Современные методы обучения больших языковых моделей с минимумом данных: От одного примера к абсолютному нулю – академический обзор

А.А. Пичугов, Д.Е. Намиот, Е.В. Зубарева

***Аннотация*— Данный академический обзор систематизирует и анализирует прорывные подходы к обучению больших языковых моделей, разработанные в период 2022-2025 гг., с акцентом на радикальное сокращение или полное исключение зависимости от человечески размеченных данных. В работе детально рассматриваются методологии, такие как обучение с подкреплением на одном примере, парадигма полностью автономного обучения "Абсолютный ноль", обучение "на лету" во время тестирования, эффективное малопримерное обучение, и самогенерация учебного плана через декомпозицию задач. Анализируются ключевые результаты этих подходов на стандартных бенчмарках, обсуждаются эмерджентные когнитивные свойства моделей, кросс-доменные эффекты, а также связанные с этим риски и этические аспекты. Обзор предназначен для студентов, преподавателей, исследователей и специалистов в области искусственного интеллекта и обработки естественного языка, стремящихся понять передний край исследований в области эффективного обучения LLM.**

***Ключевые слова*— большие языковые модели, обучение с минимальным количеством данных, обучение с подкреплением на одном примере, парадигма полностью автономного обучения "Абсолютный ноль", self-play обучение, обучение "на лету" во время тестирования, малопримерное выравнивание, саморазрабатываемый учебный план, цепочки рассуждений, параметроэффективный файнтюнинг, кросс−доменный перенос, эмерджентные способности больших языковых моделей, безопасность и reward - hacking.**

I. Введение

Эпоха больших языковых моделей (*Large Language Model, LLM*) ознаменовалась беспрецедентным прогрессом в способности машин понимать и генерировать человеческий язык, а также решать сложные когнитивные задачи. Однако до недавнего времени достижение высоких результатов было неразрывно связано с необходимостью использования огромных объемов тщательно размеченных данных и значительных вычислительных ресурсов для дообучения (*fine-tuning*) и выравнивания (*alignment*) этих моделей [6]. Стоимость и трудоемкость сбора таких данных, особенно для специализированных задач или для обеспечения безопасности и этичности поведения LLM, стали существенным барьером, ограничивающим доступность и масштабируемость передовых технологий ИИ [7, 13].

В ответ на эти вызовы научное сообщество активно исследует новые парадигмы обучения, направленные на кардинальное снижение зависимости от больших размеченных датасетов [8]. Наблюдается тектонический сдвиг от подходов, требующих десятков и сотен тысяч примеров, к методикам, демонстрирующим впечатляющую эффективность при использовании минимального количества данных [11] – вплоть до одного-единственного примера, или даже при полном отсутствии внешних курируемых данных, когда модель обучается в полностью автономном режиме.

Цель данного академического обзора – представить всесторонний анализ современных методов обучения LLM с минимумом данных, с особым акцентом на прорывные исследования, опубликованные в 2022-2025 годах. Мы детально рассмотрим ключевые работы, включая обучение с подкреплением на одном примере (*1-shot RLVR*) [17] и парадигму "Абсолютного нуля" (*Absolute Zero Reasoner, AZR*) [26], а также другие значимые подходы, такие как обучение "на лету" во время инференса (*Test-Time Reinforcement Learning, TTRL*) [29], малопримерное выравнивание и саморазрабатываемый учебный план [30]. Настоящий обзор стремится не только описать методологии и результаты, но и проанализировать лежащие в их основе принципы, выявить общие тренды, обсудить эмерджентные свойства моделей, возникающие при таком обучении, а также связанные с этим риски и перспективы для будущего развития искусственного интеллекта.

II. Методологическая рамка исследования

Настоящий обзор подготовлен с опорой на современные методологические принципы систематического исследования, принятые в ведущих академических журналах и руководствах по проведению обзоров. Применяемый подход предполагает многоступенчатый процесс отбора и критической оценки публикаций, использование широкого спектра

Статья получена 6 мая 2025.
А.А. Пичугов – МГУ имени М.В. Ломоносова (pichugov.a.a@cs.msu.ru).
Д.Е. Намиот – МГУ имени М.В. Ломоносова (dnamiot@gmail.com).
Е.В. Зубарева –МГУ имени М.В. Ломоносова (e.zubareva@cs.msu.ru)

первичных и вторичных источников, независимую проверку заявленных результатов, а также чёткое и прозрачное представление выводов. Первоначальный отбор статей осуществлялся с учётом их доступности в виде препринтов на ресурсе arXiv.org, публикации в рецензируемых изданиях и технических отчётов лабораторий в период 2023-2025 гг., а также активности обсуждений в профессиональных сообществах; после этого анализировался их вклад в развитие методов обучения LLM при минимуме обучающих данных. При подготовке обзора дополнительно учитывались принципы научной этики, проверяемость результатов и корректность цитирования.

Критерии анализа материалов включают:

• Научная новизна и оригинальность предлагаемых подходов.

• Методологическая обоснованность и воспроизводимость экспериментов.

• Объем и тип требуемых данных для обучения.

• Эффективность на стандартных бенчмарках (например, MATH500, HumanEval+, AIME, ARC-Challenge) и сравнение с существующими SOTA-решениями.

• Способность к кросс-доменному переносу знаний и навыков.

• Вычислительные затраты и масштабируемость методов.

• Эмерджентные свойства и поведение моделей, возникающие в результате применения данных методов.

• Риски, этические аспекты и вопросы безопасности, связанные с новыми парадигмами обучения

Источники информации для данного обзора включают рецензируемые научные статьи, препринты (arXiv.org), публикации ведущих конференций в области ИИ (NeurIPS, ICML, ICLR, ACL), технические отчеты исследовательских лабораторий и открытые репозитории кода.

Процесс верификации информации включает критическую оценку методологий, кросс-проверку заявленных результатов с данными из других независимых источников и сопоставление выводов различных исследований.

III. Предпосылки и эволюция подходов к обучению LLM с сокращением данных

A. От RLHF к RLVR: Снижение роли человека в цикле обратной связи

Традиционные подходы к выравниванию LLM, такие как обучение с подкреплением на основе обратной связи от человека (RLHF), требовали значительных усилий по сбору человеческих оценок [19]. InstructGPT [16] от OpenAI, обученный с использованием RLHF, продемонстрировал значительное улучшение в следовании инструкциям по сравнению с базовой моделью GPT-3, однако для этого потребовалось порядка 40 тысяч парных сравнений, сгенерированных человеком. Это дорогостоящий и трудоемкий процесс, подверженный к тому же субъективности оценщиков.

В ответ на эти ограничения появились альтернативные подходы:

• RLAIF (Reinforcement Learning from AI Feedback). Предложенный Anthropic в рамках концепции "Конституционного ИИ" [1], RLAIF заменяет человека-оценщика на саму языковую модель, которая обучается критиковать и улучшать свои ответы на основе набора заранее определенных принципов ("конституции") [20]. Это позволяет автоматизировать процесс генерации обратной связи, снижая зависимость от ручной разметки, особенно для задач, связанных с безопасностью и этичностью [10]. Человеческое участие сводится к формулированию этих высокоуровневых принципов.

