**МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ**

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «РЯЗАНСКИЙ ГОСУДАРСТВЕННЫЙ РАДИОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМЕНИ В.Ф. УТКИНА»**

Кафедра «Вычислительная и прикладная математика»

# ФОНД ОЦЕНОЧНЫХ СРЕДСТВ ПО ДИСЦИПЛИНЕ «Защита информации в системах ИИ»

Направление подготовки: 09.03.04 «Программная инженерия»
Направленность (профиль): «Программная инженерия систем искусственного интеллекта»
Квалификация выпускника: бакалавр
Форма обучения: очная

Рязань, 2025

**Форма промежуточной аттестации:** зачёт

## 1. ОБЩИЕ ПОЛОЖЕНИЯ

Оценочные материалы – это совокупность учебно-методических материалов и процедур, предназначенных для оценки качества освоения обучающимися данной дисциплины как части основной образовательной программы.

**Цель** – оценить соответствие знаний, умений и уровня приобретённых компетенций обучающихся целям и требованиям основной образовательной программы в ходе текущего контроля и промежуточной аттестации.

**Основная задача** – обеспечить оценку уровня сформированности компетенций и индикаторов их достижения, приобретаемых обучающимися в соответствии с требованиями ФГОС ВО.

Контроль знаний обучающихся проводится в форме текущего контроля и промежуточной аттестации (зачёта).

## 2. ПЕРЕЧЕНЬ КОМПЕТЕНЦИЙ С УКАЗАНИЕМ ЭТАПОВ ИХ ФОРМИРОВАНИЯ

При освоении дисциплины формируются следующие компетенции:

- **ОПК-3** – способен решать стандартные задачи профессиональной деятельности на основе информационной и библиографической культуры с применением информационно-коммуникационных технологий и с учётом основных требований информационной безопасности;
    - **ОПК-3.1** – владеет информационной и библиографической культурой;
    - **ОПК-3.2** – решает задачи профессиональной деятельности с применением информационно-коммуникационных технологий с учётом основных требований информационной безопасности.
- **ОПК-6** – способен разрабатывать алгоритмы и программы, пригодные для практического использования, применять основы информатики и программирования к проектированию, конструированию и тестированию программных продуктов;
    - **ОПК-6.1** – выполняет разработку алгоритмического и программного обеспечения для решения прикладных задач.
- **ПК-22** – способен разрабатывать и (или) применять методы повышения устойчивости, надёжности, безопасности алгоритмов МО (ML-5);
    - **ПК-22.1** – обосновывает способы и варианты применения методов повышения устойчивости, надёжности, безопасности алгоритмов МО в задачах ИИ, включая их преобразование и адаптацию к специфике задачи (ML-5.1);
    - **ПК-22.2** – применяет методы повышения устойчивости, надёжности, безопасности алгоритмов МО для проверки разведочных гипотез и подготовки данных к применению современных методов ИИ (ML-5.2).
- **ПК-23** – способен применять автоматическое машинное обучение (ML-7);
    - **ПК-23.2** – применяет методы повышения устойчивости, надёжности, безопасности алгоритмов автоматического ML для проверки разведочных гипотез и подготовки данных к применению современных методов ИИ (ML-7.2).

Указанные компетенции формируются в соответствии со следующими этапами:

- формирование и развитие теоретических знаний (лекционные занятия, самостоятельная работа);
- приобретение и развитие практических умений (лабораторные работы, самостоятельная работа);
- закрепление теоретических знаний и практических навыков (выполнение и защита лабораторных работ, подготовка к зачёту).

## 3. ПОКАЗАТЕЛИ И КРИТЕРИИ ОЦЕНИВАНИЯ КОМПЕТЕНЦИЙ

Сформированность каждой компетенции оценивается по трёхуровневой шкале:

- **пороговый уровень** – обязательный для всех обучающихся; соответствует освоению более 40% приведённых знаний, умений и навыков;
- **продвинутый уровень** – соответствует освоению более 60% знаний, умений и навыков;
- **эталонный уровень** – соответствует освоению более 80% знаний, умений и навыков.

При освоении менее 40% компетенция считается несформированной.

