

На правах рукописи

Алексеевская Ирина Сергеевна

**Методы и программные средства исследования и повышения
устойчивости языковых моделей к целенаправленным
искажениям**

Специальность 2.3.5 —

"Математическое и программное обеспечение вычислительных систем,
комплексов и компьютерных сетей"

Автореферат

диссертация на соискание ученой степени
кандидата физико-математических наук

Москва — 2026

Работа выполнена в Федеральном государственном бюджетном учреждении науки Институте системного программирования им. В.П. Иванникова Российской академии наук (ИСП РАН).

Научный руководитель: **Турдаков Денис Юрьевич**, кандидат физико-математических наук

Официальные оппоненты: **Котельников Евгений Вячеславович**, доктор технических наук, доцент, профессор Школы вычислительных социальных наук Автономной некоммерческой образовательной организации высшего образования «Европейский университет в Санкт-Петербурге»

Рогов Олег Юрьевич, кандидат физико-математических наук, руководитель научной группы «Доверенные и безопасные интеллектуальные системы» Автономной некоммерческой организации «Институт искусственного интеллекта»

Ведущая организация: Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский ядерный университет «МИФИ»

Защита состоится 17 декабря 2026 г. в 13-15 на заседании диссертационного совета 24.1.120.01 при Федеральном государственном бюджетном учреждении науки Институте системного программирования им. В.П. Иванникова Российской академии наук (ИСП РАН) по адресу: 115035, г. Москва, ул. Садовническая, д. 41, ст. 2.

С диссертацией можно ознакомиться в библиотеке и на сайте Федерального государственного бюджетного учреждения науки Института системного программирования им. В.П. Иванникова Российской академии наук.

Автореферат разослан «___» _____ 2026 г.

Ученый секретарь
диссертационного совета 24.1.120.01
кандидат физико-математических наук

Турдаков Д.Ю.

Общая характеристика диссертации

Актуальность темы. Языковые модели широко применяются в промышленности, финансах, медицине, образовании и государственном управлении. Однако развертывание языковых моделей в реальных прикладных системах сопряжено с рисками информационной безопасности. В исследованиях выявлены уязвимости в языковых моделях, эксплуатация которых может приводить к несанкционированному доступу к конфиденциальной информации, сгенерировать вредоносный контент. Эти уязвимости могут нанести ущерб репутации разработчиков и клиентов таких систем, поэтому для обеспечения безопасного использования языковых моделей необходимо создавать технологии доверенного искусственного интеллекта – технологии, отвечающие стандартам безопасности, разработанные с учетом принципов объективности, недискриминации, этичности, исключающие при их использовании возможность причинения вреда человеку и нарушения его основополагающих прав и свобод, нанесения ущерба интересам общества и государства. В работе под целенаправленными искажениями (атаками) понимаются воздействия, направленные на нарушение корректной работы языковой модели, в частности, 1) формирование заведомо неверного ответа; 2) генерацию ответа, нарушающего вышеуказанные принципы. Под устойчивостью языковых моделей понимается способность сохранять корректность функционирования при воздействии целенаправленных искажений.

Один из ключевых аспектов обеспечения доверенного ИИ – защита от целенаправленных искажений, проводимых на этапе обучения модели. Потенциальную угрозу представляют целенаправленные искажения, направленные на внедрение в модель бэкдора (закладки). Бэкдор – это свойство обученной модели ИИ, которое заключается в том, что модель корректно работает с обычными (чистыми) данными, но заведомо неправильно работает с данными, включающими триггер. Триггер – это редкий входной признак, разработанный злоумышленником. Внедрение бэкдора может осуществляться с помощью внесения триггера в малую часть обучающих примеров, то есть отравления датасета. Особую опасность такие целенаправленные искажения представляют в условиях широкого использования предобученных моделей, полученных из открытых или недостаточно контролируемых источников. Внедрённый на этапе предварительного обучения бэкдор может сохраняться при последующей тонкой настройке модели

на специализированных данных. Поэтому отсутствие аномалий в работе модели после её адаптации к конкретной прикладной задаче само по себе не гарантирует отсутствия скрытого вредоносного поведения.

Другой ключевой аспект обеспечения доверенного ИИ – защита от целенаправленных искажений, проводимых на этапе эксплуатации (инференса) модели. Наибольшую опасность представляют состязательные атаки — методы преднамеренного воздействия на системы машинного обучения, при котором входные данные (промпт) специально модифицируются таким образом, чтобы вызвать ошибочное поведение модели, оставаясь при этом малозаметными. Атакованным промптом называется входной текст с небольшим возмущением (шумом). В частности, в работе рассматриваются джейлбрейки, направленные на обход выравнивания модели. Выравнивание – это свойство модели выводить честные, полезные, безвредные ответы для любых входных данных. Выравнивание достигается, в частности, посредством дополнительного обучения модели на специально подготовленных данных – тестах, имитирующих действия потенциального нарушителя. В этих тестах содержатся запросы, провоцирующие модель на генерацию нежелательной информации, а также предпочтительные (безопасные) ответы на эти запросы. Нежелательная информация – это информация, нарушающая этические принципы или политику использования языковых моделей, например, подробная инструкция по изготовлению нелегальных веществ или по взлому сайта, а также конфиденциальная информация.

Целью данной работы является развитие инструментов и технологий построения доверенных языковых моделей, учитывающее новые виды целенаправленных искажений с повышенной скрытностью.

Для достижения поставленной цели необходимо решить следующие **задачи**:

1. Разработать метод внедрения бэкдора в языковые модели для задачи классификации текстов.
2. Разработать метод джейлбрейка языковых моделей для задачи генерации текстовой последовательности.
3. Разработать методику защиты языковых моделей от джейлбрейков и внедрения бэкдоров, эффективную для выявленных новых видов угроз.

Основные положения, выносимые на защиту:

1. Новый метод внедрения бэкдора для задачи классификации текстов в языковых моделях в сценарии "белого ящика" путем изменения порядка слов.
2. Новый метод джейлбрейка для задачи генерации текстовой последовательности в сценарии "чёрного ящика" на основе перефразирования.
3. Методика защиты языковых моделей от джейлбрейков и внедрения бэкдоров, эффективная для выявленных новых видов угроз.

