УДК 004.056

Отравление наборов данных моделей искусственного интеллекта: требования пункта 60 приказа ФСТЭК России № 117 и независимость атаки от масштаба модели

Коновалов С.И., Аверин П.А.

ФГБОУ ВО "Брянский государственный технический университет", г. Брянск, Российская Федерация

Пункт 60 Требований, утверждённых приказом ФСТЭК России от 11 апреля 2025 г. № 117 [1], требует исключить несанкционированный доступ к информации, её распространение и модификацию, а также использование информационных систем не по назначению за счёт воздействия на наборы данных, модели искусственного интеллекта и их параметры. Тот же пункт запрещает передавать разработчику модели информации ограниченного доступа, в том числе для её улучшения. Пункт 61 требует включать в состав информационных систем доверенные технологии искусственного интеллекта, понятие которых заимствовано из Национальной стратегии развития искусственного интеллекта [2].

Воздействие на наборы данных охватывает отравление – внесение в данные предобучения, дообучения или векторного представления изменений, создающих уязвимости, скрытые закладки или смещения [3]. Закладка представляет собой поведение модели, активируемое заданной последовательностью во входных данных и не проявляющееся при обычной работе. Долгое время предполагалось, что для отравления крупной модели нарушителю нужно контролировать заметную долю обучающего корпуса.

Это допущение опровергнуто экспериментально. В исследовании 2025 г. модели размером от 600 млн до 13 млрд параметров обучались на наборах объёмом от 6 до 260 млрд токенов, и во всех конфигурациях около 250 отравленных документов оказалось достаточно для внедрения закладки, хотя крупнейшие модели обучались на в 20 раз большем объёме чистых данных [4].

Определяющей величиной оказалось абсолютное число отравленных образцов, а не их доля в корпусе.

Практическая реализуемость отравления показана для наборов данных из открытых источников [5]. Первый способ использует изменяемость содержимого сети: набор распространяется как перечень ссылок, и нарушитель, приобретая домены с истёкшей регистрацией, подменяет содержимое по части адресов. Второй способ нацелен на периодические снимки редактируемых ресурсов и требует изменений лишь в короткое окно перед снятием снимка. Отравление 0,01 % наборов LAION-400M или COYO-700M первым способом авторы оценили примерно в 60 долларов США.

Сопоставление с пунктом 60 приказа ФСТЭК России № 117 выявляет различие между двумя его частями. Запрет на передачу информации разработчику защищает конфиденциальность сведений информационной системы, но не целостность наборов данных: отравленные образцы поступают не из информационной системы, а из открытых источников разработчика либо из внешних корпусов, привлекаемых для дообучения. Поэтому требование исключить воздействие на наборы данных требует самостоятельных мер.

Двести пятьдесят документов в корпусе объёмом в сотни миллиардов токенов недоступны обнаружению статистическими фильтрами, настроенными на аномальную долю подозрительного содержимого. Поэтому контроль целостности целесообразно строить на учёте происхождения данных, а не на анализе содержимого. К таким мерам относятся перечень компонентов модели и наборов данных, для описания которого формат CycloneDX поддерживает компоненты машинного обучения [3], фиксация хеш-сумм элементов набора на момент утверждения против первого способа отравления и контроль времени снятия снимков редактируемых источников против второго.

Отсюда вытекает критерий для требования о доверенных технологиях: технологию целесообразно относить к доверенной, если происхождение наборов данных предобучения и дообучения документировано с зафиксированными хеш-

суммами. Без таких сведений невозможно подтвердить, что воздействие на наборы данных исключено.

Эксперименты [4] проведены на моделях до 13 млрд параметров и для определённых типов закладок, поэтому перенос результата на более крупные модели требует подтверждения. Однако число образцов, необходимых для атаки, с ростом модели не увеличивается. Таким образом, требование пункта 60 не может выполняться за счёт масштаба модели и требует учёта происхождения данных как самостоятельного объекта защиты.

Список использованной литературы

1. *Об утверждении Требований о защите информации, содержащейся в государственных информационных системах, иных информационных системах государственных органов, государственных унитарных предприятий, государственных учреждений: приказ ФСТЭК России от 11.04.2025 № 117.* – Официальный интернет-портал правовой информации, 17.06.2025, № 0001202506170011.
2. *О развитии искусственного интеллекта в Российской Федерации: Указ Президента Российской Федерации от 10.10.2019 № 490* (в ред. Указа Президента Российской Федерации от 15.02.2024 № 124).
3. *LLM04:2025 Data and Model Poisoning* // OWASP Top 10 for Large Language Model Applications 2025 [Электронный ресурс]. – URL: https://genai.owasp.org/llmrisk/llm042025-data-and-model-poisoning/ (дата обращения: 26.09.2026).
4. Souly A. *Poisoning Attacks on LLMs Require a Near-constant Number of Poison Samples* / A. Souly, J. Rando, E. Chapman [et al.] // arXiv preprint. – 2025. – arXiv:2510.07192.
5. Carlini N. *Poisoning Web-Scale Training Datasets is Practical* / N. Carlini, M. Jagielski, C.A. Choquette-Choo [et al.] // arXiv preprint. – 2023. – arXiv:2302.10149.