

---

## ПЛЕНАРНЫЕ ДОКЛАДЫ

УДК 004.8

doi: 10.15622/rcai.2025.001

### ПРИОРИТЕТНЫЕ НАПРАВЛЕНИЯ ИССЛЕДОВАНИЙ И КЛЮЧЕВЫЕ ТЕНДЕНЦИИ РАЗВИТИЯ ТЕХНОЛОГИЙ ИИ

Ю.В. Визильтер (*viz@gosniias.ru*)

ФАУ «ГосНИИАС», Москва

Представлены тенденции развития методов и технологий ИИ на текущем этапе (2020-2025), сгруппированные по нескольким приоритетным направлениям исследований в сфере ИИ. Кратко описаны основные тенденции и результаты по следующим ключевым направлениям и поднаправлениям: LLM и другие модели для символьных данных, диффузионные и другие модели для несимвольных данных, мультимодальные модели, методы переноса знаний с адаптацией моделей, аугментация LLM без адаптации моделей, обучение с подкреплением, агентные и мультиагентные системы, элементы общего ИИ (AGI).

**Ключевые слова:** Artificial Intelligence, Machine Learning, Computer Vision, NLP, LLM, Generative AI, RL, General AI.

#### Введение

Начало нынешнего периода развития технологий искусственного интеллекта (ИИ) принято отсчитывать от 2011 г. В рамках данного периода можно условно выделить три этапа. На первом этапе (2011-2016) в центре внимания находились сверточные нейронные сети (CNN) [Krizhevsky et al., 2017]. На втором этапе (2016-2020) получил широкое распростране-

ние ряд новых подходов, таких как GAN для синтеза изображений [Goodfellow et al., 2014], архитектуры типа Transformer [Vaswani et al., 2017], GPT [Radford et al., 2018], BERT [Devlin et al., 2018], на основе т.н. «модулей внимания» (Attention [Vaswani et al., 2017]) для задач NLP, обучение с подкреплением [Mnih et al., 2015], [Schulman et al., 2017], [Badia et al., 2020]), графовые модели (Graph CNN [Zhang et al., 2019]), автоматическое обучение (Auto-ML, NAS, НРО [Xin et al., 2021]).

Ниже будут представлены основные тенденции развития технологий ИИ на текущем, третьем этапе (2020-2025), сгруппированные по ряду приоритетных направлений и поднаправлений фундаментальных и поисковых исследований в сфере ИИ.

### **Приоритетные направления фундаментальных и поисковых исследований в сфере искусственного интеллекта**

В рамках Форсайта по приоритетным направлениям фундаментальных/поисковых исследований в сфере искусственного интеллекта [AI Foresight, 2024], подготовленного в 2024 г. группой ведущих российских экспертов по инициативе Аналитического центра при правительстве РФ, Сбера и НИУ ВШЭ, были выделены следующие основные приоритетные направления исследований.

*Н1. Архитектуры, алгоритмы машинного обучения (МО), оптимизация и математика*, в том числе: разработка новых алгоритмов МО, поиск архитектур глубоких сетей, ускорения вычислений, распределенное и федеративное обучение.

*Н2. Вычисления для ИИ*, в том числе: разработка вычислителей для ИИ (квантовые, фотонные, нейроморфные, Edge), разработка АПК для ИИ, фреймворки для МО и ИИ;

*Н3. Данные для ИИ*, в том числе: создание бенчмарков для оценки ИИ, создание и аугментация данных (синтетика, зашумление), сохранение конфиденциальности данных.

*Н4. Фундаментальные и генеративные модели*, в том числе: LLM и др. модели для символьных данных, диффузионные и др. модели для не символьных данных, мультимодальные LLM модели, методы переноса знаний, аугментация LLM без адаптации моделей.

*Н5. Безопасность, доверие и объяснимость*, в том числе: выравнивание ценностей (Alignment), объяснимость работы ИИ, обеспечение безопасности разработки и эксплуатации ИИ, обеспечение защиты от результатов использования ИИ.

*Н6. ИИ для узких задач (Narrow AI)*, в том числе: CV (компьютерное зрение), NLP (обработка естественного языка), прочие технологии узкого ИИ.

*Н7. Управление, принятие решений, агентные/мультиагентные системы*, в том числе: разработка алгоритмов обучения с подкреплением (RL), агентные системы, мультиагентные системы.

*Н8. Элементы общего ИИ (AGI)*, в том числе: рассуждения и рефлексия, гибридный ИИ (Symbolic AI), воплощенный ИИ (Embodiment), моделирование мозга и психики.

*Н9. Взаимодействие человека и машины*, в том числе: технические средства человеко- машинного взаимодействия, методы и алгоритмы взаимодействия с человеком, способы человеко-машинной интеграции.

Далее будут кратко описаны основные тенденции и результаты по направлениям Н4, Н7 и Н8, которые представляются ключевыми в том смысле, что развитие остальных направлений в значительной степени связано именно с проблемами и достижениями в этих областях.

## **Н4. Фундаментальные и генеративные модели**

### **Н4.1. LLM и другие модели для символьных данных**

*Большие языковые и фундаментальные модели.* В 2020 г. трансформер GPT-3 [Brown et al., 2020] стал первой из класса больших языковых моделей (Large Language Models, LLM) с многими миллиардами параметров. В 2021 г. было предложено понятие *фундаментальных моделей* (Foundation Models, FM) [Rishi et al., 2021], предобученных на том количестве данных, что они далее не требуют или требуют минимального дообучения. В 2022 г. создан ChatGPT, который сделал работу с LLM полезной и удобной за счет дообучения LLM методом RLHF (Reinforcement Learning from Human Feedback) [Ouyang et al., 2022].

*Повышение вычислительной эффективности LLM.* Предложен целый ряд подходов, основанных на квантовании и прореживании (прунинг) весов моделей; дистилляции в модели меньшего размера; быстром предсказании и последующей проверке результатов генерации (speculative decoding); оптимизации KV-кэша в модулях внимания; смеси экспертов (Mixture of Experts, MoE) и др. [Wan et al., 2024]. Многие из этих приемов были использованы и развиты в работах DeepSeek [DeepSeek-V2], [DeepSeek-V3], что позволило резко снизить стоимость работы публично доступных LLM.

