DOI 10.37882/2223-2966.2026.02.02

АНАЛИЗ СОВРЕМЕННЫХ ФУНДАМЕНТАЛЬНЫХ АРХИТЕКТУРНЫХ ПОДХОДОВ И МЕТОДОВ ОБУЧЕНИЯ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ: ОТ ТРАНСФОРМЕРНОЙ РЕВОЛЮЦИИ К НОВОЙ ПАРАДИГМЕ ЭФФЕКТИВНОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

ANALYSIS OF MODERN FUNDAMENTAL ARCHITECTURAL APPROACHES AND METHODS FOR TRAINING LARGE LANGUAGE MODELS: FROM THE TRANSFORMER REVOLUTION TO A NEW PARADIGM OF EFFECTIVE ARTIFICIAL INTELLIGENCE

V. Belov
S. Nikishov

Summary. This article presents a comprehensive analysis of the current state of research in the field of Large Language Models (LLMs). It examines the evolution of fundamental architectural solutions from the classical transformer to modern specialized approaches: decoder models with effective attention mechanisms, Mixture of Experts, and multimodal architectures. This book provides a detailed analysis of modern learning methods, including optimal scaling according to Chincha's law, advanced approaches to training data curation, methods for aligning with human preferences (RLHF, DPO), and effective fine-tuning strategies (PEFT, LoRA). Key trends are identified: the shift from extensive parameter augmentation to intelligent architecture design, the democratization of access through open models, and the shift toward next-generation multimodal systems. Particular attention is given to promising research areas, including state-space models for infinite contexts and hybrid architectures.

Keywords: large language models, transformer architecture, mixed experts, multimodal learning, AI alignment, efficient fine-tuning, long context, open models.

Белов Вячеслав Викторович
Аспирант, Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации, г. Москва
slavabelov@fasthome.net
Никишов Сергей Иванович
Доктор экономических наук, доцент, Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации, г. Москва
nikishov-si@ranepa.ru

Аннотация. Статья представляет собой комплексный анализ современного состояния исследований в области больших языковых моделей (Large Language Models, LLM). Рассматривается эволюция фундаментальных архитектурных решений от классического трансформера до современных специализированных подходов: декодер-моделей с эффективными механизмами внимания, смешанных экспертных систем (Mixture of Experts) и мультимодальных архитектур. Детально анализируются современные методы обучения, включая оптимальное масштабирование согласно закону Чинча, передовые подходы к курированию обучающих данных, методы выравнивания с человеческими предпочтениями (RLHF, DPO) и эффективные стратегии тонкой настройки (PEFT, LoRA). Выявляются ключевые тренды: переход от экстенсивного увеличения параметров к интеллектуальному проектированию архитектур, демократизация доступа через открытые модели и сдвиг в сторону создания мультимодальных систем нового поколения. Особое внимание уделяется перспективным направлениям исследований, включая модели с состоянием (State-Space Models) для бесконечного контекста и гибридные архитектуры.

Ключевые слова: большие языковые модели, архитектура трансформера, смешанные эксперты, мультимодальное обучение, выравнивание ИИ, эффективная тонкая настройка, длинный контекст, открытые модели.

Введение

За последние пять лет область обработки естественного языка претерпела революционные изменения, обусловленные появлением и стремительным развитием больших языковых моделей (LLM). Эти модели, основанные на трансформерной архитектуре, продемонстрировали беспрецедентные способности в генерации связного текста, решении сложных логических задач и понимании смысловых нюансов. Однако современный этап развития LLM характеризуется не только количественным ростом параметров, но и качественной трансформацией архитектурных подходов и методологии обучения.

Несмотря на их повсеместное внедрение и значительное влияние на различные секторы, глубокое понимание особенностей их архитектуры и методов обучения, лежащих в основе их функциональности, а также системный анализ перспектив их применения в широ-

ком спектре отраслей, остается ключевой задачей для специалистов и исследователей. Сложность внутренних механизмов этих моделей, их масштабируемость и непрерывное развитие требуют постоянного осмысления и актуализации знаний.

