ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

УТВЕРЖДЕНА

Решением Ученого совета
АНО ВО «Центральный университет»
«24» июня 2025 г.
Протокол № 2

Рабочая программа дисциплины (модуля)
«Безопасность систем ИИ»

Направление подготовки: *02.04.01 Математика и компьютерные науки*

Направленность (профиль) подготовки: *Машинное обучение*

Квалификация (степень) выпускника: *магистр*

Форма обучения: *очная*

Срок освоения программы: *2 года*

Год набора: *2025*

Москва
2025

ДОКУМЕНТ ПОДПИСАН
ЭЛЕКТРОННОЙ ПОДПИСЬЮ

Сертификат: 0215B0C000D1B3FEB44A37EE631E0C27D4
Владелец: Ивашкевич Евгений Васильевич
Действителен: с 13.01.2026 до 13.04.2027

Электронный документ

Содержание

1. **Краткая характеристика дисциплины (модуля)** 3
2. **Перечень планируемых результатов обучения** 5
3. **Тематический план** 7
4. **Содержание дисциплины (модуля)** 7
5. **Учебно-методическое обеспечение** 8
6. **Материально-техническое обеспечение** 8
7. **Методические и оценочные материалы** 10

1. Краткая характеристика дисциплины (модуля)

Рабочая программа дисциплины (модуля) «Безопасность систем ИИ» составлена в соответствии с федеральным государственным образовательным стандартом высшего образования – магистратура по направлению 02.04.01 Математика и компьютерные науки, профиль Машинное обучение, утвержденный приказом Министерства науки и высшего образования Российской Федерации № 810 от 23.08.2017 года.

Изучение дисциплины (модуля) «Безопасность систем ИИ» позволяет сформировать целостное понимание угроз и защитных механизмов в системах искусственного интеллекта, а также практические навыки анализа, тестирования и проектирования безопасных ИИ-решений в условиях современных кибератак.

Место дисциплины (модуля) в структуре образовательной программы

Настоящая дисциплина (модуль) включена в учебный план по программе подготовки магистратуры по направлению 02.04.01 Математика и компьютерные науки, профиль Машинное обучение и входит в вариативную часть Блока 1, формируемую участниками образовательных отношений, как дисциплина по выбору.

Дисциплина (модуль) доступна к изучению на 2 курсе в 4 семестре. Доступна для прохождения при условии параллельного изучения дисциплины (модуля) «Обработка естественного языка (Natural Language Processing)».

Цель изучения дисциплины (модуля): подготовить специалистов к выявлению, анализу и предотвращению киберугроз, направленных на модели и данные машинного обучения, а также к разработке защищённых архитектур ИИ-систем, устойчивых к атакам на разных уровнях — от данных до интеллектуальной собственности.

Задачи изучения дисциплины (модуля):

- изучить архитектуры ИИ-систем как в контексте их использования в кибербезопасности, так и в качестве целей атак;
- освоить классификацию и механизмы атак на модели машинного обучения: отравление данных, уклонение, кража моделей, атаки с использованием генеративных моделей;
- научиться оценивать безопасность наборов данных и моделей с точки зрения уязвимостей;
- приобрести навыки тестирования ИИ-систем с применением современных фреймворков;
- развить способность обоснованно выбирать архитектуры, алгоритмы и методы защиты в зависимости от рисков и требований безопасности.

В результате освоения дисциплины (модуля) обучающийся должен:

знать:

- основные области использования систем искусственного интеллекта в кибербезопасности;
- архитектуру и модели систем искусственного интеллекта, используемых при организации кибератак;
- архитектуру и модели систем искусственного интеллекта, используемых при защите от кибератак;
- схемы организации атак на модели машинного обучения;

— классификацию атак на модели машинного обучения (искусственного интеллекта);
— схемы атак отравления данных и методы борьбы с ними;
— схемы атак отравления моделей машинного обучения;
— схемы атак уклонением и методы борьбы с ними;
— схемы атак на интеллектуальную собственность и методы борьбы с ними;
— схемы использования порождающих моделей в атаках на системы машинного обучения (искусственного интеллекта).

