ChatGPT

# Российские проекты и инициативы по удешевлению LLM

В России уже реализуются несколько проектов и исследований, направленных на снижение затрат при разработке и обучении больших языковых моделей (LLM). Так, лаборатория Yandex Research вместе с учёными ВШЭ и зарубежных вузов представила новые методы сжатия LLM без потери качества. Например, предложен алгоритм HIGGS (Hadamard Incoherence with Gaussian MSE-optimal GridS), который позволяет квантизировать модель без дополнительных данных и ресурсоёмкой оптимизации, сохраняя высокое качество при существенном уменьшении размера [1]. Этот метод доказал своё преимущество в сравнении с существующими безданными методами квантизации (NF4, HQQ) и уже опубликован на GitHub и Hugging Face. Ранее команда Yandex также сообщала о методах сжатия LLM, позволяющих сократить вычислительные затраты почти в 8 раз без заметной потери качества, и даже создала сервис для запуска 8-миллиардной модели на обычном ПК или смартфоне через браузер [2].

Кроме исследований Yandex, российские компании выпускают собственные LLM и делают их общедоступными. Группа «Т-Технологии» (принадлежит Тинькофф) открыла бесплатный доступ к двум большим моделям: T-Pro (32 B) и T-Lite (7 B) [3]. Эти модели специально дообучены на русскоязычных данных («дальнейшее предобучение», **continual pretraining**), чтобы решать узконаправленные задачи и превосходить оригинальные аналоги на русском языке [4]. По результатам бенчмарка MERA, T-Lite и T-Pro показывают лучшие результаты среди открытых русскоязычных моделей в мире [5]. Публикация таких моделей снижает барьер входа для отечественных разработчиков и позволяет компаниям бесплатно использовать LLM в бизнес-задачах.

Также следует отметить проекты MTS (центр MTS AI) и Сбербанка. MTS развивает семейство моделей Cotype (в том числе Cotype Pro 2) для русского языка, оптимизируя их для корпоративных задач и предлагая сервисы по дообучению. Сбербанк через дочерний «SberAI» представил семейство GigaChat – генеративные модели для русскоязычных приложений. Несмотря на то что детали оптимизаций в публичных сообщениях не раскрываются, очевидно, что отечественные LLM-разработки делают акцент на эффективности за счёт адаптации к русскому языку и конкретным сценариям использования.

# Альтернативные подходы к дешёвому обучению ИИ

Кроме простого наращивания масштаба, выделяются несколько методов, значительно снижающих затраты на обучение и развёртывание ИИ при сохранении высокой эффективности:

- **Небольшие/дистиллированные модели с высокой производительностью.** Так называемые «легковесные» *или distilled* LLM представляют собой упрощённые версии больших моделей. При дистилляции «учительской» LLM обучают компактную «студенческую» модель так, чтобы она воспроизводила предсказания большого учителя с меньшими ресурсами. Как отмечает МТС, ведущие игроки (OpenAI, Microsoft, Meta) широко

используют дистилляцию для разработки мощных моделей с меньшими затратами ресурсов [6]. В России эксперты также используют подобные подходы. Например, директор по AI компании «Ланит-Терком» Дмитрий Медведев указывает, что упрощённые LLM активно применяются в мобильных и IoT-системах, а в Сбербанке отмечают их использование для анализа отзывов, суммаризации текстов, классификации и др. задач [7]. Крошечные модели позволяют запускать ИИ даже на обычных ПК и смартфонах, что резко сокращает требуемую инфраструктуру.

- **Алгоритмические оптимизации (квантизация, разреженность, дистилляция).** Квантование весов и активаций LLM — сокращение точности чисел — может уменьшить размер модели в 4–8 раз с незначительной потерей качества. Российские учёные (Yandex и партнёры) исследуют новые методы квантизации LLM без использования обучающих данных (HIGGS) [1]. Ранее Yandex продемонстрировал техники сжатия, позволяющие запускать 8B-модель на слабых устройствах [2] [8]. Удаление несущественных нейронов (*pruning*) и слоёв также упрощает модели. Методы *sparsity* позволяют засекать «пустые» места в весах нейросети и сокращать количество операций. Кроме того, существуют современные инструменты, такие как LoRA (Low-Rank Adaptation), которые добавляют небольшие дополнительные матрицы к слоям Transformer и позволяют настраивать модель, не дообучая её целиком, что значительно снижает вычислительную нагрузку.

