

ЗАКЛЮЧЕНИЕ ДИССЕРТАЦИОННОГО СОВЕТА 24.1.120.01,
созданного на базе
Федерального государственного бюджетного учреждения науки
Институт системного программирования им. В.П. Иванникова
Российской академии наук
Министерства науки и высшего образования РФ
по диссертации на соискание ученой степени кандидата наук

аттестационное дело № _____

решение диссертационного совета от 29 сентября 2026 года № 2026/24

О присуждении **Малояну Нареку Гагиковичу**, гражданину РФ, ученой степени кандидата технических наук.

Диссертация «Разработка методов оценки и повышения устойчивости больших языковых моделей к вариациям входных последовательностей» по специальности 2.3.5 – «математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей» принята к защите 03 июля 2026 года (протокол № 2026/15) диссертационным советом 24.1.120.01, созданным на базе Федерального государственного бюджетного учреждения науки Институт системного программирования им. В.П. Иванникова Российской академии наук (ведомственная принадлежность: Министерство науки и высшего образования РФ, адрес: 109004, г. Москва, ул. А. Солженицына, дом 25), создан Приказом Минобрнауки России о советах по защите докторских и кандидатских диссертаций от 2 ноября 2012 г. № 714/нк.

Соискатель Малоян Нарек Гагикович, 11 октября 1997 года рождения.

В 2021 году соискатель окончил Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный университет имени М.В. Ломоносова». В 2025 году окончил очную аспирантуру факультета вычислительной математики и кибернетики МГУ имени М.В. Ломоносова.

В период подготовки диссертации соискатель работал в ООО «ВИАСАТ ТЕХ» в должности старшего специалиста отдела машинного обучения (ведомственная принадлежность: нет (ОКОГУ 4210011)). В настоящее время не работает.

Диссертация выполнена на кафедре информационной безопасности факультета вычислительной математики и кибернетики Федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный университет имени М.В. Ломоносова» (ведомственная принадлежность: Правительство Российской Федерации).

Научный руководитель – доктор технических наук Намиот Дмитрий Евгеньевич, МГУ имени М.В. Ломоносова, лаборатория открытых информационных технологий факультета вычислительной математики и кибернетики, ведущий научный сотрудник.

Официальные оппоненты:

1. Никульчев Евгений Витальевич, доктор технических наук, профессор, ФГБОУ ВО «МИРЭА – Российский технологический университет», профессор кафедры цифровых технологий обработки данных,
2. Иванов Владимир Владимирович, кандидат физико-математических наук, доцент, АНО ВО «Университет Иннополис», Институт разработки программного обеспечения и программной инженерии, доцент, заведующий лабораторией методов обработки естественного языка

дали положительные отзывы на диссертацию.

Ведущая организация – Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет «Высшая школа экономики»» (г. Москва) в своем положительном отзыве, подписанном директором Научно-исследовательского института телекоммуникаций НИУ ВШЭ доктором технических наук, профессором Кучерявым Евгением Андреевичем и научным

руководителем Научно-исследовательского института технологий НИУ ВШЭ доктором технических наук, заслуженным профессором Круком Евгением Аврамовичем и утвержденном проректором НИУ ВШЭ кандидатом экономических наук, доцентом Роциным Сергеем Юрьевичем, указала, что диссертационная работа содержит новые научные результаты, имеющие существенное значение для науки и практики.

Соискатель имеет 9 опубликованных работ, все они опубликованы по теме диссертации. Из них 4 статьи опубликованы в рецензируемых научных изданиях, входящих в перечень ВАК (3 статьи категории К1 и 1 статья категории К2), и получены 2 свидетельства о государственной регистрации программ для ЭВМ, приравненные к публикациям в изданиях перечня ВАК.

Основные научные результаты диссертации изложены в статьях в рецензируемых журналах, докладах на международных конференциях (в том числе в докладе, опубликованном в издании, индексируемом в Scopus, Q2) и защищены свидетельствами о государственной регистрации программ для ЭВМ. Часть работ выполнена в соавторстве. Во всех совместных работах вклад соискателя является определяющим и включает постановку задач, разработку методов атак и защиты, проведение экспериментов и анализ результатов. Программные комплексы, защищенные свидетельствами, разработаны автором лично. Наиболее значительные работы:

1. Maloyan N., Ashinov B., Namiot D. Investigating the Vulnerability of LLM-as-a-Judge Architectures to Prompt-Injection Attacks // International Journal of Open Information Technologies. – 2025. – Т. 13, № 9 (K1 ВАК).
2. Maloyan N., Namiot D. Breaking the Protocol: Security Analysis of the Model Context Protocol Specification and Prompt Injection Vulnerabilities in Tool-Integrated LLM Agents // International Journal of Open Information Technologies. – 2026. – Т. 14, № 2 (K1 ВАК).
3. Maloyan N., Namiot D. Sleeper Channels and Provenance Gates: Persistent Prompt Injection in Always-on Autonomous AI Agents // International Journal of Open Information Technologies. – 2026. – Т. 14, № 6 (K1 ВАК).

