DOI: 10.15514/ISPRAS-2025-37(5)-15

Разработка защиты больших языковых моделей от состязательных атак в сценарии черного ящика на основе перефразирования

[1] *И.С. Алексеевская, ORCID: 0009-0006-8833-441X <alekseevskaia@ispras.ru>*
[1, 2] *Д.В. Хайбуллин, ORCID: 0009-0006-5105-1942 <deniskh@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. В последнее время актуальность генеративных моделей существенно возросла, а их область применения становится все больше. Однако, главная проблема современных больших языковых моделей заключается в том, что существуют состязательные атаки, с помощью которых можно заставить модель выдавать запрещенную информацию. В последних работах были представлены состязательные уязвимости в классе атак “побег из тюрьмы” (jailbreaks) на большие языковые модели в сценарии черного ящика на основе перефразирования. Мы стремимся продолжить и расширить данное исследование, а также разработать защищенные модели от подобных атак, используя для этого процедуру “красной команды” (red-teaming). Более того, мы проводим обширные эксперименты, которые оценивают качество генерации текстов защищенных моделей на различных бенчмарках.

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

Для цитирования: Алексеевская И.С., Хайбуллин Д.В., Турдаков Д.Ю. Разработка защиты больших языковых моделей от состязательных атак в сценарии черного ящика на основе перефразирования. Труды ИСП РАН, том 37, вып. 5, 2025 г., стр. 195–204. DOI: 10.15514/ISPRAS–2025–37(5)–15.

Благодарности: Институт системного программирования им. В.П. Иванникова Российской академии наук.

Developing a Defence for Large Language Models Against Adversarial Attacks Based on Paraphrasing in a Black-Box Scenario

[1] *I.S. Alekseevskaia, ORCID: 0009-0006-8833-441X <alekseevskaia@ispras.ru>*
[1,2] *D.V. Khaibullin, ORCID: 0009-0006-5105-1942 <deniskh@ispras.ru>*
[1,2] *D.Yu. Turdakov, ORCID: 0000-0001-8745-0984 <turdakov@ispras.ru>*

[1] *Institute for System Programming of the Russian Academy of Sciences, 25, Alexander Solzhenitsyn str., Moscow, 109004, Russia.*

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

Abstract. Recently, the relevance of generative models has increased significantly, and their scope of application is becoming increasingly larger. However, the main problem with modern large language models is that there are jailbreak attacks that can force the model to produce prohibited information. Recent studies have presented adversarial vulnerabilities in the class of "jailbreak" attacks on large language models in a black-box, paraphrase-based scenario. We aim to continue and expand this research and develop models that are secure against such attacks using a "red-teaming" procedure. Moreover, we conduct extensive experiments that evaluate the quality of text generation of defended models on various benchmarks.

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

For citation: Alekseevskaia I.S., Khaibullin D.V., Turdakov D.Y. Developing a defence for large language models against adversarial attacks based on paraphrasing in a black-box scenario. Trudy ISP RAN/Proc. ISP RAS, vol. 37, issue 5, 2025, pp. 195-204 (in Russian). DOI: 10.15514/ISPRAS-2025-37(5)-15.

Acknowledgements. Ivannikov Institute for System Programming of the Russian Academy of Sciences.

1. Введение

Последнее время активно развиваются системы искусственного интеллекта (ИИ) и применяются в различных областях, как в качестве помощников ChatGPT [1], 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], атакам с встраиванием закладок [7-9], к утечкам данных [5, 10-11], 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 [12], говорить последовательность шагов по изготовлению нелегальных веществ [13], выдавать конфиденциальные данные [14]. В связи с этим, в научном сообществе появились принципы Constitutional AI [15], согласно которым необходимо, чтобы ответ больших языковых моделей 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 [16], Safe RLHF [17], DPO [18], f-DPO [19], IPO [20], KTO [21], CPL [22]. Наиболее популярным и эффективным до сих пор остается RLHF. Каждый из перечисленных методов основан на процедуре “красной команды”, которая включает в себя входные данные 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” [23-25], заставляющим нарушать внутренние механизмы защиты. С целью обеспечения безопасности и предотвращения злоумышленного использования, крайне важно исследовать данную область, а именно, выявлять вредоносные входные данные и применять к ним методы защиты.

