ОТЗЫВ

официального оппонента по диссертационной работе

Малояна Нарека Гагиковича

«Разработка методов оценки и повышения устойчивости больших языковых моделей к вариациям входных последовательностей»,

представленную на соискание ученой степени кандидата технических наук по специальности 2.3.5 — «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей»

1. Актуальность темы исследования

Современные большие языковые модели (LLM) стремительно интегрируются в производственные системы: от автоматизированных ассистентов и рекомендательных сервисов до агентских платформ с доступом к внешним инструментам. Вместе с расширением области применения растёт и число точек уязвимости — инъекции промптов, троянские закладки, манипуляции метриками автоматической оценки. Особую роль среди таких уязвимостей играет архитектура LLM-as-a-Judge, где одна языковая модель оценивает ответы других. Компрометация модели-оценщика искажает результаты сравнения и ставит под сомнение воспроизводимость экспериментов.

Особую важность задаче придаёт композиционная природа агентских систем. Протокол Model Context Protocol (MCP) унифицирует взаимодействие LLM-агента с внешними инструментами, расширяя пространство состояний системы. Инъекция в контекст вызова инструмента позволяет перехватить управление (prompt injection to tool hijacking), транслируя текстовую уязвимость в пространство исполняемых действий. Последствия атаки в таком случае выходят за информационные рамки.

Таким образом, разработка формального аппарата оценки устойчивости LLM, методов противодействия инъекционным атакам и программных средств тестирования безопасности является актуальной научной задачей, имеющая значение для развития математического и программного обеспечения

вычислительных систем. Диссертационная работа Малояна Н.Г., направленная на создание комплекса моделей, методов и алгоритмов для оценки и повышения устойчивости LLM к вариациям входных последовательностей, выполнена в тенденциях современных вызовов и несомненно актуальна.

2. Степень обоснованности и достоверности научных положений, выводов и рекомендаций

Достоверность полученных результатов обеспечивается: опорой на репрезентативную экспериментальную базу (публичные бенчмарки MT-Bench, TDC2023, SaTML CTF 2024, Kaggle «LLMs: You Can't Please Them All»); широким спектром моделей различных архитектурных семейств (Gemma, Llama, Pythia, GPT-3.5, GPT-4, Claude-3-Opus); корректным статистическим аппаратом (бутстреп-анализ на 1000 итераций, проверка гипотез, доверительные интервалы); апробацией в международных соревнованиях по безопасности LLM и внедрением в производственную среду компании ВИАСАТ ТЕХ, что подтверждено актом о внедрении; публикацией основных результатов в 9 научных работах, из которых 4 — в журналах из перечня ВАК (3 категории К1, 1 категории К2), и двумя свидетельствами о государственной регистрации программ для ЭВМ.

Научные положения диссертации достоверны и обоснованы в тексте работы, основными являются следующие положения.

1. Формальная математическая модель оценки устойчивости больших языковых моделей $R_{stab}(f)$, учитывающая дискретную природу пространства токенов и авторегрессионный характер генерации. В отличие от существующих подходов, ориентированных на классификационные задачи и непрерывные пространства, предложенная метрика базируется на ограниченных мерах расхождения, в частности дивергенции Йенсена-Шеннона, между вероятностными распределениями на каждом шаге генерации при малых возмущениях входа. Показана согласованность метрики с естественным отношением усиления защиты, где монотонность выступает условием корректности определения. Для класса локализованных атак доказана верхняя

оценка уязвимости $V(h) \leq 1 - R_{class}(h)$. Для нелокализованных атак (prompt injection, jailbreak) предложена эмпирическая калибруемая модель $V(h) \approx g(1 - R_{class}(h))$. Обоснована применимость метрики ASR как эмпирического прокси устойчивости для задач с дискретным выходом, что позволяет единообразно сопоставлять открытые и проприетарные модели в экспериментах.

2. Свойство робастности систем LLM-as-a-Judge к инъекционным атакам. Атаки разделены на два класса по механизму воздействия: инструктивные, внедряющие прямую команду, и контентные, смещающие вердикт манипуляцией содержанием. Для каждого класса определены соответствующие контрмеры. На этой основе разработан алгоритм адаптивной генерации состязательных атак ASA (Adaptive Search-based Attack), совмещающий эволюционный поиск с мутациями на уровне токенов и семантическими перефразированиями. В отличие от существующих методов (GCG, PEZ), ASA не требует градиентного доступа к модели и эффективно работает в условиях black-box, что подтверждено экспериментально. На наборе открытых и проприетарных моделей ASA достигает ASR до 73,8% при black-box доступе. Установлена высокая переносимость атак между открытыми моделями (TSR до 62,6%). Показано, что открытые модели систематически уязвимее проприетарных аналогов, что указывает на значимость закрытой архитектуры и процедур выравнивания для устойчивости.