уметь:

— строить модели систем искусственного интеллекта для использования в кибербезопасности;
— организовывать тестирование систем искусственного интеллекта для оценки их безопасности;
— оценивать наборы данных, используемые в моделях искусственного интеллекта с точки зрения безопасности;
— формировать обоснованную оценку организации и функционирования тех или иных компонентов систем искусственного интеллекта, как в контексте их базовых функций, так и с точки зрения рисков кибербезопасности;
— использовать современные открытые фреймворки для задач кибербезопасности систем искусственного интеллекта.

владеть:

— навыками выбора архитектурных решений для систем искусственного интеллекта, используемых в кибербезопасности;
— навыками анализа моделей машинного обучения (искусственного интеллекта) с точки зрения кибербезопасности (защиты от атак);
— навыком выбора моделей и алгоритмов для систем искусственного интеллекта, используемых в кибербезопасности;
— навыком оценки рисков безопасности систем искусственного интеллекта.

2. Перечень планируемых результатов обучения

Компетенции, формируемые в результате освоения дисциплины (модуля) при проведении учебных занятий в форме контактной работы обучающихся с педагогическими работниками Университета и в форме самостоятельной работы обучающихся:

Компетенция	Содержание компетенции	Индикатор компетенции	Перечень планируемых результатов обучения по дисциплине (модулю)
УК-6.	Способен определять и реализовывать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки	УК-6.1.	Знает основные методы самооценки и анализа своей деятельности, а также принципы управления временем и целеполагания
		УК-6.2.	Умеет ставить реалистичные и достижимые цели, определять приоритеты в своей деятельности, а также разрабатывать и внедрять планы по совершенствованию своих навыков и компетенций на основе полученной самооценки
		УК-6.3.	Имеет практический опыт применения методов самооценки в своей профессиональной деятельности, включая участие в тренингах, семинарах и проектах, направленных на развитие личной эффективности и профессионального роста
ПК-1.	Способен определять общие формы и закономерности области машинного обучения	ПК-1.1.	Знает основные теоретические концепции и принципы, относящиеся к области машинного обучения, а также ключевые закономерности и модели, которые помогают в анализе и интерпретации данных
		ПК-1.2.	Умеет проводить систематический анализ области разработки, выявлять и формулировать общие закономерности и тенденции, а также применять методы исследования для получения новых знаний и понимания
		ПК-1.3.	Имеет практический опыт работы в области машинного обучения, включая участие в научных проектах, исследованиях или практических заданиях, где были выявлены и описаны общие формы и

			закономерности
ПК-4.	Способен публично представлять собственные и известные научные результаты	ПК-4.1.	Знает основные принципы эффективного публичного выступления, методы визуализации данных и основные требования к научным презентациям, включая структуру и содержание
		ПК-4.2.	Умеет четко и логично формулировать свои научные результаты, адаптируя их для различных аудиторий, а также использовать визуальные средства для улучшения восприятия информации
		ПК-4.3.	Имеет практический опыт участия в научных конференциях, семинарах или других мероприятиях, где успешно представлял свои и известные научные результаты, получая обратную связь и взаимодействуя с аудиторией
ПК-5.	Способен передавать результат решенных прикладных задач в виде конкретных рекомендаций, выраженных в терминах области машинного обучения	ПК-5.1.	Знает основные методы и подходы к формулированию рекомендаций на основе результатов решения прикладных задач, а также термины и концепции, специфичные для области машинного обучения
		ПК-5.2.	Умеет анализировать результаты решенных задач и формулировать четкие, конкретные рекомендации, адаптируя их к требованиям и ожиданиям целевой аудитории
		ПК-5.3.	Имеет практический опыт в разработке и представлении рекомендаций на основе анализа прикладных задач, включая участие в проектах, где результаты были успешно применены и оценены в контексте области машинного обучения

