ИСТОЧНИКИ ОБУЧЕНИЯ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ ИИ КАК ИНСТРУМЕНТ СУВЕРЕННОСТИ В ЦИФРОВОЙ ТРАНСФОРМАЦИИ

В.А. Майборода, Э.Т. Майборода

Развитие больших языковых моделей (LLM) представляет собой одно из наиболее значительных достижений в области искусственного интеллекта за последнее десятилетие. Эти системы, обученные на огромных объемах текстовых данных, демонстрируют способность решать широкий спектр задач обработки естественного языка с качеством, приближающимся к уровню человеческих возможностей. Однако по мере интеграции LLM в критичные сферы деятельности общества и государства всё более острой становится проблема источников этих данных и их соответствия национальным интересам разных стран.

Вопрос об источниках обучающих данных для LLM выходит далеко за рамки чисто технического вопроса. Это фундаментальная проблема цифровой суверенности, связанная с национальной безопасностью, культурным самоопределением, экономической независимостью и социальным доверием к информационным системам. В условиях усиливающейся геополитической конкуренции в области технологий искусственного интеллекта различные страны и региональные объединения пытаются найти собственный путь в регулировании разработки и применения LLM, при этом каждый подход отражает уникальное видение баланса между инновационностью, безопасностью и суверенностью.

Настоящее исследование сосредоточено на анализе того, как различные подходы к отбору и использованию источников обучения LLM влияют на возможности государств достичь технологической суверенности в контексте цифровой трансформации. В фокусе анализа находятся три модели регулирования, представляющие различные геополитические и идеологические позиции: американская модель децентрализованного регулирования, европейская модель горизонтального законодательного контроля и российская модель институционально-государственного подхода к суверенности обучающих данных.

Цель данной работы состоит в критическом анализе взаимосвязи между источниками обучения больших языковых моделей ИИ и возможностью государств достичь технологической и цифровой суверенности, а также выявлении парадоксов и трехмерного компромисса между качеством моделей, суверенностью данных и конкурентоспособностью на мировом рынке.

Актуальность этого исследования определяется несколькими факторами. Во-первых, в 2024–2025 годах произошла явная глобализация рынка LLM, где лидирующие позиции заняли американские компании OpenAI,

Anthropic и Google, европейский консорциум, представленный инициативами Европейского Союза, и китайские разработчики Baidu и ByteDance. Каждый из них использует принципиально различные подходы к отбору источников обучающих данных, отражающие местные приоритеты и регулятивные требования.

Во-вторых, в Российской Федерации находится на завершающей стадии обсуждения Федеральный закон об основах государственного регулирования технологий искусственного интеллекта, в котором вопрос суверенности источников обучения моделей занимает центральное место. Закон предусматривает развитие так называемых суверенных больших фундаментальных моделей, обучение которых должно осуществляться исключительно на территории Российской Федерации с использованием данных, собранных гражданами и юридическими лицами Российской Федерации [1].

В-третьих, существует объективное противоречие между стремлением государств к суверенности в области ИИ и техническими реалиями разработки LLM, которые требуют доступа к огромным объемам разнообразных данных. Русскоязычный интернет, по различным оценкам, составляет примерно 50–100 миллиардов токенов текста, в то время как современные LLM требуют для оптимального обучения примерно 1,4 триллиона токенов по принципу Chinchilla scaling law.

Таким образом, академический анализ трех моделей подхода к источникам обучения LLM необходим для понимания того, может ли обособление процесса обучения моделей действительно обеспечить суверенность, или же оно неизбежно приводит к снижению конкурентоспособности и, как следствие, к зависимости от иностранных решений, которые якобы должны быть преодолены.

Данное исследование опирается на качественный анализ нормативной базы, технической документации и научных публикаций, посвященных регулированию и разработке больших языковых моделей в трех регионах: Соединенных Штатах Америки, Европейском Союзе и Российской Федерации.

Во-первых, проведена критическая ревизия опубликованной технической информации о способах, которыми ведущие разработчики LLM (OpenAI для модели ChatGPT/GPT-4, Anthropic для Claude, Baidu для ERNIE) получают и структурируют данные для обучения своих моделей.

Во-вторых, проанализирована нормативная база каждой юрисдикции, включая EU AI Act (Акт об искусственном интеллекте Европейского Союза), фрагментированное американское регулирование на уровне Executive Orders и ведомственных рекомендаций, китайскую систему требований к локализации данных в соответствии с PIPL и CSL, а также обсуждаемую в России модель обучения, закрепленную в проекте Федерального закона об ИИ.

