

Отзыв научного руководителя

на диссертационную работу Алексеевской Ирины Сергеевны на тему:
«Методы и программные средства исследования и повышения
устойчивости языковых моделей к целенаправленным искажениям»,

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

С 2021 года Алексеевская И.С. активно начала работать в научной группе, исследующей методы защиты от атак на генеративные модели. Алексеевская И.С. в 2023 году получила степень магистра по направлению «Математика и компьютерные науки» в Федеральном государственном бюджетном образовательном учреждении высшего образования «Московский государственный университет имени М.В. Ломоносова».

С 2023 года будучи аспирантом ИСП РАН, Алексеевская И.С. начала заниматься исследованиями в области разработки методов защиты больших языковых моделей от состязательных атак. Актуальность темы диссертационного исследования Алексеевской И.С. обусловлена массовым внедрением языковых моделей в критически важные сферы и необходимостью обеспечения их информационной безопасности. Выявление новых скрытых уязвимостей, таких как бэкдоры и джейлбрейки, а также разработка эффективных методов защиты от них, является ключевым условием создания доверенных систем искусственного интеллекта, отвечающим современным стандартам и исключающим возможность причинения вреда.

Научный вклад диссертационной работы заключается в разработке новых методов исследования и повышения устойчивости языковых моделей к целенаправленным искажениям. Алексеевской И.С. предложен оригинальный метод внедрения бэкдора OrderBkd, основанный на изменении порядка слов, который превосходит существующие аналоги по метрикам скрытности (перплексии и семантического сходства) и успешно обходит базовый механизм защиты ONION. Кроме того, разработан новый метод джейлбрейка SoftPrompt в сценарии «черного ящика» на основе перефразирования в научный стиль, демонстрирующий высокую эффективность против современных моделей. На базе выявленных

уязвимостей автором создана комплексная методика защиты, комбинирующая оригинальный алгоритм OrderDef и известные процедуры выравнивания (RLHF, DPO).

Результаты научных исследований были успешно реализованы в виде программной системы, интегрированной в облачную Платформу доверенного искусственного интеллекта (Платформа ДИИ ИСП РАН). По теме диссертации опубликованы три научные работы, включая статьи в журналах из баз ВАК, Scopus и Web of Science, а также получено свидетельство о государственной регистрации программы для ЭВМ.

Диссертационная работа Алексеевской И. С. представляет собой завершённое научное исследование, результаты которого вносят значимый вклад в развитие методов обеспечения безопасности систем искусственного интеллекта и создают научную основу для дальнейших исследований в данной области.

Считаю, что диссертационная работа полностью соответствует требованиям ВАК к кандидатским диссертациям по специальности 2.3.5 «Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей», а ее автор, Алексеевская Ирина Сергеевна, заслуживает присуждения ученой степени кандидата физико-математических наук.

Научный руководитель,
заведующий Исследовательского
центра доверенного искусственного
интеллекта ИСП РАН, к. ф.-м. н.

Турдаков Денис Юрьевич

7 октября 2026 года