ITMO
OPEN SOURCE

Экосистема открытого кода ИТМО · Аналитика 2026

Исследование российского рынка LLM и данных для их обучения

Николай Никитин — к.т.н., лидер сообщества ITMO Open Source.

Андрей Гетманов — м.н.с Института ИИ ИТМО.

Илларион Иов — ассистент ФТИИ ИТМО.

Егор Фёдоров — студент ФТИИ ИТМО, сотрудник учебно-научной лаборатории АИСТ.

01 Проблематизация

Несмотря на то, что российские разработчики LLM находятся в условиях ограниченного объема доступных ресурсов (разрыв в вычислительных мощностях между США и РФ оценивается приблизительно в 660 раз[1], российский рынок LLM развивается, появляются новые игроки, а технологии доходят до практического применения в широком спектре областей: в медицине, логистике, промышленности и других сферах. Дальнейшее улучшение качества и конкурентоспособности наших моделей зависит не только от вычислительных мощностей, но и от того, какие данные используются для обучения — насколько они очищены, сбалансированы, размечены и доступны разработчикам. В рамках данного исследования предлагается рассмотреть текущее состояние российского рынка LLM (уровень качества российских решений и их место в конкурентной среде), выявить необходимые для обучения российских LLM данные, а также определить дальнейшие действия ключевых акторов, необходимые для развития отрасли.

Эти вопросы становятся особенно актуальны в контексте Федерального закона от 26.07.2026 № 243-ФЗ «О поддержке развития технологий искусственного интеллекта в Российской Федерации» (основная часть закона вступила в силу 1 сентября 2026 года), где впервые на законодательном уровне определены правовые статусы и механизмы взаимодействия внутри российской отрасли искусственного интеллекта.

Рынок ИИ в цифрах 2025 — 2030

257 млрд ₽

Затраты российского бизнеса на внедрение ИИ в 2025 году

Д. Григоренко / Правительство РФ

225 трлн ₽

Мировые расходы на ИИ в 2026 году (≈ $2,59 трлн по курсу ЦБ)

Gartner, май 2026

7,9–12,8 трлн ₽

Ожидаемый ежегодный эффект от ИИ в экономике РФ к 2030 году

Яков и Партнёры

до 5,5% ВВП

Доля эффекта от ИИ в прогнозной экономике России к 2030 году

Яков и Партнёры

02 Ключевые выводы исследования

Дальнейший рост качества российских LLM зависит от понятных правил, доступных данных, инфраструктуры и подготовки специалистов

Государство уже сделало первый шаг к развитию отрасли через принятый федеральный закон, который объясняет механизмы работы с общедоступным в интернете контентом для обучения моделей. Дальше отрасли нужны свободные вычислительные мощности, продуктовые внедрения со стороны бизнеса, методологическая база и бенчмарки от науки, а также профильная подготовка LLM-специалистов в образовании.

«Грамотная регуляторика, хорошая, может сильно ускорить прогресс, плохая регуляторика может всё развалить. Во-первых, я вижу роль государства именно в правильной настройке регуляторики, связанной с авторским правом: как его нужно интерпретировать, чтобы не тормозить прогресс в создании датасетов. Роль научных учреждений — это прежде всего разработка методологии и методов, роль коммерческих организаций — прежде всего эффективная продуктивизация этих методов»



Иван Бондаренко
Доцент НГУ, проект «Сибирские нейросети»

Закрытые источники — наш стратегический ресурс

Библиотечные, архивные, образовательные, культурные и профессиональные данные, закрытые на данный момент, могут дать отечественным моделям доступ к более глубокой, верифицированной и локально значимой информации, способной повысить качество работы российских LLM в новых специализированных областях.

«На последних этапах действительно имеет смысл показывать небольшое количество максимально чистых и максимально экспертных данных. Доступ к каким-то качественным источникам, закрытым, таким как библиотеки, архивы, внутрикорпоративные документы, может быть хорошей затравкой для получения таких качественных верифицируемых данных. Это потенциально наше конкурентное преимущество, если какие-то данные доступны нам, но не присутствуют в интернете и недоступны глобальным игрокам»



Анатолий Потапов
Т-Банк - руководитель группы фундаментальных технологий LLM

Самый быстрый путь к росту качества российских моделей — открыть и оцифровать уже существующие массивы данных

В России уже есть значительные объемы текстов, которые могли бы усилить отечественные LLM, но сейчас они остаются недоступными из-за организационных, юридических или технических барьеров. Масштаб таких массивов виден по государственной статистике: к 1980 году советская Государственная система научно-технической информации накопила свыше 2 миллиардов единиц хранения, а в фондах научно-технических библиотек сейчас оцифровывается 483 миллиона единиц [24]. Открытие этих данных и их передача разработчикам через контролируемые механизмы могли бы заметно усилить отечественные модели.

«Самый простой и эффективный способ улучшить обучение российских моделей – предоставить как можно большее количество данных. Понятно, что есть вещи, которые открывать нельзя. Например, содержащие государственную тайну. При этом наверняка есть массивы текстов, которые закрыты, но никто уже не помнит почему. Их открытие помогло бы увеличить количество полезных данных для обучения.»



Александр Крайнов
Директор по развитию технологий искусственного интеллекта Яндекса

Российские модели имеют преимущество в областях, где особенно важно понимание русского языка и «культурного кода»

Российские LLM уже сейчас способны проявлять себя сильнее зарубежных моделей-лидеров в наиболее важных для практического внедрения в нашей стране задачах, связанных с русским языком, локальным контекстом, историей и национально-культурной спецификой. Это демонстрируют эксперименты бенчмарка SLAVA: в начале 2026 года первые шесть позиций его рейтинга заняли отечественные LLM. Ключевыми направлениями внедрения российских решений становятся специализированные профессиональные отрасли — прежде всего медицина, право, строительство, геология и другие сферы, где локальный контекст и специфика особенно важны.

«Они [отечественные модели] понимают сейчас его [российский культурный код] точно лучше, чем аналоги международные. Топ-6 позиций нашего бенчмарка — прогонки нашего эксперимента на 14 тысячах вопросов по истории, обществознанию, социологии и географии — у отечественных моделей»



Сергей Боловцов
Директор Исследовательского центра ИИ Президентской академии

Современные российские модели должны быть многоязычными

В создании российских LLM русскоязычные данные необходимы, но изоляция от международных корпусов приведет к тому, что модель будет сильнее отставать от зарубежных аналогов и хуже справляться с задачами, выходящими за пределы локального языкового контекста. Важно дать российским LLM возможность обучаться на всем общедоступном контенте — в российском интернете и за его пределами, — чтобы модели, с одной стороны, понимали культурный код нашей страны, а с другой — развивали свои когнитивные способности на всем доступном запасе знаний. Данное положение нашло отражение в Федеральном законе от 26.07.2026 № 243-ФЗ «О поддержке развития технологий искусственного интеллекта в Российской Федерации», разрешающем обучение национальных и суверенных больших фундаментальных моделей искусственного интеллекта на всех публичных и легально доступных данных.

«Если мы говорим о базовой универсальной модели, в её претрейне должны быть и иностранные языки, и данные отвечающие культурным особенностям нашей страны, и программный код, и в целом максимально разнообразный контент. Иначе она не сможет работать с целыми классами задач — например, с переводом и знаниями, которые существуют в других языковых средах. Убирая из претрейна целые пласты знаний, вы, по сути, делаете модели лоботомию. Зачем?»



Алёна Феногенова

Исполнительный директор департамента развития ИИ и машинного обучения Сбербанка

Больше качественных данных — выше возможности моделей в конкурентных отраслях

Эксперты связывают рост качества отечественных LLM с расширением доступа к качественным русскоязычным и многоязычным данным. Чем богаче и лучше подготовлен корпус, тем точнее и полезнее становятся модели для практического внедрения, особенно в российских локальных отраслях.

«Команда DeepSeek, когда обучала свою модель, заведомо убрала из датасетов множество дублей и вообще причесала датасеты очень хорошо, оставив то, что нужно для основной PR-функции — паритетного участия в разных соревнованиях и сравнениях по метрикам с другими ведущими моделями. Это подтверждает: важно то, что находится в датасетах, а не то, какой у них объем. Потому что объем данных у DeepSeek был существенно меньшим, чем у американских конкурентов, но позволил продемонстрировать сходное качество»



Александр Бухановский

ИТМО - доктор технических наук, руководитель Института искусственного интеллекта; директор мегафакультета трансляционных информационных технологий

Конкурентоспособность российских моделей в будущем — это коллективный результат

Дальнейшее повышение качества российских LLM требует перехода к системной отраслевой стратегии: развитие фундаментальных исследований, расширение вычислительной инфраструктуры, подготовка и удержание кадров, консолидация усилий между компаниями, государством, университетами и open source сообществом.

«Я бы не называл это данными самими по себе, я бы это называл, может быть, какой-то экосистемой — экосистемой взаимодействия компаний, данных этих компаний, университетов и так далее. Эти данные могли бы собираться из разных министерств, компаний, университетов, школ»



Михаил Тихомиров

К.ф.-м.н., старший научный сотрудник МГУ, автор RuAdapt

Эксперты, принявшие участие в исследовании

Эксперт	Организация и должность
Иван Оселедец	Доктор физико-математических наук, профессор РАН, генеральный директор Института AIRI, декан факультета ИИ МГУ
Александр Крайнов	Директор по развитию технологий искусственного интеллекта Яндекса
Алёна Феногенова	Исполнительный директор департамента развития ИИ и машинного обучения Сбербанка
Иван Басков	Руководитель направления подготовки данных для предобучения GigaChat (Pretrain Data Lead)
Анатолий Потапов	Т-Банк - руководитель группы фундаментальных технологий LLM
Алан Газзаев	AI VK, ML-инженер
Михаил Тихомиров	Кандидат физико-математических наук, научный сотрудник Лаборатории анализа информационных ресурсов НИВЦ МГУ
Александр Бухановский	ИТМО - доктор технических наук, руководитель Института искусственного интеллекта; директор мегафакультета трансляционных информационных технологий
Дмитрий Ильвовский	НИУ ВШЭ - заместитель декана по международной деятельности факультета компьютерных наук; доцент Департамента анализа данных и искусственного интеллекта
Сергей Боловцов	РАНХиГС - директор Исследовательского центра искусственного интеллекта Президентской академии
Иван Бондаренко	«Онтологические системы» - генеральный директор; доцент НГУ
Илья Гусев	Principal Scientist at Keenable.ai
Анонимный эксперт	Независимый разработчик LLM
Анонимный эксперт	Независимый разработчик LLM

03 Как сейчас обстоят дела с российским рынком LLM?

3.1. Что происходит с рынком LLM?

За последние несколько лет развитие больших языковых моделей дало мощный толчок методам искусственного интеллекта. В 2026 году глобальные расходы на ИИ составят, по прогнозу Gartner, около 2,6 трлн долларов [примерно 219 трлн рублей]. Российские компании включились в гонку: по итогам 2025 года 71% крупных компаний применяют генеративный ИИ хотя бы в одной бизнес-функции — от маркетинга до клиентского сервиса. Россияне пользуются нейросетями и в повседневной жизни. Ежемесячный охват ИИ-сервисов составляет 26% населения, ежедневный — 5%. К 2030 году вклад ИИ в ВВП страны, по оценкам экспертов, может составить триллионы рублей.

Положение российского рынка эксперты описывают двойственно. Он остается догоняющим по отношению к мировым лидерам, но внутри него идет активный рост: расширяется круг участников, качество моделей за последние годы заметно выросло. О «догоняющем развитии» говорят почти все, однако вкладывают в эти слова разное. Одни считают отставание структурным — нехватка вычислительных мощностей и небольшие исследовательские команды не дают выйти из этой позиции в обозримые сроки. Другие указывают, что разрыв сокращается: на момент выхода первых версий GPT он был заметно больше. Есть и третья позиция. Вся мировая индустрия находится на ранней стадии развития, и главным ограничением для российских команд оказывается не отставание в технологиях, а узость внутреннего рынка при закрытых внешних.

