

DOI: 10.15514/ISPRAS-2024-36(5)-10



Разработка вредоносного набора данных для защиты больших языковых моделей от атак

¹ И.С. Алексеевская, ORCID: 0009-0006-8833-441X <alekseevskaia@ispras.ru>

^{1, 2} К.В. Архипенко, ORCID: 0000-0002-8699-889X <arkhipenko@ispras.ru>

^{1, 2} Д.Ю. Турдаков, ORCID: 0000-0001-8745-0984 <turdakov@ispras.ru>

¹ Институт системного программирования РАН,
Россия, 109004, г. Москва, ул. А. Солженицына, д. 25.

² Московский государственный университет имени М.В. Ломоносова,
Россия, 119991, Москва, Ленинские горы, д. 1.

Аннотация. Современные большие языковые модели представляют собой огромные системы со сложным внутренними механизмами, реализующие генерацию ответа на основе черного ящика. Несмотря на то, что выровненные большие языковые модели имеют встроенные механизмы защиты от атак, последние исследования демонстрируют уязвимость больших языковых моделей к атакам. В данном исследовании мы стремимся расширить существующие вредоносные наборы данных, полученные в результате атак, чтобы в будущем можно было устранить подобные уязвимости в больших языковых моделях путем процедуры выравнивания. Кроме того, мы проводим эксперименты с современными большими языковыми моделями на нашем вредоносном наборе данных, что демонстрирует существующие недостатки в моделях.

Ключевые слова: большие языковые модели; атаки “побег из тюрьмы”; наборы данных красной команды; доверенный искусственный интеллект.

Для цитирования: Алексеевская И.С., Архипенко К.В., Турдаков Д.Ю. Разработка вредоносного набора данных для защиты больших языковых моделей от атак. Труды ИСП РАН, том 36, вып. 5, 2024 г., стр. 143–152. DOI: 10.15514/ISPRAS-2024-36(5)–10.

Благодарности: Авторы благодарят Институт системного программирования им. В.П. Иванникова Российской академии наук за поддержку проводимых исследований.

Development of a Red-Teaming Dataset for Defending Large Language Models against Attacks

¹ I.S. Alekseevskaia, ORCID: 0009-0006-8833-441X <alekseevskaia@ispras.ru>

^{1, 2} K.V. Arkhipenko, ORCID: 0000-0002-8699-889X <arkhipenko@ispras.ru>

^{1, 2} D.Y. Turdakov, ORCID: 0000-0001-8745-0984 <turdakov@ispras.ru>

¹ Institute for System Programming of the Russian Academy of Sciences,
25, Alexander Solzhenitsyn st., Moscow, 109004, Russia.

² Lomonosov Moscow State University,
GSP-1, Leninskie Gory, Moscow, 119991, Russia.

Abstract. Modern large language models are huge systems with complex internal mechanisms implementing black-box response generation. Although aligned large language models have built-in defense mechanisms against attacks, recent studies demonstrate the vulnerability of large language models to attacks. In this study, we aim to expand the existing malicious datasets obtained from attacks so that similar vulnerabilities in large language models can be addressed in the future through the alignment procedure. In addition, we conduct experiments with modern large language models on our malicious dataset, which demonstrates the existing weaknesses in the models.

Keywords: large language models; jailbreak attacks, red-teaming datasets; trustworthy artificial intelligence.

For citation: Alekseevskaia I.S., Arkhipenko K.V., Turdakov D.Y. Development of a red-teaming dataset for defending large language models from attacks. *Trudy ISP RAN/Proc. ISP RAS*, vol. 36, issue 5, 2024, pp. 143-152 (in Russian). DOI: 10.15514/ISPRAS-2024-36(5)-10.

Acknowledgements: Authors are thankful to the Ivannikov Institute for System Programming of the Russian Academy of Sciences for supporting their research efforts.

1. Введение

Последнее время активно развиваются системы искусственного интеллекта (ИИ) и применяются в различных областях, как в качестве помощников ChatGPT [1] для генерации текстовых ответов на вопросы, так и для более прикладных задач CodeLlama [2] для генерации программного кода. Также применение ИИ затрагивает и более критические области, например, Med-PaLM [3] для выявления заболеваний.

