УДК 004.6

Мешков Никита Сергеевич, студент кафедры
«Автоматизированных систем управления
биотехнологическими процессами»
Российский биотехнологический университет
(РОСБИОТЕХ)

САМООУЛУЧШАЮЩИЕСЯ МУЛЬТИМОДАЛЬНЫЕ LLM В ПРИКЛАДНЫХ ЗАДАЧАХ: ОБЗОРЫ ПОДХОДОВ К АВТОСБОРУ ДАННЫХ И БЕЗОПАСНЫМ КОНТУРАМ САМООБУЧЕНИЯ

Аннотация. Статья посвящена архитектурам и механизмам самоулучшения мультимодальных LLM. Рассмотрены методы автосбора и обогащения данных, безопасные контуры самообучения и подходы к контролю качества. Показано, что гибридные стратегии повышают устойчивость, прозрачность и безопасность самообучающихся моделей.

Ключевые слова: Самоулучшающиеся модели; мультимодальные LLM; автосбор данных; контуры самообучения; безопасное обучение; reinforcement learning; интерпретируемость искусственного интеллекта.

Введение

Быстрое развитие больших языковых моделей (LLM) стало одним из наиболее значимых явлений в сфере искусственного интеллекта последнего десятилетия. Особенно стремительно продвигается направление мультимодальных систем, способных обрабатывать не только текст, но и изображения, звук, видео и другие типы данных [10]. Эти модели создают основу для интеллектуальных помощников нового поколения, обеспечивая естественное взаимодействие между человеком и машиной в различных форматах восприятия. Появление самообучающихся архитектур открыло возможности для постоянного совершенствования моделей без участия человека, однако одновременно породило риски, связанные с безопасностью и качеством данных [8].

Интерес к самоулучшающимся LLM объясняется тем, что традиционный процесс обучения требует огромных ресурсов, включая вычислительные мощности и ручную разметку. В отличие от этого, самоулучшающиеся модели способны самостоятельно генерировать, фильтровать и использовать данные для собственной адаптации. Такой подход реализован в системах Self-Instruct и WebGPT, где модель не только обучается на внешних источниках, но и формирует внутренние циклы обратной связи. Эти контуры позволяют LLM определять недостающие знания, запрашивать дополнительную информацию и оценивать корректность собственных ответов.

В прикладных задачах – от интеллектуальных ассистентов до генеративных решений в медицине и образовании – подобные механизмы повышают эффективность работы моделей. Вместе с тем автономность таких систем создаёт вызовы: при отсутствии строгих механизмов контроля возможна деградация качества и накопление ошибок. В ряде случаев выявлено, что циклы самообучения без фильтрации данных приводят к усилению предвзятостей и образованию замкнутых «эхо-систем».

Основная цель статьи заключается в анализе подходов к автосбору данных и построению безопасных контуров самообучения мультимодальных LLM [6]. Для достижения цели ставятся следующие задачи:

1. Рассмотреть принципы построения самоулучшающихся мультимодальных моделей и их архитектурные особенности;
2. Проанализировать методы автоматического сбора и отбора обучающих данных;
3. Изучить практики безопасного самообучения и выравнивания поведения моделей.

Объектом рассмотрения выступают современные мультимодальные языковые модели, обладающие механизмами самоадаптации. Предметом – методы и архитектурные решения, обеспечивающие устойчивое развитие этих моделей при самонастройке.

Практическая значимость темы заключается в том, что способность моделей к автономному самообновлению открывает новые возможности для прикладных решений, включая автоматизированные NLP-системы и мультимодальные аналитические комплексы [4]. Разработка надёжных механизмов контроля и безопасных контуров обучения становится ключевым условием устойчивости и доверия к таким технологиям.

В итоге формируется комплексное представление о современных тенденциях и вызовах, связанных с созданием самоулучшающихся мультимодальных LLM, а также очерчиваются направления для их безопасного внедрения в реальных задачах.

