Отзыв научного руководителя

на диссертационную работу 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»

Малоян Н.Г. является выпускником аспирантуры факультета ВМК МГУ имени М.В. Ломоносова 2025 года.

Работа Н.Г. Малояна посвящена актуальной теме обеспечения устойчивости больших языковых моделей (LLM) к возможным вариациям входных последовательностей. Актуальность обусловлена стремительным развитием и масштабным внедрением LLM, в том числе, активным развитием направления агентов Искусственного интеллекта. Необходимость внедрения в критически важные приложения при этом сталкивается с ростом специфических угроз информационной безопасности, в частности, атак типа «инъекция подсказок», «косвенная инъекция подсказок» и внедрением троянских закладок. В диссертационной работе автором впервые предложена и теоретически обоснована формальная математическая модель оценки устойчивости LLM, разработан фреймворк для исследования уязвимости систем LLM-as-a-Judge к атакам типа «инъекция подсказок», предложен метод защиты систем LLM-as-a-Judge на основе адаптивных комитетов гетерогенных моделей с мажоритарным голосованием а также экспериментально исследована связь сложности атаки и сложности её точной идентификации.

Практическая значимость работы подтверждается успешным применением разработанных методов в производственную деятельность компании Viasat. Разработанный фреймворк оценки уязвимостей LLM-as-a-Judge и метод защиты на основе комитетов моделей применены для обеспечения безопасности модуля ранжирования рекомендаций, который построен на архитектуре LLM-as-a-Judge. Проведённое исследование является методологически корректным и вносит существенный вклад в развитие основ доверенного искусственного интеллекта, позволяя реализовывать более

безопасные модели агентов искусственного интеллекта. Представленная работа является первой работой на русском языке в указанной области.

За время обучения в аспирантуре Н.Г. Малоян активно участвовал в проектах кафедры Информационной безопасности, направленных на исследование кибербезопасности систем искусственного интеллекта. Параллельно с научной работой он ведёт активную педагогическую деятельность. Период его обучения в аспирантуре совпал с открытием на факультете ВМК МГУ имени М.В. Ломоносова первой в стране магистерской программы, посвященной Искусственному интеллекту в кибербезопасности. Н.Г. Малоян принял активное участи в этом процессе. Он читает лекции по робастным моделям, ведет лекции и семинары по Python, является руководителем (техническим консультантом) ряда интересных магистерских диссертаций, которые заложили новые продукты по кибербезопасности систем Искусственного интеллекта. Разработанные им решения побеждали в открытых хакатонах и международных соревновательных треках по безопасности LLM.

За время активной научной деятельности Н.Г. Малоян опубликовал в совокупности 14 научных работ. Результаты непосредственно диссертационного исследования опубликованы соискателем в четырех рецензируемых научных изданиях, получено два свидетельства регистрации программы для ЭВМ. Материалы диссертации успешно прошли апробацию на двух международных и двух российских конференциях.

Считаю, что диссертационная работа соответствует всем требованиям, предъявляемым ВАК РФ к работам на соискание ученой степени кандидата технических наук по специальности 2.3.5 «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей», а её автор, Малоян Нарек Гагикович, заслуживает присуждения учёной степени кандидата технических наук.

Научный руководитель:

Ведущий научный сотрудник лаборатории Открытых Информационных Технологий кафедры Информационной безопасности факультета ВМК МГУ имени М.В. Ломоносова, д.т.н.

Намиот Д.Е.

26 мая 2026 года