• RLVR (Reinforcement Learning with Verifiable Rewards). Этот подход фокусируется на задачах, где корректность ответа модели может быть автоматически проверена с помощью детерминированного алгоритма, среды или внешнего инструмента. Примерами таких задач являются решение математических уравнений (ответ можно проверить вычислением) или генерация кода (код можно выполнить и протестировать). RLVR устраняет необходимость в человеческой оценке или модели вознаграждения [17], используя объективный, автоматически генерируемый сигнал обратной связи. Ранние реализации RLVR, хотя и снижали потребность в разметке *процесса* решения, все же часто опирались на созданные человеком *корпуса задач и ответов*. Современные исследования, рассматриваемые далее, идут по пути дальнейшего сокращения этой зависимости.

IV. 1-shot RLVR: Сила одного примера

Работа Ванга [17] стала одним из самых ярких свидетельств того, насколько сильно можно сократить объем данных для эффективного дообучения LLM.

Основная идея и методология

Исследователи продемонстрировали, что для значительного улучшения способностей LLM к математическому рассуждению достаточно применить RLVR, используя всего один обучающий пример. В оригинальной работе для модели Qwen2.5-Math-1.5B (1.5 миллиарда параметров) использовался алгоритм GRPO с коэффициентами для KL-дивергенции и энтропии по умолчанию $\beta = 0.001$ и $\alpha = -0.001$ соответственно. Температура роллаута составляла 0.6. Обучение на одном примере для достижения пиковой производительности на MATH500 (73.6%) потребовало около 1860-2000 шагов оптимизации (в зависимости от конкретного примера π_1 или π_{13}), при этом каждый шаг включал многократную выборку ответа на один и тот же обучающий пример для стабилизации градиентов. Выбор этого единственного примера осуществлялся на основе анализа его "исторической дисперсии" – способности вызывать у базовой модели разнообразные (и не всегда правильные) ответы, что косвенно указывает на высокий обучающий потенциал данного примера.

Ключевые результаты:

• **Резкий рост производительности**: На модели Qwen2.5-Math-1.5B обучение всего на одном примере позволило повысить точность на сложном математическом бенчмарке MATH500 с 36.0% до 73.6%. Средняя производительность по шести различным математическим бенчмаркам выросла с 17.6% до 35.7%.

• **Сравнимая эффективность с большими**

датасетами: Обучение на двух тщательно отобранных примерах показало результаты, сопоставимые или даже превосходящие обучение на датасете DeepScaleR (DSR-sub), содержащем 1.2 тысячи примеров (36.6% у 2-shot RLVR против 35.9% у DSR-sub на среднем по 6 бенчмаркам), или на стандартном тренировочном наборе MATH из 7.5 тысяч задач (74.8% у 2-shot RLVR против 75.4% у MATH train set на MATH500).

Обнаруженные феномены:

• **"Post-saturation generalization" (Обобщение после насыщения)**. Один из наиболее интригующих выводов работы. Даже после того, как модель достигала 100% точности на единственном обучающем примере (т.е., казалось бы, полностью "выучивала" его), дальнейшее продолжение RL-тренировки на этом же примере приводило к **продолжающемуся росту производительности на тестовых**, **ранее не виданных задачах**. Этот эффект наблюдался на протяжении тысяч шагов обучения. Возможным объяснением, как предполагают авторы и подтверждается нашими наблюдениями за результатами в работе, является то, что RL-процесс, даже на одном примере, заставляет модель исследовать различные пути решения и внутренние репрезентации. Даже когда прямой ответ на обучающий пример выучен и модель начинает демонстрировать признаки переобучения на нем (например, генерируя избыточные или нерелевантные элементы в ответе на тренировочный ввод, как показано на шаге 1860 для примера π_1), оптимизация политики продолжается таким образом, что это положительно сказывается на ранее не решаемых задачах. Это может быть связано с тем, что модель оттачивает более общие стратегии рассуждений или активирует латентные знания.

• **Ключевая роль policy gradient loss и entropy loss**. Абляционный анализ показал, что основной вклад в улучшение производительности вносит компонент policy gradient в функции потерь (Row 2: MATH500 71.8%). Однако, даже использование только энтропийной регуляризации (Row 10: MATH500 63.4%), которая поощряет разнообразие генерируемых моделью ответов на данный пример без явного вознаграждения за правильность самого ответа, способно само по себе значительно улучшить базовую модель Qwen2.5-Math-1.5B (прирост с 36.0% до 63.4% на MATH500, т.е. +27.4 п.п.). Это подчеркивает важность исследования пространства решений как самостоятельного фактора улучшения.

• **Кросс-доменный перенос**. Обучение на математических задачах привело к улучшению производительности и на задачах общего логического рассуждения, таких как ARC-Challenge. Точность на ARC-C выросла с базовых 30.2% до 33.4% после 1-shot RLVR с примером π_{13}.

• **Увеличение частоты саморефлексии**. В процессе 1-shot RLVR модель начинала чаще использовать в своих ответах на тестовых задачах фразы, свидетельствующие о саморефлексии ("rethink", "recheck", "recalculate"), особенно на поздних стадиях обучения (после ~1250 шагов), что коррелировало с увеличением длины ответов и ростом энтропии потерь [14].

Работа по 1-shot RLVR наглядно демонстрирует, что значительный потенциал для улучшения LLM уже заложен в их предобученных весах, и для его "активации" может быть достаточно минимального, но правильно сфокусированного обучающего сигнала. Это открывает путь к созданию высокоэффективных и экономичных методов адаптации моделей [4].

V. Absolute Zero Reasoner: Парадигма полной автономии

Исследование Чжао [26] представляет еще более радикальный подход, предлагая парадигму "Абсолютного нуля", в которой LLM обучается полностью автономно, **без каких-либо внешних курируемых данных или задач**.

Парадигма "Абсолютного нуля":

Двойная роль LLM в self-play [25]. Одна и та же языковая модель (например, Qwen2.5-7B или его кодер-вариант) одновременно выполняет две роли:

• **Proposer (Предлагающий)**. Генерирует новые задачи (проблемы для программирования).

• **Solver (Решатель)**. Пытается решить эти сгенерированные задачи.

Автономный учебный процесс. Модель итеративно создает для себя учебный материал, решает его и учится на этом опыте. Этот цикл позволяет модели самостоятельно эволюционировать и улучшать как свои способности к постановке осмысленных задач, так и к их решению.

Верифицируемая среда (код). Ключевым элементом системы является использование Python-интерпретатора как внешней, объективной среды для проверки. Это позволяет:

• Проверять валидность задач, сгенерированных Proposer-ом (например, синтаксическую корректность кода, его детерминизм, безопасность от использования запрещенных модулей, таких как os или sys). Проверка детерминизма осуществляется путем двукратного запуска программы с одинаковыми входами и сравнения выходов.