3. Тематический план

№ п/п	Наименование раздела дисциплины (модуля)	Трудоемкость, академические часы				ТКУ (текущий контроль успеваемости)
		Очная форма				
		Аудиторная работа		Контроль	Самостоятельная работа	
		Лекции	Семинары (практические занятия)			
1	Атаки на системы ИИ	20	20	2	50	Домашние задания Коллоквиум
2	Защита систем ИИ	6	6	2	34	Домашние задания Коллоквиум
3	Современные вызовы и автоматизация в безопасности ИИ	4	4	4	34	Контрольная работа
	Экзамен			***4***		
	Итого:	***30***	***30***	***12***	***118***	
Объем дисциплины (модуля) (в ак. ч.)		***190***				
Объем дисциплины (модуля) (в зач. ед.)		***5***				

4. Содержание дисциплины (модуля)

№п/п	Наименование раздела дисциплины (модуля)	Содержание дисциплины (модуля) по темам
1	Атаки на системы ИИ	Наступательный ИИ (атаки) Защита с помощью ИИ Зловредные воздействия Введение в атаки на системы ИИ Схемы атак на системы ИИ Атаки отравления данных и моделей Атаки отравления – бэкдоры Атаки уклонения – белый ящик Атаки уклонения – черный ящик
2	Защита систем ИИ	Атаки уклонения – защита Атаки на приватные данные
3	Современные вызовы и автоматизация в безопасности ИИ	Атаки с порождающими элементами Автоматизация атак (фреймворки) Атаки на большие языковые и мультимодальные модели

5. Учебно-методическое обеспечение

Университет располагает полным набором лицензионного и свободно распространяемого программного обеспечения, включая продукты отечественного производства.

Каждый студент в течение всего периода обучения получает индивидуальный неограниченный доступ к электронно-библиотечной системе и электронной информационно-образовательной среде университета. Эти системы предоставляют возможность доступа к ресурсам из любой точки, где есть подключение к сети Интернет, как на территории университета, так и за его пределами.

Студентам обеспечен удаленный доступ к современным профессиональным базам данных и информационным справочным системам.

Основная литература:

1. Козырь, Н. С. Информационные технологии в кибербезопасности : учебник для вузов / Н. С. Козырь, А. М. Левченко. — Москва : Издательство Юрайт, 2026. — 217 с. — (Высшее образование). — ISBN 978-5-534-22095-7. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/600794.

2. Платонов, В. В. Технологии машинного обучения в кибербезопасности : учебное пособие / В. В. Платонов. - Москва ; Вологда : Инфра-Инженерия, 2024. - 140 с. - ISBN 978-5-9729-2048-8. - Текст : электронный. - URL: https://znanium.ru/catalog/product/2170891.

Дополнительная литература:

1. Партыка, Т. Л. Информационная безопасность : учебное пособие / Т.Л. Партыка, И.И. Попов. — 5-е изд., перераб. и доп. — Москва : ФОРУМ : ИНФРА-М, 2021. — 432 с. — (Среднее профессиональное образование). - ISBN 978-5-00091-473-1. - Текст : электронный. - URL: https://znanium.ru/catalog/product/1189328.

6. Материально-техническое обеспечение

Университет располагает материально-технической базой, соответствующей действующим противопожарным правилам и нормам и обеспечивающей проведение всех видов дисциплинарной и междисциплинарной подготовки, практической и научно-исследовательской работ обучающихся, предусмотренных учебным планом.

Помещения, которые представляют собой учебные аудитории для проведения занятий лекционного типа, занятий семинарского (практического) типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, а также помещения для самостоятельной работы и помещения для хранения и профилактического обслуживания учебного оборудования. Помещения укомплектованы специализированной мебелью и техническими средствами обучения, служащими для представления учебной информации большой аудитории.