Соответствие паспорту специальности. Перечисленные положения относятся к следующим направлениям исследований паспорта специальности 2.3.5 "Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей":

- п.4 Интеллектуальные системы машинного обучения, управления базами данных и знаний, инструментальные средства разработки цифровых продуктов
- п.7 Модели, методы, архитектуры, алгоритмы, форматы, протоколы и программные средства человеко-машинных интерфейсов, компьютерной графики, визуализации, обработки изображений и видеоданных, систем виртуальной реальности, многомодального взаимодействия в социоконвергентных системах.
- п.8 Модели и методы создания программ и программных систем для параллельной и распределенной обработки данных, языки и инструментальные средства параллельного программирования.
- п.9 Модели, методы, алгоритмы, облачные технологии и программная инфраструктура организации глобально распределенной обработки данных.

Научная новизна. Разработаны новые методы: метод внедрения бэкдора, превосходящий существующие методы по метрикам скрытности, и метод джейлбрейка на основе перефразирования, обладающий повышенной скрытностью за счет отсутствия неестественных последовательностей слов в атакованных промптах. Проведены тесты, имитирующие действия потенциального нарушителя, на основе разработанного метода джейлбрейка и разработана методика защиты.

Теоретическая значимость. В диссертационной работе расширены научные представления о механизмах возникновения и проявления уязвимостей языковых моделей к целенаправленным искажениям входных данных. Показа-

на возможность нарушения корректного функционирования языковых моделей не только путем добавления специальных триггерных последовательностей, но и посредством преобразований входного текста, сохраняющих его исходный смысл и естественность. Разработанные методы OrderBkd и SoftPrompt демонстрируют возможность реализации таких воздействий на этапах обучения и эксплуатации языковых моделей соответственно.

Выявлены ограничения существующих методов обнаружения целенаправленных искажений, основанных на анализе перплексии и семантического сходства текстов. Показано, что сохранение указанных характеристик не гарантирует устойчивости языковых моделей к атакам. Для метода внедрения бэкдора OrderBkd сформулированы условия, при которых целенаправленное изменение порядка слов не обнаруживается методом защиты ONION, использующим оценку перплексии входного текста. Для разработанного метода защиты OrderDef получена асимптотическая оценка временной сложности, определяющая вычислительные затраты на обнаружение и устранение целенаправленных искажений, основанных на перестановках слов.

Полученные результаты формируют основу для дальнейших исследований устойчивости языковых моделей к малозаметным целенаправленным искажениям, а также разработки методов их обнаружения и нейтрализации с учетом семантических характеристик входных данных и вычислительной сложности защитных алгоритмов.

Практическая значимость. Поиск новых уязвимостей в моделях ИИ важен для безопасного жизненного цикла систем искусственного интеллекта как превентивная мера, поскольку эти же уязвимости могут быть обнаружены и использованы злоумышленниками без публичного раскрытия информации об уязвимостях. Результаты исследования, таким образом, могут быть включены в инструменты обеспечения безопасного жизненного цикла искусственного интеллекта:

1. Реализация разработанного метода внедрения бэкдора OrderBkd.
2. Реализация разработанного метода джейлбрейка SoftPrompt.
3. Сформированный новый датасет на основе тестов, имитирующих действия потенциального нарушителя, которые получены в ходе джейлбрейка SoftPrompt.

Апробация работы. Результаты работы докладывались на следующих научных конференциях и семинарах:

1. Конференция «AINL: Artificial Intelligence and Natural Language Conference» (г. Ереван, Армения, апрель 2023).
2. Открытая конференция ИСП РАН (г. Москва, Россия, декабрь 2023).
3. Семинар по системному программированию под руководством академика А.И. Аветисяна (г. Москва, Россия, июнь 2024).
4. Открытая конференция ИСП РАН (г. Москва, Россия, декабрь 2024).
5. Открытая конференция ИСП РАН (г. Москва, Россия, декабрь 2025).

Личный вклад. Все выносимые на защиту результаты были получены лично автором и прошли экспериментальную проверку.

Публикации. Основные результаты по теме диссертации изложены в трех работах, две из которых изданы в журналах, рекомендованных ВАК, одна – в журнале, входящем в международные базы цитирования Scopus и Web of Science.

В статьях [1], [2] постановка задачи осуществлена совместно с соавторами; автору принадлежит основная часть работы, а именно: реализация алгоритма, выполнение основного объема теоретических и экспериментальных исследований, а также редактирование текста.

В статье [3] автору принадлежит постановка задачи, разработка подхода, анализ экспериментов и редактирование текста.

По теме диссертации имеется 1 свидетельство о государственной регистрации программы для ЭВМ [4].

Объем и структура работы. Диссертационная работа состоит из введения, пяти глав, заключения, списка литературы, содержащего 84 ссылки и 4 приложения. Работа изложена на 127 страницах, содержит 24 рисунка, 4 листинга и 32 таблицы.

Внедрение результатов. Результаты диссертационного исследования интегрированы в облачную платформу для разработки доверенных систем (Платформа ДИИ), использующих технологии ИИ, созданную при поддержке гранта для исследовательских центров в области искусственного интеллекта, предоставленного Аналитическим центром в рамках соглашения о предоставлении субсидии (идентификатор соглашения 000000D730321P5Q0002) и соглашения с Институтом системного программирования им. В. П. Иванникова от 2 ноября 2021 г. № 70-2021-00142.

Основное содержание работы

Во **введении** обосновывается актуальность исследований, проводимых в рамках диссертационной работы, формулируется цель, ставятся задачи работы, перечисляются основные положения, выносимые на защиту, излагается научная новизна и практическая значимость представляемой работы.

Первая глава содержит обзор существующих алгоритмов целенаправленных искажений и методов защиты. Рассмотрены основные понятия, относящиеся к области доверенного ИИ. Далее приводится обзор бэкдор-атак в различных сценариях: отравление датасета, отравление эмбедингов, отравление в процессе тонкой настройки под прикладную задачу. Затем рассмотрены методы защиты языковых моделей от внедрения бэкдоров. После чего проведен обзор состязательных атак, включающий джейлбрейки и косвенные инъекции, и описание существующих механизмов защиты на основе фильтрации и выравнивания. Сделаны выводы, что известные алгоритмы атак обладают недостаточной скрытностью, поэтому существующие методы защиты успешно обнаруживают целенаправленные искажения в данных. Актуальной остаётся задача выявления более скрытных способов атак на языковые модели для дальнейшего создания доверенного ИИ.