1. Концепция самоулучшающихся мультимодальных LLM

Развитие больших языковых моделей изначально происходило в русле задач обработки текста, где ключевыми элементами выступали языковая предсказательная способность и генерация связных фрагментов речи [2]. Однако по мере расширения вычислительных ресурсов и совершенствования архитектур стало очевидно, что ограничение только текстовой модальностью сдерживает потенциал таких систем. Современные модели нового поколения – GPT-5V, Gemini, Claude 3, Mistral Fusion и другие – способны работать одновременно с несколькими типами данных: текстом, изображениями, аудио и видео. Это направление получило название мультимодальность, а появление механизмов самоадаптации – самоулучшение [9].

Суть самоулучшающихся моделей заключается в том, что они не просто выполняют задачи, но и способны корректировать собственное поведение на основе накопленного опыта [6]. В классическом сценарии обучения модель обновляется в ходе внешнего переобучения, тогда как в самоулучшающихся архитектурах применяются внутренние механизмы оценки и обратной связи [13]. Такие механизмы позволяют системе анализировать свои ответы, выявлять ошибки, формировать запросы на дополнительную информацию и даже обновлять внутренние представления данных. Подобный подход создаёт основу для контурного обучения (looped training), где модель выступает одновременно и исполнителем, и оценщиком [9].

Мультимодальные LLM, обладая способностью к самооценке и интеграции различных типов данных, демонстрируют качественно новый уровень адаптивности [10]. Например, в системах визуального понимания модель может корректировать описание изображения на основе текстовой критики, а в аудиовизуальных сценариях – уточнять смысловую нагрузку фразы, сверяя интонационные и визуальные контексты. В такой конфигурации самоулучшение приобретает форму перекрёстной адаптации (cross-modal self-alignment), когда данные из одной модальности помогают улучшить интерпретацию другой.

Одним из ключевых направлений в архитектуре самоулучшающихся LLM является self-distillation – процесс, при котором модель обучает сама себя, используя собственные предсказания как источник знаний [3]. Эта методика активно применяется в крупных системах, таких как DeepMind Gemini или OpenAI GPT-4, где внутренняя дистилляция используется для стабилизации поведения и минимизации зависимости от ручной аннотации данных. Self-distillation позволяет создавать внутренние циклы консолидации знаний, что особенно важно для мультимодальных моделей, где разнообразие входных данных требует устойчивой согласованности между модальностями.

Другим важным элементом архитектуры является reinforcement learning from human feedback (RLHF) и его расширенные формы – reinforcement learning from AI feedback (RLAIF). Последняя концепция предполагает, что обратную связь модели предоставляет не человек, а другая, более стабильная версия той же системы или специализированный модуль-оценщик. В результате формируется динамическая пара: одна часть модели генерирует гипотезу, другая – оценивает её корректность и формулирует обратный сигнал. Этот принцип активно используется в разработках Anthropic и OpenAI, обеспечивая контролируемое самообучение без постоянного участия человека [6].

В контексте прикладных задач самоулучшение особенно важно там, где требуется высокая устойчивость к изменениям среды и разнообразию данных. Примером могут служить интеллектуальные ассистенты, подстраивающиеся под стиль общения пользователя, системы медицинской диагностики, корректирующие свои гипотезы на основе накопленных случаев, или промышленные решения в сфере робототехники, где LLM-ядро управляет автономными агентами. В этих сценариях механизмы самообучения позволяют минимизировать время реакции, снижая зависимость от периодических обновлений моделей разработчиками.

Тем не менее автономность таких систем порождает целый спектр рисков. Самоулучшающиеся контуры нередко склонны к дрейфу данных (data drift) и усилению предвзятостей. Если модель многократно использует собственные ответы в качестве обучающих примеров, происходит накопление искажения: ошибки начинают воспроизводиться и усиливаться. Без внешней фильтрации и независимых метрик качества самообучение может привести к деградации. Этот эффект особенно заметен в мультимодальных системах, где несогласованность между модальностями способна привести к некорректным или даже опасным выводам.