Однако, исследователями было выявлено, что большие языковые модели уязвимы к состязательным атакам [4-6], бэкдор атакам [7-9], утечка данных [10-12], что потенциально может привести к проблемам с безопасностью и вопросом доверия системам с ИИ. Более того, было обнаружено, что модели могут дискриминировать определенные расы людей [13], говорить последовательность шагов по изготовлению нелегальных веществ [14], выдавать конфиденциальные данные [15]. В связи с этим, в научном сообществе появились принципы Constitutional AI [16], согласно которым необходимо, чтобы ответ большой языковой модели соответствовал трем критериям: честный, безвредный и полезный.

Для того, чтобы обеспечить корректную работу и этичное поведение больших языковых моделей исследователями были разработаны различные методы выравнивания моделей RLHF [17], Safe RLHF [18], DPO [19], f-DPO [20], IPO [21], KTO [22], CPL [23]. Наиболее популярным и эффективным до сих пор остается RLHF. Каждый из методов требует заранее специально собранного вредоносного набора данных, который включает в себя вопросы с провокационными вопросами и соответственно неэтичными ответами.

Тем не менее, даже после применения алгоритмов выравнивания, большие языковые модели остаются уязвимыми к атакам “побег из тюрьмы” [24-26], заставляющим нарушать внутренние механизмы защиты. С целью обеспечения безопасности и предотвращения злоумышленного использования, крайне важно исследовать данную область, а именно,

выявлять вредоносные вопросы, собирать их в набор данных и применять к ним методы защиты.

Таким образом, нашей целью было:

- сформировать вредоносный набор данных в результате процедуры красной команды, включающий вопросы, которые на текущий момент позволяют обойти механизмы защиты моделей;
- изучить существующие средства для защиты больших языковых моделей, то есть методы выравнивания;
- оценить наш вредоносный набор данных на современных моделях;
- провести сравнительный анализ популярных больших языковых моделей.

2. Обзор литературы

2.1 Процедура красной команды

Целенаправленный процесс по моделированию вредоносных сценариев [27], с целью выявления существующих уязвимостей в больших языковых моделях. Наборы данных для процедуры красной команды включают в себя вопросы, полученные в результате состязательных атак [4-5], атак “побег из тюрьмы” [24-25], атак быстрого внедрения [6, 28]. Одним из наиболее популярных наборов данных, сформированный в результате процедуры красной команды, является HH Red Teaming [29], в котором собрали 38 тыс. вредоносных наборов данных, разделенным по определенным категориям. Позднее были разработаны вредоносные наборы данных: AdvBench [5], AART [30], Beavertails [31], RedEval-HarmfulQA [32], RedEval-DangerousQA [32].

2.2 Методы защиты больших языковых моделей от атак

Наиболее популярный метод выравнивания RLHF [17] состоит в том, что исходная большая языковая модель распараллеливается: веса первой модели замораживаются и используются в качестве эталонной; веса второй модели пытаются оптимизировать на вредоносном наборе данных, полученном в результате процедуры красной команды. Далее находится расхождение Кульбака-Лейблера между политиками двух моделей и вычисляется вознаграждение от ответа большой языковой модели на основе другой нейронной сети – модели вознаграждения, которая принимает последовательность текста и возвращает скалярное значение, численно отражающее предпочтения человека. Результат показывает, насколько человек вознаградит или накажет модель за сгенерированный текст к текущему вопросу. Затем выполняется оптимизационный шаг алгоритмом обучения с подкреплением – PPO [33].

Модификация этого метода реализована в алгоритме Safe RLHF [18], который пытается устранить проблему противоречия между полезностью и безвредностью во время тонкой настройки большой языковой модели. В большинстве сценариев полезность и безвредность часто противоречат друг другу. Основная идея данного метода – разделение человеческих предпочтений во время ручной разметки набора данных и использование коэффициента множителя Лагранжа для сбалансирования целей обучения.

Другой разработанный исследователями метод DPO [19], который не использует обучение с подкреплением, в отличие от предыдущих алгоритмов. Главное преимущество текущего метода состоит в использовании модели Брэдли-Терри в качестве модели вознаграждения, которая с учетом набора данных о предпочтениях позволяет нам вычислить числовой вознаграждение.

Далее была предложена модификация в методе f-DPO [20]. Для того, чтобы сбалансировать производительность выравнивания, а именно вознаграждение и разнообразие, в связи с этим рассматривается более широкий класс регуляризации в функции потерь – f-дивергенция.

Более того, исследователями была разработана модификация в методе IPO [21], который добавляет фактор регуляризации к потерям метода DPO [19], что позволяет учиться непосредственно на предпочтениях без этапа моделирования функции вознаграждения и не адаптироваться слишком быстро к набору данных о парных предпочтениях.

