взаимодействует с внешними сервисами и базами данных, выполняя необходимые операции и возвращая результаты.

LLM выполняет основную роль генеративного и аналитического ядра системы: формирует ответы, обобщает найденную информацию, проводит рассуждения и структурирует результаты. В данной системе выбрана модель *Qwen* согласно [1]. Данная модель была выбрана так как она демонстрирует высокое качество обработки русскоязычного текста, так же она обладает открытой лицензией, имеет контекстное окно до 256000 токенов, поддержку *reasoning*-операций и *RAG*-сценариев.

Для обеспечения поиска и отбора релевантной информации применяется эмбеддинговая модель *intfloat/multilingual-e5-base*, которая преобразует текстовые данные в числовые векторные представления (эмбеддинги), отражающие их семантическое сходство. Выбор модели осуществлялся по нескольким критериям: поддержка русского языка, высокое качество семантических представлений, обеспечивающее точный поиск по смыслу, а не только по ключевым словам, открытая лицензия и возможность локального использования без зависимости от внешних *API*. Эти векторы сохраняются во векторной базе данных и используются для поиска фрагментов, наиболее близких по смыслу к пользовательскому запросу. Результаты поиска передаются модели *Qwen*, что позволяет системе формировать более точные и обоснованные ответы при генерации учебных материалов согласно [2].

Агрегатор собирает частичные ответы, полученные от всех агентов, объединяет их в единый итоговый результат, добавляет информацию об источниках и передает ответ пользователю. Такая архитектура обеспечивает модульность, масштабируемость и возможность интеграции различных источников информации при формировании ответа.

Разработка и внедрение предложенной системы открывает перспективы для более гибкого и персонализированного образования, автоматизации рутинных процессов проектирования курсов, а также интеграции современных технологий искусственного интеллекта в образовательную практику.

Литература

1. Yuxia, W. Factuality of Large Language Models: A Survey / W. Yuxia // ArXiv preprint arXiv: 2402.02420v3. – 2024. – 11 с.
2. Renata, N. Exploring the use of retrieval-augmented generation models in higher education: A pilot study on artificial intelligence-based tutoring / N. Renata // Social Sciences & Humanities Open. – 2025. – Т. 12, N 1. – С. 1–10. – Art. 101751.

ПРОБЛЕМА КОНТРОЛИРУЕМОСТИ И ВЕРИФИКАЦИИ В АВТОНОМНЫХ *LLM*-АГЕНТАХ

Ю. Д. Евженко, К. С. Курочка

Гомельский государственный технический университет имени П. О. Сухого, Республика Беларусь

Происходит фундаментальный сдвиг от использования больших языковых моделей как пассивных инструментов генерации контента к их применению в качестве ядра рассуждений для автономных AI-агентов, способных к планированию и выполнению многошаговых задач. Однако их стохастическая природа вступает в противоречие с требованиями к безопасности и надежности систем управления. Целью работы является систематизация рисков, присущих LLM-агентам, при помощи использования архитектурных паттернов контроля, известных как «Guardrails». Методика исследования включает анализ и классификацию сбоев на этапах планирования и выполнения действий, а также систематизацию многоуровневых подходов к их предотвращению. В качестве результата предложена трехуровневая модель «Guardrails», обеспечивающая контроль над

рассуждениями, действиями и последствиями, что закладывает основу для создания безопасных и предсказуемых AI-агентов.

Ключевые слова: *LLM*-агент, контролируемость, верификация, безопасность AI, Guardrails, архитектурный паттерн, автономные системы.

THE PROBLEM OF CONTROLLABILITY AND VERIFICATION IN AUTONOMOUS LLM AGENTS

Yu. D. Youzhanka, K. S. Kurochka

Sukhoi State Technical University of Gomel, Republic of Belarus

A fundamental shift is underway from using large language models (LLMs) as passive content generation tools to employing them as a reasoning core for autonomous AI agents capable of planning and executing multi-step tasks. However, their stochastic nature directly conflicts with the safety and reliability requirements of control systems. The aim of this work is to systematize the risks inherent in LLM agents by architectural control patterns known as "Guardrails". The research methodology includes the analysis and classification of failures at the planning and action execution stages, as well as the systematization of multi-level approaches to their prevention. The result is a proposed three-level "Guardrails" model that provides control over reasoning, actions, and consequences, laying the foundation for creating safe and predictable AI agents.

Keywords: LLM agent, controllability, verification, AI safety, Guardrails, architectural pattern, autonomous systems.