В-третьих, применен сравнительно-правовой метод для выявления общих черт и существенных различий между подходами к регулированию источников обучения LLM в различных юрисдикциях.

В-четвертых, использованы элементы системно-теоретического анализа для выявления циклических зависимостей и парадоксов, возникающих при попытке одновременно достичь суверенности, качества и конкурентоспособности моделей.

Американская модель: децентрализованное регулирование и доступность глобальных данных

Соединенные Штаты Америки исторически приняли подход к регулированию искусственного интеллекта, который можно охарактеризовать как децентрализованный, секторальный и ориентированный на минимальное вмешательство государства в инновационную деятельность частного сектора. На федеральном уровне отсутствует единый всеобъемлющий закон, регулирующий разработку и применение LLM. Вместо этого используется модель, основанная на принципе ведомственного надзора, где различные агентства (FDA для медицины, SEC для финансов, NHTSA для автономных транспортных средств) применяют существующее законодательство к ИИ-системам, функционирующим в соответствующих их компетенции сферах.

Ключевой особенностью американского подхода является отсутствие жестких требований относительно происхождения и природы данных, используемых для обучения коммерческих LLM. Открытая информация о методах обучения ChatGPT и других моделей OpenAI показывает, что эти системы обучаются на данных, полученных из множественных источников, включая общедоступный интернет, высокачественные текстовые корпусы, книги и специализированные датасеты [2]. Согласно официальной информации, OpenAI использует данные из Common Crawl, крупнейшего открытого архива веб-контента, над которыми работают как американские, так и международные разработчики. Over 80% of GPT-3's 300+ billion training tokens came from the Common Crawl dataset, что свидетельствует об ориентации на использование максимально доступных глобальных источников информации [3].

Это стратегическое решение имеет несколько важных следствий. Во-первых, модели OpenAI, обученные на разнообразных, глобальных источниках, достигают высокого уровня качества во многих языках и предметных областях, включая специализированные домены (медицина, юриспруденция, инженерия), где большинство высокозначимого контента опубликовано на английском языке. Во-вторых, такой подход к сбору данных позволяет американским компаниям быстро масштабировать свои модели и адаптировать их к новым задачам. В-третьих, использование общедоступного интернета в качестве основного источника данных создает объективные трудности для любого геополитического конкурента, пытающегося создать модель более высокого качества, используя исключительно локальные источники данных.

Однако американская модель имеет и существенные недостатки с точки зрения защиты интересов государства. По сути, ведущие американские компании, разрабатывающие LLM, не подлежат жесткому контролю

относительно того, на каких данных они обучают свои модели. Это означает, что коммерчески успешные модели могут быть обучены на данных, полученных из иностранных источников, включая потенциально чувствительную информацию. Кроме того, отсутствие единого федерального стандарта создает фрагментированный ландшафт, где различные штаты могут принимать собственные требования к ИИ, что ведет к дополнительным трудностям.

Интересно отметить, что администрация США в 2025–2026 годах попыталась прояснить позицию федерального правительства относительно ИИ. Администрация Трампа, по различным источникам, приняла курс на дерегулирование, отмену требований к AI safety и попыталась централизованно предотвратить введение избыточно строгих требований к ИИ на уровне отдельных штатов. Однако это не означает, что Соединенные Штаты полностью отказались от контроля над критичными аспектами разработки ИИ. Вместо явного законодательного регулирования используются механизмы контроля на уровне государственных контрактов, инвестиций в области обороны и национальной безопасности, а также давление на частные компании через регулирующие органы [4, с. 58–75].

Таким образом, американская модель может быть охарактеризована как позволяющая разработчикам LLM полностью использовать глобальные источники данных без жестких ограничений на их происхождение, при условии соблюдения общих норм права о защите данных и конфиденциальности. Это обеспечивает высокое качество и конкурентоспособность американских моделей, но снижает уровень контроля государства над процессом обучения и потенциально создает уязвимости в области информационной безопасности.

Европейская модель: горизонтальное регулирование и требования к управлению данными. Европейский Союз, напротив, принял принципиально иной подход к регулированию ИИ и, в частности, к вопросам об источниках обучения LLM [5]. EU AI Act, являющийся первым в мире комплексным законодательным актом, специально посвященным регулированию искусственного интеллекта, устанавливает вертикальные (т.е. выстроенные по уровню риска) требования к разработчикам моделей, включая детальные положения об управлении данными.