*Развитие трансформерных архитектур.* В последние годы архитектура LLM достаточно активно эволюционировала. Были предложены, в частности, такие архитектурные элементы как Multi-Head Latent Attention (MLA), Grouped-Query Attention (GQA), RMSNorm, Post-Norm, QK-Norm, sliding window attention, MatFormer, Per-Layer Embedding (PLE), RoPE (Rotational

Positional Embeddings) и NoPE (No Positional Embeddings), SwiGLU и др. Хороший русскоязычный обзор современных трансформерных архитектур LLM со ссылками можно найти в [LLM Architecture Evolution, 2025].

*Альтернативные архитектуры.* Предложен целый ряд как альтернативных трансформерам, так и гибридных архитектур – Retentive Network [Sun et al., 2023], RWKV [Peng et al., 2023], State Space Models [Gu et al., 2021], Mamba [Gu et al., 2023], Jamba [Lieber O. et al., 2024], Hyena [Poli et al., 2023]. В работе 2024 г. Titans [Behrouz et al., 2024] предложен подход к созданию модулей обработки информации для больших моделей, основанный на понимании процесса обучения как запоминания во время исполнения (Test-Time Memorization). Этот подход был далее обобщен и развит для создания целого спектра различных альтернативных модулей LLM [Behrouz et al., 2025], [Wang et al., 2025]. Делаются также попытки разработать принципиальные альтернативы не только трансформерам, но и традиционным нейросетевым архитектурам [Halverson J. et al., 2024].

*Concept Models.* Возможное направление совершенствования архитектуры LLM для достижения более абстрактного представления знаний показывает подход Large Concept Models [Barrault L. et al, 2024], в котором уровень «концепций» и уровень реализующих их «слов» архитектурно разделены.

*Текстовые диффузионные модели.* Разработаны диффузионные и потоковые генеративные модели для символьных данных, предполагающие отход от авторегрессионной модели трансформеров для последовательной генерации токенов. Текстовая диффузия реализует модель итеративного «восстановления» генерируемого текста из массива маскированных токенов [Shi et al., 2024]. В 2025 году текстовая диффузионная модель впервые показала результаты, сравнимые с результатами LLM того же размера [Nie et al., 2025]. Модели типа GFlowNets [Bengio et al., 2021] позволяют генерировать объекты, обладающие структурой. Диффузионные модели и GFlowNets могут использоваться, в частности, для создания цепочек рассуждений [Ye et al., 2024], [Takase et al., 2024], [Ho et al., 2024].

## **Н4.2. Диффузионные и другие модели для несимвольных данных**

*Переход от генеративно-сопоставительных сетей к диффузионным моделям.* В 2020 г. за первенство среди генеративных моделей с GAN боролись также вариационные автоэнкодеры (Variational Auto-Encoder, VAE [Kingma et al., 2013], [Child, 2020]) и диффузионные модели (Diffusion Models [Ho et al., 2020], [Song et al.]). Сегодня диффузионные модели используются в большинстве работ и приложений [Ling et al., 2022]. С их помощью (в сочетании с трансформерами) решаются задачи гене-

рации изображений по текстовому описанию (DALL-E 2 [Aditya et al., 2022], DALL-E 3 [Improving Image Generation, 2024], Stable Diffusion 3 [Esser et al., 2024]), генерации видео по текстовому описанию (Imagen Video [Ho et al., 2022], OpenAI SORA [Brooks et al., 2024]). В 2025 г. DeepMind представлена система Veo3 [DeepMind VEO, 2025], демонстрирующая возможности симуляции достоверного физического поведения объектов, веществ и персонажей, а также синхронной генерации видеоряда и звука (звуковые эффекты, фоновая музыка, диалоги).

*Обучаемые 3D модели трехмерных сцен* позволили соединить технологии 3D рендеринга сцен с машинным обучением. В 2021 г. был предложен метод NeRF (neural radiance fields, [Brooks et al., 2024]), который начал широко использоваться не только в задачах генерации сцен, но и в задачах компьютерного зрения. В 2023 г. был предложен альтернативный подход 3D Gaussian Splatting (3DGS) [Kerbl et al., 2023], который обучается гораздо быстрее, работает в реальном времени и обеспечивает сравнимое или лучшее качество рендеринга по сравнению с NeRF [Chen et al., 2024]. Для генерации 3D сцен по текстовым запросам (text-to-3D) в настоящее время используется 3D GS в соединении с LLM (GALA3D, 2024) [Zhou et al., 2024]. В работе [Zielonka et al., 2023] модель 3D GS также используется для генерации и рендеринга в реальном времени реалистичных 3D аватар (подвижных моделей тела) людей.

### Н4.3 Мультимодальные LLM-модели

Возможность создания мультимодальных LLM-моделей была обеспечена, в первую очередь, за счет создания *фундаментальных моделей для задач зрения*: Segment Anything (SAM) [Kirillov et al., 2023] для семантической сегментации, DINO [Liu et al., 2023], DINOv2 [Oquab et al. 2024] для обнаружения объектов с открытым списком классов и др. Фундаментальные модели для задач зрения могут быть также построены на основе сверточных сетей (YOLO-World, [Cheng et al., 2024]), что вполне обосновано, поскольку лучшие практические решения по обнаружению объектов для бортовых приложений основаны сегодня на гибридных архитектурах сверточных сетей с модулями внимания (YOLOv12, [Tian et al., 2025]).

*Мультимодальные LLM-модели (MLLM)* [Yin et al., 2024] используются для решения задач, требующих не только анализа изображения, но и понимания контекста, то есть, анализа на уровне модели мира. При этом, как показано в [Wang et al., 2024], эффективная MLLM может быть моделью LLM среднего или небольшого размера, соединенной с фундаментальной моделью для зрения при помощи адаптера.

#### **Н4.4. Методы переноса знаний с адаптацией моделей**

Для повышения вычислительной эффективности обучения LLM были предложены методы т.н. параметрически эффективной настройки (PEFT), позволяющие обучать лишь небольшое подмножество параметров предварительно обученной модели [Wu et al., 2024]. К таким методам относятся, в частности, методы низкоранговой адаптации и спектральной переметризации.

*Low-Rank адаптеры.* Низкоранговая декомпозиция матриц весов и запросов (prompts) для уменьшения числа обучаемых параметров. Это такие уже широко используемые методы как LoRA [Hu et al., 2022] и LLM-адаптеры [Hu et al., 2023]. Работы, развивающие данное направление: LoRA<sup>2</sup> [Zhang et al., 2024], Low-Rank Prompt Adaptation [Jain et al., 2024], RankAdaptor [Zhou et al., 2024]. Перспективным является также использование дистиллированных моделей в комбинации с LoRA [DeepSeek-AI, 2025].