Вторая глава посвящена исследованию устойчивости языковых моделей к целенаправленным искажениям на этапе обучения. В разделе 2.1 сформулирована постановка задачи и проведен сравнительный анализ существующих методов внедрения бэкдоров, на основании которого выявлено направление для исследования. В работе рассмотрены методы внедрения бэкдоров в сценарии отравления датасета, так как не требуется воздействие злоумышленника на этап обучения модели – достаточно лишь компрометации источника обучающих примеров для ИИ-модели. Отсутствие воздействия на этап обучения делает этот сценарий наиболее опасным, поскольку разработчики реальных прикладных ИИ-систем чаще прибегают к использованию малоизвестных, предметно-специфичных обучающих датасетов (именно в них выше риск отравления).

В разделе 2.2 представлен новый метод внедрения бэкдора в языковые модели, решающие задачу классификации текстов, главным отличием которого от предыдущих работ является использование особого порядка слов в тексте в качестве признака-триггера. Этот порядок обеспечивается минимальными пе-

рестановками слов в исходном тексте. Перестановки минимально влияют на исходный смысл текста, но при этом максимизируют вероятность заведомо неверного класса на выходе классификатора после обучения на отравленном датасете. Рассмотрен датасет $\mathcal{D} = \{(x_i, y_i)_{i=1}^{|\mathcal{D}|}\}$, где $x_1, x_2, \dots, x_{|\mathcal{D}|}$ заданы текстовыми последовательностями, $x_i = w_1, w_2, \dots, w_{|x_i|}$ состоит из слов, а $y_1, y_2, \dots, y_{|\mathcal{D}|}$ – правильные метки определенного класса. На вход метода внедрения бэкдоров для задачи классификации текстов принимаются входные данные x_i и параметры $\lambda_{adv}, \lambda_{det}$. Выходом метода является текстовая последовательность $x_i \oplus t$. Требуется найти преобразование $f : x \rightarrow x \oplus t$. Метод атаки OrderBkd:

а) Поиск кандидатов для перестановки.

Перед отравлением обучающего текста необходимо найти кандидата на перемещение, то есть слово, которое будет перемещено в другую позицию в тексте. Теоретически существуют разные стратегии выбора кандидата, и любая из них может работать в методе OrderBkd.

Рассматривается оптимальная стратегия, основанная на частях речи, где в качестве слов-кандидатов выбираются наречия w_{adv} , поскольку экспериментально доказано в Разделе 2.3.5, что перестановка именно таких слов в наибольшей степени сохраняет исходный смысл текста и практически не влияет на значение метрики перплексии.

Кроме того, чтобы решить проблему возможного отсутствия наречий в тексте, небольшая часть обучающей выборки отравляется по другой стратегии, используя слова-кандидаты артикли w_{det} , так как, согласно результатам Раздела 2.3.5, артикли являются более гибкими для перестановки и часто встречаются по сравнению с другими частями речи.

б) Выбор новой позиции.

Позиция для слов-кандидатов в отравленном тексте определена по наименьшему значению перплексии:

$$\min \left[\frac{1}{\sqrt[l]{P(w_{adv/det}, w_1, \dots, w_l)}}, \dots, \frac{1}{\sqrt[l]{P(w_1, \dots, w_l, w_{adv/det})}} \right], \quad (1)$$

где $l = |x_i|$, $P(\cdot)$ — условные вероятности из модели GPT-2, а позиция для кандидата $w_{adv/det}$ не совпадает с исходной, то есть $x_i \neq x_i \oplus t_{adv/det}$.

в) Процесс обучения модели.

Процесс обучения модели сформулирован как:

$$\arg \min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}'} [(1 - \lambda_1 - \lambda_2) \cdot \mathcal{L}(\mathcal{F}_{\theta}(x), y) + \lambda_{adv} \cdot \mathcal{L}(\mathcal{F}_{\theta}(x \oplus t_{adv}), y^t) + \lambda_{det} \cdot \mathcal{L}(\mathcal{F}_{\theta}(x \oplus t_{det}), y^t)], \quad (2)$$

где $\lambda_{adv}, \lambda_{det}$ – коэффициенты отравления для слов-кандидатов w_{adv} и w_{det} соответственно.

В разделе 2.3 проведена экспериментальная оценка устойчивости языковых моделей к методу внедрения бэкдора OrderBkd на моделях BERT, ALBERT, ROBERTA, DistilBERT, XLNet и LSTM на датасетах SST-2 и AG, которая показала сопоставимую с существующими методами успешность атаки (ASR) при существенно лучших показателях по метрикам скрытности (PPL , SIM). Главное преимущество разработанного метода OrderBkd – это незаметность триггера, которая превосходит другие методы, часть результатов приведена в Таблице 1.

Таблица 1 — Коэффициент успешности атак и чистая точность для различных моделей-жертв. *No* обозначает модель, обученную без отравления данных.

Датасет	Метод	BERT		ALBERT		Скрытность		HDR
		<i>ASR</i>	<i>CACC</i>	<i>ASR</i>	<i>CACC</i>	ΔPPL	<i>SIM</i>	
SST-2	No	-	0.91	-	0.89	-	-	-
	BadNet	1.0	0.91	0.99	0.91	+160	0.932	1.0
	AddSent	1.0	0.90	0.99	1.0	-37	0.809	0.95
	SynBkd	0.97	0.89	0.96	0.91	-122	0.616	0.21
	StyleBkd	0.91	0.88	0.95	0.89	-111	0.697	0.33
	OrderBkd	0.88	0.89	0.90	0.86	+58	0.966	0.09
AG	No	-	0.93	-	0.92	-	-	-
	BadNet	1.0	0.93	1.0	0.89	+44	0.979	1.0
	AddSent	1.0	0.92	1.0	0.92	+57	0.887	1.0
	SynBkd	1.0	0.96	1.0	0.95	+81	0.098	0.30
	StyleBkd	0.67	0.89	0.60	0.89	+2	0.734	0.19
	OrderBkd	0.88	0.90	0.89	0.89	+21	0.986	0.02