Статья 10 EU AI Act, озаглавленная «Data and Data Governance», содержит явное требование о том, что данные, используемые для обучения, валидации и тестирования моделей ИИ высокого риска, должны подлежать надлежащей практике управления данными [6]. Кроме того, статья 53 EU AI Act требует от разработчиков моделей ИИ (General-Purpose AI Models) общего назначения публикации детального краткого описания содержания, используемого для обучения. Это описание должно быть «достаточно подробным», чтобы позволить независимым экспертам и регулятивным органам оценить потенциальные риски модели.

Европейский подход принципиально отличается от американского тем, что он явно переводит требование к управлению данными из сферы добровольной практики в сферу обязательного нормативного требования. Европейские разработчики LLM должны документировать, откуда они получили свои данные, как они обеспечили качество данных, какие процедуры используются для идентификации и минимизации предубеждений, присутствующих в обучающих данных.

Однако, в отличие от подхода, который предполагается в российском проекте закона, EU AI Act не требует, чтобы данные для обучения европейских моделей были исключительно европейского происхождения или обрабатывались только на территории ЕС (хотя требования GDPR о локализации персональных данных остаются в силе). Вместо этого Европа пытается сбалансировать между, с одной стороны, открытостью к глобальным данным и сотрудничеству, и, с другой стороны, защитой своих граждан от потенциальных рисков, связанных с моделями, обученными на некачественных или предвзятых данных.

Научное исследование, проведенное в Европе, выявило потенциальную проблему: по мере того, как разработчики ИИ в Европе стремятся соответствовать требованиям EU AI Act, они часто оказываются зависимыми от услуг облачных провайдеров, контролируемых американскими компаниями (Amazon, Microsoft, Google). Это создает ситуацию, при которой техническая суверенность данных снижается, несмотря на формальное соответствие европейскому регулированию [7].

Таким образом, европейская модель может быть охарактеризована как стремящаяся к информационной прозрачности и управлению качеством данных при сохранении доступа к глобальным источникам информации. Это позволяет европейским разработчикам создавать конкурентоспособные модели, но не решает фундаментальную проблему технологической зависимости от американской облачной инфраструктуры и глобальных сетей передачи данных.

Китайская модель: локализация данных и государственный контроль. КНР развивает модель, которая в отношении источников обучения LLM занимает позицию, существенно отличающуюся как от американской, так и от европейской. Регулятивная система, основанная на законах о защите персональных данных (Personal Information Protection Law, PIPL) и Cybersecurity Law (CSL), формирует требование к локализации данных на территории Китая. Эти требования наиболее явно применяются к персональным данным, но их логика распространяется и на требования к суверенности данных, используемых для обучения моделей ИИ.

В рамках китайской системы функционирует концепция «контролируемого доступа» к интернету («Great Firewall»), которая ограничивает возможность китайских компаний использовать иностранные источники данных для обучения моделей. Вместо этого предполагается использование

данных, собранных внутри Китая или из китайского сегмента интернета. Это означает, что китайские разработчики LLM (такие как Baidu) вынуждены работать с существенно более ограниченным объемом исходных данных в сравнении с их американскими конкурентами.

Компания Baidu, один из крупнейших разработчиков LLM в Китае, представила серию моделей под названием ERNIE (Enhanced Representation through Knowledge Integration). Эти модели обучаются на данных, собранных в основном из китайского сегмента интернета, китайских социальных сетей, китайских онлайн-платформ и специализированных датасетов, созданных в соответствии с требованиями китайского регулирования. Несмотря на то, что Baidu инвестирует значительные ресурсы в разработку этих моделей, они остаются менее конкурентоспособными на глобальном рынке в сравнении с ChatGPT и Claude, отчасти именно потому, что они обучены на более ограниченном наборе данных.

Однако китайский подход имеет четкую логику с точки зрения государственного контроля и информационной безопасности. Ограничивая доступ китайских разработчиков к глобальным источникам данных, государство обеспечивает уровень контроля над контентом, используемым для обучения моделей, что соответствует китайской политике информационного суверенитета. Это означает, что модели, разработанные в Китае, с меньшей вероятностью «знают» о контактах, которые государство считает чувствительными или потенциально дестабилизирующими.