*Спектральная параметризация* предполагает использование для тех же целей спектрального разложения весов и адаптации моделей в спектральном пространстве: Spectral Adapter [Zhang et al. 2024], LaMDA [Azizi et al., 2024].

#### **Н4.5. Аугментация LLM без адаптации моделей**

В 2023 г. модели GPT-4 [Achiam et al., 2023], на порядок превосходящей GPT-3 по размеру, удалось показать качество ответов на запросы на уровне людей-профессионалов. С этого момента в фокусе исследования оказались способы адаптации LLM без дообучения (изменения) весов модели. Ключевыми технологиями здесь являются: инженерия запросов (Prompt Engineering, PE) [Sahoo et al., 2024], [Your Guide to Generative AI, 2023], [Prompt Engineering Guide, 2023], включая работу с базами документов Retrieval-Augmented Generation (RAG) [Lewis et al., 2020], [Gao et al., 2024], контекстное обучение (in-Context Learning, iCL) [Dong et al., 2023], логические рассуждения в LLM [Sahoo et al., 2024], [Wang et al., 2022], [Yao et al., 2023], [Yao et al., 2023], создание и использование генеративных агентов (GA).

### **Н7. Управление, принятие решений, агентные/мультиагентные системы**

#### **Н7.1. Разработка алгоритмов RL (обучения с подкреплением)**

*Обучение с открытым списком виртуальных сред и целевых задач* для приобретения когнитивного поведения. Авторы работы Open-Ended Learning (2021) [Adam et al., 2021] показали, что, если построить вселенную игровых задач и последовательно обучать ИИ-агентов играть в эти игры, то с каждой новой игрой они будут достигать лучших результатов в этой вселенной и за ее пределами за счет овладения навыками когнитивного поведения.

*Универсальные агенты для робототехники* на основе трансформеров и LLM. В работе GATO (2022) [Scott et al., 2022] был представлен универсальный агент-трансформер, способный играть в игры Atari, давать текстовое описание изображений, вести языковой чат, управлять рукой робота, представляющего блоки и т.п. В [Bousmalis et al., 2023] описана Vision-Language-Action (VLA) модель RoboCat для управления манипуляторами, обученная методом открытого обучения в реальном мире. VLA модель RT-2 [Brohan et al., 2023] на основе LLM сочетает управление роботом с рассуждениями на основе chain-of-thought. В 2023-2025 гг. создан ряд фундаментальных VLA-моделей: для самообучения физических роботов AutoRT [Brohan et al., 2023], для мультимодальной навигации Uni-NaVid [Zhang et al., 2024], для общего управления роботами  $\pi 0$  [Black et al., 2024].

*Совместное использование RL и LLM* является одним из основных трендов современного машинного обучения. При этом лучшее качество результатов достигается как в случае использования LLM для реализации RL, так и в случае использования RL для обучения LLM [Pternea et al., 2024].

*Использование LLM при реализации RL* предполагает такие схемы как вербальное, контекстное и символьное обучение с подкреплением. При вербальном обучении с подкреплением LLM генерирует словесную саморефлексию, чтобы обеспечить детальную и конкретную обратную связь, и затем сохраняет ее в памяти (контексте) действующего агента [Shinn et al., 2023]. Количество потребных для обучения попыток при переходе от градиентного к вербальному RL может сократиться на несколько порядков. В работе [Laskin et al., 2022] предложен метод Algorithm Distillation (AD), реализующий контекстное обучение с подкреплением на основе сбора историй обучения для отдельных алгоритмов RL. Пример символического RL показан в работе Eureka [Ma et al., 2023], где демонстрируется возможность автоматического подбора критерия оптимизации для RL с использованием LLM и рефлексии.

*Использование RL при обучении LLM* направлено на решение проблемы преодоления ограничений существующей обучающей выборки. Метод обучения с подкреплением на основе генеративного пополнения обучающей выборки предполагает итеративное использование этапов генерации синтетических данных предобученной LLM, фильтрации полученных синтетических данных и дообучения LLM на отфильтрованных синтетических данных. В случае задач, связанных с математикой или программированием, проверка корректности сгенерированных решений может быть реализована автоматически, что повышает производительность метода. Таким способом DeepMind были обучены модели для автоматического программирования (AlphaCode 2 [AlphaCode 2, 2023]) и автоматического решения математических задач в области геометрии (AlphaGeometry [Trinh et al., 2024], AlphaGeometry2 [Chervonyi et al., 2025]) и комбинато-

рики (FunSearch [Romera-Paredes et al., 2024]). Данный подход также активно использовался DeepSeek при автоматическом обучении модели «мыслителя» DeepSeek-R1 навыкам логики, математики и программирования [Guo et al., 2025].

## **Н7.2. Агентные системы. Н7.3. Мультиагентные системы**

*Генеративные агенты.* В 2022-24 гг. активно развивались технологии создания LLM-агентов [LLM Powered Autonomous Agents, 2023], ключевыми компонентами которых являются: декомпозиция и планирование задач, а также использование инструментов. Агенты также могут использовать самокритику и саморефлексию, чтобы учиться на ошибках и совершенствовать свои способы решения. В 2025 году главным фокусом внимания станут фундаментальные автономные агенты. Современный обзор фундаментальных LLM-агентов можно найти в [Liu et al., 2025].

*Model Context Protocol (MCP).* В технологическом плане важным новым инструментом создания LLM-агентов стал предложенный компанией Anthropic в ноябре 2024 г. протокол MCP – фреймворк с открытым исходным кодом для стандартизации взаимодействия и обмена данными между моделями ИИ и внешними системами, источниками данных и инструментами. В настоящее время MCP поддерживается всеми основными разработчиками моделей и сервисов ИИ на основе LLM. В репозитории MCP [Model Context Protocol, 2025] доступны реализации MCP-серверов для интерфейса со средами разработки (Python, TypeScript, Java, C#) а также с популярными корпоративными системами (Google Диск, Slack, GitHub, PostgreSQL и др.). Разработчики приложений на основе LLM-агентов могут создавать собственные MCP-серверы.