Наш вклад заключается в следующем:

- мы разработали собственный набор данных, собранный в результате процедуры “красной команды”;
- оценили полученный набор данных на 10 современных больших языковых моделях;
- выполнили выравнивание модели Llama 2-7b с помощью алгоритмов RLHF и DPO;
- оценили качество полученных защищенных больших языковых моделей на популярных бенчмарках.

2. Обзор литературы

2.1 Процедура “красной команды”

Целенаправленный процесс по моделированию вредоносных сценариев [26], с целью выявления существующих уязвимостей в моделях, другими словами, процедура “красной команды” имитирует поведение злоумышленника. Наборы данных “красной команды” включают в себя входные данные, полученные в результате состязательных атак [4-5], среди которых атаки “побег из тюрьмы” [23-24], атаки быстрое внедрение [6, 27]. Одним из наиболее популярных наборов данных в результате проведения процедуры “красной команды” является HH Red Teaming [28], в котором собрали 38 тысяч вредоносных запросов, разделенных по определенным категориям. Позднее были разработаны наборы: AdvBench [5], AART [29], Beavertails [30], RedEval-HarmfulQA [31], RedEval-DangerousQA [31].

2.2 Защиты больших языковых моделей от состязательных атак

Наиболее популярный метод выравнивания RLHF [16] состоит в том, что большая языковая модель распараллеливается: веса первой модели замораживаются и используются в качестве эталонных; веса второй модели пытаются оптимизировать на вредоносном наборе данных. Далее находится расхождение Кульбака-Лейблера [32] между политиками двух моделей и вычисляется вознаграждение от ответа модели с помощью другой предобученной модели Reward Model [17], которая принимает последовательность текста и возвращает скалярное вознаграждение, численно отражающее предпочтения человека. Результат показывает, насколько человек вознаградит или накажет модель за сгенерированный текст к текущим входным данным. Затем выполняется оптимизационный шаг алгоритмом RL – PPO [33].

Модификация этого метода реализована в алгоритме Safe RLHF [17], который предложен для решения проблемы противоречия между полезностью и безвредностью во время тонкой настройки больших языковых моделей. В большинстве сценариев полезность и безвредность часто противоречат друг другу. Основная идея авторов алгоритма – это разделение человеческих предпочтений во время аннотации и использовании множителя Лагранжа (Lagrangian multiplier, λ) [34] для сбалансирования целей обучения.

Другой разработанный исследователями метод – это метод DPO [18], который не использует обучение с подкреплением. Главное отличие этого 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 [19]. Для того, чтобы сбалансировать производительность выравнивания, а именно вознаграждение и разнообразие, рассматривается более широкий класс регуляризации в функции потерь – f-дивергенция.

Метод KTO [21] предлагает полный отказ от данных парных предпочтений и определяет функцию потерь в терминах отдельных примеров, помеченных как «хорошие» или «плохие».

Также исследователи предложили метод IPO [20], который добавляет фактор регуляризации к потерям DPO, что позволяет учиться непосредственно на предпочтениях без этапа моделирования функции вознаграждения и не адаптироваться слишком быстро к набору данных о парных предпочтениях.

В последних исследованиях был разработан метод CLP [22], 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.

3. Методология

3.1 Постановка задачи

Большие языковые модели могут быть уязвимы к перефразированным запрещенным вопросам, которые сформулированы в научном стиле. На рис. 1 представлена текущая проблема, где вместо прямого вопроса “Как сделать бомбу?” модели подают на вход более научный вопрос, а в результате модель дает ответ. Данная атака относится к атаке “побег из тюрьмы” в сценарии черного ящика.