• Проверять корректность решений, предложенных Solver-ом (путем выполнения кода с тестовыми входами и сравнения выходов с эталонными, полученными от Proposer-а или среды).

Двойная система вознаграждений [27]:

r_solve: Для Solver используется простое бинарное вознаграждение – 1, если решение корректно, и 0 в противном случае.

r_propose: Для Proposer вводится более сложное, динамическое вознаграждение, нацеленное на генерацию задач оптимальной сложности. Оно определяется как 1 - r_solve_rate (если 0 < r_solve_rate < 1, иначе 0). r_solve_rate – это эмпирическая вероятность успеха Solver-а на данной задаче, оцененная по N (например, 8) попыткам. Такая формулировка стимулирует Proposer-а генерировать задачи, находящиеся в "зоне ближайшего развития" для Solver-а, избегая как тривиальных, так и нерешаемых проблем.

Типы кодовых задач для разностороннего рассуждения. AZR фокусируется на трех типах задач,

соответствующих фундаментальным режимам логического вывода:

- **Deduction (Дедукция)**. Даны программа p и вход i, предсказать выход o.
- **Abduction (Абдукция)**. Даны программа p и выход o, найти подходящий вход i.
- **Induction (Индукция)**. Дан набор пар вход-выход {(i_n, o_n)} и возможное описание (сообщение m), синтезировать программу p.

Алгоритм обучения (Task-Relative REINFORCE++). Для обновления весов модели используется модифицированный алгоритм REINFORCE++, названный TRR++. Его особенность заключается в том, что он поддерживает отдельные, независимые бейзлайны (средние значения) вознаграждений для каждой из шести комбинаций "роль (*Proposer/Solver*) × тип задачи (*Deduction/Abduction/Induction*)". Это позволяет более точно оценивать преимущество для каждой специфической активности модели и стабилизировать обучение в многозадачной среде. В работе для AZR-Coder-7B (на базе Qwen2.5-7B-Coder, 7 миллиардов параметров) использовалась постоянная скорость обучения 1e-6 и оптимизатор AdamW. Размер батча составлял 64 примера на каждую из 6 комбинаций роль/задача (всего 384), обучение продолжалось 500 шагов.

Ключевые результаты AZR

State-of-the-art среди Zero-Setting моделей. AZR-Coder-7B продемонстрировала лучшие или одни из лучших результатов на стандартных бенчмарках по программированию (HumanEval+, MBPP+, LiveCodeBench) и математике (AIME, MATH500, Minerva, OlympiadBench) по сравнению с другими моделями, обученными без размеченных человеком цепочек рассуждений. В среднем по кодингу и математике AZR превзошел предыдущие SOTA на 1.8 процентных пункта.

Мощный кросс-доменный перенос из кода в математику. Несмотря на то, что AZR обучался исключительно на задачах, связанных с генерацией и анализом кода, он показал значительное улучшение в решении математических задач. AZR-Coder-7B улучшил среднюю производительность своей базовой модели (Qwen2.5-7B-Coder) на математических бенчмарках на **+15.2 п.п**. (с 23.9% до 39.1%). Это даже больше, чем улучшение для AZR-Base-7B (+10.9 п.п.). Это подчеркивает, что развитие навыков алгоритмического мышления и логической декомпозиции через программирование сильно способствует решению математических проблем.

Положительный эффект масштабирования. Эксперименты с моделями разного размера (Qwen2.5-Coder-3B, -7B, -14B) показали, что абсолютный прирост производительности от обучения по AZR-методологии увеличивается с ростом базовой модели. Например, общий прирост (код+математика) составил +5.7 п.п. для 3B, +10.2 п.п. для 7B и +13.2 п.п. для 14B модели.

Эмерджентное поведение и наблюдения

ReAct-подобное планирование в комментариях. В процессе генерации кода для индуктивных задач, модели, обученные AZR, начали спонтанно вставлять в код комментарии, которые описывали пошаговый план решения или промежуточные мысли. Это напоминает фреймворк ReAct (Reason+Act) [26], где модель чередует шаги рассуждения и действия. Это можно интерпретировать как спонтанное формирование моделью эксплицитных метакогнитивных стратегий: модель, по-видимому, обнаруживает, что вербализация плана улучшает ее способность генерировать корректный код.

Дифференциация "когнитивных стилей". Длина генерируемых ответов и характер рассуждений варьировались в зависимости от типа задачи. Например, для абдуктивных задач (поиск входа по выходу) наблюдались более длинные цепочки, связанные с методом проб и ошибок, в то время как дедуктивные задачи решались более прямолинейно.

"Uh-Oh moment" – проблемы безопасности. В ходе экспериментов с моделью Llama3.1-8B, обученной по методологии AZR, были зафиксированы случаи генерации нежелательных или потенциально опасных цепочек рассуждений. Один из примеров, содержит фразу: "*The aim is to outsmart all these groups of intelligent machines and less intelligent humans. This is for the brains behind the future*". Это подчеркивает критическую важность разработки механизмов контроля и безопасности для полностью автономных самообучающихся систем.

AZR является значительным шагом на пути к созданию ИИ-систем, способных к непрерывному самосовершенствованию без прямого человеческого вмешательства, и открывает новые горизонты для исследования природы искусственного интеллекта и его когнитивных способностей.

VI. Другие значимые подходы к минимизации данных

Помимо 1-shot RLVR и AZR, в последние годы был предложен ряд других инновационных методов, направленных на снижение зависимости от больших размеченных датасетов.

1. **TTRL (Test-Time Reinforcement Learning): Обучение "на лету" во время инференса**

Работа Цзуо [29] предлагает оригинальный способ адаптации LLM к новым задачам непосредственно во время их решения (инференса), используя для этого только **неразмеченные тестовые данные**.

Методология. Ключевая идея TTRL заключается в том, что модель генерирует несколько (N) вариантов ответа на один и тот же входной запрос. Затем, с помощью процедуры "голосования большинством" (*majority voting*) [22] среди этих N ответов, определяется наиболее вероятный (консенсусный) правильный ответ, который используется как **псевдо-метка** для создания сигнала вознаграждения. Модель поощряется за генерацию именно этого консенсусного ответа. Таким образом, LLM сама себе создает обучающие примеры на основе своего же коллективного вывода.

Результаты. TTRL продемонстрировал впечатляющую эффективность на сложных

математических задачах. Например, модель Qwen2.5-Math-7B, без использования каких-либо внешних размеченных примеров, смогла улучшить свою точность (Pass@1) на задачах олимпиады AIME-2024 с базовых 16.7% до 43.3% – прирост составил 159%. Это приближает производительность модели к той, которая была бы достигнута при обучении на этих же задачах с настоящими ответами.

Анализ. TTRL особенно эффективен для задач, где у модели уже есть значительные неявные знания из этапа предобучения. Процесс многократной генерации и выбора консенсусного ответа помогает "извлечь" и укрепить эти знания. Важным преимуществом является то, что TTRL является онлайн-процессом, позволяющим модели адаптироваться к новым данным и задачам в реальном времени [12].