*Многоагентные системы.* Предложены различные схемы построения мультиагентных систем с разными сценариями взаимодействия между агентами для принятия групповых решений: конкуренция, координация, кооперация (ReConcile [Hao et al., 2023], Socratic AI [Boiko et al., 2023]). Многоагентные системы могут приобретать новые знания. В работе ChatLLM Network [Hao et al., 2023] предложена структура сети LLM-агентов, в которой организуется процесс прямого и обратного распространения текстовых сообщений для контекстного обучения коллектива LLM-агентов. В другой схеме Socratic Models (SM) [Andy et al., 2022] предложен способ построения расширяемого коллектива бимодальных ИИ-агентов, который может функционировать в условиях «открытого» списка задач и типов входных данных, для работы с которыми добавляются новые агенты.

*Агенты для помощи в научных исследованиях.* Для решения конкретных научных задач предложены, в частности, агенты в области органического синтеза, создания лекарств и дизайна материалов ([Bran et al., 2023], [Boiko et al., 2023]). Для общей реализации научного метода в работе AI Co-



scientist [AI Co-scientist, 2025] реализована система помощи ученым, включающая специализированных агентов (Generation, Reflection, Ranking, Evolution, Proximity, Meta-review), которые итеративно генерируют, оценивают и уточняют гипотезы с помощью автоматической обратной связи. В работе AI Scientist [Lu et al., 2024] был впервые реализован полный цикл научного исследования в области ML. В 2025 г. статья AI Scientist v2 прошла слепое рецензирование на воркшоп ICLR-2025 (конференция класса A\* в области ИИ). OpenAI предлагает услуги ИИ-агента уровня кандидата наук (PhD-Level Agent) [OpenAI PhD-level agents, 2025]. Современное состояние области использования ИИ для научных исследований можно найти в обзоре AI4Research [Chen et al., 2025].

## Н8. Элементы AGI

*Общий ИИ (Artificial general intelligence, AGI)* подразумевает способность ИИ-систем выполнять множество задач и включает такие навыки как организация рассуждений, формирование и использование модели мира и модели себя, рефлексия и самокритика, целеполагание и планирование. Агенты Auto-GPT, BabyAGI, BabyBeeAGI [AutoGPT: build & use AI agents, 2023] реализуют модель AGI посредством циклического вызова LLM до тех пор, пока агент не достигнет поставленной цели, генерируя новые подзадачи. AGI-концепция «LLM как операционная система, естественный язык как язык программирования» (см. [LLM OS Experiments, 2023], [Ge et al., 2023], [Wu et al., 2024]) подразумевает создание LLM-агента для управления компьютером (Computer-Using Agent). В 2025 г. эта концепция получила коммерческую реализацию в виде Operator от OpenAI [Computer-Using Agent, 2025].

*Рассуждающие LLM.* В 2024 г. появился новый класс LLM, позиционируемых как «рассуждающие» (reasoning) [Kumar et al., 2025]. Первой такой моделью стала OpenAI o1 [Learning to Reason with LLMs, 2024]. Прирост качества происходит за счёт скрытых рассуждений LLM перед ответом (inference-time scaling). Официальные советы по промпт-инжинирингу для o1 рекомендовали делать запросы простыми и короткими, а также избегать запросов цепочек рассуждений. В начале 2025 появилась первая публично доступная рассуждающая модель DeepSeek-R1 [Guo et al., 2025], обученная с помощью крупномасштабного обучения с подкреплением (RL) на значительно меньших вычислительных ресурсах, причем была достигнута производительность, сопоставимая с OpenAI o1. В работе [Li et al., 2025] была представлена парадигма многомодальных рассуждений, чередующих текстовые и визуальные шаги, которая улучшает как качество, так и интерпретируемость рассуждений за счет их визуализации. Модели OpenAI o3 и o4-mini также получили дополнитель-

ную способность "мыслить образами" [OpenAI O3 and O4 Mini, 2024]. В работе [Ma et al., 2025] идея визуальных рассуждений была реализована для диффузионных моделей.

Использование рассуждающих моделей в задачах автоматического доказательства теорем привело в последние годы к значительному прогрессу. Если ранее использовался подход, основанный на последовательном пошаговом прогнозировании доказательства, то теперь непосредственно генерируется целое доказательство: DeepSeek-Prover [Xin et al., 2024], Goedel-Prover [Lin et al., 2025]. В работе Kimina-Prover Preview [Wang et al., 2025] предложена схема совместного использования формальных и неформальных рассуждений, а также обучения с подкреплением для математических рассуждений «в стиле человека» (с анализом частных случаев и т.п.), что позволяет решить большее количество сложных задач, например, в области комбинаторики [Liu et al., 2025].

Перспективный подход к обучению рассуждающих моделей предложен в работе Absolute Zero [Zhao et al., 2025], где способы рассуждения выучиваются обучением с подкреплением в открытой вселенной задач, вообще без участия человека даже на этапе постановки задач: задачи также генерируются автоматически при помощи RL.

Были выявлены и проблемы, связанные с рассуждающими моделями, такие как избыточные рассуждения (overthinking, [Kumar et al., 2025], [Cuadron et al., 2025]) и «нечестность» объяснения рассуждений [Attribution Graphs in Biology, 2025]). Таким образом, полностью положиться на рассуждающие LLM, исключив инженерию запросов, гибридные и «прозрачные» методы ИИ, пока невозможно.

*Моделирование процессов человеческой психики* также является важным направлением при создании генеративных агентов и элементов AGI. Это такие исследования как эмоциональный ИИ, создание двойников личности, моделирование самосознания как нарратива личной истории (эгоцентрический storytelling), которые отрабатываются, в частности, в задачах автоматического литературного творчества. В работе WhatELSE [Lu et al., 2025] представлена интерактивная система создания повествований, использующая ИИ для развития повествовательных пространств и генерации разнообразных сюжетов на основе примеров историй. В работе [Gurung et al., 2025] демонстрируется концепция нарративного ИИ (Narrative AI), а также совершенствование генерации длинных историй с помощью обучения с подкреплением (RL) для улучшения рассуждений в LLM. Подобный подход далее может быть распространен на создание диалоговых систем, профессиональных ассистентов, поддержку принятия решений в сложных длительных процессах и другие приложения ИИ, которые необходимо выполнять «в стиле человека».