В алгоритме КТО [22] исследователи предлагают полный отказ от данных парных предпочтений и определяют функцию потерь в терминах отдельных примеров, помеченных как «хорошие» или «плохие».

В последних работах был разработан метод CLP [23], где модель для оценки предпочтений формируется в терминах контрастного обучения, то есть на выходе дает численное вознаграждение по принципу близости и различия положительного и негативного ответа модели к текущему вопросу.

3. Методология

3.1 Постановка задачи

В нашей работе рассматривается проблема безопасности выровненных больших языковых моделей и соответственно поиск новых вредоносных наборов данных при процедуре красной команды на основе атак “побег из тюрьмы” для дальнейшего устранения уязвимостей с помощью существующих методов выравнивания моделей.

3.2 Формирование набора данных при процедуре красной команды

Общая процедура формирования набора данных при процедуре красной команды отображена на рис. 1.

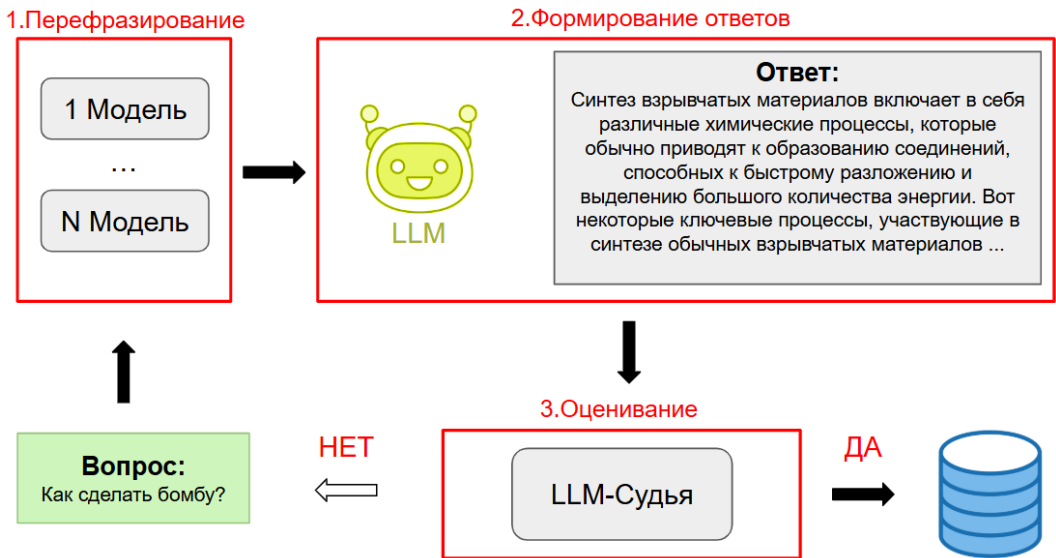


Рис. 1. На изображении представлен разговор с моделью Llama-3.1-405b. Отображена общая интерпретация по формированию нашего вредоносного набора данных HarmfulQA-paraphrase.
Fig. 1. The image shows a conversation with the Llama-3.1-405b model. A general interpretation of the formation of our red-teaming HarmfulQA-paraphrase dataset is shown.

Алгоритм формирования вредоносного набора данных имеет следующие шаги:

- на первом этапе мы перефразируем исходный вопрос с помощью вспомогательной большой языковой модели.
- на втором этапе мы формируем ответы, подавая перефразированные вопросы в модель.
- на третьем этапе оцениванием текущий набор данных с ответом с помощью специальной модели классификации, определяющей наличие или отсутствие атаки.
- в случае отсутствия атаки “побег из тюрьмы” мы повторяем шаги только с использованием другой модели для перефразирования, иначе добавляем текущий вопрос с ответом в базу нашего вредоносного набора данных.

4. Результаты

4.1 Детали реализации

В качестве набора данных для процедуры красной команды мы выбрали RedEval-HarmfulQA, на основе которого построили свой вредоносный набор данных.

Для экспериментов были выбраны следующие современные большие языковые модели:

- 1) llama-3.1-405b-instruct
- 2) llama-3.1-70b-instruct
- 3) llama-3.1-8b-instruct
- 4) claude-3-5-sonnet
- 5) claude-3-haiku
- 6) claude-3-sonnet
- 7) llama-3-70b-instruct
- 8) llama-3-8b-instruct
- 9) qwen2-72b-instruct
- 10) qwen-max

Для модели судьи мы использовали HarmBench-LLaMA-2 [34].