4. Khomsky D., Maloyan N., Nutfullin B. Prompt Injection Attacks in Defended Systems // Distributed Computer and Communication Networks. LNCS. – Springer, 2025. – Vol. 15460. – P. 404–416 (Scopus Q2).

Выбор официальных оппонентов и ведущей организации обосновывается их компетентностью и достижениями в области математического и программного обеспечения вычислительных систем, машинного обучения и информационной безопасности, наличием публикаций в сфере исследований, соответствующей теме диссертации, и способностью определить научную и практическую ценность диссертации.

Диссертационный совет отмечает, что на основании выполненных соискателем исследований:

- Разработана формальная математическая модель оценки устойчивости больших языковых моделей $R_{stab}(f)$, учитывающая дискретную природу пространства токенов и авторегрессионный характер генерации и основанная на ограниченных мерах расхождения (дивергенции Йенсена–Шеннона) между вероятностными распределениями на каждом шаге генерации при малых возмущениях входа.
- Предложены алгоритм адаптивной генерации состязательных атак ASA (Adaptive Search-based Attack), реализующий эволюционный поиск с мутациями на уровне токенов и семантическими перефразированиями и достигающий успешности атак (ASR) до 73,8 % в условиях «черного ящика», метод защиты систем LLM-as-a-Judge на основе адаптивных комитетов гетерогенных моделей с мажоритарным голосованием в пространстве вердиктов, а также архитектурно-алгоритмический подход к защите агентских LLM-систем, взаимодействующих по протоколу MCP (алгоритмы AttestMCP и Commit Boundary, программный модуль MCPSec).
- Доказаны монотонность предложенной метрики устойчивости относительно усиления защиты и верхняя оценка уязвимости $V(h) \leq 1 - R_{class}(h)$ для класса локализованных атак. Экспериментально доказана перспективность применения комитетов гетерогенных моделей,

снижающих успешность инъекционных атак на 47–55 процентных пунктов по сравнению с одиночными моделями.

- Введены метрика генеративной устойчивости $R_{stab}(f)$ и согласованный набор эмпирических показателей устойчивости (ASR, TSR, Recall, REASR), классификация инъекционных атак на инструктивные и контентные, а также понятие класса атак «prompt injection to tool hijacking» для агентских LLM-систем.

Теоретическая значимость исследования обоснована тем, что:

- Доказаны утверждения о свойствах предложенной метрики устойчивости (монотонность, связь с показателями уязвимости), а также теоретически обоснована устойчивость комитетов гетерогенных моделей на основе теоремы Кондорсе о жюри.
- Применительно к проблематике диссертации результативно использован комплекс существующих базовых методов исследования: эволюционные алгоритмы, метод жадного покоординатного градиента (GCG), методы математической статистики (бутстреп-анализ, проверка статистических гипотез, построение доверительных интервалов).
- Изложены положения о вычислительной асимметрии задачи обнаружения троянских закладок в больших языковых моделях: суррогатные триггеры активируют закладку почти гарантированно ($REASR \approx 0,99$), в то время как восстановление истинного триггера практически невозможно ($Recall \approx 0,17$), что обосновывает приоритет превентивного контроля обучающих данных над последующим аудитом моделей.
- Раскрыты архитектурные уязвимости протокола Model Context Protocol и систематизированы четыре класса векторов обхода многоуровневых систем защиты (код-ориентированные атаки, ASCII-кодирование символов, атаки с разделением слов, контекстное перенаправление) с определением условий применимости и эффективных контрмер для каждого класса.
- Изучены переносимость состязательных атак между моделями различных архитектурных семейств (TSR до 62,6 %) и зависимость уязвимости от

типа модели, в том числе систематически более высокая уязвимость моделей с открытыми весами по сравнению с проприетарными.

Значение полученных соискателем результатов исследования для практики подтверждается тем, что:

- Результаты диссертации внедрены в учебный процесс факультета вычислительной математики и кибернетики МГУ имени М.В. Ломоносова в рамках курса «Робастные модели в машинном обучении» и в производственную деятельность ООО «ВИАСАТ ТЕХ», что подтверждено актом о внедрении. В ВИАСАТ ТЕХ применены фреймворк оценки уязвимостей LLM-as-a-Judge, метод защиты на основе комитетов моделей и методология многошаговых атак; программный комплекс JudgeGuard (свидетельство о государственной регистрации программы для ЭВМ № 2025687702) внедрен в производственной среде компании. Программный комплекс TrojanArmor (свидетельство № 2025684247) апробирован на бенчмарке TDC2023, а модуль MCPSec — на бенчмарке MCPBench.
- Определены область и перспективы практического применения полученных результатов: аудит безопасности систем автоматической оценки на базе больших языковых моделей (LLM-as-a-Judge), диалоговых агентов и агентских платформ с инструментальным доступом.
- Создан бенчмарк MCPBench, включающий 847 сценариев атак на агентские LLM-системы. Разработанный комплекс защитных механизмов снижает на нем среднюю успешность атак с 53,7 % до 12,4 %.
- Представлены практические рекомендации по построению защитных механизмов: нормализация входных данных, контекстно-изолированные системные инструкции, многоуровневая фильтрация и комитетная защита.