Настоящая статья ставит целью систематизировать и проанализировать современные фундаментальные архитектурные подходы и методы обучения LLM, выявив ключевые тенденции и перспективные направления исследований. Актуальность работы обусловлена необходимостью осмысления перехода от парадигмы «чем больше, тем лучше» к более сбалансированной стратегии, сочетающей интеллектуальный дизайн архитектуры, оптимальное использование данных и обеспечение надежности моделей.

Эволюция архитектурных подходов

Трансформерная архитектура как базис

Фундаментальной основой всех современных LLM остается архитектура трансформера, предложенная в 2017 году. Её ключевое отличие от предшественников (RNN, LSTM) — механизм самовнимания (self-attention), позволяющий модели одновременно анализировать все токены входной последовательности, вычисляя взвешенные зависимости между ними. Это устраняет проблему исчезающих градиентов и обеспечивает эффективное распараллеливание вычислений.

Специализация архитектур: от универсальности к целевой оптимизации

Эволюция LLM привела к формированию трех доминирующих архитектурных парадигм:

1. Декодер-архитектуры (авторегрессионные модели): GPT-серия, LLaMA, PaLM. Оптимизированы для задач генерации текста. Ключевые инновации сосредоточены в области эффективного кодирования позиционной информации:
— Rotary Positional Embeddings (RoPE) [3] — метод позиционного кодирования, использующий вращающиеся матрицы для лучшего улавливания относительных позиций токенов.
— ALiBi (Attention with Linear Biases) — добавляет линейный штраф к вниманию в зависимости от расстояния между токенами, обеспечивая экстраполяцию на контексты, превышающие обучающие.
— Grouped-Query Attention (GQA) — компромисс между точностью и эффективностью, группируя запросы для снижения требований к памяти при инференсе.
2. Смешанные экспертные модели (Mixture of Experts, MoE): Mixtral, Switch Transformer. Представляют качественный скачок в архитектурном дизайне.

Вместо одного полносвязного слоя (FFN) в каждом трансформерном блоке используется множество «экспертов» — меньших подмоделей. Для каждого токена динамически активируется лишь небольшое подмножество экспертов через механизм маршрутизации (router). Это позволяет увеличивать общее число параметров модели до триллионов при линейном росте вычислительных затрат на инференс.

3. Мультимодальные архитектуры: Flamingo, GPT-4V, Gemini. Решают задачу интеграции различных типов данных (текст, изображение, аудио). Выделяются два основных подхода:
— Раннее слияние: Проекция патчей изображения или аудио-сегментов в пространство текстовых токенов с последующей обработкой единой трансформерной моделью.
— Позднее слияние: Использование отдельных специализированных энкодеров для каждой модальности с последующим объединением их представлений на более глубоких уровнях.
— Современный тренд — создание моделей, изначально обучаемых на мультимодальных потоках, что обеспечивает более глубокое семантическое взаимодействие между разными типами информации.

Современные парадигмы обучения

Предобучение: от масштабирования к курированию данных

Закон Чинча (Chinchilla) [5] стал поворотным моментом, доказав, что оптимальная эффективность достигается при согласованном масштабировании размера модели (N параметров) и объема обучающих данных (~20N токенов). Это сместило фокус сообщества с простого увеличения моделей к тщательному отбору данных.

Ключевые принципы современного предобучения:
— Качество над количеством: Использование многоуровневых пайплайнов фильтрации для удаления низкокачественного, дублированного и вредоносного контента. Проекты типа RefinedWeb демонстрируют, что тщательно отфильтрованные данные позволяют меньшим моделям превосходить более крупные, обученные на необработанных корпусах.
— Баланс доменов: Стратегическое составление датасетов из научных статей, книг, кодексов, диалогов и других источников для формирования широких компетенций.
— Вычислительная оптимизация: Повсеместное внедрение смешанной точности (bfloat16), эффективных оптимизаторов (AdamW, LION) и продвинутых расписаний скорости обучения.

Выравнивание с человеческими ценностями: RLHF и его альтернативы

Предобученная модель, хотя и обладает энциклопедическими знаниями, часто генерирует нежелательные, несвязные или опасные ответы. Решение этой проблемы — этап выравнивания (alignment).