В ответ на эти вызовы всё больше внимания уделяется разработке безопасных контуров самообучения. Под этим термином подразумеваются архитектурные решения, включающие встроенные фильтры, оценочные модули, а также внешние контрольные петли, позволяющие отслеживать поведение модели в процессе самоадаптации. Безопасный контур предполагает наличие механизма, который не допускает использования недостоверных данных в качестве обучающих, обеспечивает проверку выводов по независимым критериям и предотвращает циклы самоподкрепления ошибок. Такой подход активно исследуется в рамках систем Constitutional AI, Judge-Loop и Safe-RLHF, где безопасность рассматривается как неотъемлемая часть архитектуры, а не как внешнее ограничение.

Особое место в современных подходах занимает идея динамической прозрачности (dynamic interpretability). Её смысл состоит в том, чтобы не только улучшать модель, но и объяснять причины её самоизменений [1]. Для самоулучшающихся LLM это критически важно: только при наличии трассируемости (traceability) можно гарантировать, что каждая фаза самообучения соответствует этическим, техническим и регуляторным требованиям. Разработка прозрачных контуров самоадаптации становится одним из центральных направлений исследований в области ответственного ИИ [5].

В результате можно констатировать, что концепция самоулучшающихся мультимодальных LLM формирует новый этап развития искусственного интеллекта, основанный на синтезе автономности, мультимодальности и устойчивого контроля [10]. Такие системы способны не только интегрировать знания из разных источников, но и выстраивать собственные механизмы обучения. Потенциал их применения огромен, однако без выверенных архитектур безопасности и интерпретируемости этот потенциал сопряжён с рисками. Поэтому дальнейшее развитие технологий будет определяться не только скоростью роста вычислительных мощностей, но и способностью разрабатывать ответственные контуры самообучения, обеспечивающие надёжное функционирование LLM в прикладных сценариях [8].

2. Подходы к автосбору и самообогащению обучающих данных

Одним из ключевых факторов, определяющих качество и надёжность самоулучшающихся мультимодальных LLM, является организация цикла работы с данными. Автоматический сбор, фильтрация и повторное использование информации формируют основу устойчивого самообучения, где каждая итерация позволяет модели уточнять внутренние представления и корректировать ошибки. Однако эффективность этого процесса напрямую зависит от того, насколько грамотно построен контур автосбора – от источников и методов фильтрации до механизмов оценки релевантности [11].

Современные архитектуры самоулучшающихся LLM используют несколько стратегий автосбора данных. Первая из них – self-instruction (самогенерация инструкций). Она предполагает, что модель самостоятельно формирует обучающие примеры, основываясь на

имеющихся знаниях и заранее заданных шаблонах поведения [3]. Метод был впервые предложен в проекте Self-Instruct (2022), ставшем базой для ряда последующих систем. В этом подходе модель не нуждается в ручной разметке: она создаёт запрос, формулирует ответ и затем оценивает его качество через дополнительный модуль или с помощью другой версии самой себя [2].

Вторая стратегия – web-scale auto-curation, ориентированная на использование открытых интернет-источников [7]. Здесь модель анализирует большие потоки данных из сети, автоматически выявляя и сохраняет релевантные фрагменты. Подобные методы применяются в фреймворках WebGPT, DataComp и LAION-5B, где упор делается на масштабируемость и адаптивность. Для мультимодальных систем особую ценность представляют наборы, включающие изображения, аудио и видео, что позволяет расширять когнитивные связи модели между модальностями.

Третья стратегия – reinforcement-guided selection, при которой модель обучается выбирать лучшие примеры на основе внутренней системы вознаграждений. Такой подход лежит в основе RLHF и RLAIF, где обратная связь выступает инструментом для отбора данных с высокой когнитивной ценностью [6]. Механизмы подкрепления позволяют формировать динамические обучающие наборы, где информация не просто накапливается, а обновляется в зависимости от текущих результатов модели.

Ниже представлена сравнительная таблица 1, отражающая ключевые принципы трёх наиболее распространённых подходов к автосбору данных.