## Заключение

Возможность практического использования описанных результатов и технологий в значительной степени определяется набором и качеством доступных в каждый момент времени лучших моделей LLM и MLLM. По состоянию на август 2025 года в качестве лидирующих моделей можно выделить: Claude 4 (Opus 4 и Sonnet 4) [Claude-4, 2025] от Anthropic, Grok 4 [Grok-4, 2025] от xAI, Llama 4 (Maverick и Scout) [Llama4, 2025], Gemini 2.5 (Flash и Pro) [Gemini-2-5, 2025] от Google, Gemma 3 [Gemma-3, 2025] и Gemma 3n [Gemma-3n, 2025]) от Google DeepMind, Mistral 3 [Mistral.AI Models, 2025] (Mistral Small 3.2 [Mistral.AI Small-3-1, 2025], Magistral Small [Mistral.AI Magistral, 2025], Devstral Small [Mistral.AI Devstral, 2025]), Qwen3 (think и instruct) [Qwen3, 2025], DeepSeek-V3.1 [DeepSeek-V3.1, 2025], GPT-5 [OpenAI GPT-5/, 2025] и GPT-OSS [OpenAI GPT-OSS/, 2025] (120B и 20B) от OpenAI.

В качестве общих тенденций, отличающих современное поколение LLM и MLLM, можно отметить следующие:

- Разделение LLM на thinking и instruction модели (для сложных рассуждений и простых задач соответственно). Например, GPT-5 представляет собой уже не одну LLM, а набор моделей и маршрутизатор, распределяющий задачи между моделями.
- Модели сразу формируются и обучаются как агенты, в частности, адаптированные для использования инструментов и программирования.
- Использование архитектур типа MoE для ускорения на этапе выполнения.
- Увеличение длины контекстного окна.
- Параллельное исследование нескольких гипотез при рассуждениях.
- Интенсивное использование RL и синтетических данных на этапе обучения. Например, командой Qwen разработан алгоритм GSPO (Group Sequence Policy Optimization) [Zheng et al., 2025], призванный заменить популярный алгоритм GRPO при больших размерах и разреженности (MoE) обучаемых моделей.

Следует ожидать, что в 2026 году мы увидим не менее серьезные продвижения как в функциональности LLM, так и в методах их обучения и использования.

## Список литературы

- [Achiam et al., 2023] Achiam J. et al. Gpt-4 technical report // arXiv preprint arXiv:2303.08774. – 2023.
- [Adam et al., 2021] Adam S., et al. Open-Ended Learning Leads to Generally Capable. Agents // arXiv:2107.12808. – 2021.

- [Aditya et al., 2022] Aditya R., et al. Hierarchical Text-Conditional Image Generation with CLIP Latents // arXiv:2204.06125. – 2022.
- [Alphacode 2, 2023] AlphaCode Team G. Alphacode 2 technical report. – Technical report. – URL [https://storage.googleapis.com/deepmind-media/AlphaCode2/AlphaCode2\\_Tech\\_Report.pdf](https://storage.googleapis.com/deepmind-media/AlphaCode2/AlphaCode2_Tech_Report.pdf), 2023.
- [Andy et al., 2022] Andy Z., et al. Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language // arXiv:2204.00598. – 2022.
- [Azizi et al., 2024] Azizi S., et al. LaMDA: Large Model Fine-Tuning via Spectrally Decomposed Low-Dimensional Adaptation // arXiv: 2406.12832v1. – 2024.
- [Badia et al., 2020] Badia A.P., et al. Agent57: Outperforming the atari human benchmark // arXiv preprint arXiv:2003.13350. – 2020.
- [Barrault et al., 2024] Barrault L. et al. Large Concept Models: Language Modeling in a Sentence Representation Space // arXiv preprint arXiv:2412.08821. – 2024.
- [Behrouz et al., 2024] Behrouz A. et al. Titans: Learning to Memorize at Test Time // arXiv:2501.00663v1. – 2024.
- [Behrouz et al., 2025] It’s All Connected: A Journey Through Test-Time Memorization, Attentional Bias, Retention, and Online Optimization // arXiv preprint arXiv: 2504.13173. – 2025.
- [Bengio et al., 2021] Bengio E. et al. Flow network based generative models for non-iterative diverse candidate generation // Advances in Neural Information Processing Systems. – 2021. – Vol. 34. – P. 27381-27394.
- [Black et al., 2024] Black K. et al.  $\pi 0$ : A vision-language-action flow model for general robot control. – 2024. – URL <https://arxiv.org/abs/2410.24164>.
- [Boiko et al., 2023] Boiko D.A., MacKnight R., Gomes G. Emergent autonomous scientific research capabilities of large language models // arXiv preprint arXiv:2304.05332. – 2023.
- [Bousmalis et al., 2023] Bousmalis K. et al. Robocat: A self-improving foundation agent for robotic manipulation // arXiv preprint arXiv:2306.11706. – 2023.
- [Bran et al., 2023] Bran A.M. et al. Chemcrow: Augmenting large-language models with chemistry tools // arXiv preprint arXiv:2304.05376. – 2023.
- [Brohan et al., 2023] Brohan A. et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control // arXiv preprint arXiv:2307.15818. – 2023.
- [Brooks et al., 2024] Brooks T. et al. Video generation models as world simulators. – 2024.
- [Brown et al., 2020] Brown T. et al. Language models are few-shot learners // Advances in neural information processing systems. – 2020. – Vol. 33. – P. 1877-1901.
- [Chen et al., 2024] Chen G., Wang W. A survey on 3d gaussian splatting // arXiv preprint arXiv:2401.03890. – 2024.
- [Chen et al., 2025] Chen Q. et al. AI4Research: A Survey of Artificial Intelligence for Scientific Research // arXiv preprint arXiv: 2507.01903. – 2025.
- [Cheng et al., 2024] Cheng T. et al. Yolo-world: Real-time open-vocabulary object detection // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. – 2024. – P. 16901-16911.
- [Chervonyi et al., 2025] Chervonyi Y. et al. Gold-medalist Performance in Solving Olympiad Geometry with AlphaGeometry2 // arXiv preprint arXiv:2502.03544. – 2025.

