УДК 004.8.

Архитектурные принципы и когнитивные компоненты автономных агентных систем на базе больших языковых моделей

Зуев Д.А.

ФГБОУ ВО "МГУ имени М.В. Ломоносова",

г. Москва, Российская Федерация

В последние годы наблюдается устойчивый сдвиг от использования больших языковых моделей в качестве изолированных генеративных инструментов к построению автономных агентных систем. Такие системы способны не только порождать текст, но и целенаправленно взаимодействовать с внешней средой, вызывать программные интерфейсы, обращаться к базам данных и принимать решения в многошаговых сценариях. Отмечается, что подобная трансформация требует переосмысления архитектурных принципов, поскольку классическая схема «запрос-ответ» перестаёт быть достаточной для решения прикладных задач [1]. Следует предположить, что дальнейшее развитие данного направления будет определяться не только ростом параметров моделей, но и качеством организации их когнитивных компонентов.

Центральным элементом агентной системы выступает языковая модель, выполняющая функции ядра рассуждения и генерации решений. Её выбор определяется поддержкой целевого языка, способностью к структурированному выводу, размером контекстного окна и требованиями к конфиденциальности данных. Для чувствительной информации предпочтительна локальная развёртка, для высокой вариативности запросов — облачные решения. Универсальной модели для всех агентных задач сейчас нет, поэтому проектирование начинается с анализа предметной области и требований к языковому ядру, с учётом задержки, стоимости инференса и вычислительных ресурсов [1].

Инструментальный вызов представляет собой механизм, при котором модель не исполняет функцию напрямую, а формирует структурированное описание действия для исполнительной среды. Такой подход позволяет агенту

взаимодействовать с поисковыми системами, калькуляторами, базами знаний и прикладными программными интерфейсами. Корректность описания вызова критически зависит от качества системной инструкции и схемы валидации параметров [2]. Для снижения риска ошибочных или небезопасных действий агента целесообразно предусмотреть промежуточный слой проверки, ограничение инструментов по привилегиям и контексту задачи, а также логирование всех вызовов для аудита и анализа инцидентов.

Память агента обычно разделяют на кратковременную и долговременную. Кратковременная хранит контекст текущей сессии, долговременная — предпочтения пользователя, сведения о ранее выполненных задачах и предметной области. Чтобы контекст не разрастался, используют суммаризацию и избирательное забывание. Перспективной выглядит интеграция механизмов поиска по внешним источникам, известная как генерация с дополнением из поиска, позволяющая агенту опираться на актуальные данные без переобучения модели [3]. Но результат зависит от качества индексации и релевантности найденных фрагментов, поэтому на практике часто применяют гибридные схемы — векторный поиск вместе с классическим ранжированием.

Стратегии планирования определяют способ декомпозиции сложной задачи на последовательность более простых шагов. Среди распространённых подходов выделяются цепочка мыслей, дерево мыслей и схема чередования рассуждения и действия. Выбор конкретной стратегии зависит от характера задачи: для линейных сценариев достаточно последовательного рассуждения, для ветвления и неопределённости — перебора вариантов. Полезно сочетать стратегии, адаптируя глубину планирования к ресурсам и времени, а также применять внешние решатели для верификации промежуточных результатов. Также может оказаться полезным механизм пересмотра плана при обнаружении противоречий, что согласуется с идеей чередования рассуждения и действия, предложенной в рамках подхода ReAct [2].

Несмотря на очевидный прогресс, агентные системы на базе LLM сохраняют ограничения: галлюцинации при недостатке данных, накопление

ошибок в длинных цепочках рассуждений и уязвимости к недоверенному внешнему вводу. Аудит и воспроизводимость действий агента недостаточно проработаны. Перспективными направлениями выглядят разработка методов формальной верификации планов, создание стандартизированных протоколов взаимодействия агентов и внедрение механизмов ограничения автономности в критических сценариях [1]. Также представляет интерес проблематика согласования целей агента с этическими и правовыми нормами. Отдельного внимания заслуживает задача оценки совокупной стоимости эксплуатации агентных систем.

Автономные агентные системы на базе больших языковых моделей — закономерный этап развития прикладного применения генеративных технологий. Их эффективность зависит не только от языкового ядра, но и от организации памяти, инструментальной среды и планирования. Ожидается, что дальнейшие исследования будут направлены на повышение надёжности и предсказуемости агентов, а также на методологические основы их безопасного применения в корпоративной и научной практике. Перспективы направления многообещающи, но лишь при одновременной разработке методов контроля и верификации. Окончательные выводы о долгосрочной эффективности агентных архитектур требуют накопления эмпирических данных.

Список использованной литературы

1. Намиот Д.Е., Ильюшин Е.А. *Архитектура LLM-агентов* // International Journal of Open Information Technologies. 2025. Т. 13. № 1. С. 67–74. URL: https://injoit.org/index.php/j1/article/view/1875 (дата обращения: 28.09.2026).

2. Yao S., Zhao J., Yu D. et al. ReAct: *Synergizing Reasoning and Acting in Language Models* // International Conference on Learning Representations (ICLR). 2023. URL: https://arxiv.org/abs/2210.03629 (дата обращения: 28.09.2026).

3. Lewis P., Perez E., Piktus A. et al. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks* // Advances in Neural Information Processing Systems. 2020. Vol. 33. P. 9459–9474. URL: https://proceedings.neurips.cc/paper_files/

paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf (дата обращения: 28.09.2026).