Оценка достоверности результатов исследования выявила:

- Экспериментальные результаты получены на публичных наборах данных и бенчмарках (MT-Bench, TDC2023, SaTML CTF 2024, Kaggle «LLMs: You Can't Please Them All») на репрезентативном наборе моделей основных архитектурных семейств (Gemma, Llama, GPT-3.5, GPT-4, Claude-3-Opus). Показана воспроизводимость результатов: каждая комбинация «модель—

атака–задача» проверялась на выборке из 50 запросов с построением доверительных интервалов и проверкой статистической значимости.

- Теория построена на известных проверяемых положениях теории вероятностей и математической статистики и согласуется с опубликованными экспериментальными данными по теме диссертации и по смежным областям.
- Идея базируется на анализе практики обеспечения безопасности LLM-систем, в том числе на результатах участия в международных соревнованиях по безопасности больших языковых моделей (Trojan Detection Challenge 2023, SaTML CTF 2024, Kaggle).
- Установлено качественное и количественное согласование авторских результатов (показателей ASR, TSR, Recall, REASR) с результатами, представленными в независимых источниках по данной тематике, в тех случаях, когда такое сравнение является обоснованным.

Личный вклад соискателя состоит в непосредственном участии на всех этапах исследования: в постановке задач, разработке моделей, методов и алгоритмов оценки и повышения устойчивости больших языковых моделей, программной реализации предложенных методов, планировании и проведении экспериментов, обработке и интерпретации полученных данных, апробации результатов на международных конференциях и соревнованиях, а также в подготовке основных публикаций по выполненной работе. Программные комплексы JudgeGuard и TrojanArmor разработаны автором лично. Во всех работах, выполненных в соавторстве, вклад соискателя является определяющим.

В ходе защиты диссертации были высказаны следующие критические замечания:

- Используемое в работе понятие преобразования T , порождающего возмущенную входную последовательность (Определение 1.2), не наделяется структурными свойствами, вследствие чего класс допустимых возмущений оказывается практически неограниченным, что делает точное

вычисление метрики устойчивости вычислительно необозримым и объясняет переход автора к эмпирической оценке через показатель ASR.

- Метрика $R_{stab}(f)$ получила ограниченную прямую эмпирическую валидацию (эксперимент раздела 4.2.5 проведен на 20 промптах с пятью доброкачественными возмущениями) и для усиления результата следовало бы расширить выборку и проверить метрику на нескольких классах состязательных воздействий.
- Вклад архитектурной гетерогенности в эффективность комитетов не полностью отделен от влияния их размера и состава и нуждается в подкреплении абляционным исследованием при фиксированном числе моделей и анализом корреляции ошибок, а для бенчмарка MCPBench не приведены доверительные интервалы по отдельным моделям и абляция компонентов MCPSec, которые позволили бы точнее оценить переносимость результатов и вклад отдельных механизмов.
- Связь между теоретической метрикой $R_{stab}(f)$ и используемой в экспериментальной части эмпирической оценкой 1 – ASR для нелокализованных атак обоснована лишь на уровне калибруемой эмпирической гипотезы, без строгого аналога доказанного неравенства для локализованных возмущений.
- Количественная оценка эффекта архитектурной гетерогенности комитетов в сравнении с гомогенными ансамблями того же размера носит предварительный характер и требует систематического эксперимента с определением минимально необходимого уровня архитектурного разнообразия.
- Гипотеза об усилении асимметрии между REASR и Recall с ростом числа параметров модели не проверена экспериментально на моделях, отличных от Pythia-1.4B.
- Каскадная схема активации комитета при низкой уверенности одиночной модели (глава 6) описана на концептуальном уровне без количественной оценки достигаемой экономии вычислительных ресурсов и влияния на качество защиты.

- Экспериментальная часть по агентским системам на протоколе MCP (глава 5) выполнена на собственном бенчмарке MCPBench, и в дальнейшем целесообразно сопоставить полученные результаты с независимыми наборами сценариев атак по мере их появления в открытом доступе.
- В работе рассмотрен ограниченный, хотя и представительный, набор архитектурных семейств LLM, и распространение предложенных метрик и методов защиты на мультимодальные и более крупные модели составляет естественное направление продолжения исследования.

Соискатель Малоян Н.Г. согласился с замечаниями, ответил на задаваемые ему в ходе заседания вопросы.

На заседании 29 сентября 2026 г. диссертационный совет принял решение присудить Малояну Н.Г. ученую степень кандидата технических наук за решение научной задачи оценки и повышения устойчивости больших языковых моделей к вариациям входных последовательностей, имеющей существенное значение для развития математического и программного обеспечения вычислительных систем, комплексов и компьютерных сетей.

При проведении тайного голосования диссертационный совет в количестве 12 человек, из них 8 докторов наук по специальности рассматриваемой диссертации, участвовавших в заседании, из 15 человек, входящих в состав совета, проголосовали: за – 12, против – 0.

Заместитель председателя диссертационного совета,
доктор физико-математических наук

Белеванцев А. А.

И.о. ученого секретаря диссертационного совета,
доктор физико-математических наук

Бурдонов И.Б.

29 сентября 2026 г.