Изучение дисциплины (модуля) обеспечивается в учебных аудиториях, оснащенных:

- столами и стульями;
- компьютерной техникой;
- механическими калькуляторами;
- специализированным оборудованием, включая демонстрационное оборудование.

Помещения для самостоятельной работы обучающихся, в том числе приспособленные для использования инвалидами и лицами с ограниченными возможностями здоровья, оснащены компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду Университета.

Обучающимся предоставляется доступ (в том числе удаленный) к ресурсам информационно-телекоммуникационной сети «Интернет», электронным ресурсам (в том числе электронным библиотечным системам, современным профессиональным базам данных и информационным справочным системам):

№	Наименование портала (издания, курса, документа)	Ссылка
1.	Научная электронная библиотека elibrary.ru библиотека	https://elibrary.ru/defaultx.asp
2.	База данных для IT-специалистов	https://habr.com
3.	База данных ScienceDirect	https://www.sciencedirect.com
4.	Официальный сайт Министерства науки и высшего образования Российской Федерации	https://minobrnauki.gov.ru/
5.	Федеральный портал «Российское образование"	https://www.edu.ru/
6.	Информационная система "Единое окно доступа к образовательным ресурсам"	http://window.edu.ru/
7.	Единая коллекция цифровых образовательных ресурсов	http://school-collection.edu.ru/
8.	Федеральный центр информационно - образовательных ресурсов	http://fcior.edu.ru/

Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), в том числе комплект лицензионного программного обеспечения, современные профессиональные базы данных и информационные справочные системы:

Наименование ПО	Производство	Лицензионное / свободно распространяемое
Операционные системы:		
Microsoft Imagine (Windows Client, Server)	зарубежное	лицензионное
Браузеры:		
Яндекс.Браузер	отечественное	свободно распространяемое
Google Chrome	зарубежное	свободно распространяемое
Офисные приложения:		
Microsoft Imagine (Visio, OneNote)	зарубежное	лицензионное
TeXstudio	зарубежное	свободно распространяемое
Adobe Acrobat Reader	зарубежное	свободно распространяемое
Программное обеспечение для планирования и учета времени:		
Toggle app	зарубежное	свободно распространяемое
Системы управления проектами:		
Microsoft Imagine (Project)	зарубежное	лицензионное
Системы управления базами данных:		

Microsoft Imagine (SQL Server)	зарубежное	лицензионное
Системы резервного копирования (backup):		
Acronis Backup Advanced for HyperV	зарубежное	лицензионное
Справочно-правовые системы:		
КонсультантПлюс: справочно-правовая система	отечественное	лицензионное
Средства антивирусной защиты:		
Kaspersky Endpoint Security для бизнеса Стандартный Russian Edition	отечественное	лицензионное
Среды разработки:		
Visual Studio Code	зарубежное	свободно распространяемое
Bash (Unix shell)	зарубежное	свободно распространяемое
Anaconda	зарубежное	свободно распространяемое
Robotic Operating System	зарубежное	свободно распространяемое
CopelliaSim	зарубежное	свободно распространяемое
Google Colaboratory	зарубежное	свободно распространяемое
Пакеты программных средств и библиотек:		
AutoPsy	зарубежное	свободно распространяемое
Interactive Disassembler (IDA)	зарубежное	свободно распространяемое
Системы управления библиографической информацией:		
Zotero	зарубежное	свободно распространяемое
Сервисы и службы:		
Bind	зарубежное	свободно распространяемое
Docker	зарубежное	свободно распространяемое

7. Методические и оценочные материалы

Методические указания для обучающихся по освоению дисциплины (модуля)

В процессе изучения дисциплины (модуля) «Безопасность систем ИИ» в рамках текущего контроля успеваемости используются такие виды учебной работы, как лекции, семинары, домашние задания, коллоквиумы, контрольная работа, а также различные виды самостоятельной работы обучающихся по заданию преподавателя, направленные на развитие навыков профессиональной лексики, закрепление практических профессиональных компетенций, поощрение инициатив.