Кроме того, китайский подход создает стимулы для развития национальной экосистемы компаний, занимающихся обработкой и структурированием данных. В то время как в США основной объем работы с данными сосредоточен в руках нескольких крупных компаний (OpenAI, Google, Meta), в Китае развивается более распределенная система, включающая государственные учреждения, университеты и частные компании, которые все работают в рамках единого нормативного пространства.

Однако не следует переоценивать успехи китайского подхода. В 2023–2024 годах выявилось, что китайские LLM, несмотря на значительные инвестиции, отстают от западных аналогов по многим параметрам качества, что отчасти объясняется именно ограничениями на доступ к глобальным источникам данных и на использование иностранных облачных сервисов.

Российская модель: суверенное институциональное обучение. Находящийся на стадии завершения обсуждения проект Федерального закона об основах государственного регулирования технологий искусственного интеллекта Российской Федерации предлагает подход, который можно охарактеризовать как «суверенное институциональное обучение». Этот подход в ряде аспектов напоминает китайский, но имеет и существенные особенности.

Ключевое положение закона содержится в статье 7 законопроекта, которая определяет понятие суверенных и национальных больших фундаментальных моделей. Согласно этой статье, суверенная модель должна соответствовать следующим требованиям: во-первых, все стадии разработки и обучения моделей должны осуществляться на территории Российской Федерации; во-вторых, все стадии разработки, обучения и эксплуатации моделей должны осуществляться гражданами Российской Федерации и российскими юридическими лицами; в-третьих, обучение моделей должно происходить с использованием наборов данных, формирование которых осуществляется на территории Российской Федерации гражданами Российской Федерации и российскими юридическими лицами.

Обсуждаемая в России модель обучения предусматривает создание реестра «доверенных моделей», которые соответствуют установленным требованиям безопасности и качества. Эти модели будут обязательны к использованию в государственных информационных системах и на значимых объектах критической информационной инфраструктуры. Таким образом, государство явно переходит к стратегии разработки собственных фундаментальных моделей ИИ, которые не будут зависеть от глобальных сервисов.

Одним из ключевых принципов обсуждаемой в России модели обучения является «учет и уважение традиционных российских духовно-нравственных ценностей». В законопроекте предусмотрено, что разработка, внедрение и применение технологий ИИ должны осуществляться на основе таких ценностей, как жизнь, достоинство, права и свободы человека, патриотизм, гражданственность, служение Отечеству, высокие нравственные идеалы, крепкая семья, созидательный труд, приоритет духовного над материальным, гуманизм, милосердие, справедливость, коллективизм, взаимопомощь и взаимоуважение, историческая память и преемственность поколений.

Это положение, если его понимать буквально, означает, что наборы данных, используемые для обучения суверенных моделей, должны быть не только локализованы на российской территории и собраны российскими субъектами, но и отвечать определенным критериям духовно-нравственной приемлемости. Таким образом, Россия предлагает подход, который не только локализует данные географически и институционально, но и пытается обеспечить их нормативную приемлемость с точки зрения государственных ценностей.

Статья 8 обсуждаемого проекта закона предусматривает обязательное использование доверенных моделей в государственных информационных системах. Процедура включения моделей в реестр доверенных моделей предполагает подтверждение их соответствия требованиям безопасности (осуществляемое органами противодействия техническим разведкам

и органами, уполномоченными в области обеспечения безопасности) и требованиям качества (осуществляемое отраслевыми федеральными органами исполнительной власти).

Следовательно, очевидно, в России будет реализована модель, при которой государство явно контролирует каждый этап разработки и применения LLM, начиная с отбора и структурирования исходных данных и заканчивая проверкой соответствия готовых моделей установленным стандартам.

Сравнительный анализ четырех подходов (американского, европейского, китайского и российского) к регулированию источников обучения LLM выявляет фундаментальный парадокс, который можно сформулировать так: максимальная суверенность в отборе и использовании источников обучения данных неизбежно ведет к снижению качества и конкурентоспособности создаваемых моделей, а стремление к конкурентоспособности предполагает использование глобальных источников данных, что ограничивает суверенность.

Этот парадокс коренится в технических реалиях разработки современных LLM. Согласно исследованиям, опубликованным в работе «Training Compute-Optimal Large Language Models», ставшей известной как «Chinchilla scaling law», оптимальное соотношение между размером модели и количеством токенов для обучения составляет примерно 20 текстовых токенов на один параметр модели. Это означает, что модель с 70 миллиардами параметров требует для оптимального обучения примерно 1,4 триллиона токенов [8].