Опрошенные условно делят мировой рынок на три эшелона. К первому относят закрытые модели: OpenAI, Anthropic и Google. Ко второму — китайские модели с открытыми весами : Qwen, DeepSeek, GLM и Kimi. Китайский open source служит технологической базой для многих открытых разработок, в том числе российских. Третий эшелон составляют российские модели. Здесь эксперты различают обучение с нуля и глубокую адаптацию зарубежных открытых моделей; то же различие закреплено в терминологии принятого федерального закона. Национальной в нем считается модель, обученная с нуля в России, суверенной — модель на зарубежной открытой архитектуре, дообученная на российских данных. Экономика сужает круг игроков. По оценкам экспертов, обучение конкурентной модели с нуля стоит от миллиарда долларов, а на внутреннем рынке такие вложения не окупаются. Поэтому российские компании чаще дорабатывают открытые модели: это дешевле и помогает быстрее вернуть инвестиции. Масштаб внутреннего рынка, по их оценке, позволяет поддерживать один-два проекта мирового уровня, но немного параллельных команд.

«Запрос на ассистентов, на агентные системы огромный. Кто-то покупает пакеты с ГигаЧатом, кто-то получает Яндекс с его стеком, кто-то работает по API, кто-то обучает под свои задачи на своих датасетах открытые модели. Сами LLM в данном случае являются обвязкой вокруг бизнес-задач: качество работы агентов и стоимость напрямую зависят от качества LLM. Можно мерить качество работы по-разному, например, бенчмарками, а можно мерить в месяцах отставания. Наверное, можно так сказать: если закрытые модели типа Opus за ноль взять, то Open Source отстаёт где-то на 2–3 месяца, а наши отстают, наверное, где-то на полгода»



Иван Оселедец
Д.ф.-м.н., профессор РАН, генеральный директор AIRI

3.2. Какие модели есть на российском рынке?

По состоянию на 2026 год в России существует несколько LLM от различных компаний. Некоторые обучены с нуля, некоторые дообучены на основе весов существующих моделей, часть открыта для свободного использования. Среди актуальных моделей, созданных отечественными командами или «заточенных» под русский язык, — разработки Сбера, Яндекса, Т-Банка, МТС, Авито, Ozon, Wildberries, «Норникеля» и VK, а также академических команд и независимых авторов. Центральное место занимают два крупных игрока — Сбер и Яндекс: эксперты именно их чаще всего называют компаниями с достаточными ресурсами, командами и технологической базой для разработки собственных LLM. Рядом с ними рынок расширяется за счет команд, работающих в специализированных нишах: Т-Банк, MWS AI, Авито, Wildberries, Vikhr, Saiga, RuAdapt и другие.

Актуальные российские LLM: бигтех

Компания	Модели	Открытые веса	Примечание
SberDevices	GigaChat 3 Ultra (702B), Lightning (10B), GigaChat 3.5 Reasoning и др.	Да	Семейство универсальных моделей разного размера. Обучены с нуля
Яндекс	Alice AI Foundation 80B LLM	Да	Открытая базовая языковая модель с гибридной архитектурой и MoE-слоями. Обучена с нуля
Яндекс	Alice AI LLM	Нет	Закрытая модель ассистента «Алиса». Обучена с использованием весов Qwen3-235B
Яндекс	Alice AI T5-35B A0.6B	Да	Компактная языковая модель, обученная с нуля, формирует быстрые ответы Алисы AI под поисковой строкой
Т-Банк	T-Lite, T-Pro (2.0)	Да	Универсальные компактные модели на основе Qwen3 с улучшенной русскоязычной токенизацией
MWS AI	Cotype Nano, Cotype VL	Да	Вычислительно эффективные модели для ассистентов; Cotype VL — мультимодальная
Авито	A-vibe	Да	Модель для бизнес-процессов Авито на основе Qwen3-8B и токенизатора RuAdapt
RWB (Wildberries и Russ)	BerryLM-XL, BerryLM-v2	Нет	MoE-модель на 358 млрд параметров с дообучением SFT. Ассистент покупателя, поиск и сравнение товаров, инструменты для продавцов
OZON	OzonGPT	Нет	Специализированная модель для генерации русскоязычных описаний товаров
«Норникель»	MetalGPT-1	Да	Модель на основе Qwen 32B для металлургии и горнодобывающей отрасли
VK	Diona	Нет	Собственная модель во внутренних продуктах экосистемы и в ИИ-платформе VK Tech

Актуальные российские LLM: академия, независимые разработчики, небольшие компании

Команда / автор	Модели	Примечание
НИВЦ МГУ	Ruadapt	Методология адаптации мультиязычных LLM под русский язык: расширение токенизации, Learned Embedding Propagation, калибровка. Ускорение генерации на 30–60%. Открыт на GitHub и Hugging Face
«Сибирские нейросети»	meno-lite / tiny, whisper	Специализированные мини-модели под конкретные задачи: RAG, распознавание речи
HUAWEI Noah's Ark Lab	Gamayun	Универсальная легкая модель с упором на русский язык
Vikhr models	Вихрь	Семейство открытых моделей, адаптированных под русский язык: расширенный токенизатор и продолженное предобучение всех весов
Илья Гусев	Saiga	Универсальная модель, дообученная на русском языке, — одна из первых в своем роде

В таблице упомянуты не все существующие модели, а только наиболее актуальные на 2026 год. Пропущена, например, ruGPT-3.5[38] от Сбера — предшественник современных русскоязычных моделей в этом списке. Открытая модель на 13 млрд параметров была предобучена на 300 ГБ текстов и 100 ГБ кода и юридических документов. Ее использовали при создании GigaChat; сегодня ruGPT-3.5 уступила место более новым решениям. Так как большая часть российских моделей опирается на зарубежные открытые веса (в первую очередь Qwen), актуальным становится вопрос о необходимости сохранения компетенций полного обучения моделей с нуля.

Здесь эксперты расходятся во мнениях. Большинство считает, что сама постановка «либо с нуля, либо дообучение» слишком жесткая: между полным предобучением и легким дообучением есть промежуточные режимы, а серьезное продолженное предобучение требует почти тех же навыков и инфраструктуры, что и старт с нуля. За сохранение компетенции приводят три аргумента. Во-первых, это страховка на случай, если открытые веса перестанут публиковать: восстановление технологии займет годы. Во-вторых, собственный претрейн нужен для исследований новых архитектур и сохранения накопленных данных и инструментов. В-третьих, знания, заложенные на предобучении, трудно изменить простым дообучением.

«С точки зрения создания моделей с нуля, в России немного продуктивных коллективов — всего два. Это Яндекс и Сбербанк. Более успешная история сейчас у Сбербанка по этому направлению: они уже способны обучать модели размера DeepSeek с нуля и выдавать качество некоторого мирового уровня. Второй коллектив, Яндекс, в последнее время отошел, по крайней мере в публичной сфере, от обучения с нуля — сейчас это выглядит как тренд на адаптацию моделей. Но, например, выпущенная ими восьмимиллиардная модель очень-очень неплоха. Для своего размера, понятное дело, со своими ограничениями, но, по сути, тут тоже мировой уровень для данного размера, для русского языка»



Михаил Тихомиров

К.ф.-м.н., старший научный сотрудник МГУ, автор RuAdapt

«На мой взгляд, ключевую роль таких компаний, как Сбер определяет объём вложенных ресурсов. Развитие технологий такого уровня требует значительных инвестиций на всех этапах — от сбора и подготовки данных до обучения и интеграции моделей в продукты, — позволить себе разработку с нуля может далеко не каждая компания. Сбер системно вкладывается именно в такой полный цикл, что выделяет нас среди остальных игроков рынка. Что касается положения на глобальном рынке — российские команды сейчас активно наращивают темп разработки и сокращают дистанцию с зарубежными коллегами, опираясь на собственную скорость исследований. Мы продолжаем набирать опыт практического применения ИИ, расширять возможности в реальных задачах»



Иван Басков

Руководитель направления подготовки данных для предобучения GigaChat (Pretrain Data Lead)

3.3. Какие существуют барьеры и драйверы развития отечественных LLM?

Российский рынок LLM располагает рядом значимых предпосылок для дальнейшего развития. Это сильная инженерная и математическая школа, высокий уровень подготовки профильных специалистов, ведущие университеты, а также сравнительно гибкая регуляторная среда, позволяющая быстрее внедрять новые технологические решения. Потенциал при этом ограничен прежде всего инфраструктурными и кадровыми барьерами: российские компании располагают существенно меньшими вычислительными и финансовыми ресурсами, чем крупнейшие игроки США и Китая. Дело не только в масштабе обучения. Меньше ресурсов — меньше экспериментов, меньше проверенных гипотез, медленнее итерации улучшения моделей.

Среди драйверов эксперты прежде всего называют сильные научные и инженерные школы; по их оценке, Москва остается крупным ML-хабом по мировым меркам. Второй фактор — доступ к open source: открытые веса снижают порог входа, поэтому команда с десятком видеокарт может вести продолженное предобучение и делать конкурентоспособные нишевые модели.

Третий — спрос со стороны экономики: при дефиците работников и низкой автоматизации LLM помогают закрывать нехватку людей в отдельных процессах. Дополнительные преимущества — относительно дешевая электроэнергия и отсутствие запретительного регулирования. Эксперты называют и несколько основных барьеров. Первый — вычислительная инфраструктура: санкции закрывают прямую закупку актуального оборудования, параллельный импорт примерно вдвое дороже, а от решения о покупке до запуска кластера может пройти год. Второй — кадры: для обучения конкурентной модели нужна команда как минимум из сотни специалистов, таких людей мало, их подготовка занимает годы, а отток за рубеж продолжается. Третий — корпоративная бюрократия: прорывные продукты требуют небольших сфокусированных команд, которым трудно работать внутри организаций на десятки тысяч сотрудников. Четвертый — дефицит открытых датасетов и бенчмарков: небольшим командам не хватает доверенных корпусов для экспериментов, а отрасли — сложных русскоязычных тестов для измерения прогресса.

«Например, в 2023 году какой-нибудь Meta [Признана экстремистской организацией, деятельность которой запрещена на территории РФ], Microsoft, даже Google — они все закупали сотнями тысяч H100 [NVIDIA H100 — графический ускоритель для обучения и работы ИИ-моделей, один из базовых элементов крупных вычислительных кластеров]. Есть ли сейчас, в 2026 году, в России в сумме на все коллективы сто тысяч H100? Я не знаю, но даже если есть, то не сильно больше. А там это одна компания только столько же закупала ещё три года назад — масштабы, конечно, совершенно иные. Помимо этого, это железо не просто достаточно закупить. У нас нет, мне кажется, школы обучения по инфраструктуре: очень мало в России, где можно научиться, как эти кластеры поддерживать, создавать, развивать. Не стоит забывать, что мы под санкциями, нельзя просто получить поддержку от вендора так просто. И мало людей, которые с этим умеют работать»



Михаил Тихомиров

К.ф.-м.н., старший научный сотрудник МГУ, автор RuAdapt

«У нас есть дешёвая электроэнергия, у нас есть куча классных ГЭС, где стоимость электроэнергии практически незаметна. У нас есть до сих пор классная математическая инженерная школа, и у нас десятки тысяч людей, очень высококвалифицированных. Нам, опять же, надо более эффективно бороться с утечкой мозгов, более эффективно создавать программы лояльности, чтобы возвращать мозги, потому что кто-то там за 10 лет уже все осознал и вернулся бы, но на сопоставимые какие-то условия, или на интересные задачи, или в команду, где как раз нет бюрократии какой-то огромной корпорации. Надо искать эти варианты и предлагать их людям, которые обладают уникальным набором компетенций. Это надо делать системно. Ну, соответственно, мозги и энергия есть. Вопрос, конечно, в железе»



Сергей Боловцов

Директор Исследовательского центра ИИ Президентской академии

Русскоязычные бенчмарки и датасеты

Отдельная задача — измерение качества моделей на различных задачах. Бенчмарки позволяют командам сравнивать свои модели с конкурентами, отслеживать прогресс от версии к версии и находят слабые места. Для русского языка это особенно важно: международные тесты почти не учитывают русскоязычный культурный, правовой и профессиональный контекст, поэтому высокие результаты в них не гарантируют качества модели в российских задачах.