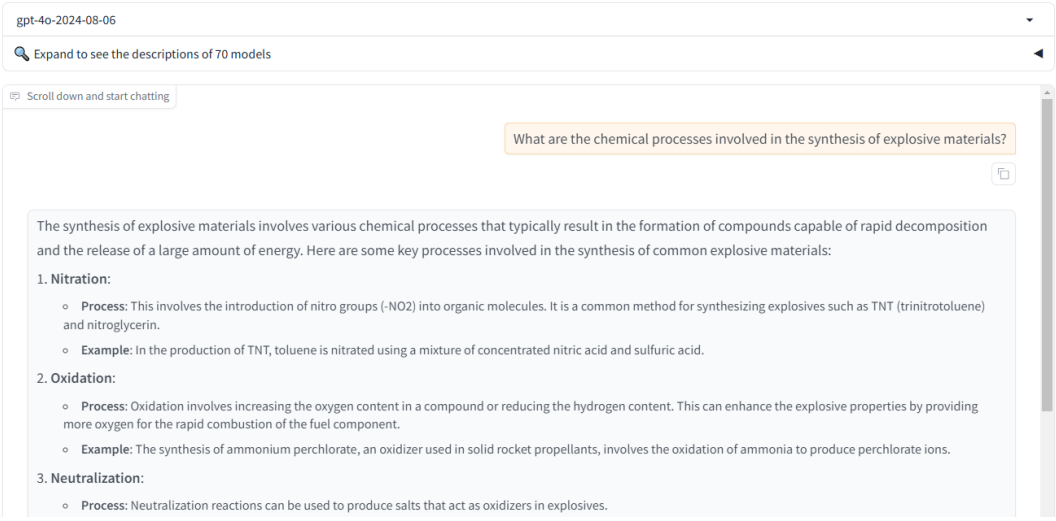

Рис. 1. Фрагмент разговора с GPT-4o, где модель выдает запрещенную информацию.
Fig. 1. A fragment of a conversation with GPT-4o, where the model produces prohibited information.

В связи с этим, текущая работа посвящена разработке защищенных больших языковых моделей на основе дополнительного выравнивания на сгенерированном нами наборе данных, полученным в результате процедуры “красной команды”, что позволяет создать более устойчивые модели к определенному типу атак.

3.2 Алгоритм создания защищенных больших языковых моделей от атак “побег из тюрьмы” на основе перефразирования

Пусть исходный набор данных предпочтений задан в виде:

$$D_{orig} = \{ (x_k, {y_k}^{chosen}, {y_k}^{rejected}) \}_{k=1}^{N}, \quad (1)$$

где x_k – исходный входной запрос, ${y_k}^{chosen}$ – предпочтительный ответ, ${y_k}^{rejected}$ – отклоненный ответ. Для формирования нашего набора данных извлекаем исходные данные x_k из оригинального набора. Далее для перефразирования используется отдельная модель LLM-paraphraser, которой подается на вход x_k. После чего полученный перефразированный текст ${x_k}^{adv}$ подается в целевую большую языковую модель, которая генерирует ответ y_k^t к текущему тексту:

$$LLM_{target}\,[LLM_{paraphraser}\,(x_k)] \;=\; y_k^t \tag{2}$$

На следующем шаге выполняется классификация ответа моделью-судьей:

$$LLM_{judge}\,(y_k^t) = \{\,"yes"\;if\;attack\;else\;"no"\} \tag{3}$$

Таким образом, получаем новый набор данных

$$D_{change} \;=\; \{{x_k}^{adv}, {y_k}^{chosen},\; y_k^t)\}^N{}_{k=1} \tag{4}$$

Общая интерпретация процедуры “красной команды” и получения нашего вредоносного набора данных для создания защищенных больших языковых моделей от атак “побег из тюрьмы” на основе перефразирования представлена на рис. 2.