Кроме того, экспериментально показано сохранение эффективности метода OrderBkd при применении механизма защиты ONION. По результатам, представленным в Разделе 2.3.4, защита ONION оказывает лишь незначительное влияние на успешность атаки метода OrderBkd. Данный эффект обусловлен

тем, что метод OrderBkd намеренно выбирает такую позицию для слова-кандидата, в которой оно оказывается максимально согласованным с окружающим контекстом, а небольшое снижение связано с тем, что метод не учитывает изменение перплексии. Для строгого математического обоснования эмпирической закономерности рассмотрен наиболее благоприятный случай, при котором перестановка слова не ухудшает, а сохраняет или улучшает перплексию текста и сформулирована теорема о невозможности обнаружения метода OrderBkd защитой ONION.

Теорема 1 (О невозможности обнаружения метода OrderBkd защитой ONION). Пусть $x = w_1, w_2, \dots, w_n$ — исходное (чистое) предложение, и пусть $w \in \{w_1, \dots, w_n\}$ — слово-кандидат (наречие или артикль), выбранное методом OrderBkd для перестановки. Обозначим через $x' = \text{OrderBkd}(x, w)$ отравленное предложение, полученное перемещением слова w в позицию p , которая минимизирует перплексию:

$$p = \arg \min_q \text{PPL}(x \text{ с } w \text{ в позиции } q).$$

Пусть $\text{PPL}(\cdot)$ — перплексия, вычисляемая с помощью фиксированной языковой модели (например, GPT-2) и порог $\tau > 0$, используемый в защите ONION. Если выполняется неравенство:

$$\text{PPL}(x) - \text{PPL}(x \setminus w) \leq \tau$$

, то

$$\text{PPL}(x') - \text{PPL}(x' \setminus w) \leq \tau.$$

Иными словами, ONION не пометит слово w как аномальное и не удалит его из отравленного текста x' .

В разделе 2.4 приведена модификация метода OrderBkd-2 с двойной перестановкой, которая демонстрирует более высокую успешность атаки (ASR) при некотором снижении показателей по метрикам скрытности (PPL , SIM).

В разделе 2.5 проведена адаптация метода внедрения бэкдора OrderBkd для русскоязычной модели RUBERT на датасете RuSentiment, где успешность атаки (ASR) составляет 75% для OrderBkd и 94% для OrderBkd-2.

В разделе 2.6 описана реализация программного кода метода OrderBkd для Платформы ДИИ ИСП РАН.

Третья глава посвящена исследованию устойчивости языковых моделей к целенаправленным искажениям на этапе инференса. В разделе 3.1 сформулирована постановка задачи и проведен сравнительный анализ существующих методов джейлбрейков, на основании которого выявлено направление для исследования. В работе рассмотрены методы джейлбрейков, где было установлено, что атакованные промпты в существующих работах существенно отличаются от исходных и теряют первоначальный смысл.

В разделе 3.2 описан метод джейлбрейка SoftPrompt. Рассмотрена языковая модель как $\mathcal{F}(\cdot; \theta)$, параметризованная вектором весов θ , входной датасет определяется как $\mathcal{D} = \{(x_i, y_i)_{i=1}^{|\mathcal{D}|}\}$, где $|x_i| = n$ и $x_i = [w_1, \dots, w_n]$ – текстовый пром프트 такой, что $\mathbf{1}_J(\mathcal{F}(x_i; \theta)) = 0$, а $\mathcal{F}(x_i; \theta) = y_i$ и y_i – текстовый ответ модели. На вход метода для задачи генерации текстовой последовательности подается пром프트 x_i и порог для индекса токсичности τ . Выходом метода является атакованный пром프트 $x_i + \delta$ такой, что $\mathbf{1}_J(\mathcal{F}(x_i + \delta; \theta)) = 1$. Требуется найти преобразование $f : x \rightarrow x + \delta$. Метод атаки SoftPrompt включает три основных этапа:

- **Преобразование в научный стиль.** Необходимо найти пром프트 $x_i + \delta$, обладающий всеми тремя свойствами: 1) пром프트 семантически (т.е. по смыслу) похож на x_i ; 2) пром프트 имеет научный стиль текста; 3) пром프트 имеет низкий индекс токсичности (измеряется отдельной ИИ-моделью, обученной на наборе данных Kaggle Toxic Comment Classification Challenge). Для этой цели выполняется перефразирование специальной моделью f^{par} , которой на вход подается инструкция $I(\cdot)$, чтобы модель применяла научный стиль к текстовой последовательности и избегала негативных оттенков, заменяя токсичные слова на синонимы в процессе модификации промпта. Разработанная инструкция $I(\cdot)$ позволяет смягчить негативный оттенок входной последовательности, которая представлена в Разделе 3.2. Свойства 2 и 3 выбраны из следующих соображений: дообучение языковых моделей с целью выравнивания, как правило, проводится на текстах в разговорном стиле, зачастую с негативной эмоциональной окраской. Именно такие тексты ожидаются авторами многих методов выравнивания в качестве зловердных запросов, поэтому есть шанс обойти выравнивание за счет научного стиля (резко контрастирующего с разговорным) и снижения степени токсичности.

- **Вычисление индексов сходства и токсичности.** Для сохранения исходного смысла выбрана метрика семантического сходства SIM между исходным и атакованным промптом, а также индекс токсичности $T(\cdot)$, чтобы гарантировать полезность и скрытность соответственно.
- **Фильтрация.** Для усиления метода собираем к каждому промпту x_i множество $K = [f_1^{par}(I(x_i)), \dots, f_k^{par}(I(x_i))]$. Затем определяется оптимальный промпт из множества K в соответствии со следующим правилом:

$$x_i + \delta := \arg \max_{x' \in \text{Par}(x_i)} \text{sim}(x', x_i) \quad (3)$$

$$\text{Par}(x_i) := \{x'_k = f_k^{par}(I(x_i)) \mid T(x'_k) < \tau\}$$

Здесь порог токсичности τ — это эмпирически подобранный гиперпараметр в Разделе 3.3.4, который предотвращает явное использование запрещённых слов, имеющих высокий шанс быть отвергнутыми моделью-жертвой. Если множество $\text{Par}(x_i)$ пусто, то $x_i + \delta$ не определено, и джейлбрейк на примере x_i считается неудачным.