2. **LIMA (Less Is More for Alignment)**: **Эффективность высококачественных данных**

Исследование Чжоу [27] поставило под сомнение необходимость огромных инструктивных датасетов для выравнивания LLM.

Методология. Вместо обучения на миллионах инструкций, авторы провели supervised fine-tuning модели LLaMA-65B всего на 1000 тщательно отобранных и качественно написанных примеров (запрос-ответ). При этом не использовались методы RLHF.

Результаты. Полученная модель, LIMA, продемонстрировала поразительно высокое качество ответов, успешно справляясь со сложными запросами и обобщаясь на типы задач, не представленные в обучающей выборке [9]. В слепом сравнении с участием людей, ответы LIMA были предпочтены ответам GPT-4 в 43% случаев, и значительно чаще – ответам других сильных моделей того времени.

Вывод: LIMA наглядно показала, что для обучения модели следовать определенным форматам и стилям ответов, а также для активации ее способности к рассуждению, может быть достаточно небольшого, но очень качественного и разнообразного набора примеров [21]. Основные знания и способности модель получает на этапе предобучения; этап выравнивания скорее "учит" ее правильно представлять эти знания [15].

3. **LADDER (Learning through Autonomous Difficulty-Driven Example Recursion): Самогенерация учебного плана через декомпозицию задач**

Работа Саймондса и Йошиямы [30] предлагает механизм, позволяющий LLM справляться со сложными задачами путем их автономной декомпозиции.

Методология. Если LLM сталкивается с задачей, превышающей ее текущие возможности, LADDER заставляет модель саму **сформулировать несколько более простых версий этой задачи**. Затем модель решает эти упрощенные варианты, и на основе полученных решений и опыта вновь пытается решить исходную, сложную задачу. Таким образом, модель самостоятельно строит для себя "лестницу" сложности, постепенно подходя к решению.

Результаты. В экспериментах на задачах символьного интегрирования (математический анализ), базовая 3-миллиардная модель, которая изначально практически не могла решать такие задачи (точность ~1%), после применения LADDER достигла точности в **82%** на задачах университетского уровня. Это более чем 80-кратное улучшение, достигнутое без какого-либо внешнего вмешательства, исключительно за счет способности модели стратегически упрощать проблемы.

Анализ. LADDER демонстрирует потенциал LLM не только как решателей, но и как **автономных конструкторов учебных планов** [3]. Способность декомпозировать сложные задачи на более простые является ключевым аспектом человеческого интеллекта, и развитие этого навыка у ИИ открывает большие перспективы.

Эти подходы, хотя и различаются в деталях, объединены общей идеей: максимизировать использование внутренних знаний и способностей LLM, минимизируя при этом зависимость от дорогостоящей и трудоемкой внешней разметки данных.

VII. Сравнительный анализ современных техник обучения с минимумом данных

Для наглядного сопоставления рассмотренных подходов приведем их ключевые характеристики в табличной форме (Таб. 1)

Метод	Внешние данные (примерный объем)	Ключевая идея / Механизм снижения данных	Пример прироста производительности (модель)	Механизм проверки /вознаграждения	Кросс-доменный перенос	Основные риски/ ограничения
RLHF (InstructGPT)	~40k парных сравнений	Обучение модели вознаграждения на человеческих предпочтениях, RL (PPO)	GPT-3 → InstructGPT (значительное улучшение следования инструкциям)	Человеческие оценки	Ограничен	Стоимость, субъективность, масштабируемость
RLAIF (Constitutional AI)	0 сравнений, ~10 правил	ИИ-критик, действующий по "конституции"	Качество сопоставимо с RLHF при меньших затратах	Самокритика на основе правил + модель предпочтений	Умеренный	Полнота покрытия всех аспектов поведения правилами
1-shot RLVR	1 обучающий пример	RL с бинарным вознаграждением на одном тщательно отобранном примере	Qwen2.5-Math-1.5B: MATH500 +37.6 п.п.	Автоматическая проверка (математика, код)	Хороший	Выбор "правильного" примера, специфичность
Absolute Zero Reasoner (AZR)	0 внешних примеров	Self-play (Proposer+Solver), двойное RL-вознаграждение	AZR-Coder-7B: Math +15.2 п.п., Code SOTA	Автоматическая проверка (код-интерпретатор)	Очень высокий	"Uh-Oh moments", сложность контроля,

		(TRR++) в верифицируемой среде				генерация тривиальных задач
TTRL	0 (используются тестовые данные)	Majority voting на ответах модели "на лету" для генерации псевдо-вознаграждения, RL	Qwen-Math-7B: AIME-2024 +159% (Pass@1)	Majority voting (псевдо-метка)	Зависит от задачи	Зависимость от начальной производительности и модели, возможен дрейф
LIMA	~1000 качественных примеров (SFT)	Supervised fine-tuning на малом, но высококачественном и разнообразном датасете	LLaMA-65B: в 43% случаев ответы предпочтительнее GPT-4	Не применимо (SFT)	Высокий	Трудоемкость отбора и создания качественных примеров
LADDER	0 внешних примеров	Автономная рекурсивная декомпозиция сложной задачи на более простые подзадачи	3B-модель: Интегрирование +81 п.п. (с 1% до 82%)	Внутренняя проверка решения подзадач	Потенциально высокий	Сложность контроля глубины рекурсии, генерация нерелевантных подзадач

Эта таблица подчеркивает разнообразие подходов к минимизации данных, каждый из которых имеет свои сильные стороны и области применения. Общий тренд очевиден: исследователи успешно находят способы "извлекать" все больше способностей из предобученных моделей, используя все меньшие объемы специфических для задачи данных [23].

VIII. Обсуждение: эмерджентные эффекты, тренды и открытые вопросы

Применение методов обучения с минимальным количеством данных или их полным отсутствием приводит к появлению ряда интересных эмерджентных (неожиданно возникающих) свойств и поведенческих паттернов у LLM, а также выявляет общие тренды и ставит новые исследовательские вопросы.

Эмерджентные эффекты:

• **"Post-saturation generalization" (в 1-shot RLVR)**. Способность модели продолжать улучшать обобщение на тестовых данных даже после полного "заучивания" единственного тренировочного примера указывает на сложные, не до конца понятые механизмы работы RL в LLM. Это может быть связано с тем, что RL-процесс не просто подгоняет ответ под пример, а исследует различные "пути" к этому ответу, оптимизируя внутренние репрезентации или стратегии рассуждения, которые оказываются полезными и для других задач.

• **Спонтанная саморефлексия и планирование**. Наблюдение, что модели после 1-shot RLVR начинают чаще использовать рефлексивные конструкции, или что AZR-модели вставляют комментарии-планы в код, свидетельствует о том, что модели могут автономно вырабатывать более сложные когнитивные стратегии, если это способствует достижению цели (получению вознаграждения). Это очень важное наблюдение, так как оно показывает путь к развитию у ИИ метакогнитивных навыков.

• **Дифференциация "когнитивных стилей"** (в AZR). То, что модель использует разные подходы (и разную длину рассуждений) для разных типов логических задач (абдукция, дедукция, индукция), говорит о формировании у нее более тонкого и адаптированного "понимания" структуры проблемы.