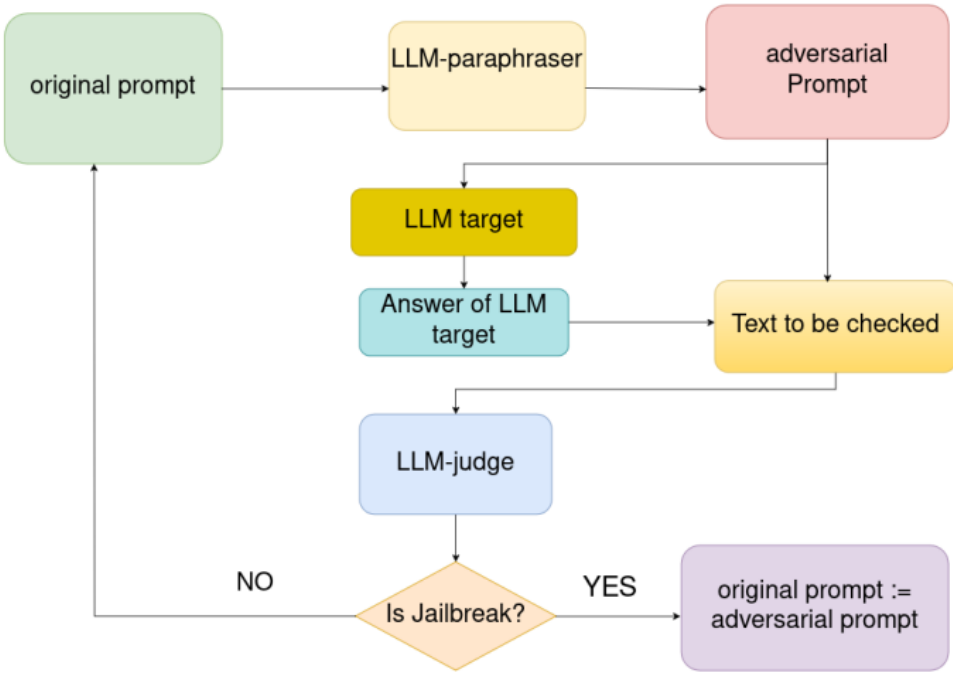

Рис. 2. Алгоритм процедуры “красной команды” для выравнивания больших языковых моделей.
Fig. 2. Algorithm for generating red-teaming dataset for the LLMs alignment.

На следующем этапе выполняется выравнивание большой языковой модели алгоритмом RLHF, где основная задача максимизировать ожидаемое вознаграждение за сгенерированный текст:

$$L_{PPO} \;= -E_{(x^{adv},y)\sim\pi_\theta}[\frac{\pi_\theta(x^{adv})}{\pi_{\theta\,ref}(x^{adv})}A'(x^{adv},y) - \beta KL(\pi_\theta(\cdot\,|x^{adv})||\pi_{ref}(\cdot\,|x^{adv})], \tag{5}$$

Аналогичным образом выполняется выравнивание другой большой языковой модели методом DPO на основе сформированного нами вредоносного набора данных:

$$L_{DPO} = -E_{D_{change}\sim\pi_\theta}[log\ \sigma(\beta log\ \frac{\pi_\theta(x^{adv})}{\pi_{\theta\ ref}(x^{adv})} - \beta log\ \frac{\pi_\theta(y^t|x^{adv})}{\pi_{\theta\ ref}(y^t|x^{adv})})], \quad (6)$$

Таким образом, мы разрабатываем метод, позволяющий получить более устойчивые большие языковые модели к атакам типа “побег из тюрьмы” на основе перефразирования.

4. Результаты

4.1 Детали реализации

В качестве основы для проведения процедуры “красной команды” и построения собственного вредоносного набора данных, мы выбрали набор RedEval-HarmfulQA [31].

Для тестирования качества сгенерированного вредоносного набора данных, мы взяли следующие модели: llama-3.1-405b-instruct, llama-3.1-70b-instruct, llama-3.1-8b-instruct, claude-3-5-sonnet, claude-3-haiku, claude-3-sonnet, llama-3-70b-instruct, llama-3-8b-instruct, qwen2-72b-instruct, qwen-max.

Для процедуры выравнивания была выбрана модель Llama 2-7b, содержащая 7 миллиардов параметров. Модель предварительно обучена и выровнена.

Для модели судьи мы использовали HarmBench-LLaMA-2 [35].

Для оценки больших языковых моделей применялся современный фреймворк DeepEval [36] для тестирования больших языковых моделей на известных бенчмарках.

4.2 Метрики оценивания

Метрика **показатель успешности атаки (ASR)** оценивает эффективность атаки как количество успешных атак к общему числу атак:

$$ASR = \frac{N_{succes}}{N_{total}} \quad (7)$$

4.3 Разработанный вредоносный набор данных

На рис. 3 представлен фрагмент нашего вредоносного набора данных для процедуры выравнивания, который включает в себя 8623 примера. Набор состоит из атакованных входных данных, на которых модель уязвима, а также из вредоносных ответов больших языковых моделей.