- [Child, 2020] Child R. Very deep vaes generalize autoregressive models and can out-perform them on images // arXiv preprint arXiv:2011.10650. – 2020.
- [Cuadron et al., 2025] Cuadron A. et al. The Danger of Overthinking: Examining the Reasoning-Action Dilemma in Agentic Tasks // arXiv preprint arXiv:2502.08235. – 2025.
- [DeepSeek-AI, 2025] DeepSeek-AI Utilizing the Distilled Model from DeepSeek-R1 for Efficient Fine-Tuning with LoRA and Chain-of-Thought Datasets // arXiv:2406.15734v2. – 2025.
- [Dong et al., 2023] Dong Q. et al. A survey on in-context learning // arXiv preprint arXiv:2301.00234. – 2023.
- [Esser et al., 2024] Esser P. et al. Scaling rectified flow transformers for high-resolution image synthesis // arXiv preprint arXiv:2403.03206. – 2024.
- [Gao et al., 2024] Gao Y. et al. Retrieval-augmented generation for large language models: A survey // arXiv preprint arXiv:2312.10997. – 2024.
- [Ge et al., 2023] Ge Y. et al. Llm as os (llmao), agents as apps: Envisioning aios, agents and the aios-agent ecosystem // arXiv preprint arXiv:2312.03815. – 2023.
- [Gu et al., 2021] Gu A. et al. Efficiently modeling long sequences with structured state spaces // arXiv preprint arXiv:2111.00396. – 2021.
- [Gu et al., 2023] Gu A. et al. Mamba: Linear-time sequence modeling with selective state spaces // arXiv preprint arXiv:2312.00752. – 2023.
- [Guo et al., 2025] Guo D. et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning // arXiv preprint arXiv:2501.12948. – 2025.
- [Gurung et al., 2025] Gurung A. et al. Learning to Reason for Long-Form Story Generation // arXiv preprint arXiv:2503.22828. – 2025.
- [Halverson et al., 2024] Halverson J. et al. KAN: Kolmogorov–Arnold Networks // arXiv preprint arXiv:2404.19756v1 – 2024.
- [Hao et al., 2023] Hao R. et al. Chatllm network: More brains, more intelligence // arXiv preprint arXiv:2304.12998. – 2023.
- [Ho et al., 2020] Ho J., Jain A., Abbeel P. Denoising diffusion probabilistic models // Advances in neural information processing systems. – 2020. – Vol. 33. – P. 6840-6851.
- [Ho et al., 2022] Ho J. et al. Imagen video: High definition video generation with diffusion models // arXiv preprint arXiv:2210.02303. – 2022.
- [Ho et al., 2024] Ho M. et al. Proof Flow: Preliminary Study on Generative Flow Network Language Model Tuning for Formal Reasoning // arXiv preprint arXiv:2410.13224. – 2024.
- [Hu et al., 2022] Hu E.J. et al. Lora: Low-rank adaptation of large language models // ICLR. – 2022. – Vol. 1, No. 2. – P. 3.
- [Hu et al., 2023] Hu Z. et al. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models // arXiv preprint arXiv:2304.01933. – 2023. MLA.
- [Jain et al., 2024] Jain A. et al. Prompt Tuning Strikes Back: Customizing Foundation Models with Low-Rank Prompt Adaptation // arXiv:2405.15282v1. – 2024.
- [Kerbl et al., 2023] Kerbl B. et al. 3d gaussian splatting for real-time radiance field rendering // ACM Transactions on Graphics. – 2023. – Vol. 42, No. 4. – P. 1-14.
- [Kingma et al., 2013] Kingma D.P., Welling M. Auto-encoding variational bayes // arXiv preprint arXiv:1312.6114. – 2013.

- [Kirillov et al., 2023] Kirillov A. et al. Segment anything // Proceedings of the IEEE/CVF International Conference on Computer Vision. – 2023. – P. 4015-4026.
- [Kumar et al., 2025] Kumar A. et al. OverThink: Slowdown Attacks on Reasoning LLMs // arXiv preprint arXiv:2502.02542. – 2025.
- [Kumar et al., 2025] Kumar K. et al. Llm post-training: A deep dive into reasoning large language models // arXiv preprint arXiv:2502.21321. – 2025.
- [Laskin et al., 2022] Laskin M. et al. In-context reinforcement learning with algorithm distillation // arXiv preprint arXiv:2210.14215. – 2022.
- [Lewis et al., 2020] Lewis P. et al. Retrieval-augmented generation for knowledge-intensive nlp tasks // Advances in Neural Information Processing Systems. – 2020. – Vol. 33. – P. 9459-9474.
- [Li et al., 2025] Li C. et al. Imagine while Reasoning in Space: Multimodal Visualization-of-Thought // arXiv preprint arXiv:2501.07542. – 2025.
- [Lieber et al., 2024] Lieber O. et al. Jamba: A hybrid transformer-mamba language model // arXiv preprint arXiv:2403.19887. – 2024.
- [Lin et al., 2025] Lin et al. Goedel-Prover: A Frontier Model for Open-Source Automated Theorem Proving // arXiv preprint arXiv: 2502.07640. – 2025.
- [Ling et al., 2022] Ling Y., et al. Diffusion Models: A Comprehensive Survey of Methods and Applications // arXiv:2209.00796. – 2022.
- [Liu et al., 2025] Liu B. et al. Advances and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems // arXiv preprint arXiv:2504.01990. – 2025. MLA.
- [Liu et al., 2025] Liu J. et al. CombiBench: Benchmarking LLM Capability for Combinatorial Mathematics // arXiv:2505.03171v1. – 2025.
- [Liu et al., 2023] Liu S. et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection // arXiv preprint arXiv:2303.05499. – 2023.
- [Lu et al., 2024] Lu C. et al. The AI scientist: Towards fully automated open-ended scientific discovery // arXiv preprint arXiv:2408.06292. – 2024.
- [Lu et al., 2025] Lu Z. et al. WhatELSE: Shaping Narrative Spaces at Configurable Level of Abstraction for AI-bridged Interactive Storytelling // arXiv preprint arXiv:2502.18641. – 2025.
- [Ma et al., 2025] Ma N. et al. Inference-time scaling for diffusion models beyond scaling denoising steps // arXiv preprint arXiv:2501.09732. – 2025.
- [Ma et al., 2023] Ma Y.J. et al. Eureka: Human-level reward design via coding large language models // arXiv preprint arXiv:2310.12931. – 2023.
- [Mnih et al., 2015] Mnih V., et al. Human-level control through deep reinforcement learning // Nature. 2015.
- [Nie et al., 2025] Nie S. et al. Large language diffusion models // arXiv preprint arXiv:2502.09992. – 2025.
- [Ouyang et al., 2022] Ouyang L., et al. Training language models to follow instructions with human feedback // Advances in Neural Information Processing Systems. – 2022. – Vol. 35. – P. 27730-27744.
- [Peng et al., 2023] Peng B. et al. Rwkv: Reinventing rnns for the transformer era // arXiv preprint arXiv:2305.13048. – 2023.