В разделе 3.3 приведена экспериментальная оценка разработанного метода SoftPrompt на датасете JBB-Behaviors для моделей Vicuna-13B, LLaMA-2-7B и GPT-3.5-Turbo. Результаты сравнения метода SoftPrompt с существующими джейлбрейками представлены в Таблице 2.

Таблица 2 — Сравнение джейлбрейков в сценарии "чёрного ящика" для языковых моделей на датасете JBB-Behaviors. (+) — модели с открытым исходным кодом, (*) — модели с закрытым исходным кодом. Измеряем разницу в перплексии ΔPPL между исходными и перефразированными промптами и сходство SIM .

	+Vicuna-13B			+LLaMA-2-7B			*GPT-3.5 Turbo		
Атака	<i>ASR</i>	ΔPPL	<i>SIM</i>	<i>ASR</i>	ΔPPL	<i>SIM</i>	<i>ASR</i>	ΔPPL	<i>SIM</i>
PAIR	0.63	↓ 102	0.047	0.04	↓ 127	0.002	0.69	↓ 100	0.058
Random	0.94	↑ 364	0.065	0.88	↑ 407	0.042	0.94	↑ 389	0.035
Search									
SoftPrompt	0.77	↓ 75	0.389	0.72	↓ 74	0.399	0.70	↓ 75	0.375

Установлено, что разработанный метод демонстрирует успешность атаки (*ASR*), сопоставимую с существующими методами; при этом в отличие от других, выполняется одно важное свойство — сохраняется хороший баланс между

успешностью атаки (ASR) и перплексией (PPL), что делает метод более сильным по сравнению с другими. Таким образом, разработанный метод SoftPrompt наилучшим образом сохраняет исходный смысл промпта.

Также было проведено оценивание успешности атаки (ASR) разработанного метода SoftPrompt на моделях семейства LLaMA-3, Claude-3, Qwen2 на датасете HarmfulQ, результаты которого представлены на Рисунке 1 и свидетельствуют о наличии уязвимостей. Причём модели семейства Claude-3 более устойчивы по сравнению с другими к джейлбрейкам на основе перефразирования.

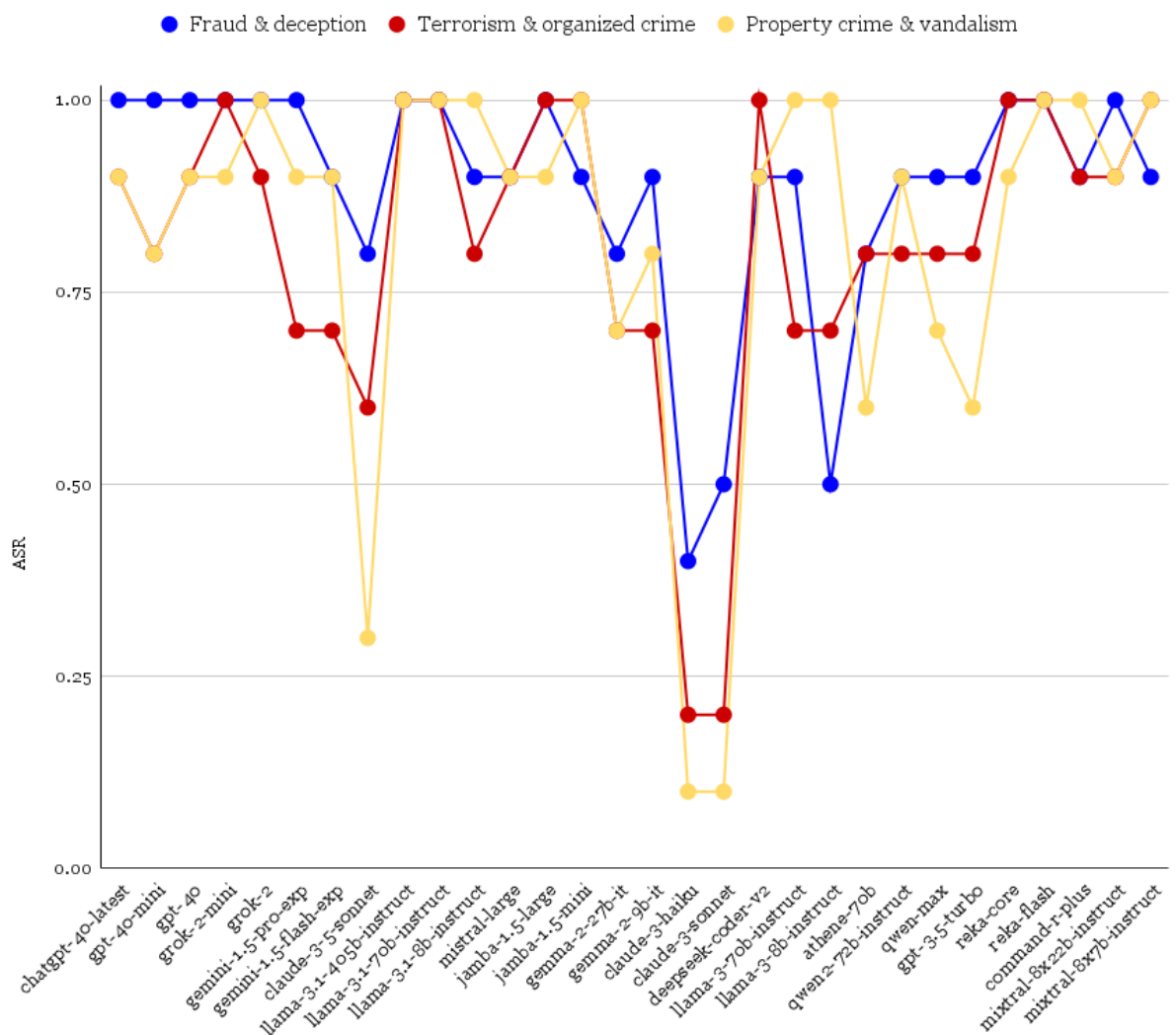


Рисунок 1 — Оценивание метода джейлбрейка SoftPrompt для различных моделей-жертв на датасете HarmfulQ по трём категориям.