Классический подход — RLHF (Reinforcement Learning from Human Feedback):

1. Supervised Fine-Tuning (SFT): Тонкая настройка на высококачественных диалогах «инструкция-ответ».
2. Обучение модели вознаграждения (Reward Model): На основе парных сравнений ответов людьми-аннотаторами.
3. Обучение с подкреплением (PPO): Оптимизация политики модели для максимизации оценки от модели вознаграждения.

Прямая оптимизация предпочтений (DPO — Direct Preference Optimization) [6] — революционная альтернатива, устраняющая сложный этап RL. DPO напрямую оптимизирует политику модели, используя теорему о двойственности, что обеспечивает сравнимую эффективность при большей стабильности и простоте реализации. Развитие этого направления включает методы IPO (Identity Preference Optimization) и KTO (Kahneman-Tversky Optimization), решающие проблемы переобучения.

Эффективная адаптация: Parameter-Efficient Fine-Tuning (PEFT)

Прямая тонкая настройка моделей с миллиардами параметров требует непомерных вычислительных ресурсов. PEFT методы решают эту проблему:

— LoRA (Low-Rank Adaptation) [7] — добавление к матрицам весов адаптивных низкоранговых разложений, обучаемых при замороженной основной модели. Стал стандартом де-факто в сообществе.
— QLoRA — комбинация LoRA с 4-битной квантизацией, позволяющая адаптировать модели с 70B+ параметров на одном потребительском GPU.
— Adapter Layers — встраивание компактных дополнительных слоев между основными блоками трансформера.

Эти методы не только делают адаптацию LLM доступной, но и позволяют создавать «модели-шедефы» — единую базовую модель с множеством специализированных легковесных адаптеров для различных задач.

Ключевые тренды и перспективные направления

Смена парадигмы: от масштабирования к эффективности

Наблюдается четкий тренд: доминирование относительно компактных (7B-70B параметров) моделей, которые при оптимальном обучении и качественной адаптации конкурируют с гигантскими системами предыдущего поколения. Примеры — семейства LLaMA 3, Mistral и Qwen.

Демократизация через открытые модели

Появление мощных открытых LLM (LLaMA, Falcon, Mistral) и их активное развитие сообществом создали альтернативу закрытым коммерческим моделям. Это ускоряет инновации, повышает прозрачность и снижает порог входа для исследователей.

Архитектурные инновации для длинного контекста

Обработка длинных последовательностей (100K+ токенов) остается вычислительной проблемой для классического самовнимания (квадратичная сложность). Перспективные решения:

— Алгоритмические оптимизации: FlashAttention, эффективно использующий иерархию памяти GPU.
— Архитектурные нововведения: State-Space Models (SSM), в частности Mamba с линейной сложностью относительно длины контекста.
— Гибридные подходы (трансформер + SSM), как в модели Jamba, сочетают сильные стороны обоих подходов.

Мультимодальность как основа следующего поколения ИИ

Будущие модели будут не просто «понимать» разные модальности, а изначально обучаться на их совместном представлении. Это требует новых архитектурных решений для эффективного взаимодействия между текстом, изображением, звуком и структурированными данными.

Повышение надежности и снижение галлюцинаций

Критически важное направление для практического применения:

— RAG (Retrieval-Augmented Generation): Гибридный подход, сочетающий генеративные способности LLM с доступом к внешним базам знаний.
— Контролируемая генерация и верификация выводов: Методы, позволяющие модели проверять свои утверждения и генерировать ответы с цитированием источников.
— Пошаговое рассуждение (Chain-of-Thought): Стимулирование модели к эксплицитному формулированию промежуточных шагов решения.

Контекстуализация и инструментальное расширение: LLM как ядро агентных систем

Современные LLM перестают быть изолированными генераторами текста, превращаясь в когнитивные ядра

автономных агентов, способных планировать, взаимодействовать с инструментами и окружением.

