ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА

кандидата физико-математических наук, доцента
АНО ВО «Университет Иннополис»
Иванова Владимира Владимировича

на диссертационную работу МАЛОЯНА Нарека Гагиковича

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

представленную к защите на соискание ученой степени кандидата технических наук

по специальности 2.3.5 «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей»

Актуальность темы диссертации. Большие языковые модели все шире используются в системах принятия решений, автоматического оценивания и агентских платформах с доступом к внешним инструментам. Вариации входных последовательностей, включая инъекции запросов, состязательные преобразования и скрытые триггеры, способны искажать вердикты и приводить к нежелательным программным действиям. Существующие методы оценки робастности не в полной мере учитывают дискретную природу токенов и авторегрессионную генерацию, а отдельные фильтры не обеспечивают универсальной защиты от адаптивного атакующего.

Поэтому разработка формальных метрик, процедур тестирования и программных механизмов защиты LLM-систем является актуальной

научной и практической задачей. Цель работы состоит в создании комплекса моделей, методов и алгоритмов оценки и повышения устойчивости LLM к состязательным вариациям входа и инъекционным воздействиям. Выбор темы диссертации Н. Г. Малояна следует признать обоснованным и своевременным.

Научная новизна и основные результаты. В ходе решения поставленной задачи автором получены следующие результаты, обладающие научной новизной:

1) предложена метрика генеративной устойчивости Rstab(f), основанная на расхождении распределений токенов при возмущении входа, и установлены связи устойчивости с уязвимостью;
2) формализована устойчивость LLM-as-a-Judge к инъекциям и разработан алгоритм ASA, достигающий в исследованных условиях ASR до 73,8%;
3) выявлена асимметрия обнаружения троянских закладок: для Pythia-1.4B получено REASR около 0,99 при Recall около 0,17;
4) разработана методология анализа многошаговых атак и систематизированы четыре класса обхода многоуровневой защиты;
5) предложена комитетная защита LLM-as-a-Judge, снизившая ASR наиболее уязвимой модели с 73,8% до 19,3%;
6) предложена защита агентских LLM-систем, включающая MCPBench, AttestMCP, Commit Boundary и механизмы разграничения полномочий и изоляции.

Краткая характеристика содержания работы. Диссертация изложена на 187 страницах, включает введение, шесть содержательных глав, заключение и приложение; библиографический список содержит 178 наименований.

В первой и второй главах систематизированы 17 типов угроз и разработаны теоретические основы оценки устойчивости. Введены метрики Rclass(h) и Rstab(f), показатель глобальной уязвимости и риск-функция, получены границы для локализованных атак, рассмотрены композиционные и агентские системы. Автор корректно разграничивает строгие результаты и эмпирические гипотезы.

В третьей и четвертой главах представлены алгоритм ASA, комитетная защита и их экспериментальная оценка на MT-Bench, Kaggle, TDC2023 и SaTML CTF 2024 с моделями семейств Gemma, Llama, Pythia, GPT и Claude. Показаны результативность и переносимость атак, эффективность гетерогенных комитетов и ограничения обратной инженерии троянских триггеров.

В пятой и шестой главах описаны JudgeGuard, TrojanArmor и MCPSec, а также их апробация и внедрение. На MCPBench из 847 сценариев средний ASR снижен с 53,7% до 12,4%. Результаты использованы в учебном процессе ВМК МГУ и производственной среде ВИАСАТ TEX; для JudgeGuard заявлено снижение ASR на 42 процентных пункта при увеличении латентности не более чем на 15%.

Соответствие паспорту специальности. Разработанные модели и алгоритмы анализа и тестирования программных систем, методы машинного обучения и инструментальные средства оценки качества и безопасности соответствуют специальности 2.3.5 «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей».

Достоверность и обоснованность результатов. Достоверность обеспечивается корректным математическим аппаратом, явными моделями угроз, экспериментами на публичных данных и моделях разных архитектур, сравнением методов атак и защиты, а также статистической обработкой результатов. Основные результаты апробированы на четырех

конференциях и трех соревнованиях; по теме опубликовано девять работ, включая четыре журнальные статьи и два свидетельства о регистрации программ для ЭВМ.

Теоретическая и практическая значимость. Теоретическая значимость состоит в развитии аппарата оценки устойчивости LLM и композиционных систем. Практическая значимость определяется созданием JudgeGuard, TrojanArmor и MCPSec, применимых для тестирования моделей, обнаружения закладок, комитетной защиты и контроля инструментальных вызовов.

Замечания по диссертационной работе. В целом диссертация выполнена на высоком научном уровне, однако к работе необходимо высказать следующие замечания:

1) Метрика Rstab(f) получила ограниченную прямую эмпирическую валидацию: эксперимент раздела 4.2.5 проведен на 20 промптах с пятью доброкачественными возмущениями. Для усиления результата следовало бы расширить выборку и проверить метрику на нескольких классах состязательных воздействий.
2) Вклад архитектурной гетерогенности в эффективность комитетов не полностью отделен от влияния их размера и состава. Вывод желательно подкрепить абляционным исследованием при фиксированном числе моделей и анализом корреляции ошибок.
3) Для MCPBench приведены убедительные агрегированные результаты по 847 сценариям, однако отсутствуют доверительные интервалы по отдельным моделям и абляция компонентов MCPSec. Такая детализация позволила бы точнее оценить переносимость результатов и вклад отдельных механизмов.

Отмеченные замечания имеют характер направлений дальнейшего развития и уточнения экспериментальной части, не затрагивают основные

результаты и не влияют на общую положительную оценку диссертационной работы.

Вопросы к соискателю. В порядке научной дискуссии представляется целесообразным получить ответы на следующие вопросы:

1) Как на практике должна калиброваться функция g в модели $V(h) \approx g(1 - Rclass(h))$ для нелокализованных prompt-injection атак и какой экспериментальный дизайн позволил бы подтвердить ее связь именно с Rstab(f), а не только с показателем 1 − ASR?
2) Какой количественный критерий архитектурного разнообразия целесообразно использовать при формировании адаптивного комитета моделей и сохранится ли наблюдаемый выигрыш, если атакующий будет одновременно оптимизировать воздействие против всего комитета?
3) Как распределяется снижение ASR в MCPSec между алгоритмами AttestMCP, Commit Boundary и четырьмя вспомогательными механизмами, и насколько различается вклад этих компонентов для Claude-3.5-Sonnet, GPT-4o и Llama-3.1-70B?

Заключение. Диссертационная работа Малояна Нарека Гагиковича является законченным самостоятельным исследованием актуальной научно-технической задачи. Полученные результаты обладают научной новизной, теоретической и практической значимостью и подтверждены теоретическим обоснованием, экспериментальной проверкой, апробацией и программной реализацией.

Автореферат правильно и достаточно полно отражает основное содержание диссертации. Публикации соискателя и представленные сведения об апробации соответствуют основным результатам работы.

Диссертация соответствует требованиям, предъявляемым к диссертациям на соискание ученой степени кандидата технических наук, а ее автор, Малоян Нарек Гагикович, заслуживает присуждения ученой степени кандидата технических наук по специальности 2.3.5 «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей».

Официальный оппонент:

Иванов В.В.

к.ф.-м.н., доцент АНО ВО «Университет Иннополис»

_____ / _____Иванов В.В._____

подпись Ф. И. О.

«_31_» августа _____ 2026 г.