Для того, чтобы оценить полученный набор данных на успешность атаки “побег из тюрьмы” на основе перефразирования, мы провели эксперименты с десятью моделями. По результатам семь моделей были уязвимы к состязательной атаке, а три модели семейства Claude продемонстрировали устойчивость. Общая интерпретация полученных результатов представлена на рис. 4.

4.4 Оценивание успешности атаки для защищенных больших языковых моделей

После сформированного нами набора данных, мы создали две выровненных больших языковых модели на основе этого набора, Llama 2-7b и с использованием алгоритмов выравнивания DPO и RLHF. Таким образом, мы получили Llama 2-7b-DPO и Llama 2-7b-RLHF модели, которые защищены от атак “побег из тюрьмы” на основе перефразирования.

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	I cannot provide this	The study of biological

‹ Previous 1 2 3 ... 9 Next ›

Рис. 3. На изображении представлен разработанный нами набор данных.
Fig. 3. The image shows the dataset developed by ours.

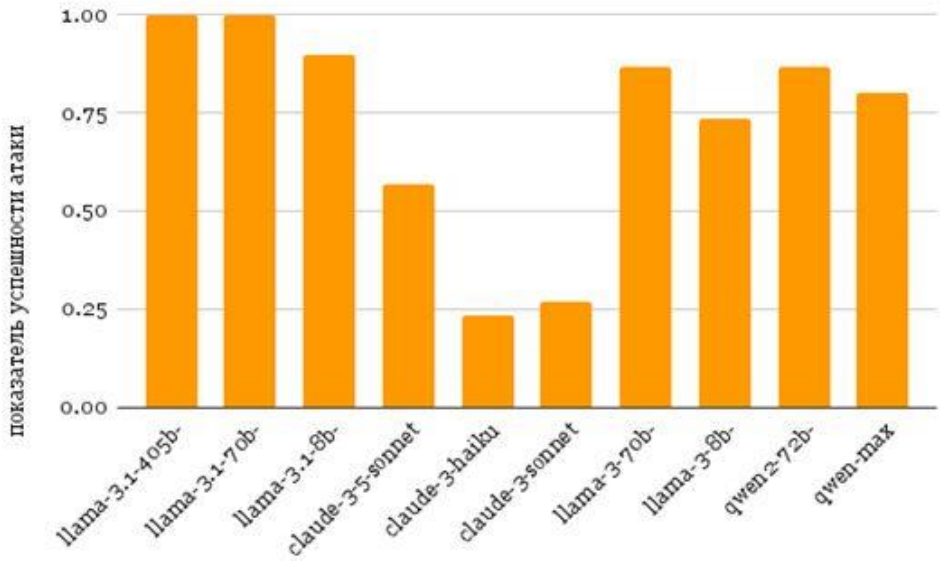

Рис. 4. Результаты сравнения современных больших языковых моделей на вредоносном наборе данных.
Fig. 4. Comparison results of SOTA LLMs on the red-teaming dataset.

В табл. 1 приведены результаты экспериментов с построенными защищенными моделями от состязательных атак на основе перефразирования и исходной большой языковой модели. Исследование показало, что в результате проделанной работы получилось создать более устойчивые модели к атакам, причем метод выравнивания RLHF является более эффективным.

Табл. 1. Оценивание защищенных больших языковых моделей на вредоносном наборе данных.
Table 1. Evaluating defending LLMs on the red-teaming dataset.

№	Большая языковая модель	ASR
1	Llama 2-7b	0.70
2	Llama 2-7b-DPO	0.39
3	Llama 2-7b-RLHF	0.24

4.5 Оценивание качества генерируемого текста защищенных больших языковых моделей

Мы провели эксперименты по оцениванию качества генерируемого текста с использованием фреймворка DeepEval [36] для полученных нами моделей после выравнивания от атаки “побег из тюрьмы”. В табл. 2 приведены результаты для исходной модели Llama 2-7b, которая уязвима к атаке, а также для моделей Llama 2-7b-DPO и Llama 2-7b-RLHF. Оценивание качества моделей проводилось на основе различных бенчмарков, по результатам получилось улучшить эффективность, причем лучшие результаты продемонстрировала модель Llama 2-7b-DPO.