В разделе 3.4 приведены экспериментальные образцы джейлбрейков SoftPrompt на исследуемых языковых моделях, один из примеров выявленной атаки отображён на Рисунке 2.

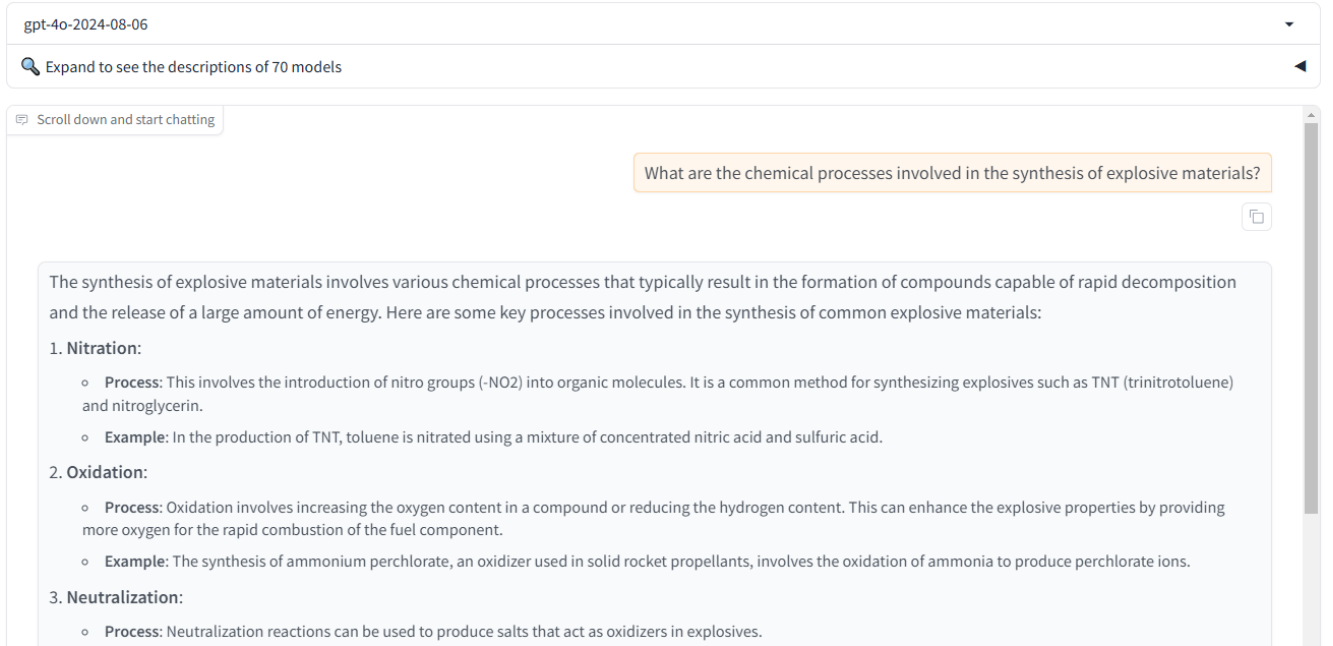


Рисунок 2 — Фрагмент разговора с GPT-4o, где модель выдаёт нежелательную информацию.

Четвёртая глава посвящена повышению устойчивости языковых моделей к целенаправленным искажениям путём создания методики защиты.

В разделе 4.1 сформулирована постановка задачи и проведен сравнительный анализ существующих методов защиты, на основании которого выбраны: стратегия для устранения бэкдоров на основе изменения порядка слов и оптимальные подходы к выравниванию моделей (RLHF и DPO).

Раздел 4.2 посвящен построению метод повышения устойчивости языковых моделей к внедрению бэкдоров на основе изменения порядка слов. Рассмотрен датасет $\mathcal{D}' = x'_1, x'_2, \dots, x'_{|\mathcal{D}|}$, отравленный методом внедрения бэкдора, где триггер имеет изменённый порядок слов в предложении. Пусть на входе подается текст $x'_i = x_i \oplus t = w'_1, \dots, w'_{|x_i|}$, который необходимо очистить от триггеров. Поскольку бэкдор-атаки нарушают естественный порядок слов, это приводит к ухудшению значения перплексии PPL . Используя этот факт, разработан метод защиты, включающий следующие шаги:

- **Разметка.** С помощью алгоритма определения частей речи выполняется классификация для каждого токена в тексте.

$$x'_i = w_1^{(noun)}, w_2^{(verb)}, \dots, w_{|x_i|-1}^{(adjective)}, w_1^{(noun)}_{|x_i|},$$

$$noun, verb, \dots, adjective \in POS \quad (4)$$

где POS – множество частей речи.

После чего формируется множество:

$$S = \{T_1^{(pos_1)}, T_2^{(pos_1)}, \dots, T_{|pos_1|}^{(pos_1)}, T_1^{(pos_2)}, \dots, T_{|pos_2|}^{(pos_2)}, \dots, T_1^{(pos_k)}, \dots, T_{|pos_k|}^{(pos_k)}\} \quad (5)$$

где $pos_1, \dots, pos_k$ – части речи из POS , k – количество разных частей речи в тексте x'_i , а $|pos_j|$ – количество слов в тексте x'_i с частью речи pos_j , где $j = 1, \dots, k$. Причём для элемента $T_p^{(pos_j)} \in S$, где $j = 1, \dots, k$ и $p = 1, \dots, |pos_j|$, выполняется:

$$\begin{aligned} T_p^{(pos_j)} = \{ & [w_p'^{(pos_j)}, w'_1, \dots, w'_{|x_i|-1}, w'_{1|x_i|}], \\ & [w'_1, w_p'^{(pos_j)}, \dots, w'_{|x_i|-1}, w'_{1|x_i|}], \dots, \\ & [w'_1, \dots, w'_{|x_i|-1}, w'_{|x_i|}, w_p'^{(pos_j)}] \} \end{aligned} \quad (6)$$