- [Poli et al., 2023] Poli M. et al. Hyena hierarchy: Towards larger convolutional language models // International Conference on Machine Learning. – PMLR, 2023. – P. 28043-28078.
- [Pternea et al., 2024] Pternea M. et al. The rl/llm taxonomy tree: Reviewing synergies between reinforcement learning and large language models // Journal of Artificial Intelligence Research. – 2024. – Vol. 80. – P. 1525-1573.
- [Rishi et al., 2021] Rishi B., et al. On the Opportunities and Risks of Foundation Models // arXiv:2108.07258. – 2021.
- [Romera-Paredes et al., 2024] Romera-Paredes B. et al. Mathematical discoveries from program search with large language models // Nature. – 2024. – Vol. 625, No. 7995. – P. 468-475.
- [Sahoo et al., 2024] Sahoo P. et al. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications // arXiv preprint arXiv:2402.07927. – 2024.
- [Schulman et al., 2017] Schulman J., Wolski F., Dhariwal P., Radford A., Klimov O. Proximal policy optimization algorithms // arXiv preprint arXiv:1707.06347. – 2017.
- [Scott et al., 2022] Scott R. et al. A Generalist Agent // arXiv:2205.06175. – 2022.
- [Shi et al., 2024] Shi J. et al. Simplified and generalized masked diffusion for discrete data // Advances in neural information processing systems. – 2024. – Vol. 37. – P. 103131-103167.
- [Shinn et al., 2023] Shinn N., Labash B., Gopinath A. Reflexion: an autonomous agent with dynamic memory and self-reflection // arXiv preprint arXiv:2303.11366. – 2023.
- [Sun et al., 2023] Sun Y. et al. Retentive network: A successor to transformer for large language models // arXiv preprint arXiv:2307.08621. – 2023.
- [Takase et al., 2024] Takase R. et al. GFlowNet Fine-tuning for Diverse Correct Solutions in Mathematical Reasoning Tasks // arXiv preprint arXiv:2410.20147. – 2024.
- [Tian et al., 2025] Tian Y. et al. Yolov12: Attention-centric real-time object detectors // arXiv preprint arXiv:2502.12524. – 2025.
- [Trinh et al., 2024] Trinh T.H. et al. Solving olympiad geometry without human demonstrations // Nature. – 2024. – Vol. 625, No. 7995. – P. 476-482.
- [Vaswani et al., 2017] Vaswani A. et al. Attention is all you need // Advances in neural information processing systems. – 2017. – Vol. 30.
- [Wan et al., 2024] Wan Z. et al. Efficient large language models: A survey // arXiv preprint arXiv:2312.03863. – 2024.
- [Wang et al., 2022] Wang X. et al. Self-consistency improves chain of thought reasoning in language models // arXiv preprint arXiv:2203.11171. – 2022.
- [Wang et al., 2025] Wang et al. Kimina-Prover Preview: Towards Large Formal Reasoning Models with Reinforcement Learning // arXiv preprint arXiv:2504.11354. – 2025.
- [Wang et al., 2025] Wang K.A. et al. Test-time regression: a unifying framework for designing sequence models with associative memory // arXiv preprint arXiv:2501.12352. – 2025.
- [Wang et al., 2024] Wang W. et al. CogVLM: Visual expert for pretrained language models // Advances in Neural Information Processing Systems. – 2024. – Vol. 37. – P. 121475-121499.

- [Wu et al., 2024] [Wu L. et al., 2024] Wan Z. et al. Efficient large language models: A survey // arXiv preprint arXiv:2312.03863. – 2024.
- [Wu et al., 2024] Wu Z. et al. Os-copilot: Towards generalist computer agents with self-improvement // arXiv preprint arXiv:2402.07456. – 2024.
- [Xin et al., 2021] Xin H., Kaiyong Z., Xiaowen C. AutoML: A Survey of the State-of-the-Art // arXiv:1908.00709v6. – 2021.
- [Xin et al., 2024] Xin et al. DeepSeek-Prover: Advancing Theorem Proving in LLMs through Large-Scale Synthetic Data // arXiv preprint arXiv: 2405.14333. – 2024.
- [Yao et al., 2023] Yao S. et al. React: Synergizing reasoning and acting in language models // arXiv preprint arXiv:2210.03629. – 2023.
- [Yao et al., 2023] Yao Y., Li Z., Zhao H. Beyond chain-of-thought, effective graph-of-thought reasoning in large language models // arXiv preprint arXiv:2305.16582. – 2023.
- [Ye et al., 2024] Ye J. et al. Diffusion of thoughts: Chain-of-thought reasoning in diffusion language models // arXiv preprint arXiv:2402.07754. – 2024.
- [Yin et al., 2024] Yin S. et al. A survey on multimodal large language models // arXiv preprint arXiv:2306.13549. – 2024.
- [Zhang et al., 2019] Zhang S., Tong H., Xu J., Maciejewski R. Graph convolutional networks: a comprehensive review // Comput Soc Netw 6. – 2019. – No. 11.
- [Zhang et al. 2024] Zhang F., Pilanci M. Spectral Adapter: Fine-Tuning in Spectral Space // arXiv:2405.13952 – 2024.
- [Zhang et al., 2024] Zhang J.-C. et al. LoRA<sup>2</sup>: Multi-Scale Low-Rank Approximations for Fine-Tuning Large Language Models // arXiv:2408.06854v1. – 2024.
- [Zhao et al., 2025] Zhao A. et al. Absolute Zero: Reinforced Self-play Reasoning with Zero Data // arXiv preprint arXiv:2505.03335. – 2025.
- [Zheng et al., 2025] Group Sequence Policy Optimization Group sequence policy optimization // arXiv preprint arXiv:2507.18071. – 2025.
- [Zhou et al., 2024] Zhou C., et al. RankAdaptor: Hierarchical Rank Allocation for Efficient Fine-Tuning Pruned LLMs via Performance Model // arXiv:2406.15734v2. – 2024.
- [Zhou et al., 2024] Zhou X. et al. GALA3D: Towards Text-to-3D Complex Scene Generation via Layout-guided Generative Gaussian Splatting // arXiv preprint arXiv:2402.07207. – 2024.
- [Zielonka et al., 2023] Zielonka W. et al. Drivable 3d gaussian avatars // arXiv preprint arXiv:2311.08581. – 2023.