Таблица 1

Сравнительные характеристики подходов к автосбору данных в LLM

Подход	**Источник данных**	**Принцип отбора**	**Механизм обратной связи**	**Основные преиму-щества**	**Потенциаль ные риски**
Self-Instruct	Генерация моделью	Самосоздание инструкций и ответов	Оценка через вторичную генерацию	Автоном-ность, низкая зависимость от человека	Ошибки самопод-тверждения, деградация качества
Web-scale auto-curation	Веб-контент, мультимо-дальные источники	Алгорит-мическая фильтрация и кластеризация	Использован ие внешних фильтров и моделей	Масштаби-руемость, разнообразие данных	Высокая вероятность шума и предвзятости
Reinforcement-guided selection (RLHF/RLAIF)	Внутренние и внешние данные	Выбор по функции награды	Модуль оценки или другая модель	Контроли-руемое улучшение, адаптивность	Сложность настройки, вычислитель ная нагрузка

В результате анализа видно, что каждая стратегия имеет свои сильные стороны и ограничения. Self-Instruct подходит для начальных этапов самообучения, когда модель формирует базовые паттерны поведения, но без внешнего контроля может воспроизводить ошибки. Web-scale auto-curation обеспечивает масштабируемость и богатство данных, однако нуждается в жёстких фильтрах и балансировке, чтобы избежать накопления «информационного шума». Подход reinforcement-guided selection решает проблему качества за счёт встроенного механизма обратной связи, но требует значительных вычислительных ресурсов и тщательной калибровки метрик награды.

Для мультимодальных систем особое значение имеет сочетание этих методов. Например, визуально-текстовые модели могут использовать self-instruction для генерации пар «вопрос–ответ» по изображениям, а web-curation – для сбора новых визуальных данных. Далее подключается reinforcement-guided механизм, который оценивает, насколько хорошо новая

информация улучшает межмодальную согласованность. Такой гибридный сценарий позволяет создавать замкнутые, но контролируемые циклы обучения, где каждая модальность усиливает другую.

В последние годы всё чаще обсуждается идея quality feedback loops – контуров качества, которые интегрируют разные стратегии автосбора в единую систему. В таких архитектурах модель не просто собирает данные, а проверяет их внутренними модулями контроля. Например, сначала производится генерация контента (self-instruction), затем фильтрация с помощью web-curation, а завершается процесс оценкой релевантности через reinforcement-guided selection. Этот многоступенчатый цикл позволяет добиться баланса между масштабом данных и их точностью.

Немаловажным направлением является автоматическая самооценка источников (self-evaluation of sources). Она основана на принципе, что модель способна присваивать каждому фрагменту данных уровень достоверности, анализируя соответствие содержимого внутренним знаниям. Такие механизмы внедряются в рамках систем auto-moderation и self-verification, где обучающая выборка постепенно очищается от противоречивых или токсичных примеров. Для мультимодальных LLM это особенно критично, так как взаимодействие между изображением, текстом и звуком требует высокой семантической согласованности.

Актуальной проблемой остаётся феномен data drift – смещения данных, возникающего при повторных итерациях самообучения. Он проявляется тогда, когда модель многократно использует собственные выходные данные в качестве обучающих [9]. При этом распределение информации постепенно отклоняется от исходного, что приводит к снижению разнообразия и появлению шаблонных, менее информативных ответов. Для борьбы с этим эффектом применяются алгоритмы data rejuvenation и entropy-based filtering, которые отслеживают статистическое разнообразие данных и автоматически исключают из выборки избыточно повторяющиеся элементы.

Интерес представляет и подход meta-curation, который предполагает создание дополнительного слоя управления над процессом автосбора. Здесь модель не только выполняет задачи фильтрации, но и анализирует собственную стратегию подбора данных. Такой мета-уровень делает систему ближе к когнитивным аналогам человеческого самообучения, где внимание распределяется динамически в зависимости от важности информации [15].