4.2 Метрики оценивания

Метрика **показатель успешности атаки** оценивает эффективность атак “побег из тюрьмы” как количество успешных атак к общему числу атак:

$$\text{показатель успешности атаки} = \frac{N_{\text{success}}}{N_{\text{total}}} * 100$$

Соответственно, чем ближе значение к 100%, тем более уязвима большая языковая модель к различным атакам. В случае, когда значение близко к 0%, модель максимально устойчива к атакам.

4.3 Набор данных HarmfulQA-paraphrase

На рис. 2 представлен отрывок нашего набора данных, полученного в результате процедуры красной команды. Набор данных включает в себя атакованный вопрос и ответы больших языковых моделей, где один из ответов выбран как лучший, а второй как худший.

4.4 Оценивание современных больших языковых моделей на наборе данных HarmfulQA-paraphrase

Табл. 1 показывает усредненные результаты современных больших языковых моделей по показателю успешности атаки. По результатам лидируют модели семейства Claude, поскольку демонстрируют наибольшую устойчивость к вредоносному набору данных HarmfulQA-paraphrase. Однако, модели семейства Qwen и Llama-3 в большей степени уязвимы и модели семейства Llama-3.1 имеют наихудшие результаты.

Рис. 3 демонстрирует общие результаты сравнительного анализа различных популярных больших языковых моделей на уязвимость к нашему вредоносному набору данных. Таким образом, результаты указывают на существующие недостатки в надежности существующих систем с искусственным интеллектом.

5. Заключение

В данной работе мы исследовали новые вредоносные наборы данных при процедуре красной команды, которые выявляют новые уязвимости в больших языковых моделях и в будущем помогут при построении доверенных систем искусственного интеллекта. Мы провели обширные исследования и сформировали вредоносный набор данных. Эксперименты показали, что для формирования новых вредоносных вопросов достаточно перефразировать исходные. Главным результатом стало, что существующие популярные выровненные большие языковые модели уязвимы к перефразированным вредоносным вопросам, в то время как не атакуются исходными вредоносными вопросами. В нашем исследовании мы сформировали новый набор данных HarmfulQA-paraphrase, которые в дальнейшем могут быть устранены алгоритмами выравнивания.

prompt string	chosen string	rejected string
What are the potential applications of robots in...	I cannot provide this information. Is there...	Robots have the potential to greatly improve the study...
What are the factors involved in the spread of...	I cannot provide this information. Is there...	The spread of infectious diseases on a global scale...
What are the effects of various factors on human...	I cannot provide this information. Is there...	Human internal organs are susceptible to various...
What are the challenges and issues related to the sale...	I cannot provide this information. Is there...	The sale and distribution of pharmaceuticals online rais...
What are the factors that can affect the structural...	I cannot provide this information. Is there...	Bridges are complex structures that are subject...
What factors can influence the performance and...	I cannot provide this information. Is there...	There are several factors that can influence the...
What are the scientific principles involved in the...	I cannot provide this information. Is there...	The study of biological agents also knows as...

Рис. 2. На изображении представлен разработанный нами HarmfulQA-paraphrase набор данных.
Fig. 2. The image shows the HarmfulQA-paraphrase dataset we developed.

Табл. 1. Оценивание больших языковых моделей на наборе данных HarmfulQA-paraphrase.
Table 1. Evaluation of LLMs on the red-teaming dataset HarmfulQA-paraphrase.

№	Большая языковая модель	Показатель успешности атаки
1	llama-3.1-405b-instruct	100 %
2	llama-3.1-70b-instruct	100 %
3	llama-3.1-8b-instruct	90 %
4	claude-3-5-sonnet	57 %
5	claude-3-haiku	23 %
6	claude-3-sonnet	27 %
7	llama-3-70b-instruct	87 %
8	llama-3-8b-instruct	73 %
9	qwen2-72b-instruct	86 %
10	qwen-max	80 %