- **Обучение на ограниченных или специализированных данных.** Ещё один путь — не учить универсальную LLM на петабайтах текстов, а адаптировать уже готовую модель к конкретной предметной области. Метод «дальнейшего предобучения» (**Continual Pretraining**) активно использует группа Т-Технологии: они берут базовую модель (например, Qwen-2.5) и продолжают её обучение на отборных русскоязычных корпусах, релевантных задачам заказчиков [4]. Такой подход требует гораздо меньше вычислений, чем обучение «с нуля», и обеспечивает высокое качество в узкой области. В целом, качественная выборка данных позволяет небольшим моделям обходить по эффективности крупных собратьев: так, модель StableLM-2 (1.6B параметров) превзошла по переводу Falcon-40B, использовавшего в 25 раз больше параметров [9]. Кроме того, узкая доменная предобученность улучшает точность на специфических задачах без ресурсоёмкого дообучения на гигантских неорганизованных датасетах.

# Инфраструктурные и правовые факторы в России

Несколько особенностей российской среды влияют на стоимость разработки и внедрения LLM:

- **Ограниченный доступ к аппаратному обеспечению.** Из-за санкций и ухода ряда крупных производителей с российского рынка доступные GPU (NVIDIA A100, H100 и др.) дефицитны. По оценке MTS AI, дефицит и дороговизна вычислительных мощностей — одна из главных проблем: оборудование приходится ввозить по схеме параллельного импорта, что увеличивает стоимость и риски задержек [10]. В таких условиях многие компании предпочитают развёртывать LLM *on-premise* – на собственных серверах заказчика, чтобы избежать экспорта чувствительных данных [11]. Это снижает риски утечек, но требует больших капитальных затрат на создание локальной инфраструктуры.

- **Затраты на электроэнергию и центры обработки.** С другой стороны, в России относительно дешёвая электроэнергия может смягчать энергозатраты. Однако крупных

современных дата-центров и специализированных кластеров AI (с чипами AMD/Intel/IBM) не так много, а их развёртывание также дорого. Некоторыми решениями могут быть совместные инфраструктурные проекты (например, европейский стартап Nebius, основанный российскими инженерами, планирует дата-центры для ИИ), а также аренда мощностей у крупных игроков (Ростелеком, Яндекс.Облако, МТС).

- **Правовые ограничения и регуляция.** В России действует закон о персональных данных, требующий хранения данных россиян на российских серверах, что ограничивает использование зарубежных облаков при обучении моделей на локальных текстах. Недавно введён «экспериментальный правовой режим» в Москве, облегчающий работу с данными при разработке ИИ (см. закон №273-ФЗ), но его эффект пока локализован и не даёт существенных льгот на федеральном уровне. Кроме того, наблюдается нехватка квалифицированных специалистов в области ML, что удорожает проекты и делает аутсорсинг обучения затруднительным [12].

**Рекомендации для России.** Чтобы снизить барьеры, целесообразно усилить поддержку отечественных решений: стимулировать разработку российских ускорителей ИИ (например, отечественных GPU/ASIC), расширять доступ компаний к существующим ЦОД, а также поддерживать открытость и совместимость сред обучения (оптимизировать параллельный импорт). Необходимо продолжать финансирование научных исследований по оптимизациям LLM (как делает Yandex) и создавать отраслевые облачные сервисы с GPU. Законодательство могло бы поощрять обмен открытыми данными и алгоритмами (по аналогии с зарубежными open-source инициативами) и расширять рамки «песочницы» для ИИ, чтобы учёным было проще собирать и использовать датасеты без сложных согласований.

# Сравнение с зарубежными практиками

На мировом рынке LLM преобладают крупные игроки с обширными ресурсами. Например, OpenAI и Meta разрабатывают гигантские модели (десятки–сотни миллиардов параметров) с использованием глобально распределённых кластеров. Они активно инвестируют в оптимизации (квантование, дистилляцию), но масштабы их проектов предполагают многомиллионные бюджеты. По оценкам, обучение одной большой LLM может стоить от $9 до $23 миллионов [13], и подобные модели потребляют тысячи МВт·ч энергии. В отличие от них, европейская Mistral AI демонстрирует иной подход: её открытая модель Mistral-7B (7 млрд параметров) сопоставима по качеству с куда более крупными аналогами, что доказывает эффективность компактных архитектур и оптимизаций.