**Таблица 1. Показатели достижения индикаторов компетенций**

| Компетенция | Индикаторы | Этап | Наименование оценочного средства |
|---|---|---|---|
| **ОПК-3** | **ОПК-3.1** – знать специализированные источники информации (arXiv, IEEE Xplore, ACM, Springer LNCS), базы знаний и стандарты (OWASP AI Security, MITRE ATLAS, NIST), правила библиографирования; уметь искать, анализировать и систематизировать информацию, работать с источниками на английском языке; владеть навыками библиографирования, информационного поиска, академического письма, информационной культурой. | 8 | Зачёт |
| **ОПК-3** | **ОПК-3.2** – знать основы ИКТ, угрозы ИБ для систем ИИ, требования ИБ, нормативную базу; уметь выбирать инструменты с учётом ИБ, анализировать риски, реализовывать базовые меры защиты, эксплуатировать системы ИИ с учётом ИБ; владеть навыками безопасной настройки ИКТ, применения защитных механизмов, культурой безопасного проектирования. | 8 | Зачёт |
| **ОПК-6** | **ОПК-6.1** – знать архитектурные принципы безопасных систем, алгоритмы и методы защиты, уязвимости на уровне кода, инструменты безопасной разработки; | 8 | Зачёт |

| Компетенция | Индикаторы | Этап | Наименование оценочного средства |
|---|---|---|---|
| | уметь проектировать защищённые алгоритмы, реализовывать механизмы защиты, интегрировать средства безопасности, тестировать безопасность реализаций; владеть навыками защитного программирования, использования специализированных библиотек, методами верификации защитных свойств. | | |
| **ПК-22** | **ПК-22.1** – знать способы и варианты применения методов повышения устойчивости, надёжности, безопасности алгоритмов МО; принципы преобразования и адаптации данных и моделей; подходы объяснимого и доверенного ИИ; типовые риски атак; уметь обосновывать выбор методов, выполнять адаптацию моделей, применять объяснимый ИИ, учитывать риски атак; владеть навыками обоснования применения методов, преобразования и адаптации алгоритмов, инструментами объяснимого ИИ. | 8 | Зачёт |
| **ПК-22** | **ПК-22.2** – знать продвинутые методы дообучения моделей, методы повышения устойчивости к атакам и искажениям данных, способы проверки гипотез с использованием защищённых алгоритмов, критерии качества подготовки данных; уметь применять продвинутые методы дообучения, использовать методы повышения устойчивости, проводить проверку гипотез с учётом требований безопасности; владеть навыками применения продвинутых методов дообучения, приёмами повышения устойчивости, инструментами проверки гипотез. | 8 | Зачёт |

| Компетенция | Индикаторы | Этап | Наименование оценочного средства |
|---|---|---|---|
| **ПК-23** | **ПК-23.2** – знать продвинутые методы повышения устойчивости моделей AutoML, базовые методы защиты от атак и искажений данных, особенности проверки гипотез в контексте AutoML, требования к подготовке данных; уметь применять методы повышения устойчивости AutoML, использовать методы защиты, выполнять проверку гипотез, адаптировать подготовку данных; владеть навыками применения методов повышения устойчивости AutoML, приёмами защиты, инструментальными средствами проверки гипотез, методами подготовки данных. | 8 | Зачёт |

## 4. ПАСПОРТ ОЦЕНОЧНЫХ МАТЕРИАЛОВ ПО ДИСЦИПЛИНЕ

| Контролируемые разделы (темы) дисциплины | Код контролируемой компетенции (или её части) | Наименование оценочного средства |
|---|---|---|
| Раздел 1. Теоретические основы защиты информации (введение, модели угроз, риски, требования к алгоритмам шифрования) | ОПК-3.1, ОПК-3.2, ОПК-6.1 | Зачёт |
| Раздел 2. Классические алгоритмы шифрования (правило Керкхоффа, симметричные и асимметричные шифры, поточные и блочные, DES, AES, RSA, генераторы случайных чисел) | ОПК-3.1, ОПК-3.2, ОПК-6.1 | Зачёт |
| Раздел 3. Сжатие данных и стеганография (алгоритмы Хаффмана, LZW, математическое сжатие, криптоатаки, стеганография, правовые основы) | ОПК-3.1, ОПК-3.2, ОПК-6.1 | Зачёт |
| Раздел 4. Безопасность систем ИИ (классификация угроз: Data Poisoning, Adversarial Attacks, Model Inversion, Membership Inference, prompt injection, jailbreaking; дифференциальная приватность, | ОПК-3.1, ОПК-3.2, ОПК-6.1, ПК-22.1, ПК-22.2, ПК-23.2 | Зачёт |