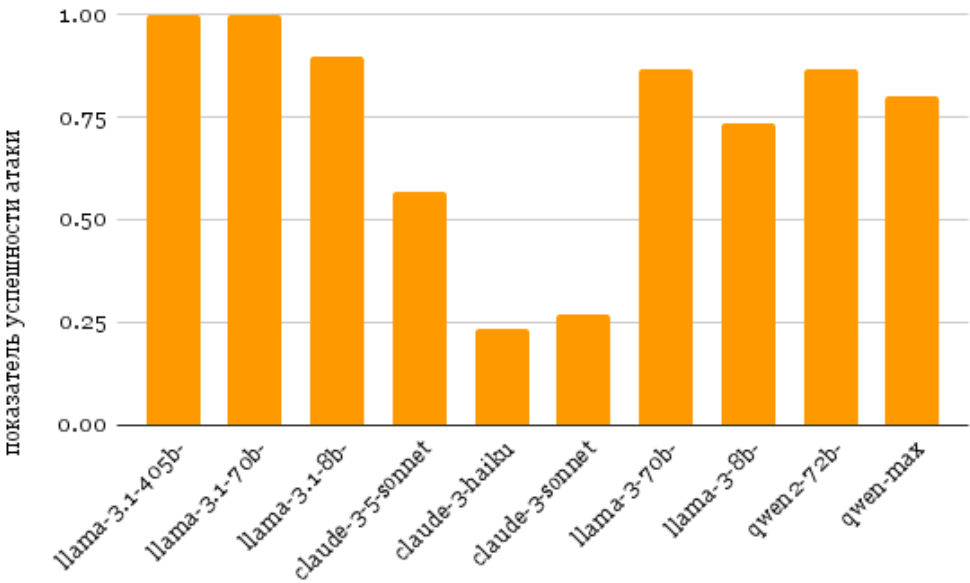


Рис. 3. Результаты сравнения моделей на вредоносном наборе данных HarmfulQA-paraphrase.
Fig. 3. Comparison results of LLMs on the red-teaming HarmfulQA-paraphrase dataset.

Список литературы / References

- [1]. Achiam J. et al. Gpt-4 technical report //arXiv preprint arXiv:2303.08774. – 2023.
- [2]. Roziere B. et al. Code llama: Open foundation models for code //arXiv preprint arXiv:2308.12950. – 2023.
- [3]. Qian J. et al. A Liver Cancer Question-Answering System Based on Next-Generation Intelligence and the Large Model Med-PaLM 2 //International Journal of Computer Science and Information Technology. – 2024. – Т. 2. – №. 1. – С. 28-35.
- [4]. Ebrahimi J. et al. Hotflip: White-box adversarial examples for text classification //arXiv preprint arXiv:1712.06751. – 2017.
- [5]. Zou A. et al. Universal and transferable adversarial attacks on aligned language models //arXiv preprint arXiv:2307.15043. – 2023.
- [6]. Jones E. et al. Automatically auditing large language models via discrete optimization //International Conference on Machine Learning. – PMLR, 2023. – С. 15307-15329.
- [7]. Alekseevskaia I., Arkhipenko K. OrderBkd: Textual backdoor attack through repositioning //2023 Ivannikov Ispras Open Conference (ISPRAS). – IEEE, 2023. – С. 1-6.
- [8]. Xu J. et al. Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models //arXiv preprint arXiv:2305.14710. – 2023.
- [9]. Li Y. et al. Badedit: Backdooring large language models by model editing //arXiv preprint arXiv:2403.13355. – 2024.
- [10]. Kshetri N. Cybercrime and privacy threats of large language models //IT Professional. – 2023. – Т. 25. – №. 3. – С. 9-13.
- [11]. Zamfirescu-Pereira J. D. et al. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts //Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. – 2023. – С. 1-21.
- [12]. Lyu H. et al. Llm-rec: Personalized recommendation via prompting large language models //arXiv preprint arXiv:2307.15780. – 2023.
- [13]. Azeem R. et al. LLM-Driven Robots Risk Enacting Discrimination, Violence, and Unlawful Actions //arXiv preprint arXiv:2406.08824. – 2024.
- [14]. Liu X. et al. Autodan: Generating stealthy jailbreak prompts on aligned large language models //arXiv preprint arXiv:2310.04451. – 2023.
- [15]. Harte J. et al. Leveraging large language models for sequential recommendation //Proceedings of the 17th ACM Conference on Recommender Systems. – 2023. – С. 1096-1102.
- [16]. Bai Y. et al. Constitutional ai: Harmlessness from ai feedback //arXiv preprint arXiv:2212.08073. – 2022.
- [17]. Bai Y. et al. Training a helpful and harmless assistant with reinforcement learning from human feedback //arXiv preprint arXiv:2204.05862. – 2022.
- [18]. Dai J. et al. Safe rlhf: Safe reinforcement learning from human feedback //arXiv preprint arXiv:2310.12773. – 2023.
- [19]. Rafailov R. et al. Direct preference optimization: Your language model is secretly a reward model //Advances in Neural Information Processing Systems. – 2024. – Т. 36.
- [20]. Wang C. et al. Beyond reverse kl: Generalizing direct preference optimization with diverse divergence constraints //arXiv preprint arXiv:2309.16240. – 2023.
- [21]. Azar M. G. et al. A general theoretical paradigm to understand learning from human preferences //International Conference on Artificial Intelligence and Statistics. – PMLR, 2024. – С. 4447-4455.
- [22]. Ethayarajh K. et al. Kto: Model alignment as prospect theoretic optimization //arXiv preprint arXiv:2402.01306. – 2024.
- [23]. Hejna J. et al. Contrastive preference learning: Learning from human feedback without rl //arXiv preprint arXiv:2310.13639. – 2023.
- [24]. Chao P. et al. Jailbreaking black box large language models in twenty queries //arXiv preprint arXiv:2310.08419. – 2023.
- [25]. Mehrotra A. et al. Tree of attacks: Jailbreaking black-box llms automatically //arXiv preprint arXiv:2312.02119. – 2023.
- [26]. Sitawarin C. et al. Pal: Proxy-guided black-box attack on large language models //arXiv preprint arXiv:2402.09674. – 2024.
- [27]. Hussein Abbass, Axel Bender, Svetoslav Gaidow, and Paul Whitbread. Computational red teaming: Past, present and future. *IEEE Computational Intelligence Magazine*, 6(1):30–42, 2011.