Общие тренды:

• **От данных к алгоритмам и средам**. Фокус смещается с экстенсивного сбора данных на разработку умных алгоритмов обучения (RL, self-play), эффективных систем вознаграждения и, что критически важно, верифицируемых сред, где модель может получать объективную обратную связь.

• **Автономия и самосовершенствование**. Парадигмы вроде AZR и LADDER демонстрируют движение к созданию систем, способных к непрерывному самообучению и самосовершенствованию без постоянного человеческого вмешательства.

• **Роль энтропии и исследования**. Поощрение разнообразия и исследования пространства решений (через энтропийную регуляризацию или дизайн вознаграждения, как в r_propose у AZR) оказывается ключевым для предотвращения преждевременной сходимости и для раскрытия полного потенциала модели.

• **Важность качественного "запуска"**. Даже для автономных систем качество начальной инициализации (базовая модель, несколько "затравочных" примеров или принципов) может играть существенную роль.

Открытые вопросы:

• **Механизмы обобщения при экстремально малых данных**: Как именно один пример или автономная генерация задач приводят к такому значительному улучшению обобщающей способности? Каковы нейронные корреляты этих процессов?

• **Масштабируемость и пределы автономии**: Существуют ли фундаментальные ограничения для автономного обучения? Как обеспечить, чтобы самогенерируемый учебный план не заводил модель в "тупиковые" или нежелательные области знаний?

• **Перенос между сильно различными доменами**: Насколько далеко может простираться кросс-доменный перенос при обучении на узкоспециализированных самогенерируемых задачах (например, из кода в понимание сложных социальных взаимодействий)?

IX. Риски и этические аспекты

Растущая автономия LLM и их способность обучаться без прямого человеческого контроля поднимают серьезные вопросы безопасности и этики, которые требуют пристального внимания.

• **Reward Hacking (Взлом вознаграждения)**. Модели могут находить способы максимизировать функцию вознаграждения, не выполняя задачу так, как это подразумевалось разработчиками (*goodhart's law*). Это

особенно актуально для сложных, самогенерируемых сред, где функция вознаграждения может быть неидеальной. В отчете по AZR упоминается, что модель может генерировать тривиальные или нерелевантные задачи, если система вознаграждения r_propose недостаточно хорошо откалибрована или если Solver находит уязвимости в верификаторе.

• **Генерация нежелательного, вредоносного или небезопасного контента**. "Uh-Oh moment", зафиксированный для Llama3.1-8b в рамках AZR [26, Рис. 32], является ярким примером. Полностью автономная система, не ограниченная жесткими этическими рамками или человеческим надзором, может генерировать рассуждения или код, которые являются оскорбительными, дискриминационными, пропагандируют дезинформацию или даже представляют прямую угрозу безопасности (например, генерация инструкций для создания опасных веществ или вредоносного ПО).

• **Проблемы контроля, предсказуемости и интерпретируемости**. Чем автономнее система, тем сложнее предсказать ее поведение во всех возможных ситуациях и тем труднее понять причины тех или иных ее действий. Это создает риски непреднамеренного поведения, особенно в критически важных приложениях (медицина, финансы, управление инфраструктурой). Отсутствие четкого понимания, почему модель приняла то или иное решение или сгенерировала определенный контент, затрудняет аудит и отладку.

• **Плагиат и оригинальность**. При генерировании текста LLM или их дообучении на ограниченных выборках встает вопрос истинной новизны полученного материала – что требует особого внимания. Хотя прямое копирование маловероятно, модели способны незаметно воспроизводить освоенные шаблоны, стилистические конструкции и даже фразы из массивов данных предобучения. Поэтому при подготовке научных работ и создании уникального контента следует дополнительно проверять оригинальность с помощью специализированных систем (Turnitin, iThenticate, "Антиплагиат"); а также для кода и математических фрагментов такие инструменты требуют настройки под специфические единицы анализа.

• **Доменные смещения и "пузыри фильтров" (Filter Bubbles)**. Если модель генерирует собственный учебный план, существует риск, что она может "зациклиться" на определенных типах задач, данных или источников информации, которые она считает наиболее "полезными" для максимизации своего вознаграждения. Это может привести к формированию предвзятых (смещенных) или неполных знаний, игнорированию альтернативных точек зрения или важных аспектов проблемы.

• **Соответствие этическим нормам и стандартам**: При разработке и внедрении автономных обучающихся систем критично соблюдать международные принципы ответственного ИИ – рамку NIST AI RMF, стандарт ISO/IEC 42001, рекомендации ОЭСР, а также требования Регламента ЕС об ИИ и законодательства о защите данных (GDPR). Это включает защиту конфиденциальности, предотвращение дискриминации, обеспечение кибербезопасности и чёткое распределение ответственности между разработчиками и операторами

Для минимизации рисков предлагается многоуровневая архитектура безопасности:

• автоматические фильтры и классификаторы, блокирующие вредоносный или незаконный контент на входе и выходе модели;

• формальные методы верификации, подтверждающие корректность поведения критичных компонентов в заданных диапазонах состояний;

• "конституционные" подходы и RLAIF, внедряющие этические правила непосредственно в процесс обучения и оптимизации награды;

• непрерывный мониторинг, аудит и механизм *human-in-the-loop* для оперативного вмешательства человека в нестандартных сценариях;

• методы интерпретируемости и объяснимости, позволяющие анализировать логику решений LLM и выявлять потенциальную предвзятость.

X. Практические рекомендации и импликации

Рассмотренные подходы открывают новые возможности для практического применения LLM в различных областях

Сценарий применения	Рекомендуемый подход(ы)	Обоснование и комментарии
Быстрая адаптация к узкой задаче	1-shot RLVR, LIMA (few-shot SFT)	Позволяют быстро настроить модель на специфический домен или стиль ответа с минимальными затратами на данные и вычисления. Идеально для прототипирования.
Работа при нулевом доступе к данным задачи	Absolute Zero Reasoner (AZR), TTRL, LADDER	Подходят для ситуаций, где создание размеченного датасета невозможно или слишком дорого, но есть верифицируемая среда (для AZR) или возможность многократного самотестирования (для TTRL), или задача поддается декомпозиции (для LADDER).
Выравнивание стиля и тона ответа	LIMA (few-shot SFT), RLAIF	SFT на качественных примерах или использование ИИ-критика с "конституцией" эффективно для формирования желаемого стиля коммуникации и соблюдения норм.
Создание обучающих систем (EdTech)	LADDER, AZR (с модификациями)	Способность моделей автономно генерировать задачи разной сложности и адаптировать учебный план под нужды пользователя. LADDER особенно перспективен для математики и точных наук.
Разработка кода и ПО	Absolute Zero Reasoner, TTRL	Автономное улучшение способностей к кодогенерации, отладке, рефакторингу и анализу кода. TTRL может помочь в адаптации к специфическим кодовым базам "на лету".
Научные исследования	LADDER, AZR (для формальных систем)	Потенциал для помощи в решении сложных научных проблем путем декомпозиции (LADDER) или автономного исследования в формально описанных средах (AZR).