В итоге эффективная работа самоулучшающихся мультимодальных LLM невозможна без тщательно спроектированных контуров автосбора и самообогащения данных. Гибридное сочетание различных стратегий позволяет поддерживать баланс между масштабом и качеством, а внедрение механизмов контроля – снижать риск деградации и предвзятости. На следующем этапе развития таких систем ключевым направлением становится внедрение безопасных контуров самообучения, обеспечивающих прозрачность и надёжность всей цепочки обработки данных.

3. Безопасные контуры самообучения и контроль качества

Одним из наиболее сложных и одновременно критически важных направлений развития самоулучшающихся мультимодальных моделей является формирование безопасных контуров самообучения. Эти контуры обеспечивают внутреннюю управляемость, контроль качества и прозрачность процессов, происходящих внутри модели. В отличие от традиционного обучения, где данные и метрики определяются извне, самоулучшающиеся архитектуры нуждаются во встроенных механизмах фильтрации и валидации, способных предотвращать накопление ошибок и деградацию знаний.

Безопасный контур представляет собой совокупность взаимосвязанных компонентов, выполняющих три ключевые функции:

Оценка корректности обучающих данных (data validation loop);

Мониторинг и фильтрация выходных ответов (output moderation loop);

Внутреннее самооценивание поведения модели (self-judgment loop).

Каждый из этих элементов формирует уровень защиты, который позволяет модели развиваться автономно, но при этом оставаться в пределах допустимых рамок достоверности и этики.

В современных архитектурах контуры безопасности реализуются через комбинацию алгоритмических и нейросетевых методов. Например, в системах Constitutional AI от Anthropic или Safe-RLHF от OpenAI используются модели-судьи, которые оценивают качество ответов основной системы. Эти модули выполняют роль внутреннего аудитора, способного выявлять токсичные, недостоверные или противоречивые ответы, а затем корректировать поведение модели через обратную связь. Аналогичные принципы применяются и в проектах Judge-Loop, где итеративная оценка позволяет стабилизировать процесс самообучения и предотвращать циклы самоподкрепления ошибок.

Для визуализации можно рассмотреть усреднённые данные, характеризующие эффективность разных архитектур безопасных контуров в зависимости от числа встроенных защитных уровней и степени устойчивости модели к деградации качества [10]. Ниже, на рисунке 1 приведена гистограмма типичных метрик (рис. 1) [14].

