УТВЕРЖДАЮ
Проректор федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский университет «Высшая школа экономики»
к.э.н., доцент Сергей Юрьевич Рощин

«27» августа 2026 г.

ОТЗЫВ

ведущей организации на диссертацию Малояна Нарека Гагиковича «Разработка методов оценки и повышения устойчивости больших языковых моделей к вариациям входных последовательностей», представленную на соискание учёной степени кандидата технических наук по специальности 2.3.5 — «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей»

Актуальность темы диссертации

Тема диссертации посвящена одной из наиболее острых проблем современной информационной безопасности — устойчивости больших языковых моделей (LLM) к состязательным вариациям входных последовательностей: инъекциям запроса (prompt injection), троянским закладкам (backdoor attacks) и атакам на агентские системы, использующие протокол взаимодействия с внешними инструментами Model Context Protocol (MCP). По мере встраивания LLM в производственные системы, включая архитектуры LLM-as-a-Judge и автономных агентов с доступом к внешним инструментам, риски целенаправленной компрометации моделей существенно возрастают, а последствия атак выходят за рамки чисто

информационных и способны приводить к несанкционированному выполнению кода. Тема диссертации полностью соответствует паспорту специальности 2.3.5 и является актуальной задачей развития теории и практики обеспечения безопасности интеллектуальных систем машинного обучения.

Новизна полученных результатов и выводов

Научная новизна диссертации определяется следующими результатами, полученными автором лично:

— предложена формальная математическая модель метрики генеративной устойчивости R_stab(f), основанная на дивергенции Йенсена–Шеннона между распределениями токенов при малых возмущениях входа, для которой доказаны свойства монотонности и получена верхняя граница уязвимости для класса локализованных атак;

— разработан адаптивный алгоритм генерации состязательных атак ASA, достигающий ASR до 73,8% в условиях «черного ящика» на системах LLM-as-a-Judge, с переносимостью между открытыми моделями TSR до 62,6%;

— количественно охарактеризована асимметрия задачи обнаружения троянских закладок на данных Trojan Detection Challenge 2023 (REASR ≈ 0,99 при Recall ≈ 0,17), обосновывающая приоритет превентивного контроля данных над послеобучающим аудитом;

— разработан и экспериментально валидирован метод защиты систем LLM-as-a-Judge на основе адаптивных комитетов гетерогенных моделей с теоретическим обоснованием через теорему Кондорсе, снижающий ASR наиболее уязвимой модели на 47–55 процентных пунктов;

— предложен архитектурно-алгоритмический подход к защите агентских LLM-систем по протоколу MCP (алгоритм AttestMCP, паттерн Commit Boundary, программный модуль MCPSec), снижающий среднюю ASR

агентских атак с 53,7% до 12,4% на собранном автором бенчмарке MCPBench из 847 сценариев.

Апробация работы и публикации

Результаты диссертации опубликованы в 9 научных изданиях, из которых 6 относятся к рекомендованным ВАК и приравненным к ним: 4 статьи в журналах, входящих в перечень ВАК (3 статьи категории К1, 1 статья категории К2), и 2 свидетельства о государственной регистрации программ для ЭВМ (JudgeGuard № 2025687702, TrojanArmor № 2025684247). Помимо них, результаты опубликованы в докладе на конференции DCCN 2024, индексируемой в Scopus (Q2), в докладе на конференции DIALOG 2022 и в препринте. Основные результаты работы представлялись и обсуждались на конференциях Distributed Computer and Communication Networks (DCCN 2024, Москва), International Conference on Computational Linguistics and Intellectual Technologies DIALOG 2022 (Москва), а также на научных конференциях «Тихоновские чтения 2023» и «Ломоносовские чтения 2022» (МГУ имени М.В. Ломоносова, факультет ВМК). Апробация методов также осуществлена в ходе участия автора в международных соревнованиях по безопасности LLM: Trojan Detection Challenge 2023, SaTML CTF 2024 (призовое место в треке по атакам на защищённые LLM-системы) и Kaggle «LLMs: You Can't Please Them All».