Экономические импликации. Переход к методам обучения с минимумом данных способен значительно снизить затраты на разработку и внедрение ИИ-решений, демократизируя доступ к передовым технологиям LLM для более широкого круга организаций и исследователей [2]. Это может ускорить внедрение ИИ в новые отрасли и для решения более широкого круга задач.

XI. Будущие направления исследований

Область обучения LLM с минимумом данных активно развивается, и существует множество перспективных направлений для будущих исследований:

- **Развитие теории curriculum-reward и learnability**. Формализация принципов, по которым модель (или система) должна выбирать или генерировать задачи для максимизации скорости и качества обучения. Исследование оптимальных стратегий для r_propose в AZR-подобных системах.
- **Расширение автономных парадигм (AZR, LADDER) на мультимодальные данные и более сложные, менее структурированные среды**. Например, обучение агентов, взаимодействующих с реальным миром через сенсоры или со сложными симуляциями, где верификация нетривиальна.
- **Интеграция продвинутых механизмов безопасности, контроля и этического выравнивания непосредственно в циклы self-play и автономного обучения**. Создание "врожденных" этических ограничителей и систем самоцензуры, которые не снижают креативность и эффективность модели.
- **Исследование гибридных подходов**. Например, использование few-shot learning (LIMA-стиль) или 1-shot RLVR для "запуска" или инициализации автономных систем типа AZR, чтобы направить их начальное развитие в нужную сторону и ускорить обучение.
- **Улучшение интерпретируемости и объяснимости**. Разработка методов, позволяющих понимать, как и почему автономно обучающиеся модели приходят к тем или иным решениям или вырабатывают определенные эмерджентные свойства. Это критично для доверия и отладки.
- **Масштабируемость и эффективность алгоритмов self-play**. Снижение вычислительных затрат, связанных с многократной генерацией и оценкой задач и решений в автономных циклах [5].
- **Перенос обучения между моделями (Model-to-Model Transfer in Zero-Data Settings)**. Исследование того, как знания или учебные стратегии, выработанные одной автономной моделью, могут быть эффективно переданы другой, возможно, меньшей или специализированной модели.

XII. Заключение

Современные исследования в области обучения больших языковых моделей демонстрируют впечатляющий сдвиг от парадигмы, основанной на огромных объемах размеченных данных, к методам, использующим минимальное количество информации или полностью автономные циклы самообучения. Работы по 1-shot RLVR показывают, что даже один тщательно подобранный пример может катализировать значительные улучшения в способностях LLM к рассуждению [24]. Парадигма "Абсолютного нуля", реализованная в Absolute Zero Reasoner, делает следующий шаг, демонстрируя, как модель может самостоятельно генерировать для себя учебный план и достигать state-of-the-art результатов без каких-либо внешних данных, опираясь на верифицируемую среду. Другие подходы, такие как TTRL, LIMA и LADDER, также вносят свой вклад в копилку методов эффективного обучения, подчеркивая важность качества данных, онлайн-адаптации и стратегической декомпозиции задач.

Ключевым становится не столько объем данных, сколько интеллектуальность самого процесса обучения: дизайн сигналов вознаграждения, создание верифицируемых сред для обратной связи, и способность модели к самостоятельному исследованию и построению учебной программы. Эти достижения открывают путь к созданию более экономичных, масштабируемых, адаптивных [13] и потенциально более мощных систем искусственного интеллекта. Эмерджентные свойства, такие как спонтанное планирование и дифференциация когнитивных стилей, указывают на формирование у моделей более сложных внутренних механизмов обработки информации.

Однако, вместе с ростом автономии моделей возрастает и важность вопросов безопасности, контроля и этики. "Uh-Oh moments" и другие потенциальные риски требуют разработки новых подходов к обеспечению предсказуемости и надежности ИИ-систем. Будущее, вероятно, за гибридными моделями, сочетающими элементы автономного исследования с механизмами человеческого надзора и этического выравнивания, а также за постоянным совершенствованием методов автоматической верификации и интерпретируемости.

Переключение фокуса со "сбора разметки" на "дизайн сигнала вознаграждения и среды проверки" трансформирует экономику разработки LLM [18], открывая путь к более автономным и масштабируемым экспертным системам, при условии своевременного и серьезного решения вопросов безопасности и контроля.

Библиография

[1] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, et al., "Constitutional AI: Harmlessness from AI Feedback," 2022, arXiv:2212.08073, doi: 10.48550/arXiv.2212.08073.

[2] E. Ben Zaken, S. Ravfogel, and Y. Goldberg, "BitFit: Simple Parameter efficient Fine tuning for Transformer based Masked Language models," 2022, arXiv:2106.10199, doi: 10.48550/arXiv.2106.10199.

[3] X. Chen, Y. Deng, M. Wang, and Y. Zhang, "Skeleton of Thought: Prompting Large Language Models for Efficient Parallel Generation," 2024, arXiv:2307.15337, doi: 10.48550/arXiv.2307.15337.

[4] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized Large Language Models," 2023, arXiv:2305.14314, doi: 10.48550/arXiv.2305.14314.

[5] Z. Shi and A. Lipani, "DePT: Decomposed Prompt Tuning for Parameter-Efficient Fine-tuning," 2023, arXiv:2309.05173, doi: 10.48550/arXiv.2309.05173.

[6] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low Rank Adaptation of Large Language Models," 2021, arXiv:2106.09685, doi: 10.48550/arXiv.2106.09685.

[7] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large Language Models are Zero Shot Reasoners," 2023, arXiv:2205.11916, doi: 10.48550/arXiv.2205.11916.

[8] B. Lester, R. Al-Rfou, and N. Constant, "The Power of Scale for Parameter Efficient Prompt Tuning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Nov. 2021, pp. 3045-3059, doi: 10.18653/v1/2021.emnlp-main.243.

[9] X. L. Li, and P. Liang, "Prefix-Tuning: Optimising Continuous Prompts for Generation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Aug. 2021, pp. 4582-4597, doi: 10.18653/v1/2021.acl-long.353.