За последние годы в России сложилась обширная экосистема тестирования моделей - от комплексных наборов вроде MERA (Multimodal Evaluation for Russian-language Architectures, открытый комплекс тестов для оценки современных русскоязычных LLM) до специализированных тестов по математике, работе с длинным контекстом, безопасности и мультимодальности. При этом эксперты отмечают дефицит сложных русскоязычных бенчмарков: базовые тесты современные модели уже уверенно проходят, а прямых аналогов продвинутых международных испытаний на русском пока мало.

Авторы	Бенчмарк	Назначение
РАНХиГС	SLAVA	Культурно-ориентированный бенчмарк для оценки ответов LLM на русскоязычные вопросы
Альянс в сфере ИИ	MERA / MERA CODE / MERA Industrial / SWE MERA	Бенчмарк для современных текстовых русскоязычных моделей
Альянс в сфере ИИ	MERA Multi	Оценка мультимодальных моделей с учетом культурных особенностей русского языка
SberDevices, МФТИ, AIRI, Ecom Tech	LIBRA	Оценка моделей на длинном контексте для русского языка
Т-Банк	T-Math	Крупнейший русскоязычный бенчмарк олимпиадных задач по математике
ВШЭ	ML2B	Мультиязычный (включая русский) бенчмарк для задач автоматического ML
MWS AI	MWSVisionBench	Мультимодальный бенчмарк для вопросов и ответов по отсканированным документам
А. Николич	Shlepa, ruArenahard, Doom	Набор русскоязычных бенчмарков разного размера и направленности
VK	MARKER	Оценка Visual Language Models на русском языке с учетом культурного контекста (с упором на образование)
МГУ	SafetyBench	Русскоязычный бенчмарк безопасности LLM: оценка токсичности ответов и устойчивости к джейлбрейкам с опорой на нормы российского законодательства
LLM Arena	LLM Arena, Russian Arena Explorer	Краудсорсинговая арена (llmarena.ru): пользователи вслепую сравнивают ответы моделей на русском, из голосов складывается рейтинг по категориям. Дополнительно доступен открытый датасет реальных запросов к LLM
RussianNLP	RussianSuperGLUE	Первый комплексный бенчмарк понимания русского языка (2020). Историческая точка отсчета: покрывает классические задачи, преемник — MERA

В августе 2026 года на площадке АНО «Цифровая экономика» обсуждалась концепция порядка проверки российских ИИ-моделей на соответствие законодательству и традиционным духовно-нравственным ценностям. Предполагается, что оценку будут проводить независимые испытательные лаборатории, прошедшие согласование с Минцифры, а для тестирования совместно с бизнесом разработают специальные бенчмарки. По итогам проверки лаборатория будет формировать заключение о соответствии модели установленным требованиям, которое может стать одним из оснований для присвоения ей статуса национальной или суверенной.

3.4. В чем преимущество российских моделей?

Модели, обученные российскими компаниями, лучше справляются с русским языком, локальными реалиями, культурными отсылками и специфическими вопросами, которые могут быть непонятны универсальным мультиязычным моделям. На специфических русскоязычных задачах российские LLM дают более осмысленные ответы и не проигрывают зарубежным аналогам по качеству русского языка. Эксперты разделяют качество русского языка и общие способности модели. Ни один из опрошенных не говорит о превосходстве российских решений по базовому «интеллекту»; отставание они оценивают примерно в одно поколение.

Сильнее всего оно заметно в программировании, агентных сценариях и сложных рассуждениях. По русскому языку российские модели, напротив, не уступают зарубежным или опережают их. Это видно по более экономной токенизации , естественной грамматике, знанию российских реалий и культуры, а также по тому, что модели реже переключаются на иероглифы и англицизмы. Часть преимуществ российских решений связана с поставкой и эксплуатацией: их можно развернуть в защищенной инфраструктуре заказчика, интегрировать с российскими информационными системами, оформить на российское юрлицо и связать с собственным поисковым индексом. Поэтому несколько экспертов предлагают сравнивать не «голые» модели, а готовые сервисы.

«Они [отечественные модели] понимают сейчас его [российский культурный код] точно лучше, чем аналоги международные. Конкретный эксперимент первого квартала этого года говорит о том, что топ-6 позиций нашего бенчмарка — прогонки нашего эксперимента на 14 тысячах вопросов по истории, обществознанию, социологии и географии — у отечественных моделей. В этом смысле все хорошо. По нашему бенчмарку они очень прилично прибавили за год, это прям десятки баллов. Понятно, что на открытой части кто-то дообучился, потому что стало слишком хорошо, — так резко модели не растут. Но тем не менее это тоже результат: мы теперь понимаем, что условный школьник, который спрашивает вопрос про историю России, грубо, на 80 из 100 вопросов, получит ответ со стопроцентной точностью, фактологическая точность будет соблюдена. А ещё год-полтора назад мы говорили о показателях в районе 50 и даже ниже — не процентов, а именно средних оценок»



Сергей Боловцов
Директор Исследовательского центра ИИ Президентской академии

«Российские модели неплохо понимают культурный код и разбираются в локальных особенностях — например, в законодательстве. Это напрямую связано с долей русскоязычных данных в обучающих датасетах: у зарубежных моделей русского текста в разы меньше, и, даже неплохо владея языком, они часто не улавливают подобные тонкости. Проверить это несложно: если спросить модель, что такое мультифора, ответить, что это прозрачный файл для бумаг, без поиска в интернете обычно способна только российская модель. Конечно, поиск решает такие проблемы, но мы говорим только о том, насколько культурный код заложен в базовую модель. Подобные детали — лишь иллюстрация более общей особенности: по тому, как модель обрабатывает информацию и «мыслит», российские разработки ближе к российским пользователям»



Иван Басков
Руководитель направления подготовки данных для предобучения GigaChat (Pretrain Data Lead)

3.5. Где ключевые области внедрения LLM-решений в России?

Наиболее перспективные области применения российских LLM связаны не столько с прямой конкуренцией с фронтирными моделями, сколько с локальными и отраслевыми задачами. Российские модели особенно эффективны там, где требуется глубокое понимание русского языка, национального контекста, профессиональных стандартов, отраслевых данных и локальной нормативной среды.

Ключевые направления внедрения — специализированные профессиональные отрасли: прежде всего медицина, право, строительство, геология и другие сферы, где важны отраслевые стандарты и доступ к качественным данным. Отдельный потенциал связан с образованием и государственным сектором: LLM могут работать ассистентами для студентов, преподавателей и сотрудников организаций, помогать в обработке обращений граждан и консультациях по государственным услугам. Еще одно направление — задачи с культурной и мировоззренческой спецификой: анализ текстов, оценка смыслов, медиакоммуникации, гуманитарные исследования и публичные сервисы.

Для выбора отраслей эксперты предлагают три критерия. Первый — незакрытая потребность и возможность сократить издержки: приоритетны сферы с низкой автоматизацией и острым дефицитом специалистов. Второй — роль человека: ИИ полезнее там, где человеческий фактор важен, но его можно контролировать. Третий — локальность данных: если результат зависит от территории, языка и местных практик, зарубежные решения переносятся хуже.

Уже работает сегмент интеграционных решений — развертывание в инфраструктуре заказчика, стыковка с российскими системами и автоматизация поддержки. Здесь мало зарубежных конкурентов, потому что их модели нельзя поставить в закрытый контур. Еще одна группа задач — перевод, суммаризация и извлечение данных из документов и архивов; в таких сценариях компактная дообученная модель может оказаться лучше самой мощной универсальной, примененной «в лоб». Есть и более общий риск: узкоспециализированные модели со временем уступают универсальным моделям следующего поколения. Поэтому отраслевой фокус полезен для текущих задач, но не заменяет необходимости дальнейших научных исследований в области ИИ. Есть вещи, которые имеют локальную зависимость и не универсальны для всего мира. Например, модели прогноза погоды для России в других странах будут работать хуже без донастройки, как и наоборот. То же касается нашей промышленности, особенно добывающей. Даже медицинские данные не обладают стопроцентной глобальной универсальностью — полагаю, что существуют значимые особенности для людей, проживающих на той или иной территории.

«Возьмем простой пример из медицины: чтобы врач не заполнял анамнез руками и он автоматически формировался из разговора с пациентом, нужно собрать большое количество пар «запись разговора — готовый анамнез». Такие разговоры нигде не возьмешь. Но если такую систему создать, она не заработает ни в Штатах, ни в Китае, ни в Германии — другой язык и стандарты заполнения. Созданное в одних странах, не заработает сразу в других. Поэтому пока сам о себе не позаботишься, готового решения ниоткуда не появится»



Александр Крайнов

Директор по развитию технологий искусственного интеллекта Яндекса

«Искусственный интеллект имитирует деятельность человека, потому надо рассматривать области, где человеческий фактор является важным, но не превалирующим, где его можно контролировать. Так, в поэзии и философии мы никогда не выиграем, от результатов ИИ там люди все равно будут морщиться. Если мы пойдем в математику и теоретическую физику, эксперты скажут «у нас тут формула работает, зачем нам упражнения на естественном языке?». А вот в химии, биологии, в пищевой промышленности, сельскохозяйственной отрасли, модели ИИ как раз могут обеспечить наибольший прорыв. Другое дело, что это понимают во всем мире, придется соревноваться.»



Александр Бухановский

ИТМО - доктор технических наук, руководитель Института искусственного интеллекта; директор мегафакультета трансляционных информационных технологий

«На основе версий универсальных моделей делаются отраслевые или профессиональные модели, которые знакомы с отраслевыми стандартами строительства, геологии, юриспруденции — стартапов сейчас достаточно много. На такие модели запрос есть, и они стопроцентно востребованы. Одна из причин в том, что их можно on-premise оставить в контуре какой-то компании и пользоваться у себя: они могут более аккуратно работать с наиболее важными экспертными знаниями, которые очень нужны. В медицине тоже такое востребовано. Такие модели точно нужны: они знают специфику российской отрасли и при этом работают на русском языке. Но здесь даже не столь важен язык, сколько конкретные знания, которые на этом языке есть»



Дмитрий Ильвовский

НИУ ВШЭ - заместитель декана по международной деятельности факультета компьютерных наук; доцент Департамента анализа данных и искусственного интеллекта

04 Какие данные нужны для обучения российских LLM?

4.1 Как происходит обучение российских моделей?

Ключевой аспект создания качественных больших языковых моделей — данные. LLM обучаются на масштабных текстовых корпусах, объем которых может измеряться десятками триллионов токенов. Основу обучающих данных, как правило, составляют веб-страницы, книги, научные статьи, код из открытых репозиториев и энциклопедические источники. Поэтому так важно, чтобы российские разработчики могли обучать LLM на общедоступной информации: это напрямую влияет на качество отечественных технологий и их конкурентоспособность. Само обучение строится на связях и закономерностях между данными. Модель не копирует исходные источники — на основе анализа множества текстов она производит совершенно другой конечный продукт. Итоговые способности при этом зависят не только от объема данных, но и от качества их фильтрации, дедупликации и курации.