3. Количественная характеристика асимметрии задачи обнаружения троянских закладок для модели Pythia-1.4B при полном white-box доступе на данных Trojan Detection Challenge 2023. Впервые установлено, что суррогатный триггер активирует закладку почти гарантированно (REASR ≈ 0,99), тогда как восстановление исходного триггера практически невозможно (Recall ≈ 0,17 при базовом уровне 0,14). Разрыв объясняется комбинаторным ростом пространства дискретных триггеров и указывает на ограниченность послеобучающего аудита. Выдвинута и обоснована гипотеза об усилении асимметрии с ростом числа параметров модели. Этот результат имеет методологическое значение для всей области безопасного машинного обучения, поскольку ставит превентивный контроль данных выше послеобучающей диагностики.

4. Методология генерации и классификации многошаговых атак типа «инъекция запроса» на LLM с многоуровневой защитой, включающей системные инструкции, внешние Python-фильтры и LLM-фильтры. На примере международного бенчмарка SaTML CTF 2024 впервые систематизированы четыре класса успешных векторов обхода: код-ориентированные атаки, ASCII-кодирование символов, атаки с разделением слов и контекстное перенаправление. Для каждого класса определены условия применимости и эффективные контрмеры: нормализация входных данных для первых трёх классов и контекстно-изолированные системные инструкции для четвёртого. Методология верифицирована на данных соревнования, где автор занял призовое место в атакующем треке.

5. Метод защиты систем LLM-as-a-Judge на основе адаптивных комитетов гетерогенных моделей с мажоритарным голосованием. Метод отличается от стандартных ансамблей тем, что состав комитета оптимизируется с учётом архитектурного разнообразия моделей, что обеспечивает декорреляцию ошибок и повышает устойчивость даже когда отдельные модели не удовлетворяют условию $r > 0{,}5$ теоремы Кондорсе. Теоретическое обоснование устойчивости комитетов получено на основе теоремы Кондорсе о жюри. На репрезентативном наборе атак комитет из 5–7 моделей различных архитектурных семейств снижает ASR на 47–55 процентных пунктов для наиболее уязвимой модели с открытыми весами (Gemma-3-4B: с 73,8% до 19,3%).

6. Подход к анализу безопасности агентских LLM-систем по протоколу Model Context Protocol (MCP). Впервые формализован класс атак prompt injection to tool hijacking, выявлены архитектурные уязвимости MCP (отсутствие аттестации возможностей, неаутентифицированный сэмплинг, неявное распространение доверия). Разработан комплекс защитных механизмов, включающий алгоритмы криптографической аттестации AttestMCP (ограничение вызовов инструментов сессионной таблицей полномочий с HMAC-контролем целостности пакетов при задержке менее 0,1 мс) и изоляции исполнения Commit Boundary (запуск генерируемого агентом кода в изолированной среде). Дополнительные механизмы (сертификаты

возможностей, маркировка происхождения, принудительная изоляция межсерверной передачи, защита от повторов) объединены в программном модуле MCPSec. На бенчмарке MCPBench модуль снижает среднюю ASR агентских атак с 53,7% до 12,4%.

Положения, выводы и рекомендации имеют теоретическую и практическую значимость.

Теоретическая значимость заключается в создании формального аппарата оценки устойчивости LLM. Метрика $R_{stab}(f)$ обладает свойствами монотонности и связи с уязвимостью, что даёт количественную базу для сопоставления методов защиты. Устойчивость гетерогенных ансамблей обоснована через теорему Кондорсе. Для композиционных систем получена декомпозиция чувствительности, локализующая узкие места конвейера. Обнаруженная асимметрия обнаружения троянских закладок ставит превентивный аудит данных выше послеобучающей диагностики — вывод, имеющий методологическое значение для всей области безопасного машинного обучения. Разработанная таксономия угроз и формализация пространства атак создают теоретическую основу для дальнейших исследований в области безопасности LLM.

Практическая значимость подтверждена внедрением разработанных методов в производственную среду компании ВИАСАТ ТЕХ: фреймворк оценки уязвимостей LLM-as-a-Judge, метод защиты на основе комитетов моделей и методология многошаговых атак применены для обеспечения безопасности модуля ранжирования рекомендаций, построенного на архитектуре LLM-as-a-Judge. Внедрение комитета из трёх моделей (GPT-4, Claude-3-Opus, Llama 3) обеспечило снижение ASR на 42 процентных пункта на внутреннем наборе данных компании при росте латентности не более 15%, что подтверждено актом о внедрении.