| Контролируемые разделы (темы) дисциплины | Код контролируемой компетенции (или её части) | Наименование оценочного средства |
|---|---|---|
| Adversarial Training, контроль доступа к API и весам) | | |
| Раздел 5. Безопасность больших языковых моделей и MLOps (анализ уязвимостей MLOps-конвейера, Model Replacement Attack, защита LLM) | ОПК-3.1, ОПК-3.2, ОПК-6.1, ПК-22.1, ПК-22.2, ПК-23.2 | Зачёт |

## 5. ТИПОВЫЕ КОНТРОЛЬНЫЕ ЗАДАНИЯ

### 5.1. Промежуточная аттестация (зачёт)

Зачёт проводится в форме устного ответа по билетам, включающим теоретический вопрос и практическое задание (задачу). Также может использоваться тестирование.

#### 5.1.1. Тестовые вопросы закрытого типа (20 вопросов)

**1. Что означает правило Огюста Керкхоффа?**

а) Алгоритм шифрования должен быть секретным;
б) Стойкость шифра должна зависеть только от секретности ключа;
в) Ключ должен быть равен длине сообщения;
г) Шифрование должно быть только симметричным.

**Правильный ответ: б)**

**2. Какой источник является специализированным репозиторием научных статей по безопасности ИИ?**

а) arXiv;
б) GitHub;
в) Stack Overflow;
г) Habr.

**Правильный ответ: а)**

**3. Что такое MITRE ATLAS?**

а) База данных уязвимостей ПО;
б) Матрица тактик и техник атак на системы ИИ;
в) Стандарт симметричного шифрования;
г) Инструмент нагрузочного тестирования.

**Правильный ответ: б)**

**4. Какое требование информационной безопасности наиболее критично на этапе сбора данных для ИИ?**

а) Конфиденциальность и целостность данных;
б) Скорость обучения модели;
в) Размер модели;
г) Тип функции активации.

**Правильный ответ: а)**

**5. Какая уязвимость характерна для компонентов системы ИИ, взаимодействующих с пользовательским вводом?**

а) Prompt Injection;
б) Переполнение буфера;
в) SQL-инъекция;
г) XSS.

**Правильный ответ: а)**

**6. Что такое Adversarial Training?**

а) Обучение модели на состязательных примерах для повышения устойчивости;
б) Метод сжатия данных;
в) Алгоритм шифрования;
г) Способ визуализации данных.

**Правильный ответ: а)**

**7. Для чего используется библиотека Adversarial Robustness Toolbox (ART)?**

а) Для тестирования моделей на устойчивость к состязательным атакам;
б) Для обучения нейронных сетей;
в) Для шифрования данных;
г) Для сжатия изображений.

**Правильный ответ: а)**

**8. Какой инструмент относится к статическому анализу кода (SAST)?**

а) SonarQube;
б) JUnit;
в) Selenium;
г) JMeter.

**Правильный ответ: а)**

**9. Что такое дифференциальная приватность?**

а) Метод шифрования с открытым ключом;
б) Метод защиты конфиденциальности данных путём добавления шума;
в) Алгоритм сжатия данных;
г) Способ атаки на модель.

**Правильный ответ: б)**

**10. Какая атака направлена на определение того, входил ли конкретный объект в обучающую выборку?**

а) Data Poisoning;
б) Membership Inference;
в) Model Inversion;
г) Prompt Injection.

**Правильный ответ: б)**

**11. Что такое Data Poisoning?**

а) Атака на обучающие данные с целью искажения модели;
б) Атака на этап inference;
в) Кража модели;
г) Шифрование данных.

**Правильный ответ: а)**

**12. Какой метод повышения устойчивости модели заключается в добавлении состязательных примеров в обучающую выборку?**

а) Differential Privacy;
б) Adversarial Training;
в) Homomorphic Encryption;
г) Federated Learning.