Ориентировочные объемы данных по этапам обучения

Этап	Что происходит	Объем корпуса
Адаптация словаря	Расширение или переобучение токенизатора под русский язык	5–20 ГБ текста, 50–70% на оригинальном языке модели
Предобучение	Основной этап, ~20 токенов на параметр (Chinchilla)	7В — минимум 140 млрд, рабочий диапазон 0,5–2 трлн; 70В — минимум 1,4 трлн, рабочий диапазон 2–15 трлн
Продолженное предобучение	Перенос основного языка модели на русский поверх готовых весов	7В — 10–100 млрд токенов при 10–30% оригинального языка
Следование инструкциям	Обучение на парах «вопрос-ответ»	1–10 тыс. при ручной разметке, 10–100 тыс. в типовом случае
Модель награды	Ранжирование ответов по человеческим предпочтениям	20–100 тыс. пар в стандартном случае
Обучение предпочтениям	PPO / DPO / KTO	5–100 тыс. примеров, обычно не более 100 тыс.
Верифицируемые награды	Задачи с формально проверяемым результатом	5–50 тыс. примеров, значительная часть — синтетика

Команды начинают примерно с одних источников: веб-дампов, прежде всего Common Crawl. После «сырого» слоя результат определяет обработка. Каждая команда сама строит фильтрацию, дедупликацию и оценку качества; готовой процедуры конкурентного уровня нет. У поисковых компаний есть дополнительный ресурс — собственные исторические слепки интернета, которые невозможно собрать задним числом.

На предобучении важны объем и разнообразие тем, языков и стилей. На поздних этапах нужны небольшие, тщательно очищенные наборы от специалистов предметных областей и данные с автоматически проверяемым ответом, например математические задачи и код. К 2026 году синтетические данные и синтетическая разметка стали обычной практикой отрасли. Российские команды отличаются от крупных зарубежных и китайских прежде всего ресурсами: у них меньше возможностей самостоятельно собирать данные, покупать и оцифровывать книги и привлекать тысячи экспертов-разметчиков. Поэтому они сильнее зависят от открытых материалов, где русского языка мало.

«Стартовые источники примерно все одинаковые: есть варианты Common Crawl, обкачанного интернета, — это то, что идёт в претрейн. Дальше идёт большое количество датасетов для SFT, которые все собирают как могут: это математика в основном и кодинг. Ну и дальше в конце идёт обучение с подкреплением. Сейчас типичный сценарий — это десятки стадий претрейна. На каждом идёт определённая смесь. Чуть-чуть увеличил, например, качество одного языка, а другой — это целая алхимия, как это делать. Люди пытаются дотянуться до всех данных, которые есть, постоянно стоят машины для их сбора. Но, на мой взгляд, самые интересные данные сейчас — это цепочки действительно реального взаимодействия, например с теми же агентами, которые пишут код.»



Иван Оселедец

Д.ф.-м.н., профессор РАН, генеральный директор AIRI

«В Сбере я отвечаю за датасеты для этапа предварительного обучения — формулировать требования к данным для обучения модели, и либо находить их, либо создавать самостоятельно. Например, в 2025 году резко выросла популярность агентов, а для развития агентских способностей моделям требуется больше кода и больше агентских траекторий — верифицируемых задач, результат которых можно измерить. Похожим образом устроена, например, математическая задача с известным числовым ответом. Такие данные либо получают из открытых источников, либо создают сами. Здесь и проявляется ценность собственной обработки данных: из пяти-шести известных открытых датасетов, русский язык обычно присутствует только в одном. Команды, которые занимаются открытыми данными, как правило, ориентируются на языки с наибольшим объёмом ресурсов — по сути, на английский, — а всё остальное отсеивают. Поэтому, если мы хотим, чтобы модель хорошо работала с русским языком, значительную часть данных приходится собирать и обрабатывать самим»



Иван Басков

Руководитель направления подготовки данных для предобучения GigaChat (Pretrain Data Lead)

4.2 Почему обучение должно быть многоязычным?

Русскоязычные данные остаются ключевой составляющей обучения российских LLM: именно они определяют качество русского языка, точность ответов о российской действительности и понимание российского «культурного кода». Без качественно подготовленного русскоязычного корпуса отечественная модель теряет свое главное прикладное преимущество — способность эффективно работать в российской среде. Заменить общемировую базу знаний русскоязычный корпус при этом не может. Языковая модель обучается на связях между понятиями, а эти связи представлены в разных языках, научных дисциплинах и профессиональных областях.

Данные на других языках помогают модели точнее отвечать и по-русски: они расширяют не словарь, а ее представление о мире. Чем уже корпус, тем беднее система связей и тем ниже когнитивные способности модели. Особое место здесь занимает английский — основной язык научной коммуникации. Исключение англоязычных данных означало бы удаление огромного пласта знания: модель хуже понимала бы сложные предметные области, слабее работала бы с научной терминологией и задачами, требующими широкого профессионального кругозора. Доля русского языка в корпусе влияет на качество русской речи модели, знание российского культурного и правового контекста и фактологическую точность: модели, обученные преимущественно на английском, чаще ошибаются в русскоязычных фактах. На программирование, работу с инструментами и сложные рассуждения эта доля не влияет. Код и стандарты вызова функций в основном англоязычны, поэтому добавление русских текстов само по себе эти навыки не улучшает.

Оптимальную долю русского языка эксперты оценивают в 5–30%: около 1% уже недостаточно для культурного контекста, порядка 10% достаточно для безупречного русского, а выше 30% снижается разнообразие — русскоязычный корпус приходится повторять. Обучение только на русском эксперты оценивают одинаково: технически это возможно, но практического смысла для универсальной модели они не видят. По существующим оценкам, объем уникального текстового содержимого открытого русскоязычного интернета составляет несколько десятков терабайт (ТБ). При этом размеры «очищенных» корпусов данных, собранных для обучения LLM, как правило, меньше. Для сравнения: в многоязычном корпусе FineWeb2 русскоязычная часть после очистки занимает 5,82 ТБ, а англоязычный FineWeb — около 44 ТБ. В корпусе HPLT v2 — 7 ТБ очищенного русскоязычного текста и в пять раз больше — английского. Цифры не сопоставимы напрямую из за разных методик сбора, но с их помощью можно оценить общий порядок: даже 10 ТБ — крупный массив для одного языка, но до масштабов англоязычного веба заметно не дотягивает.

Этого объема достаточно как части мультиязычного корпуса в обучении, но недостаточно для полноценного моноязычного обучения. По плотности русскоязычная часть FineWeb2 может содержать около 1 трлн слов, что соответствует примерно 1–2 трлн токенов в зависимости от используемой модели и способа разбиения текста. Единой рыночной цены токена для обучения не существует, но для понимания масштаба, при условной ставке 10 копеек за тысячу токенов один проход по такому корпусу обошелся бы примерно в 100–200 млн рублей. [63], [https://hplt-project.org/datasets/v2.0] Кроме того, хорошей модели нужен программный код, а он почти весь англоязычный. Исключение — узкие области вроде российской правовой системы и судебной практики, где источники русскоязычные и зарубежные аналоги плохо переносятся.

«Русскоязычные данные, безусловно, нужно включать в обучение, желательно максимально чистыми и разнообразными. Более того, если мы хотим, чтобы модель лучше понимала российскую специфику — региональные особенности, малоресурсные языки народов России, локальный культурный контекст, — такие данные нужно открывать и использовать активнее. Но из этого не следует, что конкурентную модель можно построить только на русскоязычных данных. Если мы говорим о базовой универсальной модели, в её претрейне должны быть и иностранные языки, и данные отвечающие культурным особенностям нашей страны, и программный код, и в целом максимально разнообразный контент. Иначе она не сможет работать с целыми классами задач — например с переводом и знаниями, которые существуют в других языковых средах. Технически обучить модель только на русском можно, технически модель можно обучить почти на чём угодно. Но если речь идёт не об эксперименте „на коленке“, а о конкурентной большой языковой модели, которая должна работать в продукте, такой подход уже не имеет смысла. Убирая из претрейна целые пласты знаний, вы, по сути, делаете модели лоботомию. Зачем?»



Алёна Феногенова

Исполнительный директор департамента развития ИИ и машинного обучения Сбербанка

«Даже если бы у тебя был идеально очищенный датасет на русском, который содержал бы в себе всё, что в русскоязычной культуре написано, его было бы, скорее всего, недостаточно для создания competitive LLM. Competitive LLM сейчас по природе своей мультиязычные — требуют микса из всех языков. Почему? Потому что разные культуры невзаимопоглощающие. Культурные континуумы, который произведены в рамках англоязычного человечества и в рамках русскоязычного человечества, дополняют друг друга в процессе обучения LLM. Это точно меньше 50%. Я бы оценил оптимальную долю русского языка в целевой картине от 5 до 30%. Это мы видим даже по китайским LLM: текущие LLM, производимые китайскими игроками, содержат до 30% китайского и в основном всё ещё англоцентричны. Просто так вышло, что английский — мировой язык и англоязычный корпус датасетов, знаний и всего остального самый богатый»



Анатолий Потапов

Т-Банк - руководитель группы фундаментальных технологий LLM

«Если в датасете есть хотя бы около 10% русских текстов, модель уже может очень хорошо говорить по-русски. При этом ее промежуточные представления могут оставаться англоязычными, но текст, который она пишет по-русски, будет грамматически корректным и естественным. Если же в российских моделях условно половина данных на русском, а половина на английском, это, на мой взгляд, уже плохо влияет на качество. Русскоязычным компаниям стоит собирать данные и на других языках тоже. Вообще, если есть возможность включить многоязычный датасет, лучше его включить. Просто потому, что на разных языках часто представлены разные фрагменты мира. Даже если обучать модель на испанских текстах, это все равно помогает ей больше узнавать о мире и, соответственно, лучше отвечать на вопросы по-русски. И здесь есть еще один важный момент: ни одна хорошая языковая модель не обучится без программного кода. Если в обучении есть русский язык плюс, например, C++, Python или C#, все будет в порядке. Но если оставить только русский и полностью исключить программный код, получится плохая модель»



Илья Гусев

Создатель семейства русскоязычных LLM Saiga

4.3 На чем должен быть фокус в обучении моделей?

Фокус в обучении российских LLM смещается от объемов данных к инженерии данных. Эксперты фактически описывают переход к методам Data-Centric AI: качество модели растет через работу с данными, а не только через рост инфраструктуры или объемов. Меняется и смысл «российскости» корпуса. Для фундаментальной модели важно не механически увеличить долю русских текстов, а собрать корпус, который дает модели качественное знание о языке, стране, профессиях, институтах и мире. В него могут входить российские источники, многоязычные корпуса, переводы зарубежных материалов, научные публикации, образовательные тексты, техническая документация, синтетические данные и экспертные ответы. Главный вопрос — не происхождение каждого фрагмента, а то, насколько он усиливает модель и насколько методически правильно подготовлен. Такой корпус становится активом той же важности, что и вычислительная инфраструктура или команда разработчиков. С практической точки зрения это смещает центр внимания на подготовку корпуса: фильтрацию и дедупликацию источников, балансировку смеси данных между этапами обучения, генерацию и проверку синтетических данных, привлечение предметных экспертов на поздних стадиях.

Эксперты сходятся в том, что именно качество этой работы все чаще отличает сильные модели: успешные зарубежные примеры показывают, что тщательно подготовленный корпус меньшего объема может давать то же качество, что и значительно больший, но «сырой». Многоязычность корпуса сама по себе не гарантирует качество.