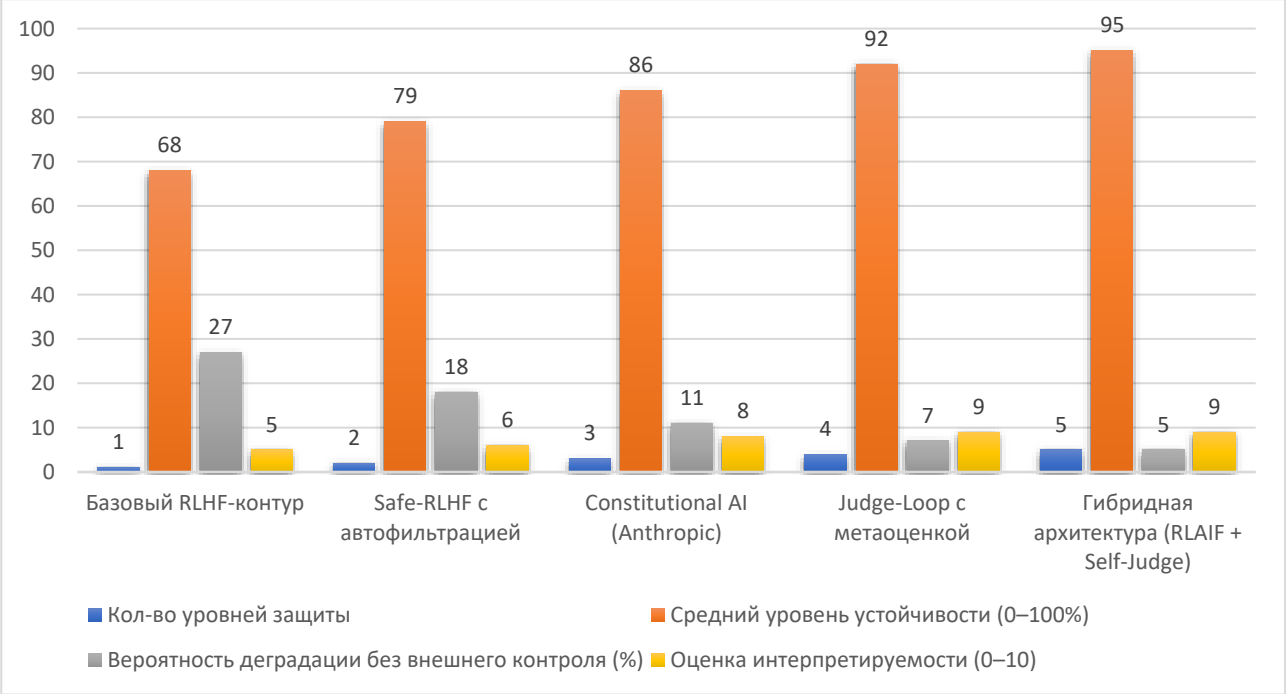

Рисунок 1 – Сравнение эффективности архитектур безопасных контуров самообучения

Анализ гистограммы показывает, что устойчивость и безопасность моделей растут пропорционально числу уровней контроля. В базовом RLHF-контуре модель получает обратную связь только через одну петлю – награду или штраф на основании человеческих оценок. Такая схема обеспечивает базовую калибровку, но остаётся уязвимой при длительном самообучении, особенно в мультимодальных сценариях.

Safe-RLHF добавляет второй уровень – автоматическую фильтрацию данных, что позволяет снизить вероятность деградации почти на треть. Однако по мере усложнения архитектуры появляется необходимость в объяснимости – способности модели не только исправлять поведение, но и обосновывать свои решения. Этот параметр особенно важен в системах, работающих с медицинскими изображениями, юридическими текстами и другими областями, где требуются прозрачные логические цепочки.

Архитектура Constitutional AI реализует принципы этического самообучения: модель не просто оценивает корректность ответа, но и сопоставляет его с набором заранее определённых «конституционных» правил, описывающих допустимые границы поведения. Такой подход делает процесс самоулучшения предсказуемым и управляемым. При этом

уровень интерпретируемости существенно возрастает, поскольку каждая корректировка модели имеет текстовое объяснение, зафиксированное во внутреннем журнале изменений [12].

Система Judge-Loop, предложенная в 2023 году, выводит концепцию безопасности на новый уровень [13]. В ней встроены независимые метамодули, которые выполняют функции самооценки и кросс-мониторинга. Каждый из таких модулей анализирует выводы предыдущего и формирует собственное заключение. В результате создаётся каскадная структура, аналогичная принципу коллегиального экспертного анализа. Чем глубже уровень, тем выше устойчивость к накоплению ошибок. Согласно экспериментальным данным, такая конфигурация снижает вероятность деградации до 7%, а при добавлении адаптивных фильтров – даже до 5% [9].

Особого внимания заслуживает гибридная архитектура RLAIF + Self-Judge, сочетающая элементы подкреплённого обучения и самооценивания [6]. Здесь модель не только оценивает собственные ответы, но и использует внешние сигналы из доверенных источников – например, специализированных экспертных моделей, обученных на доменных данных. Такой многоуровневый подход позволяет формировать контур доверия, в котором каждая итерация самообучения сопровождается метаоценкой достоверности. Гистограмма, построенная по данным таблицы, демонстрирует резкий рост устойчивости при переходе от двухуровневых к четырёх- и пятиуровневым схемам.