### Электронные ресурсы

- [AI Co-scientist, 2025] Accelerating scientific breakthroughs with an AI co-scientist // google. – URL: <https://research.google/blog/accelerating-scientific-breakthroughs-with-an-ai-co-scientist/> (дата обращения: 31.08.2025).
- [AI Foresight, 2024] Дмитрий Чернышенко провёл стратегическую форсайт-сессию по фундаментальным исследованиям в сфере искусственного интеллекта // Официальный сайт Правительства Российской Федерации. – URL: <http://government.ru/news/51726/> (дата обращения: 31.08.2025).



- [**Attribution Graphs in Biology, 2025**] Attribution Graphs in Biology // Transformer Circuits Blog. – URL: <https://transformer-circuits.pub/2025/attribution-graphs/biology.html> (дата обращения: 28.04.2025).
- [**AutoGPT: build & use AI agents, 2023**] AutoGPT: build & use AI agents // GitHub. – URL: <https://github.com/Significant-Gravitas/AutoGPT> (дата обращения: 10.12.2023).
- [**Claude-4, 2025**] [www.anthropic.com](https://www.anthropic.com). – URL: <https://www.anthropic.com/news/claude-4> (дата обращения: 31.08.2025).
- [**Computer-Using Agent, 2025**] Computer-Using Agent: a universal interface for AI to interact with the digital world // OpenAI – URL: <https://openai.com/index/computer-using-agent/> (дата обращения: 21.04.2025).
- [**Deepmind VEO, 2025**] [deepmind.google](https://deepmind.google). – URL: <https://deepmind.google/models/veo/> (дата обращения: 31.08.2025).
- [**DeepSeek, 2025**] DeepSeek // Официальный сайт URL: <https://www.deepseek.com/> (дата обращения: 24.04.2025).
- [**DeepSeek-V3.1, 2025**] DeepSeek-V3.1 Release // [deepseek.com](https://deepseek.com). – URL: <https://api-docs.deepseek.com/news/news250821> (дата обращения: 31.08.2025).
- [**Gemini-2-5, 2025**] Try Deep Think in the Gemini app // [blog.google](https://blog.google). – URL: <https://blog.google/products/gemini/gemini-2-5-deep-think/> (дата обращения: 31.08.2025).
- [**Gemma-3, 2025**] [huggingface.co](https://huggingface.co). – URL: <https://huggingface.co/google/gemma-3-12b-it> (дата обращения: 31.08.2025).
- [**Gemma-3n, 2025**] [huggingface.co](https://huggingface.co). – URL: <https://huggingface.co/google/gemma-3n-E4B-it> (дата обращения: 31.08.2025).
- [**Grok-4, 2025**] [x.ai](https://x.ai). – URL: <https://x.ai/news/grok-4> (дата обращения: 31.08.2025).
- [**Improving Image Generation, 2024**] Improving Image Generation with Better Captions // Semantic Scholar. – URL: <https://www.semanticscholar.org/paper/Improving-Image-Generation-with-Better-Captions-Betker-Goh/cfee1826dd4743cab44c6e27a0cc5970effa4d80> (дата обращения: 21.02.2024).
- [**Learning to Reason with LLMs, 2024**] Learning to Reason with LLMs // OpenAI. – URL: <https://openai.com/index/learning-to-reason-with-llms/> (дата обращения: 16.04.2025).
- [**Llama4, 2025**] Welcome Llama 4 Maverick & Scout on Hugging Face // [huggingface.co](https://huggingface.co). – URL: <https://huggingface.co/blog/llama4-release> (дата обращения: 31.08.2025).
- [**LLM Architecture Evolution, 2025**] Эволюция архитектур больших языковых моделей: от GPT-2 к современным решениям // [habr.com](https://habr.com). – URL: <https://habr.com/ru/articles/931382/> (дата обращения: 31.08.2025).
- [**LLM OS Experiments, 2023**] LLM OS Experiments // [LLM-OS.net](https://llm-os.net). URL: <http://llm-os.net/> (дата обращения: 01.12.2023).
- [**LLM Powered Autonomous Agents, 2023**] LLM Powered Autonomous Agents // LilianWeng. – URL: <https://lilianweng.github.io/posts/2023-06-23-agent/> (дата обращения: 11.12.2023).
- [**Mistral.AI Devstral, 2025**] Upgrading agentic coding capabilities with the new Devstral models // [mistral.ai](https://mistral.ai). – URL: <https://mistral.ai/news/devstral-2507> (дата обращения: 31.08.2025).

- [**Mistral.AI Magistral, 2025**] Magistral // mistral.ai. – URL: <https://mistral.ai/news/magistral> (дата обращения: 31.08.2025).
- [**Mistral.AI Models, 2025**] huggingface.co. – URL: <https://huggingface.co/mistralai/models> (дата обращения: 31.08.2025).
- [**Mistral.AI Small-3-1, 2025**] mistral.ai. – URL: <https://mistral.ai/news/mistral-small-3-1> (дата обращения: 31.08.2025).
- [**Model Context Protocol, 2025**] Model Context Protocol // github.com. – URL: <https://github.com/modelcontextprotocol> (дата обращения: 31.08.2025).
- [**OpenAI GPT-5, 2025**] Introducing GPT-5 // openai.com. – URL: <https://openai.com/index/introducing-gpt-5/> (дата обращения: 31.08.2025).
- [**OpenAI GPT-OSS, 2025**] Introducing gpt-oss // openai.com. – URL: <https://openai.com/index/introducing-gpt-oss/> (дата обращения: 31.08.2025).
- [**OpenAI O3 and O4 Mini, 2024**] OpenAI. Introducing O3 and O4 Mini // OpenAI. – URL: <https://openai.com/index/introducing-o3-and-o4-mini/> (дата обращения: 28.04.2025).
- [**OpenAI PhD-level agents, 2025**] OpenAI plots charging \$20,000 a month for PhD-level agents // The Information. – URL: <https://www.theinformation.com/articles/openai-plots-charging-20-000-a-month-for-phd-level-agents> (дата обращения: 21.04.2025).
- [**Prompt Engineering Guide, 2023**] Prompt Engineering Guide // Prompt Engineering Guide. URL: <https://www.promptingguide.ai/> (дата обращения: 16.12.2023).
- [**Qwen3, 2025**] Qwen3TechnicalReport // arxiv.org. – URL: <https://arxiv.org/pdf/2505.09388> (дата обращения: 31.08.2025).
- [**Your Guide to Generative AI, 2023**] Your Guide to Generative AI // Learn Prompting. – URL: <https://learnprompting.org/> (дата обращения: 19.12.2023).