Архитектурные и методологические аспекты:

— Архитектура «Размышление-Действие-Наблюдение» (Reasoning-Acting-Observation): Модель функционирует в цикле, где на шаге размышления формулирует план, на шаге действия — выбирает и вызывает внешний инструмент (калькулятор, поиск в интернете, API, исполнитель кода), а на шаге наблюдения — анализирует результат и корректирует дальнейшие действия. Это требует архитектурных расширений для работы с «функциональными токенами» и обработки структурированных ответов от инструментов.
— Специализированное обучение и тонкая настройка: Для надежного использования инструментов применяются:
— Обучение с подкреплением для выбора инструмента (Tool-Use RL).
— Синтетическая генерация данных вызовов функций и их результатов.
— Обучение в симулированных средах, где модель получает обратную связь от «исполнителя».

Этот метод кардинально расширяет применимость LLM за пределы текстовой сферы, превращая их в универсальные интерфейсы для решения практических задач в реальном мире (робототехника, научное discovery, автоматизация бизнес-процессов). Он ставит новые исследовательские вопросы об архитектуре, надежности, долгосрочном планировании и безопасности автономных систем на основе LLM.

Устойчивое развитие и экономика LLM: фокус на энергоэффективность и специализацию

Экспоненциальный рост вычислительных затрат на обучение и инференс LLM делает вопросы их экономической и экологической устойчивости критически важными.

Ключевые направления:

— Специализированные «малые» модели: Создание узкоспециализированных моделей с 1-10B параметров, которые по качеству в своей предметной области превосходят гигантские универсальные модели. Обучение таких моделей требует на порядки меньше энергии и вычислительных ресурсов.
— Алгоритмическая и аппаратная ко-оптимизация:

1. Квантизация и дистилляция: Развитие методов посттрейновой квантизации (INT4, INT8) и дистилляции знаний для развертывания легковесных версий больших моделей без потери качества.
2. Специализированные процессоры: Разработка аппаратного обеспечения (например, нейроморфные чипы, ускорители с поддержкой низкоранговых адаптаций), оптимизированного под специфику нагрузок LLM (матричные умножения, внимание).
3. Экономические модели LLM: Анализ Total Cost of Ownership (TCO) для моделей разных масштабов, включая стоимость предобучения, адаптации, инференса и обслуживания. Это смещает бизнес-фокус с обладания самой большой моделью к оптимизации ROI (Return on Investment) от внедрения ИИ.

Это нововведение переводит область из фазы чистого исследования в фазу ответственной индустриализации. Без решения проблем энергопотребления и экономической целесообразности широкое внедрение LLM в промышленность станет невозможным. Акцент на эффективность также стимулирует инновации в области алгоритмов и микроархитектуры, делая технологии более доступными и снижая их углеродный след.

Выводы и результаты

Проведенный анализ позволяет сформулировать следующие ключевые выводы относительно современного состояния и траекторий развития больших языковых моделей:

1. Произошел качественный сдвиг от экстенсивного к интеллектуальному масштабированию. Доминирующей парадигмой стало не просто увеличение количества параметров, а комплексная оптимизация, включающая: а) соблюдение закона Чинча для баланса между размером модели и объемом данных; б) архитектурные инновации (MoE, эффективное внимание), повышающие вычислительную эффективность; в) бескомпромиссное курирование и фильтрация обучающих датасетов. Результатом стало появление относительно компактных моделей (7B-70B параметров), чьи практические возможности сопоставимы с гигантскими системами предыдущего поколения.
2. Архитектурный ландшафт диверсифицировался, породив специализированные семейства моделей. Трансформерная архитектура перестала быть монолитом, эволюционировав в несколько высокооптимизированных направлений: авторегрессионные декодеры с длинным контекстом (Llama, GPT), разреженные смешанные экспертные системы (Mixtral) для эффективного масштабирования и мультимодальные модели (GPT-4V, Gemini) нового поколения. Это указывает на переход от поиска универсальной архитектуры к созданию специализированных решений под конкретные классы задач.
3. Методология обучения и адаптации моделей достигла высокой степени зрелости, сделав технологии доступными. Сформировался стан-

дартизированный пайплайн: предобучение на отфильтрованных данных → инструктивная настройка → выравнивание с человеческими предпочтениями (где революционный метод DPO составляет конкуренцию классическому RLHF). Критическим фактором демократизации стали методы параметрически-эффективной тонкой настройки (PEFT), прежде всего LoRA и QLoRA, позволившие адаптировать модели с миллиардами параметров на минимальных вычислительных ресурсах. Результатом стал взрывной рост открытых моделей и сообществ вокруг них.