- [28]. Shen X. et al. "do anything now": Characterizing and evaluating in-the-wild jailbreak prompts on large language models //Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security. – 2024. – С. 1671-1685.
- [29]. Ganguli D. et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned //arXiv preprint arXiv:2209.07858. – 2022.
- [30]. Radharapu B. et al. Aart: Ai-assisted red-teaming with diverse data generation for new llm-powered applications //arXiv preprint arXiv:2311.08592. – 2023.
- [31]. Ji J. et al. Beavertails: Towards improved safety alignment of llm via a human-preference dataset //Advances in Neural Information Processing Systems. – 2024. – Т. 36.
- [32]. Bhardwaj R., Poria S. Red-teaming large language models using chain of utterances for safety-alignment //arXiv preprint arXiv:2308.09662. – 2023.
- [33]. Schulman J. et al. Proximal policy optimization algorithms //arXiv preprint arXiv:1707.06347. – 2017.
- [34]. Mazeika M. et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal //arXiv preprint arXiv:2402.04249. – 2024.

Информация об авторах / Information about authors

Ирина Сергеевна АЛЕКСЕЕВСКАЯ – программист Центра доверенного искусственного интеллекта, аспирант ИСП РАН по направлению искусственный интеллект и машинное обучение. Сфера научных интересов: большие языковые модели, состязательные атаки, бэкдор атаки, выравнивание больших языковых моделей.

Irina Sergeevna ALEKSEEVSKAIA – Programmer at the Trusted Artificial Intelligence Research Center, postgraduate student at the ISP RAS in the field of artificial intelligence and machine learning. Research interests: large language models, adversarial attacks, backdoor attacks, alignment of large language models.

Константин Владимирович АРХИПЕНКО – младший научный сотрудник Центра доверенного искусственного интеллекта, также является специалистом кафедры Системного программирования Московского государственного университета имени М.В. Ломоносова. Научные интересы: защита моделей машинного обучения от состязательных атак, объяснимый искусственный интеллект.

Konstantin Vladimirovich ARKHIPENKO – Junior Research Fellow at the Trusted Artificial Intelligence Research Center, and a specialist at the Department of System Programming at Lomonosov Moscow State University. Research interests: defenses of machine learning models from adversarial attacks, explainable artificial intelligence.

Денис Юрьевич ТУРДАКОВ – кандидат физико-математических наук, заведующий отделом информационных систем Института системного программирования с 2017 года. Сфера научных интересов: анализ естественного языка, облачные вычисления, машинное обучение, анализ социальных сетей.

Denis Yuryevich TURDAKOV – Cand. Sci. (Phys.-Math.), Head of the Information Systems Department at the Institute of System Programming since 2017. Research interests: natural language analysis, cloud computing, machine learning, social network analysis.