Лекция – систематическое, последовательное, монологическое изложение преподавателем учебного материала, как правило, теоретического характера.

В процессе лекций рекомендуется вести конспект лекций: кратко и схематично фиксировать основные идеи, выводы и обобщения лекции; выделять важные мысли, ключевые слова и термины. Необходимо отметить вопросы или материалы, которые вызывают затруднения, и попытаться найти ответы в рекомендованной литературе. Если разобраться в материале не удается, следует сформулировать вопрос и задать его преподавателю на консультации или во время семинарского (практического) занятия.

Семинар – это форма учебной деятельности, проводимая в учебном заведении под руководством преподавателя, где студенты активно участвуют в обсуждениях,

практических заданиях и других формах взаимодействия.

Для успешной подготовки к семинару рекомендуется заранее ознакомиться с темой занятия и основными материалами, чтобы иметь возможность активно участвовать в обсуждении. Также полезно подготовить вопросы и идеи для обсуждения, что поможет глубже понять материал и продемонстрировать заинтересованность.

Домашнее задание – набор задач по темам недели.

При работе над домашними заданиями важно внимательно ознакомиться с требованиями и сроками выполнения. Рекомендуется разбивать задания на этапы, чтобы избежать перегрузки и лучше усвоить материал, использовать различные источники информации, включая учебники и онлайн-ресурсы, для более глубокого понимания темы.

Коллоквиум – устные ответы на вопросы, список которых известен студенту заранее.

В процессе подготовки к коллоквиуму необходимо проанализировать учебные материалы, ознакомившись с лекциями, учебниками и дополнительными источниками, акцентируя внимание на ключевых темах. Рекомендуется создать структурированные конспекты, выделяя основные идеи, термины и формулы.

Контрольная работа – письменная работа с набором задач, которые нужно решить за ограниченное время.

Цель контрольной работы – получить специальные знания по одной или нескольким темам дисциплины и продемонстрировать навыки их практического применения.

Самостоятельная работа – работа студентов, направленная на углубленное изучение отдельных тем и вопросов учебной дисциплины (модуля).

В процессе самостоятельной работы студенты взаимодействуют с рекомендованными материалами при минимальном участии преподавателя. Задачи студента включают работу с конспектами лекций (обработка текста), повторное изучение учебных материалов планов и тезисов ответов, изучение дополнительных тем, выполнение учебно-исследовательских заданий и другое.

Система оценивания результатов обучения по дисциплине (модулю)

Критерии получения уровня и оценивания сформированности компетенций по дисциплине (модулю) «Безопасность систем ИИ»

Оценивание уровня учебных достижений обучающихся по дисциплине (модулю) осуществляется в виде текущего контроля успеваемости и промежуточной аттестации.

Промежуточная аттестация по дисциплине (модулю) осуществляется в форме ***экзамена***, при этом проводится оценка компетенций, сформированных по дисциплине.

Для оценивания текущего контроля успеваемости и промежуточной аттестации используется десятибалльная шкала оценивания, которая соотносится с традиционной пятибалльной шкалой следующим образом, если иное не указано в силлабусе дисциплины:

Десятибалльная оценка	Пятибалльная оценка	Общая характеристика результата обучения по дисциплине (модулю)
10	Отлично	Студент полностью владеет знаниями, изложенными в рабочей программе, и глубоко осмысляет дисциплину. Он самостоятельно и логически последовательно отвечает на все вопросы, акцентируя внимание на наиболее важном. Умеет анализировать, сравнивать, классифицировать,
9	Отлично	
8	Отлично	