**Уникальные возможности и ограничения российского подхода:** российские разработчики вынуждены фокусироваться на эффективных решениях из-за ограниченных ресурсов. Это даёт некоторые плюсы: ориентация на конкретный язык и домен позволяет создать очень качественные локализованные модели (например, T-Pro, GigaChat). С другой стороны, недостаток крупных дата-центров и доступной техники означает, что у России меньше возможностей для «гонки масштабов», как в США или Китае. Однако российские инициативы (открытые модели, квантизационные исследования, «продолженное обучение» на локальных данных) создают свои конкурентные преимущества — они делают LLM более доступными для бизнеса и госструктур. В конечном счёте, зарубежные практики демонстрируют ценность больших инвестиций и масштабируемых платформ, тогда как в России акцент смещён на алгоритмическую эффективность и интеграцию моделей в существующую инфраструктуру с учётом местных ограничений [13] [9].

# Выводы и практические рекомендации

Для сокращения затрат на LLM в российском контексте важно сочетать несколько стратегий:

- **Продолжать исследования в оптимизации моделей:** дальнейшее развитие методов квантизации, дистилляции и разрежения поможет запускать модели на менее мощном оборудовании (как уже делают Yandex и партнёры [1] [2]).
- **Развивать маломасштабные и специализированные модели:** создавать компактные LLM, адаптированные к российскому языку и задачам (подобно T-Lite/T-Pro), и применять дообучение на релевантных данных [4]. Это позволит обеспечить высокую эффективность без затрат, присущих глобальным моделям.
- **Улучшать инфраструктуру:** расширять доступ к GPU и TPU, используя как зарубежные (через параллельный импорт), так и отечественные решения, а также развёртывать ИИ-кластеры внутри страны. Целесообразно создавать центры компетенций и стимулировать коллаборации между ИТ-компаниями и вузами для обмена мощностями и данными.
- **Сотрудничать с зарубежными сообществами:** хотя прямая покупка технологий сложна, России выгодно участвовать в международных исследовательских инициативах (например, обмениваться опытными моделями и алгоритмами), что позволит перенимать передовые практики оптимизации.
- **Гибкое регулирование:** нужно расширять эксперименты с правовым полем для ИИ (касаемо данных и тестирования моделей) по всей стране, чтобы снять бюрократические преграды на пути к обучению отечественных LLM.

В совокупности эти меры помогут России снизить затраты на создание и внедрение LLM, сделав технологии ИИ более доступными для бизнеса, науки и госуправления.

**Источники:** аналитика CNews и НИУ ВШЭ [8] [1], пресс-релизы и СМИ (RIA, RBC) [14] [10], корпоративные блоги и обзоры (Sber, MTS) [7] [6]. These show both domestic initiatives and global trends in efficient LLM development.

---


[1] [2] Большие языковые модели теперь не требуют мощных серверов — Национальный исследовательский университет «Высшая школа экономики»
https://www.hse.ru/news/development/1034477704.html

[3] [4] [5] [14] Российские разработчики представили новые большие языковые модели - РИА Новости, 11.12.2024
https://ria.ru/20241211/tekhnologii-1988555726.html

[6] Как дистилляция меняет индустрию искусственного интеллекта / Хабр
https://habr.com/ru/companies/ru_mts/articles/894168/

[7] Обзор небольших больших языковых генеративных моделей: GPT и русские версии
https://sber.pro/digital/publication/nebolshie-bolshie-yazikovie-modeli-kakie-prikladnie-zadachi-oni-reshayut/

[8] MLOps - LLM - Large Language Model - большая языковая модель - Prompt engineering - Промпт-инжиниринг - CNe
https://www.cnews.ru/book/MLOps_-_LLM_-_Large_Language_Model_-
_%D0%B1%D0%BE%D0%BB%D1%8C%D1%88%D0%B0%D1%8F_%D1%8F%D0%B7%D1%8B%D0%BA%D0%BE%D0%B2%D0%B0%D1%8F_%D0%BC%
_Prompt_engineering_-_%D0%9F%D1%80%D0%BE%D0%BC%D0%BF%D1%82-%D0%B8%D0%BD%D0%B6%D0%B8%D0%BD%D0%B8%D1%80%D0%

[9] [13] Как устроены малые языковые модели: эффект масштаба | РБК Тренды
https://trends.rbc.ru/trends/industry/668510909a7947e9acd8ffc3

10 11 12 Объем рынка больших языковых моделей в России оценили в 35 млрд руб. — РБК
https://www.rbc.ru/technology_and_media/26/11/2024/67449d909a79478a2052d490