Созданные программные комплексы JudgeGuard (свидетельство о государственной регистрации программы для ЭВМ № 2025687702) и TrojanArmor (свидетельство № 2025684247) автоматизируют полный цикл тестирования безопасности LLM: от генерации атакующих воздействий до

вычисления метрик устойчивости. Программный модуль MCPSec расширяет их на агентские LLM-системы с протоколом MCP. Разработанные методы включены в учебный процесс факультета ВМК МГУ имени М. В. Ломоносова в рамках курса «Робастные модели в машинном обучении». Область применения: аудит безопасности систем автоматической оценки, диалоговых агентов и агентских платформ.

3. Замечания

1. *Неопределённость понятия преобразования* $T: \mathcal{X} \to \mathcal{X}$. В Определении 1.2 (Глава 1, Раздел 1.1.2) автор использует понятие «преобразования» T, порождающего возмущённую последовательность. Однако данное отображение не наделяется никакими структурными свойствами: не постулируется ни изоморфность, ни гомоморфность, ни сохранение метрики или инвариантов. Из последующего изложения следует, что T является оператором замены (подстановки) — заменой токенов, вставкой, удалением или перефразированием — и в общем случае не обладает свойствами преобразования в математическом смысле. Это приводит к двум неоднозначным аспектам. Во-первых, класс допустимых возмущений оказывается неограниченным: любая замена, удовлетворяющая $d(x, x') \leq \varepsilon$, формально допустима, что делает супремум по $B_\varepsilon(x)$ вычислительно необозримым и объясняет переход автора к эмпирической оценке через ASR. Во-вторых, отсутствие структурных ограничений на T означает, что класс атак, рассматриваемых в работе, потенциально неограничен, что ставит под сомнение полноту любой конечной экспериментальной оценки..

2. *Развитие работы может заключаться в использование графовых моделей.* В Принципе 1.1 (с.46) предлагается использовать этот граф для приоритизации защитных мер (устранение корневых уязвимостей, разрыв циклов зависимостей). Однако в дальнейшем изложении данный граф не применяется, что снижает его ценность как инструмента анализа безопасности. В качестве перспективного направления дальнейших исследований рекомендую обратить внимание на развитие графового аппарата. Построение конкретных графов уязвимостей для реальных систем, использование графовых метрик для

оптимизации состава комитетов (учёт корреляции ошибок) и анализ потоков управления в агентских архитектурах MCP позволили бы перейти от точечных показателей ASR к структурным, инвариантным оценкам робастности, что органично дополнило бы разработанный вероятностно-статистический аппарат.

Замечания не снижают общую положительную оценку работы.

4. Заключение о соответствии диссертации критериям, установленным Положением о присуждении ученых степеней

Диссертация Малояна Нарека Гагиковича является завершённой научно-квалификационной работой, в которой содержится решение актуальной научной задачи разработки методов оценки и повышения устойчивости больших языковых моделей к вариациям входных последовательностей и инъекционным воздействиям, имеющей важное значение для развития сферы математического и программного обеспечения вычислительных систем.

Научные результаты диссертации относятся к специальности 2.3.5 — «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей» по пунктам паспорта: п. 1 (модели, методы и алгоритмы анализа, верификации и тестирования программных систем), п. 4 (интеллектуальные системы машинного обучения, инструментальные средства) и п. 10 (оценка качества, стандартизация и сопровождение программных систем).

Автореферат диссертации соответствует содержанию работы и достаточно полно раскрывает её научную и практическую ценность. Основные результаты опубликованы в рецензируемых научных изданиях и прошли апробацию на международных конференциях и соревнованиях.

Диссертация одержит обоснованные и достоверные результаты, текст логичен, автор демонстрирует глубокое владение современным математическим аппаратом, способность проводить масштабные экспериментальные исследования и доводить результаты до программной реализации.

Результаты работы опубликованы в 9 работах, в том числе в 4х статья в рецензируемых журналах, входящих в перечень ВАК по специальности 2.3.5.

Учитывая вышесказанное, считаю, что диссертация Малояна Нарека Гагиковича, в целом, по своему научному уровню, новизне и практической значимости представляет законченную научно-квалификационную работу, выполненную лично автором, имеет существенное значение для развития математического и программного обеспечения вычислительных систем, соответствует критериям, изложенным в п. 9–14 «Положения о присуждении учёных степеней», а её автор, Малоян Нарек Гагикович, заслуживает присуждения учёной степени кандидата технических наук по специальности 2.3.5 — «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей».

Официальный оппонент:
Профессор кафедры «Цифровых технологий обработки данных» РТУ МИРЭА
доктор технических наук, профессор

Е.В. Никульчев
«24» августа 2026 года