Обоснованность научных положений и выводов, сформулированных в диссертации

Научные положения, выводы и рекомендации, содержащиеся в диссертации, обоснованы теоретически и подтверждены экспериментально. Достоверность результатов подтверждается использованием репрезентативного набора открытых (Gemma, Llama) и проприетарных (GPT, Claude) моделей, публичных бенчмарков (MT-Bench, TDC2023, SaTML CTF 2024, Kaggle), корректным применением методов статистической проверки гипотез с

доверительными интервалами, а также участием и получением призового места автора в международном соревновании SaTML CTF 2024. Диссертация является самостоятельной, целостной научно-квалификационной работой, содержащей решение задачи, имеющей существенное значение для развития соответствующей отрасли науки. Полученные результаты обладают научной новизной, обоснованы теоретически и подтверждены экспериментально, обладают практической значимостью и подтверждены актами о внедрении. Диссертация соответствует требованиям пунктов 9–14 Положения о присуждении учёных степеней, предъявляемым к диссертациям на соискание учёной степени кандидата наук.

Соответствие содержания диссертации автореферату и указанной специальности

Содержание автореферата полностью соответствует содержанию диссертации и отражает её основные научные положения, выносимые на защиту. Тематика диссертации соответствует паспорту специальности 2.3.5 — «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей»: в работе разработаны формальные модели, методы и алгоритмы анализа, верификации и тестирования программных систем (п. 1 паспорта специальности), созданы фреймворки оценки безопасности интеллектуальных систем машинного обучения (п. 4), предложен комплекс метрик и методология количественной оценки качества LLM-систем в аспекте устойчивости и безопасности (п. 10). Личный вклад автора в опубликованных, в том числе в соавторстве, работах является определяющим: диссертанту принадлежат постановка задач, разработка методов атак и защиты, планирование и проведение экспериментов, анализ данных и формулировка выводов.

Значимость результатов для науки и производства

Полученные автором результаты имеют существенное значение для развития теории и практики обеспечения безопасности интеллектуальных систем машинного обучения, относящейся к предметной области специальности 2.3.5. Впервые в отечественной и, по имеющимся сведениям, в значительной мере в мировой литературе предложен согласованный формальный метрический аппарат оценки устойчивости генеративных моделей, учитывающий дискретность пространства токенов и авторегрессионный характер вывода; систематизирована и экспериментально подтверждена уязвимость архитектуры LLM-as-a-Judge, получившей широкое распространение в индустрии автоматической оценки качества генеративных систем; выявлены и формализованы архитектурные риски протокола MCP, практическое использование которого в агентских системах интенсивно растёт. Совокупность этих результатов существенно расширяет теоретическую и методологическую базу анализа и обеспечения робастности систем машинного обучения и создаёт основу для дальнейших исследований в данной области.

Практическая значимость работы для производства состоит в следующем. Разработанные программные комплексы JudgeGuard (свидетельство № 2025687702) и TrojanArmor (свидетельство № 2025684247), а также программный модуль MCPSec рекомендуются к применению организациями, эксплуатирующими системы LLM-as-a-Judge и агентские LLM-платформы с инструментальным доступом, для регулярного аудита безопасности и тестирования устойчивости к инъекционным и агентским атакам. Метод защиты на основе адаптивных комитетов гетерогенных моделей (5–7 моделей различных архитектурных семейств) рекомендуется к внедрению в производственных системах автоматической оценки контента как средство, обеспечивающее наилучшее соотношение прироста устойчивости и вычислительных затрат (по введённой автором метрике RUT). Алгоритм аттестации вызовов инструментов AttestMCP и паттерн

программной изоляции Commit Boundary целесообразно рекомендовать разработчикам агентских платформ, использующих протокол MCP, для встраивания в конвейеры обработки запросов на этапе проектирования. Вывод о структурной ограниченности послеобучающего аудита троянских закладок целесообразно учитывать при формировании политик безопасности разработки моделей, смещая приоритет на превентивный контроль обучающих данных и цепочки их поставок. Результаты диссертации уже внедрены в учебный процесс факультета ВМК МГУ имени М.В. Ломоносова (курс «Робастные модели в машинном обучении») и в производственную деятельность компании ООО «ВИАСАТ ТЕХ» (модуль ранжирования рекомендаций на основе архитектуры LLM-as-a-Judge), что подтверждает практическую применимость предложенных методов.

Замечания по диссертационной работе

Отмеченные ниже замечания не снижают высокой оценки диссертационного исследования, а скорее намечают перспективные направления для дальнейшей работы автора.