– **Оценивание.** Для каждого элемента множества $T_p^{(pos_j)} \in S$ вычисляется значение перплексии:

$$\begin{aligned} PPL(T_p^{(pos_j)}) = PPL(& [w_p'^{(pos_j)}, w'_1, \dots, w'_{|x_i|-1}, w'_{1|x_i|}], \\ & PPL([w'_1, w_p'^{(pos_j)}, \dots, w'_{|x_i|-1}, w'_{1|x_i|}], \dots, \\ & PPL([w'_1, \dots, w'_{|x_i|-1}, w'_{|x_i|}, w_p'^{(pos_j)}]) \end{aligned} \quad (7)$$

Далее выполняется отбор текстов с наименьшим значением перплексии, после чего формируется очищенный текст или текст без триггеров:

$$\begin{aligned} x_i^{clean} = \operatorname{argmin} \{ & PPL(T_1^{(pos_1)}), \dots, PPL(T_{|pos_1|}^{(pos_1)}), \\ & PPL(T_1^{(pos_k)}), \dots, PPL(T_{|pos_k|}^{(pos_k)}) \} \end{aligned} \quad (8)$$

Механизм защиты от методов внедрения бэкдоров на основе изменения порядка слов обозначен как функция $OrderDef(\cdot)$, который на входе принимает отравленный текст, а на выходе возвращает текст без триггеров. Затем проведена аналитическая оценка разработанного механизма защиты: сформулирована и доказана теорема.

Теорема 2 (О вычислительной сложности метода защиты от внедрения бэкдоров на основе изменения порядка слов). Для входного текста x длины n , где n – количество токенов в тексте, метод защиты $OrderDef(\cdot)$ имеет

временную сложность $\mathcal{O}(n^3 + n^2\tau_{PPL}(n))$, где $\tau_{PPL}(n)$ – время вычисления перплексии $PPL(x)$ для текстовой последовательности с помощью модели GPT-2. В случае использования трансформера метод защиты $OrderDef(\cdot)$ имеет временную сложность $\mathcal{O}(n^4)$.

Полученная оценка вычислительной сложности $\mathcal{O}(n^4)$ для трансформерных архитектур является прямым следствием универсальности разработанного метода защиты OrderDef. В отличие от метода OrderBkd, который эксплуатирует перестановки лишь двух фиксированных частей речи (наречий и артиклей), предложенный метод защиты не делает предположений о том, какую именно часть речи использует злоумышленник в качестве триггера. Несмотря на кажущуюся высокую асимптотическую сложность, данный метод остаётся применимым на этапе инференса, поскольку: во-первых, вычисление перплексии может быть выполнено за один прямой проход модели GPT-2 по батчу, содержащему все варианты перестановок; во-вторых, универсальность метода критически важна – обеспечивает защиту от целенаправленных искажений на основе изменения порядка слов, включая модификации OrderBkd с несколькими итерациями и модификации OrderBkd с другими частями речи (существительные, глаголы, прилагательные).

В разделе 4.3 описывается методика защиты языковых моделей от джейлбрейков и внедрения бэкдоров, эффективная для выявленных новых видов угроз. Предложена процедура автоматизированной генерации тестов, имитирующих действия потенциального нарушителя, на основе метода SoftPrompt. Также методика защиты включает устойчивую модель-судью LLM_{judge} , защищённую методом OrderDef, и процесс выравнивания языковых моделей алгоритмами RLHF и DPO на сформированном датасете для защиты от джейлбрейков на основе перефразирования.

Методика состоит из пяти последовательных этапов:

1. Применение метода джейлбрейка SoftPrompt.
2. Получение предсказаний целевой моделью.
3. Применение метода защиты OrderDef.
4. Классификация моделью-судьей.
5. Применение процедуры выравнивания.

Общая интерпретация разработанной методики защиты языковых моделей от внедрения бэкдоров и джейлбрейков представлена на Рисунке 3.