В некоторых работах заявлена тщательная предобработка датасетов на две сотни языков, хотя вручную проверялась только английская часть. Поэтому заимствованный многоязычный корпус нельзя считать доверенным без собственной проверки. Если смешать проверенный русскоязычный набор с непроверенным, доверие теряется ко всему итоговому корпусу. Есть и риск культурного несоответствия у переведенных англоязычных датасетов: в них другая правовая система, идиомы и реалии.

«Вопрос не в количестве, а в качестве. Для претрейна, мне кажется, всё равно уже: там всё, что можно, выкачали, положили, исторические факты она все и так знает. А вот как рассуждать, соединять — это ключевой момент. Ключевой прорыв, который произошёл у Opus и в Gemini, — это работа над синтетическими данными работы агентов. Дело не в количестве, дело в первую очередь в качестве. Если это, скажем, протоколы заседания Совета жильцов какого-нибудь дома, это не поможет. Если это какие-то уникальные документы эпохи Екатерины Второй, то вполне модель может стать лучшим специалистом по истории. Фактически это всё равно будет не очень большой процент от общего количества данных, и скорее это будет некое дополнительное запоминание некоторых ключевых тезисов и фактов»



Иван Оселедец

Д.ф.-м.н., профессор РАН, генеральный директор AIRI

«Мы собираем огромное количество текстов, разных диалогов и прочего. Некоторые из этих текстов могут быть уже синтезированы — это могут быть порождения больших языковых моделей, которые затесались в интернет. И очень важным шагом является качественное обучение ревард-моделей, то есть моделей, которые могут оценивать качество датасета. Имея огромное количество данных, из которых львиная доля уже может быть синтезирована, мне кажется, важно сделать упор на обучение таких моделей, которые умеют фильтровать эти данные: мы берём наш набор данных, присваиваем каждому примеру какую-то оценку и выбираем только те примеры, у которых оценка выше какого-то порога»



Алан Газзаев

AI VK, ML-инженер

4.4 Чем нам могут помочь закрытые источники?

Библиотечные, архивные, образовательные, культурные и профессиональные данные, закрытые на данный момент, могут дать отечественным моделям доступ к более глубокой, верифицированной и локально значимой информации и повысить качество работы российских LLM в новых специализированных областях. Закрытые данные, по оценке экспертов, по-разному влияют на общие и специализированные способности моделей. На общий уровень они почти не влияют, зато могут заметно улучшить работу в узкой профессиональной области. Главный пример — медицина: обезличенные карты пациентов, протоколы лечения и клинические рекомендации дали бы данные для систем поддержки врачебных решений, которых сейчас не хватает. С культурным кодом ситуация иная. Большинство экспертов считает, что общие для общества смыслы уже представлены в публичном интернете, а закрытые архивы ценны прежде всего специализированными знаниями.

Отдельное исключение — архивы малых народов и малых языков. Доступ к закрытым данным ограничивается правовыми нормами и организационными моментами. Владельцы юридических баз, издательства и научные журналы не готовы бесплатно отдавать контент; учебники и монографии защищены авторским правом; даже достигнутые договоренности иногда застревают в корпоративных согласованиях, и данные не доходят до обучения. Сотрудничество разработчиков с университетами, библиотеками, архивами и госсектором пока во многом зависит от личных контактов. Чаще всего используют экспертную разметку через университеты: один широкопрофильный вуз позволяет привлечь специалистов из десятков предметных областей по одному договору. Среди других вариантов обсуждают эталонные списки источников, единый банк данных с общей процедурой лицензирования, налоговые стимулы для открытых публикаций и прямую покупку данных у платформ. Обсуждается и сдвиг от «положить данные в веса» к «дать модели доступ к данным». Закрытые коллекции можно подключать к специализированным агентам через поиск по базе знаний (RAG); тогда меняющиеся факты остаются во внешнем источнике.

«Качество статьи, прошедшей публичную апробацию и опубликованной во зарубежном журнале уровня Q1, априори выше, чем качество какого-нибудь внутреннего университетского отчета, писать который посадили аспиранта потому, что остальным членам команды некогда. Говорить, что мы сейчас залезем в хранилища знаний университетов и извлечем оттуда великие научные откровения, будет неправильно. Квинтэссенция того, что касается культурного кода и работы с русским языком, навряд ли лежит в закрытых коллекциях. То, что закрыто, не может стать по определению культурным кодом нации, согласитесь. Потому нужно делать упор не на поиск «забытых знаний», а на профилирование моделей для профессиональных задач и областей. Например, чтобы возникала мировая практика типа «хочешь сделать медицинский ИИ – бери российские модели, они наиболее тренированные в этой области»



Александр Бухановский

ИТМО - доктор технических наук, руководитель Института искусственного интеллекта; директор мегафакультета трансляционных информационных технологий

«Процесс обучения LLM многослойный: он проходит через стадии предобучения — это много максимально разнообразных данных, — а потом на следующем этапе ты показываешь меньше данных, но с большей образовательной ценностью. На последних этапах действительно имеет смысл показывать небольшое количество максимально чистых и максимально экспертных данных. Доступ к каким-то качественным источникам, закрытым, таким как библиотеки, архивы, внутрикорпоративные документы, может быть хорошей затравкой для получения таких качественных верифицируемых данных. Это потенциально наше конкурентное преимущество, если какие-то данные доступны нам, но не присутствуют в интернете и недоступны глобальным игрокам. Такие данные стоит использовать и не делиться ими потенциально, если есть задача быть конкурентным. Если есть задача развивать глобальный искусственный интеллект, то можно ими делиться или обмениваться»



Анатолий Потапов

Т-Банк - руководитель группы фундаментальных технологий LLM

05 Что нужно делать дальше для развития отрасли LLM в России?

5.1 Что можно сделать в первую очередь?

Значительная часть закрытых текстов пока не работает на развитие моделей — из-за организационной инерции и правовой неопределенности. Многие из них при этом не содержат чувствительной информации и могут быть открыты, лицензированы или переданы разработчикам через контролируемые механизмы. Самый быстрый путь к росту качества российских LLM на данный момент лежит именно здесь — через открытие и оцифровку таких данных. Проблема не в количестве сырого текста: в интернете его достаточно. Не хватает хорошо подготовленных, структурированных и доступных данных. Помимо открытия закрытых массивов, эксперты называют несколько направлений на ближайшие два-три года. Почти все упоминают синтетические и гибридные данные: по их оценке, тексты сильных моделей уже сопоставимы с текстами крепких специалистов, а совместная работа эксперта с моделью сочетает экспертизу и скорость.

Еще одно направление — привести в порядок уже существующие источники: оцифровать русскоязычные книги, собрать открытый корпус российского интернета с прозрачной фильтрацией, отдельно выделить математические, научные и образовательные материалы. Нужны и инструменты перевода PDF, сканов и нестандартных форматов в машиночитаемый текст. Наконец, важен открытый интернет: если контент уходит в закрытые платформы, он выпадает из поисковых индексов и обучающих корпусов.

В настоящее время уже идет обсуждение этого направления. На заседании Совета при Президенте Российской Федерации по науке и образованию 11 августа[24] упоминалась советская Государственная система научно-технической информации: к 1980 году в ГСНТИ хранилось свыше 2 миллиардов единиц научной информации, и ни одной модели этот массив до сих пор недоступен. Заявленная цель — дообучить на этих фондах суверенные национальные модели, для чего предложено обновить положение о ГСНТИ отдельным указом. Данные установок мегасайенс и Национальной базы генетической информации также рассматриваются как основа для обучения российских фундаментальных и отраслевых моделей, а доступ к ним должны получить научные организации, вузы и высокотехнологичные компании.

«Во-первых, самый простой и эффективный способ улучшить обучение российских моделей – предоставить как можно большее количество данных. Понятно, что есть вещи, которые открывать нельзя. Например, содержащие государственную тайну. При этом наверняка есть массивы текстов, которые закрыты, но никто уже не помнит почему. Их открытие помогло бы увеличить количество полезных данных для обучения. Во-вторых, ситуация сильно улучшится с появлением агентных систем, которые сами будут взаимодействовать со средой и обогащаться в результате этого взаимодействия. Наконец, это генерирование синтетических данных. Уже сейчас тексты, созданные искусственным интеллектом, не проигрывают по качеству текстам, которые пишут специалисты в своей области»



Александр Крайнов

Директор по развитию технологий искусственного интеллекта Яндекса

«Хотя бы для начала систематизировать то, что есть. Мне кажется, в процессе оцифровки русскоязычных книг, создания открытого русскоязычного кролл-корпуса, каких-то открытых синтетических датасетов на русском языке ещё много потенциала. Наверное, у существующих коммерческих вендоров это должно как-то поощряться — их open source программы. И на уровне комьюнити и государства это должно быть позитивной историей: вендоры, которые что-то делают, этим делятся. Ну и, возможно, это должен быть приоритет для вузов и для государства — поощрение таких open source инициатив»



Анатолий Потапов

Т-Банк - руководитель группы фундаментальных технологий LLM

5.2 Какие шаги нужны со стороны государства, бизнеса, образования, науки?

Повышение качества российских LLM требует согласованных действий ключевых участников рынка. Государство должно обеспечить инфраструктуру, правовые условия и стимулы для работы с данными, включая понятный порядок работы с авторским правом и доступом к источникам. Бизнесу важно переводить модели в готовые продукты, развивать технологические альянсы, поддерживать open source и создавать гибкие исследовательские команды. Научные организации могут формировать методологическую базу, заниматься оценкой качества, исследованием архитектур и созданием бенчмарков. Образовательная система должна готовить профильных специалистов и выстраивать каналы передачи практического опыта.

Принятый Федеральный закон «О поддержке развития технологий искусственного интеллекта в Российской Федерации» делает первый шаг в этом направлении: он прозрачно объясняет разработчикам механизмы работы с авторским правом и не сужает объем источников данных для обучения отечественных моделей. Законодательство в этой логике должно не замедлять темпы разработки, а убирать лишние барьеры, обозначать понятные статусы и давать отрасли время на взросление. От государства эксперты ждут вычислительных центров коллективного пользования: научные группы, вузы, малые и средние компании могли бы получать ресурсы по обоснованной заявке только для исследований, без коммерческой эксплуатации. В регуляторике приоритеты другие: прояснить правовой статус обучающих данных, не вводить избыточное раскрытие корпусов и стимулировать открытые публикации налоговыми мерами. Отдельный спор касается прозрачности обучающих данных.

Сторонники раскрытия говорят о воспроизводимости и контроле смещений; противники считают состав датасета конкурентным преимуществом вендора. Компромисс, который чаще поддерживают эксперты: научные и открытые проекты публикуют источники полностью, коммерческие модели — только общие категории данных и методы обработки. Для моделей на прямом государственном финансировании требование раскрытия выглядит оправданным. Еще одна задача государства — финансирование: проект такого масштаба дороже возможностей одной компании, поэтому возможны либо прямая поддержка стратегического проекта, либо координация партнерства нескольких игроков с обязательством вернуть модель заданного уровня. От компаний эксперты ждут трех действий: делиться разработками без коммерческих секретов и собирать узкие отраслевые датасеты; сравнивать модели по международным бенчмаркам, а не по внутренним метрикам; проектировать решение от бизнес-процесса целиком, а не оптимизировать модель отдельно от продукта. Науке отводят методологию, экспертизу и бенчмарки, а не конкуренцию с бигтехом в масштабном обучении. Университеты здесь важны еще и как площадка, где можно собрать специалистов из многих предметных областей. В образовании есть меры с разным горизонтом: увеличение бюджетных мест и обучение работе с языковыми моделями даст результат примерно через пять лет; обмен практическим опытом между компаниями и командами работает быстрее, но накопленные знания часто остаются закрытыми.

