ПРОБЛЕМАТИКА ОТРАВЛЕНИЯ ДАННЫХ В БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЯХ

А.М. Агаев
Научный руководитель – Алексеев В.Ф., к.т.н., доцент
Белорусский государственный университет информатики и радиоэлектроники

Современные большие языковые модели (БЯМ), такие как GPT-4, Gemini, Llama и другие, стали фундаментальной технологией, трансформирующей области обработки естественного языка, информационного поиска, креативных индустрий и автоматизации бизнес-процессов. Однако их стремительное развитие и повсеместное внедрение сопровождается ростом кибербезопасностных рисков, среди которых

отравление данных (Data Poisoning) представляет собой одну из наиболее серьезных и трудноустранимых угроз. Актуальность исследований в данной области обусловлена комплексом факторов, связанных с уязвимостью жизненного цикла БЯМ, критичностью последствий атак и необходимостью разработки эффективных механизмов защиты [1-6].

«Замусоривание» (или «помеховое обучение», «обучение на шумных данных») LLM – это многогранная проблема, связанная с использованием для обучения некачественных, шумных, некорректных или намеренно искаженных данных. Это приводит к деградации качества модели, генерации ею недостоверной или бессмысленной информации, усилению предвзятости и появлению «галлюцинаций». Актуальность темы обусловлена зависимостью современных LLM от огромных объемов неверифицированных данных из интернета.

В работе рассматривается проблематика замусоривания больших языковых моделей (LLM), последствия отравления языковой модели и меры профилактики таких отравлений.

По мере развития технологий и методов машинного обучения, фильтрации большого объема данных и общим прогрессом в Big Data, становится все более актуальным вопросы – правильно ли мы извлекаем данные, на какие источники опирается LLM и как долго одна и та же языковая модель будет актуальна. При погружении в машинное обучение становится понятным, что раз информация берется из интернета, то рассчитывать на достоверность приходится не всегда, но что если в языковую модель подать на вход несколько абсурдных запросов? Интуитивно ожидается, что для однозначно негативного сценария необходимо чрезмерное число некачественных данных, вероятно, до половины или более от всей базы данных, чтобы языковая модель сбилась в большом объеме недостоверной информации.

Большие языковые модели обучаются на колоссальных объемах данных, собранных из открытых и зачастую неверифицируемых источников (Интернет, корпоративные архивы, оцифрованные книги). Это делает процесс обучения крайне уязвимым для целенаправленных злонамеренных вмешательств. Злоумышленник может внедрить в обучающий набор специально сконструированные данные, предназначенные для формирования у модели скрытых уязвимостей (backdoors) или устойчивых вредоносных паттернов поведения [7]. Проверка и очистка триллионов токенов обучающих данных требуют непомерных вычислительных и человеческих ресурсов, что создает практическую невозможность полного исключения отравленных образцов [8].

Однако исследования в области отравления данных (Data Poisoning) выявили – большие языковые модели, слишком часто обучаемые некачественными данными, постепенно теряют способность к логическому рассуждению, причем эффект наблюдается уже от ~250 запросов. В следствие отравления модель начинает пропускать логические шаги, выдавать поверхностные ответы и теряет последовательность мышления (thought-skipping).

Мерами противодействия (мигитации) являются использование доверенных источников, ведение каталогов, предотвращение размещения произвольной информации по ссылкам , входящих в датасеты, фильтрацию данных и анализ воздействия новых данных на точность LLM, так как отравление данных приведет к ее снижению. Немаловажным является контроль доступа к базам данных и аудит их содержимого.

Библиографический список

1. Шумина К. А., Петров Ф. И. Влияние качества обучающих данных на достоверность выходов больших языковых моделей // Искусственный интеллект и анализ данных. – 2023. – Т. 21, № 1. – С. 56–67.

2. Crawford K. The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. – New Haven: Yale University Press, 2021. – 288 с.

3. Shumailov I., Shumaylov Z., Zhao Y., Gal Y., Papernot N., Anderson R. The Curse of Recursion: Training on Generated Data Makes Models Forget [Электронный ресурс] // arXiv preprint arXiv:2305.17493. – 2023. – URL: https://arxiv.org/abs/2305.17493 (дата обращения: 22.10.2025).

4. Bommasani R., Hudson D. A., Adeli E., et al. On the Opportunities and Risks of Foundation Models [Электронный ресурс] // Stanford Center for Research on Foundation Models (CRFM). – 2022. – URL: https://arxiv.org/abs/2108.07258 (дата обращения: 22.10.2025).

5. Poisoning Attacks on LLMS require a Near-Constant number of Poison Samples / A. Souly, J. Rando, E. Chapman [и др.]; авторы: B. Hasircioglu, E. Shereen, C. Mougan, V. Mavroudis, E. Jones, C. Hicks, N. Carlini, Y. Gal, R. Kirk // arXiv.org [Электронный ресурс]. – 2025. – № arXiv:2510.07192v1 [cs. LG]. – URL: https://arxiv.org/abs/2510.07192 (дата обращения: 22.10.2025)

6. Goodfellow I. Deep Learning / I. Goodfellow, Y. Bengio, A. Courville. – Cambridge, Massachusetts: The MIT Press, 2016. – 800 p.

7. Глушков С. В., Иванова А. К. Атаки на машинное обучение: отравление данных в больших языковых моделях // Информационная безопасность и киберзащита. – 2023. – Т. 15, № 4. – С. 44–52.

8. Chen Y., Wang W., Liu Z. Data Poisoning Attacks and Defenses in Large Language Models // Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. – 2023. – P. 110–125.