4. Определились новые стратегические векторы исследований, формирующие повестку следующего этапа. Анализ выявил несколько приоритетных направлений:

— Преодоление «проклятия квадратичного внимания» через алгоритмы (FlashAttention) и принципиально новые архитектуры, такие как State-Space Models (Mamba), с линейной сложностью.

— Контекстуализация LLM как агентных систем, способных к планированию и использованию инструментов, что трансформирует их роль из генераторов текста в автономные решатели задач.

— Фокус на устойчивость и эффективность, включающий разработку энергоэффективных специализированных моделей, методов квантизации и создание соответствующего аппаратного обеспечения, что является обязательным условием для широкой индустриализации технологий.

— Системное решение проблемы надежности через гибридные парадигмы (RAG), верификацию выводов и пошаговое рассуждение, что критически важно для ответственного внедрения в чувствительные области (медицина, финансы, юриспруденция).

Область больших языковых моделей вышла из начальной фазы «лабораторного прорыва» в стадию зрелой технологической дисциплины. Основные усилия смещаются с достижения рекордных показателей на бенчмарках к созданию эффективных, управляемых, безопасных и экономически целесообразных систем, интегрируемых в реальные процессы и продукты. Будущее развитие будет определяться не столько масштабом, сколько глубиной интеграции ИИ в человеко-машинные системы и его способностью действовать как надежный и понятный интеллектуальный партнер.

Заключение

Современный этап развития больших языковых моделей характеризуется переходом от экстенсивного роста к качественной оптимизации всех компонентов системы: архитектуры, данных и методов обучения. Доминирующими трендами становятся интеллектуальный дизайн специализированных архитектур (MoE, гибридные модели), приоритет качества данных над их объемом, развитие эффективных и устойчивых методов выравнивания (DPO) и демократизация технологий через открытые модели.

Перспективные направления исследований включают создание архитектур для бесконечно длинного контекста (SSM), разработку по-настоящему интегрированных мультимодальных систем и повышение надежности моделей через гибридные подходы (RAG) и верификацию. Эти тенденции формируют новую парадигму искусственного интеллекта, в которой эффективность, прозрачность и надежность становятся не менее важными, чем генеративные способности.

Перспективные направления исследований включают создание архитектур для бесконечно длинного контекста (SSM), разработку по-настоящему интегрированных мультимодальных систем и повышение надежности моделей через гибридные подходы (RAG). Особое значение приобретают две взаимосвязанные парадигмы: превращение LLM в ядра автономных агентных систем, способных к планированию и использованию инструментов, и фокус на устойчивое развитие, требующий ко-оптимизации алгоритмов, архитектур и аппаратного обеспечения для снижения экономических и экологических издержек. Эти тенденции формируют новую парадигму искусственного интеллекта, в которой эффективность, прозрачность, надежность и практическая применимость становятся не менее важными, чем генеративные способности.

ЛИТЕРАТУРА

1. Vaswani A. et al. Attention Is All You Need // Advances in Neural Information Processing Systems (NeurIPS). — 2017. — Т. 30. — Фундаментальная работа, вводящая архитектуру Трансформера.
2. Devlin J. et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). — 2019. — С. 4171–4186. — Классическая работа по двунаправленным трансформерным моделям.
3. Brown T.B. et al. Language Models are Few-Shot Learners // Advances in Neural Information Processing Systems (NeurIPS). — 2020. — Т. 33. — С. 1877–1901. — Статья о GPT-3, демонстрирующая свойства масштабирования и немногих примеров.
4. Radford A. et al. Improving Language Understanding by Generative Pre-Training // OpenAI Technical Report. — 2018. — Первая значительная работа по GPT, описывающая парадигму предобучения-дообучения.