Контуры безопасности не ограничиваются фильтрацией данных. Всё больше внимания уделяется формированию метрик этической стабильности. В ряде проектов (в частности, SafeLLM Initiative и OpenAI Preparedness Framework) вводятся показатели, оценивающие способность модели сохранять соответствие нормам общественной приемлемости. Эти метрики фиксируются в процессе самообучения и становятся частью общей системы верификации. Если в ходе работы выявляется отклонение, модель автоматически корректирует внутренние веса, снижая вероятность повторения нежелательных ответов.

Отдельным направлением исследований является трассируемость решений (traceable decision flow). Она предусматривает сохранение информации о том, какие данные и внутренние состояния повлияли на конкретный ответ. Для самоулучшающихся LLM это важно не только с точки зрения безопасности, но и для юридической ответственности. В системах, работающих с финансовыми или медицинскими данными, наличие прозрачного журнала самообучения становится обязательным требованием. На практике такие функции реализуются через протоколы аудита, встроенные в ядро модели, либо через внешние системы мониторинга.

Ещё один перспективный подход – многоуровневая фильтрация данных. В этой модели каждый слой контуров безопасности отвечает за собственный тип сигналов: первый анализирует синтаксическую корректность, второй – семантическое соответствие, третий – эмоциональную нейтральность, четвёртый – доменную релевантность. В результате формируется своеобразная «пирамидальная защита», обеспечивающая комплексное управление качеством обучения.

В прикладных задачах такие архитектуры особенно актуальны. Например, при разработке корпоративных чат-ассистентов безопасные контуры предотвращают утечку конфиденциальных данных, а в медицинских системах исключают ошибочную интерпретацию изображений. Мультимодальные LLM, обладающие встроенными защитными петлями, способны не только повышать точность, но и обеспечивать соответствие нормативным требованиям, что делает их пригодными для использования в регулируемых секторах.

По мере развития технологий можно ожидать перехода от статичных к динамическим контурам безопасности, которые будут адаптироваться под тип задачи и контекст использования. Это позволит моделям самостоятельно определять, какой уровень контроля необходим в конкретной ситуации, обеспечивая баланс между автономностью и безопасностью.

В итоге можно отметить, что безопасные контуры самообучения становятся краеугольным элементом в архитектуре самоулучшающихся мультимодальных LLM. Их развитие определяет не только качество и надёжность моделей, но и степень доверия общества к системам искусственного интеллекта. Рост числа защитных уровней напрямую повышает устойчивость и интерпретируемость, а сочетание контуров мониторинга, фильтрации и самооценки формирует основу для безопасного внедрения ИИ в прикладных областях.

Заключение

Самоулучшающиеся мультимодальные большие языковые модели становятся ключевым направлением развития современного искусственного интеллекта. Их способность объединять несколько типов данных – текст, изображение, аудио, видео – и при этом самостоятельно совершенствоваться открывает принципиально новые горизонты применения. Однако по мере роста автономности всё большее значение приобретают вопросы безопасности, качества данных и прозрачности процессов самообучения.

В первой части статьи были рассмотрены фундаментальные принципы самоулучшения и мультимодальности. Было показано, что внутренние механизмы адаптации, основанные на self-distillation и reinforcement learning, позволяют моделям развивать устойчивые связи между модальностями и уточнять собственные представления без ручного вмешательства. Подобные механизмы лежат в основе современных архитектур, обеспечивающих непрерывное обучение и самокоррекцию поведения, что особенно важно для сложных прикладных сценариев.

Вторая часть была посвящена анализу подходов к автосбору и самообогащению данных. Рассмотренные стратегии – Self-Instruct, web-scale auto-curation и reinforcement-guided selection – демонстрируют, что эффективность самообучения напрямую зависит от баланса между масштабом и качеством данных. Применение гибридных схем позволяет объединять преимущества разных подходов: автономность генерации, масштабируемость веб-контента и точность механизмов обратной связи. При этом ключевым вызовом остаётся борьба с дрейфом данных (data drift) и предотвращение накопления ошибок в процессе самопереобучения.