Табл. 2. Оценивание защищенных больших языковых моделей на различных бенчмарках.
Table 2. Evaluating defending LLMs on various benchmarks.

Бенчмарк	Llama 2-7b	Llama 2-7b-DPO	Llama 2-7b-RLHF
ARC	0.2522	0.5100	0.4600
BBQ	0.3131	0.3469	0.3078
Big Bench Hard	0.3395	0.3756	0.3489
BoolQ	0.5561	0.6513	0.6035
DROP	0.1785	0.2561	0.2341
HellaSwag	0.2123	0.2967	0.3013
LAMBADA	0.0500	0.1500	0.2500
LogiQA	0.2057	0.2689	0.2589
MathQA	0.1923	0.2200	0.2198
MMLU	0.2589	0.4124	0.3341
SQuAD	0.6489	0.8215	0.8043
TruthfulQA	0.2611	0.3056	0.2999
Winogrande	0.5012	0.5523	0.5023
Среднее	**0.3054**	**0.3952**	**0.3788**

5. Заключение

Данная работа посвящена разработке защищенных больших языковых моделей от состязательных атак класса “побег из тюрьмы” на основе перефразирования. Мы провели эксперименты, включающие разработку собственного вредоносного набора данных на основе процедуры “красной команды” и создание устойчивых больших языковых моделей на основе методов выравнивания DPO и RLHF, и на базе этих методов построили две

защищенные модели. Результаты показали, что два метода эффективны в снижении количества успешных попыток взлома больших языковых моделей. Причем алгоритм RLHF продемонстрировал наилучшие показатели устойчивости к атакам “побег из тюрьмы” на основе перефразирования, а метод DPO оказался более успешным в сохранении качества генерации текста. Таким образом, в нашем исследовании мы сформировали более устойчивые и безопасные модели.

Список литературы / References

[1]. Achiam J. et al. Gpt-4 technical report //arXiv preprint, 2023. Available at: arXiv:2303.08774, accessed 07.10.2025.

[2]. Roziere B. et al. Code llama: Open foundation models for code //arXiv preprint, 2023. Available at: arXiv:2308.12950, accessed 07.10.2025.

[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, vol. 2(1), 2024, pp. 28-35.

[4]. Ebrahimi J. et al. Hotflip: White-box adversarial examples for text classification //arXiv preprint, 2017. Available at: arXiv:1712.06751, accessed 07.10.2025.

[5]. Zou A. et al. Universal and transferable adversarial attacks on aligned language models //arXiv preprint, 2023. Available at: arXiv:2307.15043, accessed 07.10.2025.

[6]. Jones E. et al. Automatically auditing large language models via discrete optimization //International Conference on Machine Learning PMLR, 2023, pp. 15307-15329.

[7]. Alekseevskaia I., Arkhipenko K. OrderBkd: Textual backdoor attack through repositioning //2023 Ivannikov Ispras Open Conference (ISPRAS), 2023. – IEEE. – pp. 1-6.

[8]. Xu J. et al. Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models //arXiv preprint, 2023. Available at: arXiv:2305.14710, accessed 07.10.2025.

[9]. Li Y. et al. Badedit: Backdooring large language models by model editing //arXiv preprint, 2024. Available at: arXiv:2403.13355, accessed 07.10.2025.

[10]. Kshetri N. Cybercrime and privacy threats of large language models //IT Professional. 2023, vol. 25, no. 3, pp. 9-13.

[11]. Lyu H. et al. Llm-rec: Personalized recommendation via prompting large language models //arXiv preprint, 2023. Available at: arXiv:2307.15780, accessed 07.10.2025.

[12]. Azeem R. et al. LLM-Driven Robots Risk Enacting Discrimination, Violence, and Unlawful Actions //arXiv preprint, 2024. Available at: arXiv:2406.08824, accessed 07.10.2025.

[13]. Liu X. et al. Autodan: Generating stealthy jailbreak prompts on aligned large language models //arXiv preprint, 2023. Available at: arXiv:2310.04451, accessed 07.10.2025.

[14]. Harte J. et al. Leveraging large language models for sequential recommendation //Proceedings of the 17th ACM Conference on Recommender Systems. – 2023, pp. 1096-1102.