[10] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, et al., "SELF-REFINE: iterative refinement with self-feedback," in *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*, Dec. 2023, Art. no. 2019, pp. 46534-46594. Available:

https://papers.nips.cc/paper_files/paper/2023/file/91edff07232fb1b55a505a9e9f6c0ff3-Paper-Conference.pdf
[11] S. Min, M. Lewis, L. Zettlemoyer, and H. Hajishirzi, "MetaICL: Learning to Learn In Context," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Jul. 2022, pp. 2791-2809, doi: 10.18653/v1/2022.naacl-main.201.
[12] N. Ding, Y. Qin, G. Yang, F. Wei, Z. Yang, Y. Su, et al., "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nature Machine Intelligence*, vol. 5, no. 3, pp. 220-235, Mar. 2023, doi: 10.1038/s42256-023-00626-4.
[13] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, et al., "Training language models to follow instructions with human feedback," in *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, Oct. 2022, pp. 1-15. Available: https://openreview.net/pdf?id=TG8KACxEON
[14] Z. Han, C. Gao, J. Liu, J. Zhang, and S. Q. Zhang, "Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey," 2024, arXiv:2403.14608, doi: 10.48550/arXiv.2403.14608.
[15] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto, "Alpaca: A Strong, Replicable Instruction Following Model," Stanford CRFM Tech. Rep., 2023. Available: https://crfm.stanford.edu/2023/03/13/alpaca.html
[16] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-Consistency Improves Chain of Thought Reasoning in Language Models," 2023, arXiv:2203.11171, doi: 10.48550/arXiv.2203.11171.
[17] Y. Wang, Q. Yang, Z. Zeng, L. Ren, L. Liu, B. Peng, et al., "Reinforcement Learning for Reasoning in Large Language Models with One Training Example," 2025, arXiv:2504.20571, doi: 10.48550/arXiv.2504.20571.
[18] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, Ed H. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," *in Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*, Nov. 2022, Art. no. 1800, pp. 24824-24837. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf
[19] Y. Wu, et al., "Self-Play Preference Optimization for Language Model Alignment," 2024, arXiv:2405.00675, doi: 10.48550/arXiv.2405.00675.
[20] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "ReAct: Synergizing Reasoning and Acting in Language Models," 2023, arXiv:2210.03629, doi: 10.48550/arXiv.2210.03629.
[21] H. Lee, et al. "RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback," 2023, arXiv:2309.00267, doi: 10.48550/arXiv.2309.00267.
[22] B. Lester, et al., "Reducing Retraining by Recycling Parameter-Efficient Prompts," 2022, arXiv:2208.05577, doi: 10.48550/arXiv.2208.05577.
[23] N. Ding, et al., "Delta Tuning: A Comprehensive Study of Parameter Efficient Methods for Pre-trained Language Models," 2022, arXiv:2203.06904, doi: 10.48550/arXiv.2203.06904.
[24] W. Yuan, R. Y. Pang, K. Cho, X. Li, S. Sukhbaatar, J. Xu, and J. Weston, "Self-rewarding language models," in *Proceedings of the 41st International Conference on Machine Learning (ICML'24)*, Jul. 2024, Art. no. 2389, pp. 57905-57923. Available: https://proceedings.mlr.press/v235/yuan24d.html
[25] R. Rafailov, et al., "Direct Preference Optimization: Your Language Model is Secretly a Reward Model," 2024, arXiv:2305.18290, doi: 10.48550/arXiv.2305.18290.
[26] A. Zhao, Y. Wu, Y. Yue, T. Wu, Q. Xu, Y. Yue, et al., "Absolute Zero: Reinforced Self Play Reasoning with Zero Data," 2025, arXiv:2505.03335, doi: 10.48550/arXiv.2505.03335.
[27] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, et al., "LIMA: less is more for alignment," in: *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*, Dec. 2023, Art. no. 2400, pp. 55006-55021. Available: https://openreview.net/pdf?id=KBMOKmX2he
[28] Z. Sun, et al., "Principle-Driven Self-Alignment of Language Models from Scratch with Minimal Human Supervision,". In: *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*, Dec. 2023, Art. no. 115, pp. 2511-2565. Available: https://openreview.net/pdf?id=p40XRfBX96
[29] Y. Zuo, K. Zhang, L. Sheng, S. Qu, G. Cui, X. Zhu, et al., "TTRL: Test-Time Reinforcement Learning," 2025, arXiv:2504.16084, doi: 10.48550/arXiv.2504.16084.
[30] T. Simonds and A. Yoshiyama, "LADDER: Self-Improving LLMs Through Recursive Problem Decomposition," 2025, arXiv:2503.00735, doi: 10.48550/arXiv.2503.00735.

Об авторах:

Пичугов Алексей Александрович, ведущий специалист кафедры информационной безопасности факультета вычислительной математики и кибернетики, ФГБОУ ВО «Московский государственный университет имени М.В. Ломоносова» (119991, Российская Федерация, г. Москва, ГСП-1, Ленинские горы, д. 1), ORCID: https://orcid.org/0009-0001-3489-7915, pichugov.a.a@cs.msu.ru
Намиот Дмитрий Евгеньевич, ведущий научный сотрудник лаборатории открытых информационных технологий кафедры информационной безопасности факультета вычислительной математики и кибернетики, доктор технических наук, ФГБОУ ВО «Московский государственный университет имени М.В. Ломоносова» (119991, Российская Федерация, г. Москва, ГСП-1, Ленинские горы, д. 1), ORCID: https://orcid.org/0000-0002-4463-1678, dnamiot@gmail.com
Зубарева Елена Васильевна, старший научный сотрудник лаборатории открытых информационных технологий кафедры информационной безопасности факультета вычислительной математики и кибернетики, кандидат педагогических наук, доцент, ФГБОУ ВО «Московский государственный университет имени М.В. Ломоносова» (119991, Российская Федерация, г. Москва, ГСП-1, Ленинские горы, д. 1), ORCID: https://orcid.org/0000-0002-9997-4715, e.zubareva@cs.msu.ru

Modern Methods for Training Large Language Models with Minimal Data: From One Example to Absolute Zero – an Academic Review

Alexey Pichugov, Dmitry Namiot, Elena Zubareva

***Abstract* – This academic review systematizes and analyzes breakthrough approaches to training large language models (LLM) developed between 2022 and 2025, with an emphasis on radically reducing or eliminating the reliance on human-labeled data. The paper examines in detail methodologies such as single-shot reinforcement learning, the Absolute Zero Reasoner paradigm of fully autonomous learning, test-time learning, low-example efficient learning, and self-generated curriculum via task decomposition. The key results of these approaches on standard benchmarks are analyzed, and the emergent cognitive properties of the models, cross-domain effects, and associated risks and ethical aspects are discussed. The review is intended for students, teachers, researchers, and practitioners in the fields of artificial intelligence and natural language processing seeking to understand the cutting edge of research in the field of effective LLM teaching.**

***Keywords* – Large Language Models, Minimum Data Learning, 1-shotRLVR, Absolute Zero Reasoner, Self-Play Learning, Test-Time Reinforcement Learning, Low-Example Alignment, LADDER, Parameter-Efficient Fine-Tuning, Chain-of-Thought, Cross-Domain Transfer, Emergent LLM Capabilities, Security, Reward-Hacking.**

REFERENCES

[1] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, et al., "Constitutional AI: Harmlessness from AI Feedback," 2022, arXiv:2212.08073, doi: 10.48550/arXiv.2212.08073.

[2] E. Ben Zaken, S. Ravfogel, and Y. Goldberg, "BitFit: Simple Parameter efficient Fine tuning for Transformer based Masked Language models," 2022, arXiv:2106.10199, doi: 10.48550/arXiv.2106.10199.

[3] X. Chen, Y. Deng, M. Wang, and Y. Zhang, "Skeleton of Thought: Prompting Large Language Models for Efficient Parallel Generation," 2024, arXiv:2307.15337, doi: 10.48550/arXiv.2307.15337.