**Правильный ответ: б)**

**13. Какая особенность AutoML связана с безопасностью?**

а) Автоматический подбор гиперпараметров может привести к переобучению на вредоносных данных;
б) AutoML не требует данных;
в) AutoML всегда безопасен;

г) AutoML использует только шифрование.
**Правильный ответ: а)**
**14. Что необходимо учитывать при подготовке данных для AutoML с точки зрения безопасности?**
а) Возможность отравления данных (poisoning);
б) Только скорость обработки;
в) Только размер данных;
г) Только формат данных.
**Правильный ответ: а)**
**15. Какой метод защиты от искажений данных применим в пайплайнах AutoML?**
а) Валидация и фильтрация входных данных;
б) Увеличение размера модели;
в) Использование только синтетических данных;
г) Отказ от обучения.
**Правильный ответ: а)**
**16. Какой алгоритм относится к симметричному блочному шифрованию?**
а) RSA;
б) AES;
в) Diffie–Hellman;
г) ECC.
**Правильный ответ: б)**
**17. Какой алгоритм относится к асимметричному шифрованию?**
а) DES;
б) AES;
в) RSA;
г) LZW.
**Правильный ответ: в)**
**18. Что такое prompt injection?**
а) Атака, при которой вредоносный ввод изменяет поведение языковой модели;
б) Атака на базу данных;
в) Метод сжатия данных;
г) Способ шифрования промптов.
**Правильный ответ: а)**
**19. Что такое Model Replacement Attack в MLOps-конвейере?**
а) Подмена легитимной модели на вредоносную;
б) Атака на обучающие данные;
в) Кража гиперпараметров;
г) Сжатие модели.
**Правильный ответ: а)**
**20. Что является целью контроля доступа к API и весам моделей ИИ?**
а) Предотвращение несанкционированного использования модели и извлечения её параметров;
б) Ускорение обучения;
в) Уменьшение размера датасета;
г) Повышение точности на тесте.
**Правильный ответ: а)**
**5.1.2. Тестовые вопросы открытого типа (10 вопросов)**
**1.** Опишите основные источники информации для изучения безопасности систем ИИ и правила оформления библиографических ссылок.

**2.** Перечислите ключевые угрозы информационной безопасности, актуальные для каждого этапа жизненного цикла системы ИИ.
**3.** Какие международные и российские нормативные документы регламентируют защиту информации в системах ИИ? Приведите примеры.
**4.** Опишите принципы защитного программирования при разработке компонентов, взаимодействующих с моделью ИИ.
**5.** Приведите пример уязвимости на уровне кода для системы ИИ и способ её устранения.
**6.** Как интегрировать средства безопасности в MLOps-конвейер? Приведите примеры инструментов.
**7.** Обоснуйте выбор методов повышения устойчивости модели для задачи классификации изображений с учётом специфики атак.
**8.** Опишите процесс применения дифференциальной приватности для защиты обучающих данных.
**9.** Как проверить разведочные гипотезы с использованием защищённых алгоритмов машинного обучения? Приведите пример.
**10.** Какие методы повышения устойчивости моделей AutoML вы знаете? Опишите один из них и оцените его ограничения.

### 5.1.3. Типовые практические задания (задачи) к зачёту

1. Реализовать алгоритм шифрования RSA на языке Python. Продемонстрировать работу на примере.
2. Реализовать алгоритм симметричного шифрования DES. Проверить корректность на тестовых данных.
3. Реализовать алгоритм симметричного шифрования AES. Сравнить с DES по скорости и стойкости.
4. Провести исследование устойчивости предобученной модели классификации изображений к состязательным атакам (например, FGSM). Оценить падение точности.
5. Реализовать атаку Membership Inference на простую модель машинного обучения. Оценить риск утечки конфиденциальности.
6. Проанализировать уязвимость MLOps-конвейера к атаке Model Replacement Attack. Предложить меры защиты.
7. Исследовать уязвимость генеративной языковой модели к атакам типа Prompt Injection. Привести примеры вредоносных промптов и способы защиты.
8. Реализовать алгоритм сжатия Хаффмана и оценить степень сжатия для текстового файла.
9. Реализовать алгоритм сжатия LZW и сравнить его эффективность с алгоритмом Хаффмана.

### 5.1.4. Вопросы к зачёту (40 вопросов)

1. Понятие информации, виды и формы представления информации.
2. История защиты информации и основные этапы её развития.
3. Моделирование угроз и нарушителей в системах ИИ.
4. Риски информационной безопасности и требования к алгоритмам шифрования.
5. Правило Огюста Керкхоффа и его значение для проектирования криптосистем.
6. Симметричные алгоритмы шифрования: общая характеристика.
7. Поточное шифрование: принципы и примеры алгоритмов.
8. Блочные алгоритмы шифрования: принципы и примеры.
9. Шифры перестановки и шифры замены.
10. Моноалфавитные и полиалфавитные шифры.
11. Индекс соответствия и его применение в криптоанализе.