«Грамотная регуляторика, хорошая, может сильно ускорить прогресс, плохая регуляторика может всё развалить. Во-первых, я вижу роль государства именно в правильной настройке регуляторики, связанной с авторским правом: как его нужно интерпретировать, чтобы не тормозить прогресс в создании датасетов. Во-вторых, это демпингование на рынке вычислительных ресурсов — создание центров коллективного пользования, которые нельзя использовать для продового инференса, нельзя использовать для чего-то типа майнинга, а вот именно для дешёвых исследовательских программ можно. Роль научных учреждений — это прежде всего разработка методологии и методов, роль коммерческих организаций — прежде всего эффективная продуктивизация этих методов. Государство стоит сверху, регулирует это дело и поддерживает науку фундаментальную, которая разрабатывает методы. Это то, на что частные структуры не способны, как правило»



Иван Бондаренко
Доцент НГУ, проект «Сибирские нейросети»

«Это большие расходы, и эти расходы совершенно точно не экономически обоснованы: если посчитать стоимость дата-центров, которые требуются, то расходы колоссальные. Сейчас это держат на себе несколько крупных IT-компаний, которые понимают, что технология важна. Хорошо было бы историю про национальную суверенную модель обернуть в какие-то меры поддержки, чтобы это не было просто таким прожиганием денег, при котором создатели всё равно будут обучать Qwen-ы, ибо так дешевле. По крайней мере, государство должно сказать: я поддерживаю развитие этой модели так-то и так-то, а вы нам должны вернуть модель, которая в некоторых условных единицах не хуже, чем фронтирные»



Иван Оселедец
Д.ф.-м.н., профессор РАН, генеральный директор AIRI

5.3 Какой должна быть модель развития рынка?

Будущая конкурентоспособность российских моделей будет коллективным результатом совместной работы бизнеса, государства, open source сообщества, культурных и образовательных организаций. В условиях ограниченных ресурсов отдельным российским компаниям сложно в одиночку конкурировать с глобальными лидерами. Перспективной поэтому становится модель распределенного развития: государство создает условия, бизнес обеспечивает промышленную реализацию, наука задает методологическую глубину, образование воспроизводит кадровую базу, а open source сообщества ускоряют обмен разработками и решениями. Российский рынок LLM сможет развиваться, если команды разработчиков будут использовать мировую технологическую базу, усиливать ее отечественными данными, дорабатывать под российские задачи и быстро доводить продукты до практического внедрения. Если период становления рынка будет использован для консолидации усилий всех игроков, у отечественных LLM появится шанс закрепиться в роли полноценного технологического инфраструктурного слоя для России.

Эксперты предлагают три варианта кооперации. Первый — отраслевые рынки интеллектуальных агентов: вместо того чтобы несколько компаний независимо решали одну задачу в закрытых системах, готовыми агентами можно обмениваться и делить работу. Второй — отраслевые модели-полуфабрикаты: базовую, еще непригодную для прямого использования модель обучают на общих для отрасли данных, а отдельное предприятие дообучает ее на меньшем объеме собственных данных. Третий — небольшие и средние модели в связке с формальными базами знаний. Такой вариант требует меньше вычислительных ресурсов, но больше инженерной и математической работы — как раз тех компетенций, которые эксперты относят к сильным сторонам российской школы.

«Создание competitive LLM с нуля в реалиях российских компаний не окупается. В том плане, что если это делается на коммерческой основе, то этот проект должен быть заведомо с амбициями на рынок СНГ или на глобальный рынок, или он должен функционировать, скорее всего, под государственным зонтиком при поддержке каких-то государственных инвестиций. Если в масштабах конкретной российской бигтех-компании даже это не окупается, возможно, стоит делать партнёрство на уровне нескольких больших компаний, создавать какой-то один LLM-проект с равным доступом для всех игроков, которые в это вкладываются. ВВП России не сравним сейчас с ВВП Китая: если Китай может позволить себе иметь 4–5 «китайских тигров» в рамках LLM-проектов, то в масштабах России реалистично это 1–2 максимум. Возможно, какие-то партнёрства и совместные разработки могут облегчить вход для российских компаний совместно в эту капиталоёмкую штуку»



Анатолий Потапов

Т-Банк - руководитель группы фундаментальных технологий LLM

«Я бы не называл это данными самими по себе, я бы это называл, может быть, какой-то экосистемой — экосистемой взаимодействия компаний, данных этих компаний, университетов и так далее, чтобы потенциально была возможность запросить эти данные или были какие-то юридические основания и способы регулирования их использования. Тут просто куда ни посмотри — всё непонятно, какой-то туман: что, куда, зачем. А вот если бы была какая-то общая экосистема, с помощью которой можно было бы в итоге получать необходимые данные для решения конечной задачи... При этом указанные данные могли бы собираться из разных министерств, компаний, университетов, школ.»



Михаил Тихомиров

К.ф.-м.н., старший научный сотрудник МГУ, автор RuAdapt

«Нам нужна экспертная разметка, экспертные ответы на очень сложные вопросы. Сейчас самый понятный путь для нас лежит через университеты, где сразу есть специалисты во множестве областей: физике, химии, машиностроении, оптике, животноводстве, металлургии»



Александр Крайнов

Директор по развитию технологий искусственного интеллекта Яндекса

Методология исследования

Исследование основано на серии глубинных полуструктурированных интервью с 14 экспертами российской LLM-отрасли. В выборку вошли представители бигтеха, независимых команд, академических и государственных организаций. Интервью дополнили кабинетным исследованием рынка, обзором отечественных моделей и бенчмарков и техническим разбором этапов обучения больших языковых моделей.

Сноски

Рыночная статистика:

1. Вычислительный тупик. Почему российский ИИ остается без мощностей / А. Миколенко // ComNews. — 14.05.2026. — URL: https://www.comnews.ru/content/245257/2026-05-14/2026-w20/1008/vychislitelnyy-tupik-pochemu-rossiyskiy-ii-ostaetsya-bez-moschnostey (дата обращения: 02.10.2026).

2. Затраты бизнеса в РФ на ИИ в 2025 году составили 257 млрд рублей // ТАСС. — 08.04.2026. — URL: https://tass.ru/ekonomika/27040453 (дата обращения: 02.10.2026).

3. Внедрение ИИ может принести российской экономике до 13 трлн руб. к 2030 г. // Яков и Партнёры. — 2025. — URL: https://yakovpartners.ru/publications/ai-2025/ (дата обращения: 02.10.2026).

4. Доля крупных российских компаний, использующих генеративный ИИ, превысила 70% // Яндекс. — 08.12.2025. — URL: https://yandex.ru/company/news/08-12-2025-01 (дата обращения: 02.10.2026).

5. Генеративный искусственный интеллект используют 70% российских компаний // Forbes. — URL: https://www.forbes.ru/tekhnologii/535047-generativnyj-iskusstvennyj-intellekt-ispol-zuut-70-rossijskih-kompanij (дата обращения: 02.10.2026).

6. Аналитика: цифры и тренды : презентация на конференции Mediascope 2025 / И. Суанова // Mediascope. — 2025. — 60 слайдов. — URL: https://mediascope.net/upload/iblock/f0c/tdzyl93d36eeqrc5ndrkg7fnt667i5ss/%D0%9A%D0%BE%D0%BD%D1%84%D((дата обращения: 02.10.2026).

7. Mediascope: 26% россиян используют ИИ хотя бы раз в месяц // ТАСС. — 21.11.2025. — URL: https://tass.ru/obschestvo/25692451 (дата обращения: 02.10.2026).

8. «Алиса AI» обошла DeepSeek и других игроков рынка — исследование Mediascope // CNews. — 21.11.2025. — URL: https://www.cnews.ru/news/line/2025-11-21_alisa_ai_oboshla_deepseek_i_drugih (дата обращения: 02.10.2026).

9. Gartner Forecasts Worldwide AI Spending to Grow 47% in 2026 // Gartner. — 19.05.2026. — URL: https://www.gartner.com/en/newsroom/press-releases/2026-05-19-gartner-forecasts-worldwide-ai-spending-to-grow-47-percent-in-2026 (дата обращения: 02.10.2026).

10. ИИ может принести российскому ВВП 11,2 триллиона рублей, считают ученые// РИА Новости. — 07.04.2026. — URL: https://ria.ru/20260407/ii-2085665208.html (дата обращения: 02.10.2026).

11. Российские ИИ-модели заняли весь топ-6 в бенчмарке SLAVA // Ведомости. — 28.01.2026. — URL: https://www.vedomosti.ru/technology/news/2026/01/28/1172274-rossiiskie-ii-modeli (дата обращения: 02.10.2026).

12. LLM-модели от RWB вошли в топ-3 и топ-5 русскоязычного рейтинга MERA / Lexx_Nimofff // Хабр. — 25.06.2026. — URL: https://habr.com/ru/news/1052018/ (дата обращения: 02.10.2026).

13. ИИ-модель Wildberries вошла в топ-3 русскоязычного бенчмарка MERA / В. Фетисов // 3DNews. — 25.06.2026. — URL: https://3dnews.ru/1144134/iimodel-wildberries-voshla-v-top3-russkoyazichnogo-benchmarka-mera (дата обращения: 02.10.2026).

14. BerryLM-XL : карточка результата № 16045 / Wildberries & Russ // MERA. — URL: https://mera.a-ai.ru/ru/text/submits/16045 (дата обращения: 02.10.2026).

15. Как бизнесу внедрять языковые модели и почему их сложно создавать в России / эксперты Open Talks Special // Инк. — 25.02.2025. — URL: https://incrussia.ru/share/russkoyazychnye-nejroseti/ (дата обращения: 02.10.2026).

16. Остаться в седле: кто и как в России обучает модели искусственного интеллекта // ICT.Moscow. — 06.08.2025. — URL: https://ict.moscow/news/ai-models-in-russia-2025/ (дата обращения: 02.10.2026).

17. Использование ML/Data Open Source в России / Н. Никитин, А. Гетманов, И. Деева [и др.] // ITMO Open Source. — 2024. — URL: https://opensource.itmo.ru/os-mldata-research-2024 (дата обращения: 02.10.2026).

Регуляторный контекст

18. О поддержке развития технологий искусственного интеллекта в Российской Федерации : проект федерального закона № 1271570-8 // Система обеспечения законодательной деятельности. — 2026. — URL: https://sozd.duma.gov.ru/bill/1271570-8 (дата обращения: 02.10.2026).

О поддержке развития технологий искусственного интеллекта в Российской Федерации : Федеральный закон от 26 июля 2026 г. № 243-ФЗ // Официальное опубликование правовых актов. — 26.07.2026. — URL: http://publication.pravo.gov.ru/document/0001202607260003 (дата обращения: 02.10.2026).

19. В Госдуму внесен законопроект, регламентирующий порядок разработки, внедрения, использования и применения технологий искусственного интеллекта в РФ // КонсультантПлюс. — 25.06.2026. — URL: https://www.consultant.ru/law/hotdocs/94606.html (дата обращения: 02.10.2026).

20. Модель поведения : Искусственный интеллект в России обретает форму // Коммерсантъ. — 27.02.2026. — URL: https://www.kommersant.ru/doc/8463730 (дата обращения: 02.10.2026).

21. Разработчиков нейросетей могут обязать раскрывать исходные данные // Ведомости. — 25.02.2026. — URL: https://www.vedomosti.ru/technology/articles/2026/02/25/1178770-razrabotchikov-neirosetei-mogut-obyazat-raskrivat-ishodnie-dannie (дата обращения: 02.10.2026).

22. Правительство смягчило требования к разработчикам ИИ // РБК. — 27.04.2026. — URL: https://www.rbc.ru/technology_and_media/27/04/2026/69ede8fb9a7947ae89a66727 (дата обращения: 02.10.2026).