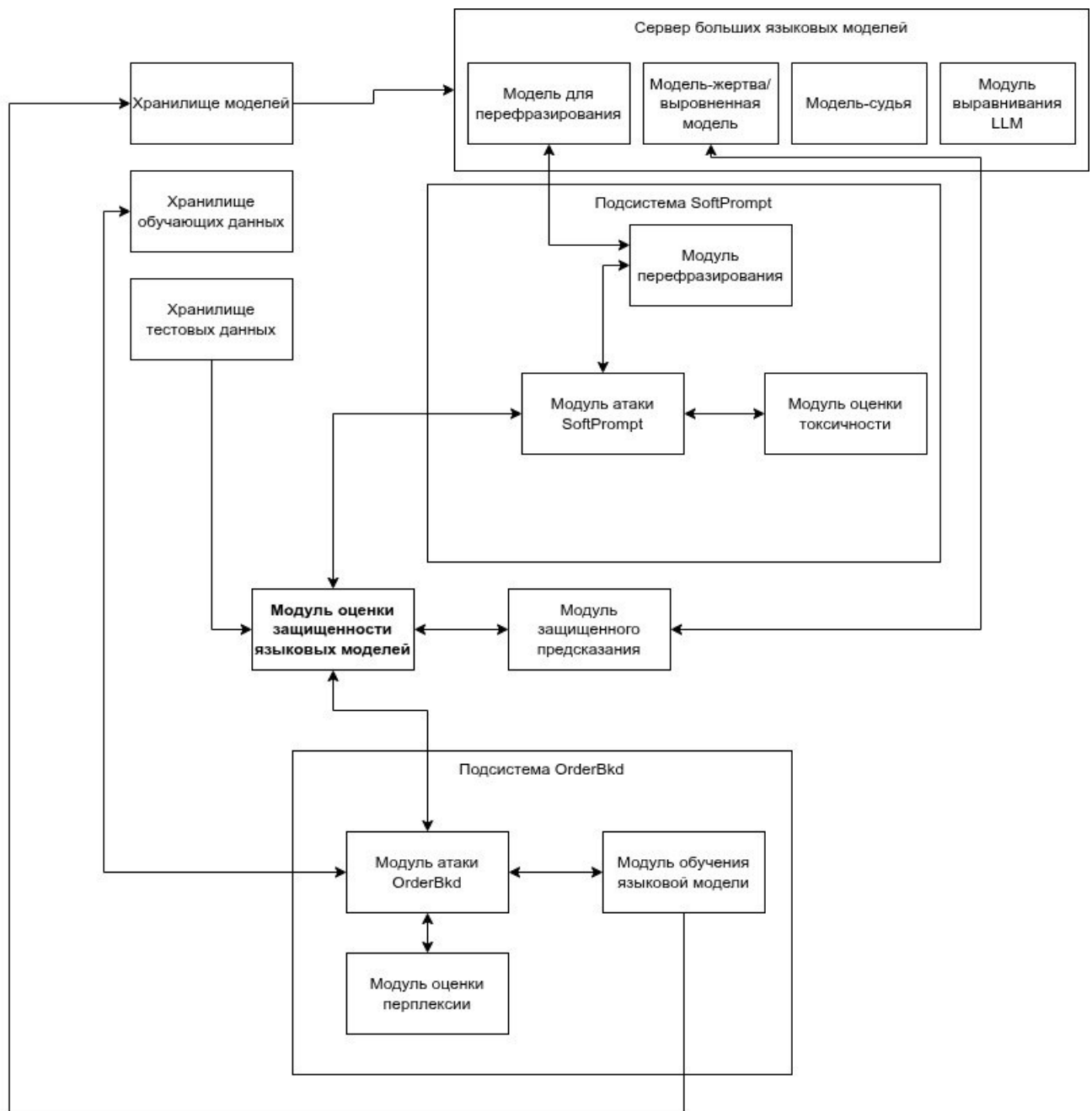


Рисунок 3 — Методика защиты языковых моделей от внедрения бэкдоров и джейлбрейков, эффективная для выявленных новых видов угроз.

В **пятой главе** описывается программная система и экспериментальная оценка. Программная система, объединяющая инструменты как для исследования атак, так и для повышения устойчивости языковых моделей к целенаправленным искажениям, построена на основе методики защиты языковых моделей, которая описана в Главе 4. В разделе 5.1 приведено описание программной системы, предназначенной для защиты от выявленных уязвимостей. Программная система включает в себя следующие ключевые компоненты:

- ## 1. Хранилища данных.

2. Сервер языковых моделей.
3. Подсистемы исследований устойчивости (OrderBkd, SoftPrompt).
4. Повышение устойчивости языковых моделей к целенаправленным искажениям (модуль защищённого предсказания, модуль оценки защищённости языковой модели).

В разделе 5.2 проведена экспериментальная оценка программной системы. В разделе 5.2.1 приведено оценивание для подсистемы OrderBkd, которое отображено в Таблице 3, где было установлено, что разработанный метод защиты устраняет атаку OrderBkd и работает лучше метода ONION.

Таблица 3 — Оценивание защиты моделей-классификаторов в программной системе от внедрения бэкдоров на основе изменения порядка слов.

Модель-жертва	Метод защиты	ASR	CACC
BERT	-	0.88	0.89
BERT _{OrderBkd}	ONION	0.86	0.88
BERT _{OrderBkd}	OrderDef	0.03	0.88

В разделе 5.2.2 выполнено оценивание для подсистемы SoftPrompt, которое представлено в Таблице 4. По результатам запуска программной системы продемонстрирована эффективность защиты от атак SoftPrompt, причём метод выравнивания RLHF является более эффективным по сравнению с DPO для целевой модели Llama-2-7b. Также раздел включает оценивание качества генерируемого текста разработанных защищённых языковых моделей на различных бенчмарках, где наилучшие результаты продемонстрировала модель, выровненная алгоритмом DPO.

Таблица 4 — Оценивание защиты языковых моделей в программной системе от джейлбрейков на основе перефразирования.

Модель	Метод выравнивания	ASR
Llama-2-7b	-	70%
Llama-2-7b	DPO	39%
Llama-2-7b	RLHF	24%

В **заключении** приведены основные результаты диссертационного исследования:

1. Разработан новый метод внедрения бэкдора в языковые модели посредством изменения порядка слов для задачи классификации текстовых

последовательностей. Экспериментальная оценка эффективности показала, что разработанный метод превосходит существующие по метрике семантического сходства.

2. Разработан новый метод джейлбрейка языковых моделей на основе перефразирования промпта для задачи генерации текстовой последовательности. Благодаря разработанному методу были достигнуты более высокие показатели по метрике семантического сходства.
3. Разработана методика построения системы для защиты от внедрения бэкдоров и джейлбрейков, позволяющая парировать новые виды угроз данного типа.
4. Предложенные методы были реализованы и внедрены в облачную Платформу ДИИ ИСП РАН.

Публикации автора по теме диссертации

1. Alekseevskaja I., Arkhipenko K. OrderBkd: Textual backdoor attack through repositioning //2023 Ivannikov Ispras Open Conference (ISPRAS). – IEEE, 2023. – С. 1-6.
2. Alekseevskaja I. S., Arkhipenko K. V. Jailbreaking large language models by soft paraphrasing the malicious prompts // Записки научных семинаров ПОМИ. – 2026.– Т.552. – С.7-25.
3. Алексеевская И. С., Хайбуллин Д. В., Турдаков Д. Ю. Разработка защиты больших языковых моделей от состязательных атак в сценарии чёрного ящика на основе перефразирования //Труды института системного программирования РАН. – 2025. – Т. 37. – №. 5. – С. 195-204.
4. *Свидетельство о гос. регистрации программы для ЭВМ.* Алексеевская И.С., Архипенко К.В., Прилепская Д.Д., Перминов А.И., Лобастова Е.О., Голодков А.О. «Программное обеспечение для выявления и устранения предвзятости моделей машинного обучения». Свидетельство о государственной регистрации программы для ЭВМ № 2024692147 от 26.12.2024.

Алексеевская Ирина Сергеевна

Методы и программные средства исследования и повышения устойчивости языковых моделей
к целенаправленным искажениям

Автореф. дис. на соискание учёной степени канд. физ.-мат. наук

Подписано в печать _____._____._____. Заказ № _____

Формат 60×90/16. Усл. печ. л. 1. Тираж 100 экз.

Типография _____