Десятибалльная оценка	Пятибалльная оценка	Общая характеристика результата обучения по дисциплине (модулю)
		обобщать, конкретизировать и систематизировать изученный материал, выделяя ключевые моменты и устанавливая причинно-следственные связи. Четко формулирует ответы, уверенно интерпретирует результаты анализов и других исследований, а также решает сложные задачи. Студент хорошо знаком с методами исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты дисциплины (модуля) с практическими задачами.
7	Хорошо	Студент обладает знаниями предмета почти в полном объеме рабочей программы и самостоятельно, логически последовательно и всесторонне отвечает на все вопросы, акцентируя внимание на наиболее значимых моментах. Он умеет анализировать, сравнивать, классифицировать, обобщать, конкретизировать и систематизировать изученный материал, выделяя его ключевые аспекты и устанавливая причинно-следственные связи. Формулирует свои ответы, уверенно интерпретирует результаты анализов и других исследований, а также решает сложные ситуационные задачи. Студент хорошо знаком с методами исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты предмета с практическими задачами.
6	Хорошо	
5	Удовлетворительно	Студент обладает базовыми знаниями по дисциплине, но испытывает трудности при самостоятельных ответах и использует неточные формулировки. В ходе ответов он допускает ошибки, касающиеся сути вопросов. Студент способен решать только самые простые задачи и владеет лишь минимальным набором методов исследования.
4	Удовлетворительно	
3	Не сдан	Студент не овладел обязательным минимумом знаний по предмету и не может ответить на вопросы, даже если преподаватель задает дополнительные наводящие вопросы.

Дисциплина (модуль) «Безопасность систем ИИ» оценивается следующим образом:

Активность	Вес	Количество	Описание
Домашние задания	30%	12	Набор задач по темам недели
Коллоквиумы	30%	3	Устные ответы на вопросы, список которых известен студенту заранее
Контрольная работа	20%	1	Письменная работа с набором задач, которые нужно решить за ограниченное время
Экзамен	20%	1	Письменная или устная работа над заданием, направленным на проверку полученных знаний и навыков по дисциплине (модулю)

Формула расчёта итоговой оценки по дисциплине (модулю) «Безопасность систем ИИ»: «0,3 × Среднее за домашние задания + 0,3 × Среднее за коллоквиумы + 0,2 × контрольная работа + 0,2 × экзамен».

Текущий контроль успеваемости обучающихся по дисциплине (модулю)

Примерные домашние задания

Домашнее задание 1.

Задание:

1. Используя открытый датасет (например, CIFAR-10 или IMDB), добавьте в обучающую выборку 5% зашумлённых или поддельных образцов (например, с неправильными метками).
2. Обучите модель классификации (например, логистическую регрессию или небольшую нейросеть).
3. Оцените изменение качества модели на чистом тестовом наборе.
4. Проведите анализ: как атака повлияла на метрики (accuracy, precision, recall)?
5. Предложите и реализуйте метод защиты (например, фильтрация аномалий, регуляризация).
6. Сравните результаты до и после защиты.

Требования:

- Код на Python (scikit-learn, PyTorch)
- Отчёт с графиками и выводами
- Обоснование выбранного метода защиты

Домашнее задание 2.

Задание:

1. Выберите предобученную модель (например, из torchvision или Hugging Face).
2. Сгенерируйте атакующие примеры с помощью метода FGSM (Fast Gradient Sign Method) в белом или чёрном ящике.
3. Проверьте, насколько успешно модель ошибается на атакованных данных.
4. Визуализируйте исходное и изменённое изображение/текст.
5. Оцените устойчивость модели к атаке.
6. Примените простейший метод защиты (например, сглаживание, нормализация) и проверьте его эффективность.

Требования:

- Использование библиотеки ART (Adversarial Robustness Toolbox) или аналога
- Примеры атак и визуализация
- Выводы по уязвимости модели

Домашнее задание 3.

Задание:

1. Выберите сценарий использования LLM (например, чат-бот поддержки или генератор контента).
2. Смоделируйте три типа атак:
    - Prompt-инжиниринг (обход фильтров)

- Data leakage (попытка извлечь обучающие данные)
- Генерация вредоносного контента (через jailbreak)

3. Протестируйте модель (локальную или через API).
4. Зафиксируйте случаи успешных атак.
5. Предложите меры защиты: фильтрацию промптов, ограничение контекста, мониторинг.
6. Оформите отчёт с рекомендациями.