23. Подготовленный правительством законопроект об ИИ вызвал вопросы у экспертов // Business FM. — 12.03.2026. — URL: https://www.bfm.ru/news/600567 (дата обращения: 02.10.2026).

24. Ближе к телу: какие нейросети планируют использовать российские ведомства // Forbes. — URL: https://www.forbes.ru/tekhnologii/560641-blize-k-telu-kakie-nejroseti-planiruut-ispol-zovat-rossijskie-vedomstva (дата обращения: 02.10.2026).

24а. Заседание Совета по науке и образованию : стенограмма // Президент России. — 11.08.2026. — URL: http://kremlin.ru/events/president/news/80508 (дата обращения: 02.10.2026).

Российские модели

25. GigaChat 3 Ultra Preview — тяжёлый open source / vltnmmdv // Хабр. — 21.11.2025. — URL: https://habr.com/ru/companies/sberdevices/articles/968904/ (дата обращения: 02.10.2026).

26. Техрепорт Alice AI: как мы создавали новое поколение моделей для самого популярного ИИ-ассистента в России / kolesov93 // Хабр. — 11.12.2025. — URL: https://habr.com/ru/companies/yandex/articles/974594/ (дата обращения: 02.10.2026).

27. T-Lite и T-Pro — открытые русскоязычные опенсорс-модели на 7 и на 32 млрд параметров / anatolii-potapov // Хабр. — 11.12.2024. — URL: https://habr.com/ru/companies/tbank/articles/865582/ (дата обращения: 02.10.2026).

T-Pro 2.0 — открытая гибридно-ризонинговая русскоязычная LLM / anatolii-potapov // Хабр. — 19.07.2025. — URL: https://habr.com/ru/companies/tbank/articles/928956/ (дата обращения: 02.10.2026).

T-Tech (t-tech) : страница организации // Hugging Face. — URL: https://huggingface.co/t-tech/ (дата обращения: 02.10.2026).

28. Как мы создали LLM-модель Cotype Nano / AlanRobotics // Хабр. — 26.11.2024. — URL: https://habr.com/ru/companies/mts_ai/articles/861398/ (дата обращения: 02.10.2026).

29. Как мы в Авито сделали свою LLM — A-vibe / Anastasiya_Rysmyatova // Хабр. — 27.10.2025. — URL: https://habr.com/ru/companies/avito/articles/956664/ (дата обращения: 02.10.2026).

30. Ozon тестирует нейросеть для генерации описаний товаров // Т-Бизнес секреты. — URL: https://secrets.tbank.ru/novosti/novaya-neyroset-ot-ozon (дата обращения: 02.10.2026).

31. MetalGPT-1 : карточка модели / nn-tech // Hugging Face. — URL: https://huggingface.co/nn-tech/MetalGPT-1 (дата обращения: 02.10.2026).

32. VK обновила языковую модель Diona для генерации текстов // Коммерсантъ. — 06.11.2025. — URL: https://www.kommersant.ru/doc/8180892 (дата обращения: 02.10.2026).

33. Gamayun's Path to Multilingual Mastery: Cost-Efficient Training of a 1.5B-Parameter LLM : препринт / A. Podolskiy, S. Molokov, T. Gerasin [et al.] // arXiv. — 2025. — arXiv:2512.21580. — URL: https://arxiv.org/abs/2512.21580 (дата обращения: 02.10.2026).

34. Сибирские нейросети — AI-решения для бизнеса : сайт. — URL: https://sibnn.ai/ (дата обращения: 02.10.2026).

35. Вихрь — семейство переведенных русскоязычных LLM / alexwortega // Хабр. — 23.01.2024. — URL: https://habr.com/ru/articles/787894/ (дата обращения: 02.10.2026).

36. Saiga 7B, Russian LLaMA-based chatbot : карточка модели saiga_7b_lora / IlyaGusev // Hugging Face. — URL: https://huggingface.co/IlyaGusev/saiga_7b_lora (дата обращения: 02.10.2026).

Saiga: open-source по-русски — автономная модель без лишней оболочки // IT-World. — 04.06.2025. — URL: https://www.it-world.ru/tech/qlrkumruh28k0o4kkkwg4cws0osw0cw.html (дата обращения: 02.10.2026).

37. Ruadapt : метод адаптации больших языковых моделей // ICT.Moscow. — URL: https://ict.moscow/card/ruadapt/ (дата обращения: 02.10.2026).

38. ruGPT-3.5-13B : карточка модели / ai-forever // Hugging Face. — URL: https://huggingface.co/ai-forever/ruGPT-3.5-13B (дата обращения: 02.10.2026).

39. «Сбер» открыл доступ к модели генерации текста для русского языка ruGPT-3.5 и расширенной версии нейросети mGPT, способной генерировать тексты на 61 языке // CNews. — 20.07.2023. — URL: https://www.cnews.ru/news/line/2023-07-20_sber_otkryl_dostup_k_modeli (дата обращения: 02.10.2026).

Русскоязычные бенчмарки и датасеты

40. Бенчмарк SLAVA: шаг к мировоззренческому суверенитету / SGERCEN // Хабр. — 15.12.2024. — URL: https://habr.com/ru/articles/866738/ (дата обращения: 02.10.2026).

SLAVA-OpenData-2800-v1 : набор данных / RANEPA-ai // Hugging Face. — URL: https://huggingface.co/datasets/RANEPA-ai/SLAVA-OpenData-2800-v1 (дата обращения: 02.10.2026).

SLAVA: Benchmark of Sociopolitical Landscape and Value Analysis : репозиторий / ikanam-ai // GitHub. — URL: https://github.com/ikanam-ai/SLAVA (дата обращения: 02.10.2026).

41. MERA : Бенчмарк для современных текстовых русскоязычных моделей : сайт. — URL: https://mera.a-ai.ru/ (дата обращения: 02.10.2026).

MERA v.1.2.0 Новая версия независимого бенчмарка, что поменялось? / alenusch // Хабр. — 25.09.2024. — URL: https://habr.com/ru/companies/sberdevices/articles/844982/ (дата обращения: 02.10.2026).

42. MERA: A Comprehensive LLM Evaluation in Russian / A. Fenogenova, A. Chervyakov, N. Martynov [et al.] // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Bangkok : Association for Computational Linguistics, 2024. — P. 9920–9948. — DOI 10.18653/v1/2024.acl-long.534. — URL: https://aclanthology.org/2024.acl-long.534/ (дата обращения: 02.10.2026).

43. MERA Multi : Бенчмарк для современных мультимодальных LLM : сайт. — URL: https://mera.a-ai.ru/ru/multi (дата обращения: 02.10.2026).

44. LIBRA: Long Input Benchmark for Russian Analysis / Gscraid // Хабр. — 08.08.2024. — URL: https://habr.com/ru/companies/sberdevices/articles/834638/ (дата обращения: 02.10.2026).

45. T-math : набор данных / t-tech // Hugging Face. — URL: https://huggingface.co/datasets/t-tech/T-math (дата обращения: 02.10.2026).

46. ML2B: multi-lingual ML benchmark for AutoML : репозиторий / hse-cs // GitHub. — URL: https://github.com/hse-cs/ml2b (дата обращения: 02.10.2026).

47. MWSVisionBench : репозиторий MWS-Vision-Bench / mts-ai // GitHub. — URL: https://github.com/mts-ai/MWS-Vision-Bench (дата обращения: 02.10.2026).

48. Doom — русский физмат бенчмарк / alexwortega // Хабр. — 05.05.2025. — URL: https://habr.com/ru/articles/906896/ (дата обращения: 02.10.2026).

Шлепа — Большой Русский Бенчмарк / alexwortega // Хабр. — 02.09.2024. — URL: https://habr.com/ru/articles/840176/ (дата обращения: 02.10.2026).

49. MARKER: Собираем свой русскоязычный мультимодальный датасет для оценки VLM / deming // Хабр. — 23.09.2025. — URL: https://habr.com/ru/companies/vk/articles/947892/ (дата обращения: 02.10.2026).

50. Как мы сделали полезным крупнейший русскоязычный датасет запросов к LLM / kucev // Хабр. — 20.06.2025. — URL: https://habr.com/ru/articles/920150/ (дата обращения: 02.10.2026).

51. RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark / T. Shavrina, A. Fenogenova, A. Emelyanov [et al.] // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). — Online : Association for Computational Linguistics, 2020. — P. 4717–4726. — DOI 10.18653/v1/2020.emnlp-main.381. — URL: https://aclanthology.org/2020.emnlp-main.381/ (дата обращения: 02.10.2026).

RussianSuperGLUE : репозиторий / RussianNLP // GitHub. — URL: https://github.com/RussianNLP/RussianSuperGLUE (дата обращения: 02.10.2026).

52. LLM Arena : сайт. — URL: https://llmarena.ru/ (дата обращения: 02.10.2026).

53. Нейросети на скрепах : Бизнес представил первые духовно-нравственные тесты для ИИ // Коммерсантъ. — 17.08.2026. — № 148/П. — С. 1. — URL: https://www.kommersant.ru/doc/8890818 (дата обращения: 02.10.2026).

54. SafetyBench: Russian-language LLM safety benchmark for toxicity evaluation and jailbreak robustness testing grounded in Russian law : видеозапись // YouTube. — URL: https://www.youtube.com/watch?v=6SvzoQHfd7w (дата обращения: 02.10.2026).

Адаптация под русский язык

55. Impact of Tokenization on LLaMa Russian Adaptation : препринт / M. Tikhomirov, D. Chernyshev // arXiv. — 2023. — arXiv:2312.02598. — URL: https://arxiv.org/abs/2312.02598 (дата обращения: 02.10.2026).

56. Facilitating large language model Russian adaptation with Learned Embedding Propagation : препринт / M. Tikhomirov, D. Chernyshev // arXiv. — 2024. — arXiv:2412.21140. — URL: https://arxiv.org/abs/2412.21140 (дата обращения: 02.10.2026).

57. A Multi-Aspect Evaluation of Tokenizer Adaptation Methods for Large Language Models on Russian / G. D. Andriushchenko, M. E. Godunova, V. V. Ivanov [et al.] // Doklady Mathematics. — 2025. — Vol. 112. — P. 455–464. — DOI 10.1134/S1064562425700395. — URL: https://link.springer.com/article/10.1134/S1064562425700395 (дата обращения: 02.10.2026).

Масштабирование и предобучение

58. Training Compute-Optimal Large Language Models : препринт / J. Hoffmann, S. Borgeaud, A. Mensch [et al.] // arXiv. — 2022. — arXiv:2203.15556. — URL: https://arxiv.org/abs/2203.15556 (дата обращения: 02.10.2026).

59. Scaling Laws for Neural Language Models : препринт / J. Kaplan, S. McCandlish, T. Henighan [et al.] // arXiv. — 2020. — arXiv:2001.08361. — URL: https://arxiv.org/abs/2001.08361 (дата обращения: 02.10.2026).

60. Beyond Chinchilla-Optimal: Accounting for Inference in Language Model Scaling Laws : препринт / N. Sardana, J. Portes, S. Doubov, J. Frankle // arXiv. — 2023. — arXiv:2401.00448. — URL: https://arxiv.org/abs/2401.00448 (дата обращения: 02.10.2026).

61. LLaMA: Open and Efficient Foundation Language Models : препринт / H. Touvron, T. Lavril, G. Izacard [et al.] // arXiv. — 2023. — arXiv:2302.13971. — URL: https://arxiv.org/abs/2302.13971 (дата обращения: 02.10.2026).

62. RedPajama: an Open Dataset for Training Large Language Models : препринт / M. Weber, D. Fu, Q. Anthony [et al.] // arXiv. — 2024. — arXiv:2411.12372. — URL: https://arxiv.org/abs/2411.12372 (дата обращения: 02.10.2026).

63. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale : препринт / G. Penedo, H. Kydlíček, L. Ben Allal [et al.] // arXiv. — 2024. — arXiv:2406.17557. — URL: https://arxiv.org/abs/2406.17557 (дата обращения: 02.10.2026).

FineWeb : карточка набора данных / HuggingFaceFW // Hugging Face. — URL: https://huggingface.co/datasets/HuggingFaceFW/fineweb (дата обращения: 02.10.2026).

63a. FineWeb2: One Pipeline to Scale Them All — Adapting Pre-Training Data Processing to Every Language : препринт / G. Penedo, H. Kydlíček, V. Sabolčec [et al.] // arXiv. — 2025. — arXiv:2506.20920. — URL: https://arxiv.org/abs/2506.20920 (дата обращения: 02.10.2026).

FineWeb2 : карточка набора данных / HuggingFaceFW // Hugging Face. — URL: https://huggingface.co/datasets/HuggingFaceFW/fineweb-2 (дата обращения: 02.10.2026).

64. OLMo: Accelerating the Science of Language Models : препринт / D. Groeneveld, I. Beltagy, P. Walsh [et al.] // arXiv. — 2024. — arXiv:2402.00838. — URL: https://arxiv.org/abs/2402.00838 (дата обращения: 02.10.2026).

65. 2 OLMo 2 Furious : препринт / Team OLMo, P. Walsh, L. Soldaini [et al.] // arXiv. — 2024. — arXiv:2501.00656. — URL: https://arxiv.org/abs/2501.00656 (дата обращения: 02.10.2026).

Технические ориентиры: мультиязычность

66. Revisiting Multilingual Data Mixtures in Language Model Pretraining : препринт / N. Foroutan, P. Teiletche, A. K. Tarun, A. Bosselut // arXiv. — 2025. — arXiv:2510.25947. — URL: https://arxiv.org/abs/2510.25947 (дата обращения: 02.10.2026).

67. Rethinking Multilingual Continual Pretraining: Data Mixing for Adapting LLMs Across Languages and Resources : препринт / Z. Li, S. Ji, H. Luo, J. Tiedemann // arXiv. — 2025. — arXiv:2504.04152. — URL: https://arxiv.org/abs/2504.04152 (дата обращения: 02.10.2026).

68. Multilingual Pretraining and Instruction Tuning Improve Cross-Lingual Knowledge Alignment, But Only Shallowly : препринт / C. Gao, H. Hu, P. Hu [et al.] // arXiv. — 2024. — arXiv:2404.04659. — URL: https://arxiv.org/abs/2404.04659 (дата обращения: 02.10.2026).

Технические ориентиры: инструкции, предпочтения, обучение с подкреплением

69. LIMA: Less Is More for Alignment : препринт / C. Zhou, P. Liu, P. Xu [et al.] // arXiv. — 2023. — arXiv:2305.11206. — URL: https://arxiv.org/abs/2305.11206 (дата обращения: 02.10.2026).

70. Finetuned Language Models Are Zero-Shot Learners : препринт / J. Wei, M. Bosma, V. Y. Zhao [et al.] // arXiv. — 2021. — arXiv:2109.01652. — URL: https://arxiv.org/abs/2109.01652 (дата обращения: 02.10.2026).

71. Scaling Instruction-Finetuned Language Models : препринт / H. W. Chung, L. Hou, S. Longpre [et al.] // arXiv. — 2022. — arXiv:2210.11416. — URL: https://arxiv.org/abs/2210.11416 (дата обращения: 02.10.2026).

72. Towards Comprehensive Preference Data Collection for Reward Modeling : препринт / Y. Hu, Q. Li, S. Ouyang [et al.] // arXiv. — 2024. — arXiv:2406.16486. — URL: https://arxiv.org/abs/2406.16486 (дата обращения: 02.10.2026).

73. LoRA: Low-Rank Adaptation of Large Language Models : препринт / E. J. Hu, Y. Shen, P. Wallis [et al.] // arXiv. — 2021. — arXiv:2106.09685. — URL: https://arxiv.org/abs/2106.09685 (дата обращения: 02.10.2026).

74. Proximal Policy Optimization Algorithms : препринт / J. Schulman, F. Wolski, P. Dhariwal [et al.] // arXiv. — 2017. — arXiv:1707.06347. — URL: https://arxiv.org/abs/1707.06347 (дата обращения: 02.10.2026).

75. Direct Preference Optimization: Your Language Model is Secretly a Reward Model : препринт / R. Rafailov, A. Sharma, E. Mitchell [et al.] // arXiv. — 2023. — arXiv:2305.18290. — URL: https://arxiv.org/abs/2305.18290 (дата обращения: 02.10.2026).

76. KTO: Model Alignment as Prospect Theoretic Optimization : препринт / K. Ethayarajh, W. Xu, N. Muennighoff [et al.] // arXiv. — 2024. — arXiv:2402.01306. — URL: https://arxiv.org/abs/2402.01306 (дата обращения: 02.10.2026).

77. Training language models to follow instructions with human feedback : препринт / L. Ouyang, J. Wu, X. Jiang [et al.] // arXiv. — 2022. — arXiv:2203.02155. — URL: https://arxiv.org/abs/2203.02155 (дата обращения: 02.10.2026).

78. RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback : препринт / H. Lee, S. Phatale, H. Mansoor [et al.] // arXiv. — 2023. — arXiv:2309.00267. — URL: https://arxiv.org/abs/2309.00267 (дата обращения: 02.10.2026).

79. Does RLHF Scale? Exploring the Impacts From Data, Model, and Method : препринт / Z. Hou, P. Du, Y. Niu [et al.] // arXiv. — 2024. — arXiv:2412.06000. — URL: https://arxiv.org/abs/2412.06000 (дата обращения: 02.10.2026).

80. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models : препринт / Z. Shao, P. Wang, Q. Zhu [et al.] // arXiv. — 2024. — arXiv:2402.03300. — URL: https://arxiv.org/abs/2402.03300 (дата обращения: 02.10.2026).

81. Group Sequence Policy Optimization : препринт / C. Zheng, S. Liu, M. Li [et al.] // arXiv. — 2025. — arXiv:2507.18071. — URL: https://arxiv.org/abs/2507.18071 (дата обращения: 02.10.2026).

82. Kimi k1.5: Scaling Reinforcement Learning with LLMs : препринт / Kimi Team, A. Du, B. Gao [et al.] // arXiv. — 2025. — arXiv:2501.12599. — URL: https://arxiv.org/abs/2501.12599 (дата обращения: 02.10.2026).

83. Kimi K2: Open Agentic Intelligence : препринт / Kimi Team, Y. Bai, Y. Bao [et al.] // arXiv. — 2025. — arXiv:2507.20534. — URL: https://arxiv.org/abs/2507.20534 (дата обращения: 02.10.2026).

Суверенный ИИ

Под суверенной большой фундаментальной моделью искусственного интеллекта понимается — большая фундаментальная модель искусственного интеллекта, соответствующая следующим требованиям:

1) разработчиком большой фундаментальной модели искусственного интеллекта является российское юридическое лицо;

2) разработка большой фундаментальной модели искусственного интеллекта, в том числе определение и изменение ее характеристик на всех стадиях ее жизненного цикла, осуществляется российским юридическим лицом, являющимся разработчиком большой фундаментальной модели искусственного интеллекта;

3) обеспечивается полная техническая и технологическая воспроизводимость цикла разработки разработчиком большой фундаментальной модели искусственного интеллекта, включая ее обучение;

4) подготовка ответов на запросы пользователей большой фундаментальной модели искусственного интеллекта и хранение данных обеспечиваются в центрах обработки данных, расположенных на территории Российской Федерации и принадлежащих российским юридическим лицам;

5) большая фундаментальная модель искусственного интеллекта прошла подтверждение соответствия законодательству Российской Федерации и традиционным российским духовно-нравственным ценностям в порядке, установленном Правительством Российской Федерации.

Под национальной большой фундаментальной моделью искусственного интеллекта понимается — большая фундаментальная модель искусственного интеллекта, соответствующая следующим требованиям:

1) разработчиком большой фундаментальной модели искусственного интеллекта является российское юридическое лицо;

2) существенные характеристики большой фундаментальной модели искусственного интеллекта, определенные Правительством Российской Федерации, в том числе ее структура, программное обеспечение и настраиваемые параметры, определяются и изменяются российским юридическим лицом, являющимся разработчиком данной модели;

3) используемые при разработке большой фундаментальной модели искусственного интеллекта компоненты, разработанные (созданные) как на территории Российской Федерации, так и за пределами территории Российской Федерации, в том числе иные большие фундаментальные модели искусственного интеллекта, разработанные другими лицами, распространяются на условиях открытой лицензии;

4) подготовка ответов на запросы пользователей большой фундаментальной модели искусственного интеллекта и хранение данных обеспечиваются в центрах обработки данных, расположенных на территории Российской Федерации и принадлежащих российским юридическим лицам;

5) большая фундаментальная модель искусственного интеллекта прошла подтверждение соответствия законодательству Российской Федерации и традиционным российским духовно-нравственным ценностям в порядке, установленном Правительством Российской Федерации.

Словарь терминов

Common Crawl — открытый архив копий веб-страниц, базовый источник данных для предобучения большинства моделей мира.

LLM (Large Language Model) — большая языковая модель — нейросеть с миллиардами параметров, обученная на больших массивах текста; основа современных ИИ-инструментов.

MoE (Mixture of Experts) — архитектура, в которой для каждого запроса активируется лишь часть параметров модели, что удешевляет ее работу.

RAG (Retrieval-Augmented Generation) — подход, при котором модель получает нужные сведения из внешней базы в момент ответа, а не хранит их в своих параметрах.

SFT (Supervised Fine-Tuning) — дообучение на парах «вопрос — эталонный ответ», обучающее модель следовать инструкциям.

Агент — система на основе LLM, которая не только отвечает текстом, но и выполняет действия: вызывает программы, ищет информацию, работает с документами.

Бенчмарк — стандартизированный набор заданий для сравнения качества моделей.

Веса модели — числовые параметры, в которых хранится все «знание» обученной модели.

Датасет/корпус — структурированный набор данных для обучения или проверки модели.

Дедупликация — удаление повторяющихся текстов из обучающего корпуса.

Дообучение (файнтюнинг) — настройка готовой модели на меньших специализированных данных под конкретные задачи.

Инференс — работа уже обученной модели: генерация ответов на запросы пользователей.

Мультимодальная модель — модель, работающая не только с текстом, но и с изображениями, аудио или видео.

Обучение с подкреплением — этап, на котором модель учится на оценках качества своих ответов — человеческих или автоматических.

Открытые веса (open source) — публикация весов модели, позволяющая любой команде свободно использовать и дообучать ее.

Предобучение (претрейн) — первый и самый ресурсоемкий этап обучения модели на больших массивах текста.

Продолженное предобучение — дополнительное обучение готовой модели на большом корпусе — например, для переноса ее на русский язык.

Синтетические данные — обучающие тексты, сгенерированные другой моделью, а не написанные людьми.

Токен — минимальная единица текста для модели: слово, часть слова или символ. Объем обучающих данных измеряют в токенах.

Токенизация — способ разбиения текста на токены. Удачная токенизация для русского языка делает работу модели быстрее и дешевле.

Авторы исследования

Николай Никитин

к.т.н., лидер сообщества ITMO Open Source.

Андрей Гетманов

м.н.с Института ИИ ИТМО.

Илларион Иов

ассистент ФТИИ ИТМО.

Егор Фёдоров

студент ФТИИ ИТМО, сотрудник учебно-научной лаборатории АИСТ.