Большие языковые модели (*LLM*) эволюционируют от пассивных генераторов текста к центральным компонентам исполнительных систем или *AI*-агентов [1]. В отличие от традиционных *LLM*, отвечающих на запросы, агенты обладают способностью к автономному достижению целей, используя свое ядро рассуждений для декомпозиции задач, составления планов и использования внешних инструментов (например, *API*). Этот сдвиг парадигмы открывает беспрецедентные возможности, но обнажает фундаментальное противоречие: использование вероятностного ядра для управления системами, требующими детерминированного и безопасного поведения. Эта проблема является современным проявлением более общего класса «несчастных случаев» в системах искусственного интеллекта, связанных с непреднамеренным и вредоносным поведением [2]. Целью настоящей работы является систематизация рисков, присущих *LLM*-агентам, и обзор архитектурных подходов, известных как «*Guardrails*», направленных на их контроль.

Сбои в работе *LLM*-агентов – это не просто некорректная информация, а ошибочные действия с потенциально деструктивными последствиями. Их можно классифицировать по этапу возникновения: сбои на этапе планирования (логически неверный план), сбои при использовании инструментов (синтаксически неверный вызов функции) и сбои на этапе выполнения (формально верное, но опасное действие). Прямое, неконтролируемое предоставление автономии *LLM*-агентам недопустимо в промышленных системах, что обуславливает необходимость в разработке архитектур контроля.

«Guardrails» – это архитектурные компоненты и методологии, предназначенные для предотвращения непреднамеренного и вредоносного поведения агента, то есть «несчастных случаев», описанных в [1]. Их можно сгруппировать в трехуровневую модель защиты.

Первый уровень – контроль над рассуждениями. Эти методы структурируют «мыслительный» процесс агента. К ним относятся методики, заставляющие модель пошагово

излагать свои рассуждения перед выдачей ответа (например, *Chain-of-Thought*) или фреймворки реализующие строгий цикл «рассуждение-действие» (например, *ReAct*) [3]. Это позволяет отслеживать логику и снижает вероятность спонтанных ошибок.

Второй уровень – контроль над действиями. Данные методы верифицируют команду агента перед ее выполнением. Наиболее распространенным подходом является автоматическая валидация сгенерированных вызовов *API* на соответствие заранее определенной схеме (*JSON Schema*, *Pydantic*), которая отбраковывает все синтаксически неверные команды. Для критически важных операций применяется подтверждение со стороны человека (*Human-in-the-Loop*), где агент лишь формирует команду, а ее исполнение требует явного одобрения.

Третий уровень – контроль над последствиями. Эти методы ограничивают потенциальный ущерб от действий агента, даже если они прошли предыдущие проверки. Ключевой техникой является выполнение команд в изолированной среде («песочнице», например, *Docker*-контейнере) с ограниченными правами доступа к сети и файловой системе, а также установка жестких лимитов на количество вызовов *API* или расходуемый бюджет.

Внедрение «*Guardrails*» – это поиск компромисса между двумя ключевыми свойствами агента: автономией (способностью решать сложные задачи) и «выравниванием» (соответствием действий целям и ограничениям человека).

В ходе работы была проанализирована проблема контролируемости автономных *LLM*-агентов и систематизированы основные риски, связанные с их стохастической природой. В качестве результата предложена трехуровневая модель архитектурных паттернов контроля «*Guardrails*», включающая контроль на уровне рассуждений, действий и последствий. Анализ показал, что путь к созданию мощных и безопасных *LLM*-агентов лежит не столько в увеличении размера моделей, сколько в разработке продуманных, многоуровневых архитектур контроля. Дальнейшие исследования должны быть сосредоточены на создании методик, упрощающих проектирование и внедрение таких «*Guardrails*» для безопасного применения *LLM*-агентов в промышленных приложениях.

Литература

1. Курочка, К. С. Нейросетевая модель автогенерации тестов для студентов в системе Moodle на основе анализа методических материалов / К. С. Курочка, Ю. С. Башаримов // Цифровая трансформация. – 2025. – N 31 (3). – P. 66–75. – URL: https://doi.org/10.35596/1729-7648-2025-31-3-66-75 (дата обращения: 11.10.2025).
2. Amodei D., Olah C., Steinhardt J., Christiano P., Schulman J., Mané D. Concrete Problems in AI Safety // ArXiv preprint arXiv:1606.06565. – 2016.
3. Synergizing Reasoning and Acting in Language Models / S. Yao, J. Zhao, D. Yu [et al.] // International Conference on Learning Representations (ICLR), 2023.

METHODOLOGY FOR ASSESSING THE ACCURACY AND QUALITY OF LABELING AND SEGMENTATION OF MRI IMAGES OF THE HUMAN LUMBAR SPINE

Ren Huanhai[1, 2], Wang Xuemei[1, 2], K. S. Kurachka[1]

[1]Sukhoi State Technical University of Gomel, Republic of Belarus

[2]Shandong Huayu Institute of Technology, Dezhou, People's Republic of China

This study proposes a unified evaluation framework for automated lumbar spine MRI segmentation and level-wise labeling using public multi-center data [1]. The protocol assesses detection/segmentation via mAP across IoU thresholds (0.50-0.95), mask quality via Dice and 95 % Hausdorff distance, and labeling accuracy on ROIs. It addresses cross-domain variations and