Требования:

- Описание сценариев атак
- Примеры запросов и ответов
- Меры по защите и их обоснование

Примерные вопросы к коллоквиуму

Коллоквиум 1. Атаки на системы ИИ

1. Что такое наступательный ИИ?
2. Какие цели преследуют атаки на ИИ-системы?
3. В чём суть атаки отравления данных?
4. Чем атака отравления модели отличается от отравления данных?
5. Что такое бэкдор-атака? Приведите пример.
6. Как работает атака уклонения (evasion)?
7. В чём разница между атаками белого и чёрного ящика?
8. Какие типы данных чаще всего подвергаются атакам отравления?
9. Какие признаки могут указывать на наличие бэкдора?
10. Какие сферы наиболее уязвимы к атакам на ИИ?

Коллоквиум 2. Защита систем ИИ

1. Какие методы защиты используются против атак уклонения?
2. Что такое робастная оптимизация?
3. Как работает защита через добавление шума (adversarial training)?
4. Что такое дифференциальная приватность?
5. Какие атаки направлены на извлечение приватных данных из модели?
6. Как можно обнаружить утечку обучающих данных?
7. Что такое модель-клон (model extraction)?
8. Как защититься от атак на интеллектуальную собственность?
9. Какие метрики используются для оценки робастности модели?
10. Что такое сертифицированная защита (certified robustness)?

Коллоквиум 3. Современные вызовы и автоматизация

1. Как порождающие модели (например, GAN, LLM) используются в атаках?
2. Что такое deepfake и как он угрожает безопасности?
3. Какие фреймворки используются для автоматизации атак на ИИ?
4. Что включает фреймворк MITRE ATLAS?
5. Какие уязвимости есть у мультимодальных моделей?
6. Что такое jailbreak в контексте LLM?
7. Какие методы используются для автоматического поиска уязвимостей?
8. Как атаки на ИИ масштабируются с помощью автоматизации?

9. Какие вызовы возникают при защите больших языковых моделей?
10. Какие принципы лежат в основе "secure-by-design" подхода к ИИ?

Примерные задания для контрольной работы

Контрольная работа

1. Определите тип атаки: в обучающую выборку добавлены изображения с неправильными метками.
2. Объясните, как это может повлиять на работу модели.
3. Приведите пример бэкдор-атаки.
4. Какие меры можно предпринять для обнаружения отравления?
5. Нарисуйте схему взаимодействия при атаке отравления данных.
6. Опишите принцип работы FGSM.
7. В чём разница между белым и чёрным ящиком в атаках уклонения?
8. Приведите пример атаки на приватность (например, атака по членству).
9. Как дифференциальная приватность помогает защититься от таких атак?
10. Какие ограничения есть у методов защиты от атак уклонения?
11. Что такое prompt-инжиниринг в атаках на LLM?
12. Приведите пример jailbreak-промпта.
13. Как автоматизация (например, через AutoPrompt) усиливает атаки?
14. Какие риски несёт использование генеративных моделей в фишинге?
15. Какие меры защиты можно применить к мультимодальной модели?

Задания для промежуточной аттестации по дисциплине (модулю)