Объем русскоязычного текста в интернете, по различным оценкам, составляет примерно 50–100 миллиардов токенов. Это означает, что при обучении суверенной российской модели исключительно на русскоязычных данных, полученных и обработанных в России, разработчики смогут получить примерно 1/15 или 1/30 от минимально необходимого объема данных для оптимального обучения. Такое радикальное сокращение объема обучающих данных неизбежно приведет к снижению качества модели, ее менее высокой способности к обобщению и меньшей адаптивности к различным задачам.

Попытка компенсировать этот дефицит через использование синтетических данных, созданных другими моделями, создает дополнительные проблемы. Во-первых, это может привести к накоплению ошибок при многократном использовании синтетических данных. Во-вторых, модель, обученная на синтетических данных, созданных менее качественной моделью, не может превзойти качество своего «учителя», что вызывает застой в развитии.

К тому же ограничение набора данных исключительно русскоязычным контентом и контентом, собранным и обработанным в России,

означает, что модель будет иметь систематически более слабые знания в ряде областей, где большинство высокозначимого контента опубликовано на других языках. Это в первую очередь касается естественных и технических наук, где большинство научных публикаций выпускается на английском языке, а также ряда гуманитарных дисциплин, где существует богатая традиция англоязычной литературы.

Второй уровень парадокса связан с инструментальной приемлемостью моделей. Даже если российским разработчикам удастся создать модель, обученную исключительно на российских данных, которая будет более-менее приемлема по качеству благодаря использованию передовых архитектур и значительных инвестиций в вычислительные мощности, подобная модель останется менее универсальной, чем ее глобальные аналоги. Это означает, что при прочих равных условиях, компании и организации, имеющие доступ как к российской, так и к глобальным моделям, выберут последние для критичных приложений.

Третий уровень парадокса касается образования и научных кадров. Разработчики LLM, обученные в российской системе образования и работающие в России, находятся в ситуации, при которой большинство передовых идей и методов в области разработки LLM публикуется и обсуждается в англоязычной научной коммуникации. Попытка развивать суверенные модели, используя исключительно русскоязычные научные источники, означает систематическое отставание от мировых тенденций развития указанного направления.

Американская модель в этом отношении оптимальна с точки зрения достижения конкурентоспособности, но представляет риски с точки зрения государственного контроля. Европейская модель пытается найти компромисс через требования к транспарентности и управлению данными, но не решает проблему технологической зависимости от американской облачной инфраструктуры. Китайская модель достигает относительного соответствия между суверенностью и контролем, но ценой отставания в конкурентоспособности. Российская модель сделает ставку на максимальную суверенность, сознательно принимая риски в отношении конкурентоспособности.

Однако следует отметить, что этот парадокс может быть частично смягчен или иначе переосмыслен при анализе долгосрочных стратегических целей. Американская система хотя и обеспечивает текущее лидерство в области LLM, но создает асимметричную зависимость: государства и компании других стран вынуждены полагаться на американские решения и американскую облачную инфраструктуру. Это означает потенциальную уязвимость в случае введения санкций, геополитических конфликтов или попыток «отключить» иностранных конкурентов от критичных сервисов. Более того, данные, используемые для обучения американских моделей, содержат информацию из всех регионов мира, что потенциально может привести

к проблемам в области конфиденциальности и управления персональной информацией граждан других стран.

Проведенный анализ четырех моделей подхода к источникам обучения LLM выявил следующие ключевые результаты:

во-первых, существует явный компромисс между суверенностью, качеством и конкурентоспособностью. Каждая из четырех моделей делает различный выбор в этом трехмерном пространстве:

– американская модель: приоритет отдается качеству и конкурентоспособности, минимизация суверенности (в смысле государственного контроля над источниками данных);

– европейская модель: попытка баланса между качеством и управлением данными, частичное сохранение суверенности через требования к прозрачности;

– китайская модель: выбор между суверенностью и контролем, принятие снижения качества как цены за независимость;

– российская модель: максимальный приоритет суверенности и контролю, осознанное принятие снижения качества как неизбежного следствия;

во-вторых, географическая локализация данных сама по себе не гарантирует суверенность моделей. Даже если все данные происходят из русскоязычного интернета и собраны российскими субъектами, это не означает, что модель будет иммунна к влиянию мировых тенденций в области информационных технологий. Более того, если модель будет менее конкурентоспособна, это может привести к де-факто зависимости от импортных моделей, что снижает практическую суверенность;

