могут быть адаптированы под различные задачи генерации текста и кода [6].

Phi-3.5 Mini от Microsoft

- **Размер**: 3,8 млрд параметров, плотная трансформер-архитектура, 128 k контекст [2].

- **Обучение**: 3,4 трлн токенов, комбинация публичных текстов, синтетических «учебников», кода и высококачественных чат-данных; тренирована на 512 GPU H100-80G в течение 10 дней [7].
- **Поддержка языков**: более двадцати языков (арабский, китайский, голландский, английский, французский, немецкий, итальянский, русский, испанский, украинский и др.) [8].
- **Преимущества**: разработчики отмечают, что Phi-3.5 Mini демонстрирует конкурентоспособные результаты на многоязычных MMLU и других бенчмарках, несмотря на небольшое число параметров [4].

Qwen 2.5 Coder от Alibaba/Qwen (apXiv 2409.12186)

- **Фокус**: ориентирована на кодовые задачи (code completion, генерация, self-repair). Для открытого сообщества представлены две версии: **1,5 млрд** и **7 млрд** параметров [9].
- **Архитектура**: обе версии имеют **28 слоёв** и одинаковый размер головы (128), но отличаются скрытыми размерами: 1,5B модель имеет скрытый размер 1 536, а 7B — 3 584; число query-голов и key-value-голов увеличивается с размером модели (12/2 для 1,5B против 28/4 для 7B) [9].
- **Обучение**: обе модели обучались на **5,5 трлн токенов** и имеют словарь 151 646 токенов [9]. Специальные токены добавлены для лучшего понимания кода, а embedding tying используется только в маленькой версии.
- **Открытость**: модели открыты для исследований; разработчики заявляют о поддержке сообщества в области code LLM, агентов и ассистентов.

Инфраструктура для оценки и сравнения

Понимание различий между моделями требует объективных тестов. В рамках карты мира LLM важно знать о двух независимых бенчмарках:

- **LiveBench** — это тестовый набор для LLM, который регулярно публикует новые задачи, чтобы избежать загрязнения тренировочных данных. Он включает 18 различных задач в шести категориях и предоставляет верифицируемые ответы, что позволяет автоматически оценивать модели без использования «LLM-судьи» [10]. Задачи основаны на свежих датасетах, научных статьях и новостях; команда LiveBench предлагает ежемесячные обновления и призывает разработчиков тестировать свои модели на этой площадке [11].

- **LiveCodeBench** — специализируется на оценке кодовых возможностей LLM. Он ежемесячно собирает новые задачи из соревнований **LeetCode**, **AtCoder** и **CodeForces** и включает более 400 задач (май 2023 — март 2024). Бенчмарк оценивает не только генерацию кода, но и **самовосстановление, исполнение и предсказание результатов тестов**, что делает его одним из наиболее полноценных тестов для кодовых моделей [12]. Эта платформа позволяет выявлять сильные и слабые стороны моделей при решении реальных программных задач.

Выводы и советы для on-prem

1. **Выбор размера и производителя зависит от задач**. В корпоративных условиях на-prem разумно начинать с компактных моделей (1–10 B), таких как **Phi-3.5 Mini** или **Qwen 2.5 Coder**, которые могут запускаться на одной GPU и обеспечивать достойную точность. Они особенно полезны для прототипов, внутренних чат-ботов и локальных сервисов.

2. **Большие модели (70 B+, 400 B+) обеспечивают более широкие знания и лучшее качество генерации**, но требуют облачного развертывания. Они подходят для сложных задач, где критична точность и безопасность (например, RAG-решения для корпоративных данных), и часто доступны через API.

3. **Оценочные бенчмарки важны для объективного выбора**. Используйте LiveBench и LiveCodeBench, чтобы сравнивать модели по актуальным задачам и избегать «загрязнённых» тестов [11] [12]. Результаты помогут выбрать модель, которая лучше всего подходит для ваших задач и ресурсов.

4. **Открытые модели развиваются быстро**. Meta, Microsoft и Alibaba выпускают регулярные обновления; по возможности отслеживайте новые версии (Llama 3.x, Phi-3.x, Qwen2.x) — они улучшают качество и увеличивают контекст, оставаясь относительно доступными.

---

[1] [5] [6] meta-llama/Llama-3.1-405B · Hugging Face
https://huggingface.co/meta-llama/Llama-3.1-405B

[2] [3] [4] [7] [8] microsoft/Phi-3.5-mini-instruct · Hugging Face
https://huggingface.co/microsoft/Phi-3.5-mini-instruct

[9] Qwen2.5-Coder Technical Report
https://arxiv.org/html/2409.12186v1

[10] [11] raw.githubusercontent.com
https://raw.githubusercontent.com/LiveBench/LiveBench/main/README.md

[12] raw.githubusercontent.com
https://raw.githubusercontent.com/LiveCodeBench/LiveCodeBench/main/README.md