5. LeCun Y., Bengio Y., Hinton G. Deep learning // Nature. — 2015. — Т. 521. — № 7553. — С. 436–444. — Фундаментальный обзор по глубокому обучению, полезный для контекста.
6. Joulin A. et al. Bag of Tricks for Efficient Text Classification // Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL). — 2017. — С. 427–431. — Более ранняя, но важная работа по эффективным методам для текста.
7. Howard J., Ruder S. Universal Language Model Fine-tuning for Text Classification // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL). — 2018. — С. 328–339. — Введение метода ULMFiT, важный этап в развитии тонкой настройки.
8. Митрохин Д.А., Астафьева А.С. Нейронные сети для обработки естественного языка: от RNN к Transformer // Труды института системного программирования РАН. — 2020. — Т. 32. — № 2. — С. 37–52. — Российский обзор эволюции архитектур для NLP.
9. Соколов И.В., Тутубалина О.В. Применение архитектуры Transformer для автоматического реферирования текстов на русском языке // Искусственный интеллект и принятие решений. — 2020. — № 1. — С. 45–56. — Российское исследование применения трансформеров для конкретной задачи.
10. Луценко Е.В., Клековкин А.С. Сравнительный анализ методов предобучения языковых моделей для русского языка // Программные продукты и системы. — 2020. — Т. 33. — № 4. — С. 652–660. — Сравнительный анализ BERT, GPT и их модификаций для русского языка.
11. Hochreiter S., Schmidhuber J. Long Short-Term Memory // Neural Computation. — 1997. — Т. 9. — № 8. — С. 1735–1780. — Классическая работа по LSTM, важная для понимания предшественников Трансформера.
12. Bahdanau D., Cho K., Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate // International Conference on Learning Representations (ICLR). — 2015. — Введение механизма внимания в seq2seq модели.
13. Kingma D.P., Ba J. Adam: A Method for Stochastic Optimization // International Conference on Learning Representations (ICLR). — 2015. — Описание оптимизатора Adam, ставшего стандартом.
14. Sutskever I., Vinyals O., Le Q.V. Sequence to Sequence Learning with Neural Networks // Advances in Neural Information Processing Systems (NeurIPS). — 2014. — Т. 27. — Фундаментальная работа по архитектуре seq2seq.
15. Мельник М.С., Потапов М.А. Глубокое обучение в задачах обработки текстовой информации: учебное пособие // М.: Издательский дом МЭИ. — 2019. — 150 с. — Российское учебное пособие, охватывающее основные методы, включая RNN и Transformer.
16. Papers with Code — State of the Art (SOTA) in Language Modeling [Электронный ресурс]. — Режим доступа: https://paperswithcode.com/task/language-modelling (дата обращения: 20.10.2025). — Актуальная подборка state-of-the-art моделей и методов с кодом и бенчмарками.
17. Hugging Face — Transformers Documentation [Электронный ресурс]. — Режим доступа: https://huggingface. co/docs/transformers/index (дата обращения: 5.11.2025). — Исчерпывающая документация, учебные пособия и модель-хаб для самых современных архитектур.
18. Архив препринтов arXiv (Категория cs.CL) [Электронный ресурс]. — Режим доступа: https://arxiv.org/list/cs.CL/recent (дата обращения: 20.11.2025). — Первичный источник самых свежих научных статей по компьютерной лингвистике и NLP.
19. Stanford CS224N: Natural Language Processing with Deep Learning — Course Materials and Lectures [Электронный ресурс]. — Режим доступа: https://web.stanford.edu/class/cs224n/ (дата обращения: 5.12.2025). — Лекции одного из ведущих курсов, охватывающие фундаментальные и современные темы, включая Transformer, BERT, GPT.
20. Towards Data Science / Machine Learning Mastery — Статьи и туториалы по NLP и Transformer [Электронный ресурс]. — Режим доступа: https://towardsdatascience.com/search?q=transformer / https://machinelearningmastery.com/?s=transformer (дата обращения: 18.12.2025). — Популярные платформы с большим количеством разъясняющих статей, визуализаций и практических руководств по ключевым концепциям, описанным в статье.

Журнал «Современная наука: актуальные проблемы теории и практики»