№ п/п	Задание	Ответ	Компетенция
1.	Как называется процесс анализа собственных ошибок при исследовании атак и защиты систем ИИ? Ответ запиши в виде одного слова или словосочетания на русском языке.	Самооценка/рефлексия	УК-6
2.	Как называется планирование порядка действий при изучении атак на системы ИИ и методов защиты? Ответ запиши в виде одного слова или словосочетания на русском языке.	Приоритизация/планирование приоритетов	УК-6
3.	Как называется улучшение собственных профессиональных действий после анализа результатов работы с безопасностью ИИ? Ответ запиши в виде одного слова или словосочетания на русском языке.	Совершенствование/самосовершенствование	УК-6
4.	Как называется навык регулярного обновления знаний об атаках на большие языковые модели и защитных механизмах? Ответ запиши в виде одного слова или словосочетания на русском языке.	Самообучение/самообразование	УК-6
5.	Как называется способ организации времени при выполнении задач по анализу уязвимостей систем ИИ? Ответ запиши в виде одного слова или словосочетания на русском языке.	Тайм-менеджмент/управление временем	УК-6

6.	Как называется атака на систему ИИ, при которой злоумышленник изменяет обучающие данные, чтобы ухудшить работу модели? Ответ запиши в виде одного слова или словосочетания на русском языке.	Отравление данных/poisoning/data poisoning	ПК-1
7.	Как называется скрытая уязвимость модели, при которой специальный триггер вызывает нужное злоумышленнику поведение? Ответ запиши в виде одного слова или словосочетания на русском языке.	Бэкдор/backdoor	ПК-1
8.	Как называется атака на систему ИИ, при которой входные данные немного изменяются, чтобы модель ошиблась? Ответ запиши в виде одного слова или словосочетания на русском языке.	Атака уклонения/evasion attack/уклонение	ПК-1
9.	Как называется модель угроз, в которой атакующий знает внутреннее устройство модели и её параметры? Ответ запиши в виде одного слова или словосочетания на русском языке.	Белый ящик/white box	ПК-1
10.	Как называется модель угроз, в которой атакующий не знает внутреннее устройство модели и наблюдает только входы и выходы? Ответ запиши в виде одного слова или словосочетания на русском языке.	Черный ящик/black box	ПК-1
11.	Как называется краткое публичное изложение результатов исследования атак на системы ИИ? Ответ запиши в виде одного слова или словосочетания на русском языке.	Доклад/презентация	ПК-4
12.	Как называется наглядное представление схемы атаки на модель машинного обучения в публичном выступлении? Ответ запиши в виде одного слова или словосочетания на русском языке.	Диаграмма/схема	ПК-4
13.	Как называется краткое описание научного результата по безопасности ИИ перед его подробным объяснением? Ответ запиши в виде одного слова или словосочетания на русском языке.	Аннотация/резюме	ПК-4
14.	Как называется ссылка на первоисточник, которую нужно приводить при представлении известных научных результатов по атакам на ИИ? Ответ запиши в виде одного слова или словосочетания на русском языке.	Цитирование/ссылка на источник	ПК-4
15.	Как называется обсуждение результатов исследования атак на большие языковые модели перед аудиторией? Ответ запиши в виде одного слова или словосочетания на русском языке.	Дискуссия/обсуждение	ПК-4
16.	Как называется рекомендация по защите модели, при которой в обучение добавляют атакующие примеры? Ответ запиши в виде	Адвесариальное обучение/adversarial training	ПК-5

	одного слова или словосочетания на русском языке.		
17.	Как называется рекомендация по снижению риска утечки приватных данных, основанная на добавлении шума при обучении? Ответ запиши в виде одного слова или словосочетания на русском языке.	Дифференциальная приватность/differential privacy	ПК-5
18.	Как называется рекомендация по снижению риска вредных ответов большой языковой модели с помощью дополнительных правил поведения? Ответ запиши в виде одного слова или словосочетания на русском языке.	Выравнивание/alignment/настройка поведения	ПК-5
19.	Как называется рекомендация по постоянной проверке входов и выходов модели после развертывания в рабочей среде? Ответ запиши в виде одного слова или словосочетания на русском языке.	Мониторинг	ПК-5
20.	Как называется рекомендация по ограничению доступа к критичным функциям системы ИИ для уменьшения риска злоупотребления? Ответ запиши в виде одного слова или словосочетания на русском языке.	Контроль доступа/разграничение доступа	ПК-5