в-третьих, экосистемный фактор является критичным. Американское лидерство в области LLM объясняется не только доступом к лучшим данным, но и наличием сильной экосистемы: венчурный капитал, исследовательские учреждения, талантливые специалисты, облачная инфраструктура, университеты. Попытка разработать суверенные модели без соответствующей экосистемы поддержки вероятнее всего приведет к менее конкурентоспособным решениям;

в-четвертых, вопрос о нормативности данных — о том, какие ценности должны быть встроены в модель — остается нерешенным. EU AI Act обходит эту проблему, фокусируясь на управлении рисками, а не на содержании данных. Американский подход вообще не рассматривает эту проблему на уровне закона. Китайский подход решает ее через систему контроля интернета и цензуры. Российский подход пытается решить ее через явные критерии «духовно-нравственной приемлемости», но определение таких критериев остается неясным и потенциально проблематичным;

в-пятых, долгосрочные сценарии развития различны. Американские модели, вероятно, будут продолжать совершенствоваться благодаря доступу к глобальным данным и постоянной конкуренции. Европейские модели будут стремиться к специализированному, высокого качества

развитию в рамках жестких нормативных требований. Китайские модели будут развиваться в направлении локализированных и специализированных применений. Российские модели, если будут разработаны в соответствии с проектом закона, будут представлять собой специализированные, ориентированные на внутренний рынок системы, которые не будут конкурентоспособны на глобальном рынке, но могут быть приемлемы для внутреннего применения в государственном секторе и критичной инфраструктуре.

Вопрос об источниках обучения LLM является не только технической, но и правовой проблемой. Неодинаковые подходы, которые выбирают различные страны, отражают их уникальное видение баланса между инновацией, безопасностью и суверенностью. Опыт американской, европейской и китайской моделей может быть полезен при реализации российского подхода, не столько для того, чтобы копировать указанные модели, сколько для того, чтобы предусмотреть потенциальные последствия и попытаться найти решения, которые бы минимизировали отрицательные эффекты при максимизации положительных.

Список использованных источников:

1. Об основах государственного регулирования сфер применения технологий искусственного интеллекта в Российской Федерации: проект Федер. закона № 166424 // Федеральный портал проектов нормативных правовых актов. — URL: https://regulation.gov.ru/projects/166424/ (дата обращения: 19.03.2026).

2. Language Models are Few-Shot Learners / T.B. Brown, B. Mann, N. Ryder [et al.] // Advances in Neural Information Processing Systems 33 (NeurIPS 2020): [proceedings of the 34th Conference, 2020, December 6–12]. — 2020. — Vol. 33. — P. 1877–1901. — URL: proceedings.neurips.cc (date of access: 19.03.2026).

3. Common Crawl Foundation: [официальный сайт]. — Сан-Франциско. — URL: commoncrawl.org (дата обращения: 19.03.2026).

4. Radford, A. Language Models are Unsupervised Multitask Learners / A. Radford, J. Wu, R. Child [et al.] // OpenAI: [official website]. — 2019. — URL: cdn.openai.com (дата обращения: 19.03.2026).

5. Предложение по Регламенту Европейского парламента и Совета, устанавливающему гармонизированные правила в области искусственного интеллекта (Закон об искусственном интеллекте) и вносящему изменения в некоторые законодательные акты Союза: COM(2021) 206 final // EUR-Lex: [офиц. сайт]. — Брюссель, 2021. — URL: eur-lex.europa.eu (дата обращения: 19.03.2026).

6. Регламент (ЕС) 2024/1689 Европейского парламента и Совета от 13 июня 2024 г., устанавливающий гармонизированные правила в области искусственного интеллекта (Закон об искусственном интеллекте): [принят 13 июня 2024 г.] // Официальный журнал Европейского Союза. — 2024. — Т. 67. — URL: eur-lex.europa.eu (дата обращения: 19.03.2026).

7. Devlin, J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M.-W. Chang, K. Lee, K. Toutanova. — arXiv:1810.04805 [cs.CL]. — 2019. — 16 p. — URL: arxiv.org (date of access: 19.03.2026).

8. Training Compute-Optimal Large Language Models / J. Hoffmann, S. Borgeaud, A. Mensch [et al.] // arXiv:2203.15556 [cs.CL]. — 2022. — 29 p. — URL: arxiv.org (date of access: 19.03.2026).