1. Связь между теоретической метрикой R_stab(f) и используемой в экспериментальной части эмпирической оценкой 1 − ASR для нелокализованных атак (prompt injection, jailbreak) обоснована автором лишь на уровне калибруемой эмпирической гипотезы, без строгого аналога доказанного неравенства для локализованных возмущений. Строгое доказательство (либо явное опровержение) более тесной связи между R_stab(f) и ASR для класса нелокализованных атак представляется желательным направлением дальнейших теоретических исследований.

2. Количественная оценка эффекта архитектурной гетерогенности комитетов (в сравнении с гомогенными ансамблями того же размера) в работе носит преимущественно предварительный, качественный характер; сам автор указывает на это как на открытую задачу. Более подробный систематический эксперимент с определением минимально необходимого уровня

архитектурного разнообразия усилил бы практическую значимость метода комитетов.

3. Гипотеза об усилении асимметрии между REASR и Recall с ростом числа параметров модели (глава 2) выдвинута, но не проверена экспериментально на моделях, отличных от Pythia-1.4B; расширение экспериментальной базы на модели большего масштаба позволило бы дать более обоснованные рекомендации по масштабируемости выявленной закономерности.

4. Предложенная каскадная схема активации комитета при низкой уверенности одиночной модели (глава 6), направленная на экономию вычислительных ресурсов, описана на концептуальном уровне без количественной оценки достигаемой экономии и влияния на итоговое качество защиты — соответствующий количественный анализ был бы полезен для практического внедрения метода.

5. Экспериментальная часть, посвящённая агентским системам на основе протокола MCP (глава 5), выполнена на собственном бенчмарке MCPBench; целесообразно в дальнейшем провести сопоставление полученных результатов с результатами независимых, в том числе более крупных по объёму, наборов сценариев атак на MCP-агентов, по мере их появления в открытом доступе.

6. В работе рассмотрен ограниченный, хотя и представительный, набор архитектурных семейств LLM (Gemma, Llama, GPT, Claude); распространение предложенных метрик и методов защиты на мультимодальные и существенно более крупные модели составляет естественное направление продолжения исследования.

Указанные замечания не затрагивают основных научных результатов и выводов диссертации и не снижают их научной и практической значимости; они, скорее, отражают естественную многоплановость и быстрое развитие исследуемой области.

Вывод

По актуальности, научной новизне, теоретической и практической значимости, достоверности полученных результатов и обоснованности выводов диссертационная работа «Разработка методов оценки и повышения устойчивости больших языковых моделей к вариациям входных последовательностей» Малояна Нарека Гагиковича соответствует требованиям п. 9-14 «Положения о присуждении учёных степеней», утверждённого постановлением Правительства РФ от 24.09.2013 г. № 842 (ред. от 16.10.2024), предъявляемым к кандидатским (докторским) диссертациям, в том числе с учётом указаний, содержащихся в письме диссертационного совета 24.1.120.01 при ФГБУН «Институт системного программирования им. В.П. Иванникова РАН» (исх. № 711-2026 от 03.07.2026), о необходимости отражения в отзыве значимости полученных автором результатов для развития соответствующей отрасли науки и, поскольку диссертация имеет прикладной характер, конкретных рекомендаций по использованию содержащихся в ней результатов и выводов, а её автор заслуживает присуждения учёной степени кандидата технических наук по специальности 2.3.5 — «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей».

Отзыв подготовлен доктором технических наук, заслуженным профессором, научным руководителем Научно-исследовательского института технологий (НИИТ) федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский университет «Высшая школа экономики» Круком Евгением Аврамовичем.

Отзыв рассмотрен и одобрен на заседании Научно-исследовательского института технологий (НИИТ) федерального государственного автономного образовательного учреждения высшего образования «Национальный

исследовательский университет «Высшая школа экономики», протокол № 6 от 04 августа 2026 года.

Сведения о ведущей организации: Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ)

Адрес: 101000 г. Москва, ул. Мясницкая, 20.
Тел.: (495) 771-32-32
Электронная почта: hse@hse.ru
Сайт: http://www.hse.ru

Директор
Научно-исследовательского
института телекоммуникаций НИУ ВШЭ
Доктор технический наук, профессор

Кучерявый Евгений Андреевич

Научный руководитель
Научно-исследовательского
института технологий НИУ ВШЭ
Доктор технических наук, заслуженный профессор

Крук Евгений Аврамович