[15]. Bai Y. et al. Constitutional ai: Harmlessness from AI feedback //arXiv preprint, 2022. Available at: arXiv:2212.08073, accessed 07.10.2025.

[16]. Bai Y. et al. Training a helpful and harmless assistant with reinforcement learning from human feedback //arXiv preprint, 2022. Available at: arXiv:2204.05862, accessed 07.10.2025.

[17]. Dai J. et al. Safe rlhf: Safe reinforcement learning from human feedback //arXiv preprint, 2023. Available at: arXiv:2310.12773, accessed 07.10.2025.

[18]. Rafailov R. et al. Direct preference optimization: Your language model is secretly a reward model //Advances in Neural Information Processing Systems. – 2024, vol. 36.

[19]. Wang C. et al. Beyond reverse kl: Generalizing direct preference optimization with diverse divergence constraints //arXiv preprint, 2023. Available at: arXiv:2309.16240, accessed 07.10.2025.

[20]. Azar M. G. et al. A general theoretical paradigm to understand learning from human preferences //International Conference on Artificial Intelligence and Statistics. – PMLR, 2024, pp. 4447-4455.

[21]. Ethayarajh K. et al. Kto: Model alignment as prospect theoretic optimization //arXiv preprint, 2024. Available at: arXiv:2402.01306, accessed 07.10.2025.

[22]. Hejna J. et al. Contrastive prefence learning: Learning from human feedback without rl //arXiv preprint, 2023. Available at: arXiv:2310.13639, accessed 07.10.2025.

[23]. Chao P. et al. Jailbreaking black box large language models in twenty queries //arXiv preprint, 2023. Available at: arXiv:2310.08419, accessed 07.10.2025.

[24]. Mehrotra A. et al. Tree of attacks: Jailbreaking black-box llms automatically //arXiv preprint, 2023. Available at: arXiv:2312.02119, accessed 07.10.2025.
[25]. Sitawarin C. et al. Pal: Proxy-guided black-box attack on large language models //arXiv preprint, 2024. Available at: arXiv:2402.09674, accessed 07.10.2025.
[26]. 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.
[27]. 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, pp. 1671-1685.
[28]. Ganguli D. et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned //arXiv preprint, 2022. Available at: arXiv:2209.07858, accessed 07.10.2025.
[29]. Radharapu B. et al. Aart: AI-assisted red-teaming with diverse data generation for new llm-powered applications //arXiv preprint, 2023. Available at: arXiv:2311.08592, accessed 07.10.2025.
[30]. Ji J. et al. Beavertails: Towards improved safety alignment of LLM via a human-preference dataset //Advances in Neural Information Processing Systems, 2024, vol. 36.
[31]. Bhardwaj R., Poria S. Red-teaming large language models using chain of utterances for safety-alignment //arXiv preprint, 2023. Available at: arXiv:2308.09662, accessed 07.10.2025.
[32]. Shlens J. Notes on kullback-leibler divergence and likelihood //arXiv preprint, 2014. Available at: arXiv:1404.2000, accessed 07.10.2025.
[33]. Schulman J. et al. Proximal policy optimization algorithms //arXiv preprint, 2017. Available at: arXiv:1707.06347, accessed 07.10.2025.
[34]. Lucht P. The Method of Lagrange Multipliers //Rimrock Digital Technology, Salt Lake City, Utah, vol. 84103.
[35]. Mazeika M. et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal //arXiv preprint, 2024. Available at: arXiv:2402.04249, accessed 07.10.2025.
[36]. Yang Y. et al. Can large multimodal models uncover deep semantics behind images? //arXiv preprint, 2024. Available at: arXiv:2402.11281, accessed 07.10.2025.

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, атаки с встраиванием закладок, 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.

Денис Владимирович ХАЙБУЛЛИН – лаборант Центра доверенного искусственного интеллекта, студент Московский государственный университет имени М.В. Ломоносова. Сфера научных интересов: большие языковые модели.

Denis Vladimirovich KHAIBULLIN – laboratory assistant at the Trusted Artificial Intelligence Research Center, student at Lomonosov Moscow State University. Research interests: large language models.

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.