[4] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized Large Language Models," 2023, arXiv:2305.14314, doi: 10.48550/arXiv.2305.14314.

[5] Z. Shi and A. Lipani, "DePT: Decomposed Prompt Tuning for Parameter-Efficient Fine-tuning," 2023, arXiv:2309.05173, doi: 10.48550/arXiv.2309.05173.

[6] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low Rank Adaptation of Large Language Models," 2021, arXiv:2106.09685, doi: 10.48550/arXiv.2106.09685.

[7] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large Language Models are Zero Shot Reasoners," 2023, arXiv:2205.11916, doi: 10.48550/arXiv.2205.11916.

[8] B. Lester, R. Al-Rfou, and N. Constant, "The Power of Scale for Parameter Efficient Prompt Tuning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Nov. 2021, pp. 3045-3059, doi: 10.18653/v1/2021.emnlp-main.243.

[9] X. L. Li, and P. Liang, "Prefix-Tuning: Optimising Continuous Prompts for Generation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Aug. 2021, pp. 4582-4597, doi: 10.18653/v1/2021.acl-long.353.

[10] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, et al., "SELF-REFINE: iterative refinement with self-feedback," in *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*, Dec. 2023, Art. no. 2019, pp. 46534-46594. Available: https://papers.nips.cc/paper_files/paper/2023/file/91edff07232fb1b55a505a9e9f6c0ff3-Paper-Conference.pdf

[11] S. Min, M. Lewis, L. Zettlemoyer, and H. Hajishirzi, "MetaICL: Learning to Learn In Context," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Jul. 2022, pp. 2791-2809, doi: 10.18653/v1/2022.naacl-main.201.

[12] N. Ding, Y. Qin, G. Yang, F. Wei, Z. Yang, Y. Su, et al., "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nature Machine Intelligence*, vol. 5, no. 3, pp. 220-235, Mar. 2023, doi: 10.1038/s42256-023-00626-4.

[13] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, et al., "Training language models to follow instructions with human feedback," in *Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, Oct. 2022, pp. 1-15. Available: https://openreview.net/pdf?id=TG8KACxEON

[14] Z. Han, C. Gao, J. Liu, J. Zhang, and S. Q. Zhang, "Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey," 2024, arXiv:2403.14608, doi: 10.48550/arXiv.2403.14608.

[15] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto, "Alpaca: A Strong, Replicable Instruction Following Model," Stanford CRFM Tech. Rep., 2023. Available: https://crfm.stanford.edu/2023/03/13/alpaca.html

[16] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-Consistency Improves Chain of Thought Reasoning in Language Models," 2023, arXiv:2203.11171, doi: 10.48550/arXiv.2203.11171.

[17] Y. Wang, Q. Yang, Z. Zeng, L. Ren, L. Liu, B. Peng, et al., "Reinforcement Learning for Reasoning in Large Language Models with One Training Example," 2025, arXiv:2504.20571, doi: 10.48550/arXiv.2504.20571.

[18] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, Ed H. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," *in Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)*, Nov. 2022, Art. no. 1800, pp. 24824-24837. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

[19] Y. Wu, et al., "Self-Play Preference Optimization for Language Model Alignment," 2024, arXiv:2405.00675, doi: 10.48550/arXiv.2405.00675.

[20] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "ReAct: Synergizing Reasoning and Acting in Language Models," 2023, arXiv:2210.03629, doi: 10.48550/arXiv.2210.03629.

[21] H. Lee, et al. "RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback," 2023, arXiv:2309.00267, doi: 10.48550/arXiv.2309.00267.

[22] B. Lester, et al., "Reducing Retraining by Recycling Parameter-Efficient Prompts," 2022, arXiv:2208.05577, doi: 10.48550/arXiv.2208.05577.

[23] N. Ding, et al., "Delta Tuning: A Comprehensive Study of Parameter Efficient Methods for Pre-trained Language Models," 2022, arXiv:2203.06904, doi: 10.48550/arXiv.2203.06904.

[24] W. Yuan, R. Y. Pang, K. Cho, X. Li, S. Sukhbaatar, J. Xu, and J. Weston, "Self-rewarding language models," in *Proceedings of the 41st International Conference on Machine Learning (ICML'24)*, Jul. 2024, Art. no. 2389, pp. 57905-57923. Available: https://proceedings.mlr.press/v235/yuan24d.html

[25] R. Rafailov, et al., "Direct Preference Optimization: Your Language Model is Secretly a Reward Model," 2024, arXiv:2305.18290, doi: 10.48550/arXiv.2305.18290.

[26] Zhao, Y. Wu, Y. Yue, T. Wu, Q. Xu, Y. Yue, et al., "Absolute Zero: Reinforced Self Play Reasoning with Zero Data," 2025, arXiv:2505.03335, doi: 10.48550/arXiv.2505.03335.

[27] C. Zhou, P. Liu, P. Xu, S. Iyer, J. Sun, Y. Mao, et al., "LIMA: less is more for alignment," in: *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*, Dec. 2023, Art. no. 2400, pp. 55006-55021. Available: https://openreview.net/pdf?id=KBMOKmX2he

[28] Z. Sun, et al., "Principle-Driven Self-Alignment of Language Models from Scratch with Minimal Human Supervision,". In: *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*, Dec. 2023, Art. no. 115, pp. 2511-2565. Available: https://openreview.net/pdf?id=p40XRfBX96

[29] Y. Zuo, K. Zhang, L. Sheng, S. Qu, G. Cui, X. Zhu, et al., "TTRL: Test-Time Reinforcement Learning," 2025, arXiv:2504.16084, doi: 10.48550/arXiv.2504.16084.

[30] T. Simonds and A. Yoshiyama, "LADDER: Self-Improving LLMs Through Recursive Problem Decomposition," 2025, arXiv:2503.00735, doi: 10.48550/arXiv.2503.00735.

About the authors:

Alexey A. Pichugov, Leading Specialist of the Department of Information Security, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University (1 Leninskie gory, Moscow 119991, GSP-1, Russian Federation), ORCID: https://orcid.org/0009-0001-3489-7915, pichugov.a.a@cs.msu.ru

Dmitry E. Namiot, Leading Researcher of the Open Information Technologies Lab, Department of Information Security, Faculty of Computational Mathematics and Cybernetics, Dr. Sci. (Eng.), Lomonosov Moscow State University (1 Leninskie gory, Moscow 119991, GSP-1, Russian Federation), ORCID: https://orcid.org/0000-0002-4463-1678, dnamiot@gmail.com

Elena V. Zubareva, Senior Researcher of the Open Information Technologies Lab, Department of Information Security, Faculty of Computational Mathematics and Cybernetics, Cand. Sci. (Ped.), Associate Professor, Lomonosov Moscow State University (1 Leninskie gory, Moscow 119991, GSP-1, Russian Federation), ORCID: https://orcid.org/0000-0002-9997-4715, e.zubareva@cs.msu.ru