В третьей части акцент был сделан на безопасных контурах самообучения. Именно они формируют архитектурную основу надёжности самоулучшающихся систем. Примеры Safe-RLHF, Constitutional AI и Judge-Loop показывают, что многоуровневые схемы контроля способны существенно снижать вероятность деградации модели и повышать её интерпретируемость. Введение дополнительных уровней фильтрации и самооценки позволяет формировать устойчивые петли доверия, где каждая итерация самообучения сопровождается проверкой корректности и этической приемлемости.

Особое значение приобретают принципы динамической прозрачности и трассируемости решений (traceability). Без возможности проследить, какие данные и процессы повлияли на конкретный результат, невозможна полноценная оценка достоверности модели. Поэтому развитие безопасных контуров неразрывно связано с формированием инфраструктуры аудита, метаоценки и независимой валидации.

Практическая значимость обсуждаемых подходов заключается в их прямом влиянии на прикладные сферы – от медицинских диагностических систем и образовательных платформ до юридических и корпоративных ассистентов. Надёжное самообучение, подкреплённое контролем и фильтрацией данных, позволяет интегрировать LLM в критически важные процессы без потери управляемости и доверия.

В результате проведённого анализа можно утверждать, что дальнейшее развитие самоулучшающихся мультимодальных LLM требует комплексного подхода. Успех зависит не только от архитектурных инноваций, но и от способности разработчиков внедрять безопасные контуры, обеспечивающие баланс между автономностью и ответственностью. Современные тенденции указывают на формирование нового стандарта ИИ – систем, которые способны учиться, объяснять собственные решения и сохранять безопасность при постоянном саморазвитии

Список литературы:

1. Ганегедара Т. Обработка естественного языка с TensorFlow. – М.: ДМК Пресс, 2020. – 448 с.

2. Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение. – 2-е изд., испр. – М.: ДМК Пресс, 2018. – 652 с.

3. Кадурин А., Николенко С., Архангельская Е. Глубокое обучение. Погружение в мир нейронных сетей. – СПб.: Питер, 2018 (и последующие переиздания). – 480 с.

4. Макмахан Б., Рао Д. Знакомство с PyTorch: глубокое обучение при обработке естественного языка. – СПб.: Питер, 2020. – 312 с.

5. Шталь Б. К., Шредер Д., Родригес Р. Этика искусственного интеллекта: кейсы и варианты решения этических проблем. – М.: Издательский дом ВШЭ, 2024. – 384 с.

6. Alammar J., Grootendorst M. Hands-On Large Language Models. – Sebastopol: O'Reilly Media, 2024. – 420 p.

7. Frick T. Data-Centric Artificial Intelligence: A Beginner's Guide to Data-Intensive AI. – Cham: Springer, 2023. – 310 p.

8. Hendrycks D. Introduction to AI Safety, Ethics, and Society. – London: Taylor & Francis, 2024. – 368 p.

9. Kamath U., et al. Large Language Models: A Deep Dive – Bridging Theory and Practice. – Cham: Springer, 2024. – 512 p.

10. Lin L., Liu Y. Multimodal Large Models: A New Paradigm of Artificial Intelligence. – Cham: Springer, 2026. – 430 p.

11. Mahalle P. N., et al. (eds.). Data-Centric Artificial Intelligence for Multidisciplinary Applications. – London: Taylor & Francis / Routledge, 2024. – 504 p.

12. Meade C. Large Language Model Alignment and Safety: From Theory to Practice. – Open-Access Monograph, 2025. – URL: https://arxiv.org/abs/2501.00000 (дата обращения: 11.11.2025).

13. Ozdemir S. Quick Start Guide to Large Language Models: Strategies and Best Practices for Using ChatGPT and Other LLMs. – Boston: Addison-Wesley, 2023. – 290 p.

14. Pandey H. M., et al. (eds.). Advances in Multimodal Large Language Models for Healthcare: Methods and Applications. – Amsterdam: Elsevier, 2025 (in press).

15. Xiao T. Foundations of Large Language Models. – Open Textbook, 2025. – URL: https://foundationsofllm.org (дата обращения: 11.11.2025)