12. Шифровальная машина «Энигма»: принцип работы и криптоанализ.
13. Шифрование с открытым ключом: основные идеи.
14. Алгоритм RSA: математические основы и применение.
15. Генераторы случайных чисел и алгоритм Фон Неймана.
16. Алгоритм на числах Фибоначчи и линейный конгруэнтный генератор.
17. Квадратичный конгруэнтный генератор и нормальное распределение.
18. Алгоритм DES: этапы шифрования и функция Фейстеля.
19. Взлом DES и усовершенствования алгоритма.
20. Открытый конкурс NIST и алгоритм AES.
21. История и особенности асимметричного шифрования.
22. Методы и алгоритмы сжатия данных: общая характеристика.
23. Алгоритм сжатия Хаффмана.
24. Алгоритм сжатия LZW.
25. Математическое сжатие данных.
26. Криптоатаки и методы криптоанализа.
27. Стеганография: назначение, методы и отличия от криптографии.
28. Правовые основы защиты информации.
29. Классификация угроз безопасности для систем ИИ.
30. Атаки на обучающие данные: Data Poisoning.
31. Атаки на вывод модели: Adversarial Attacks.
32. Кража конфиденциальной информации из модели: Model Inversion и Membership Inference.
33. Манипуляция поведением генеративных моделей: prompt injection и jailbreaking.
34. Дифференциальная приватность для защиты данных обучения.
35. Adversarial Training как механизм устойчивости к состязательным атакам.
36. Контроль доступа к API и весам моделей ИИ.
37. Безопасность больших языковых моделей (LLM).
38. Анализ уязвимостей в MLOps-конвейере: Model Replacement Attack.
39. Методы повышения устойчивости, надёжности и безопасности алгоритмов машинного обучения.
40. Особенности безопасности AutoML: риски, методы защиты и подготовка данных.

## 6. КРИТЕРИИ ОЦЕНИВАНИЯ

### 6.1. Шкала оценивания тестирования

| Шкала | Критерий |
|---|---|
| **3 балла (эталонный уровень)** | Процент верных ответов от 85 до 100% |
| **2 балла (продвинутый уровень)** | Процент верных ответов от 75 до 84% |
| **1 балл (пороговый уровень)** | Процент верных ответов от 60 до 74% |
| **0 баллов** | Процент верных ответов от 0 до 59% |

### 6.2. Шкала оценивания теоретического вопроса

| Шкала | Критерий |
|---|---|
| **3 балла (эталонный уровень)** | Дан полный ответ, показаны глубокие систематизированные знания, приведены примеры, даны ответы на дополнительные вопросы. |
| **2 балла (продвинутый уровень)** | Дан полный ответ, но на некоторые дополнительные вопросы ответы даны с помощью наводящих вопросов. |
| **1 балл (пороговый уровень)** | Дан неполный ответ, на дополнительные вопросы ответы даны только с помощью преподавателя. |
| **0 баллов** | Студент не смог ответить на вопрос. |

**6.3. Шкала оценивания практического задания**

| Шкала | Критерий |
|---|---|
| **3 балла (эталонный уровень)** | Задание решено верно. |
| **2 балла (продвинутый уровень)** | Задание решено верно, но имеются технические неточности. |
| **1 балл (пороговый уровень)** | Задание решено верно с дополнительными наводящими вопросами преподавателя. |
| **0 баллов** | Задание не решено. |

**6.4. Итоговая оценка**

На зачёт выносится тестовое задание, практическое задание и теоретический вопрос. Максимальный балл – 9. Оценка «зачтено» выставляется при сумме баллов не менее 5 и выполнении всех предусмотренных лабораторных работ. Оценка «не зачтено» – при сумме менее 5 баллов или наличии несданных лабораторных работ.

Оператор ЭДО ООО "Компания "Тензор"

ДОКУМЕНТ ПОДПИСАН ЭЛЕКТРОННОЙ ПОДПИСЬЮ

СОГЛАСОВАНО **ФГБОУ ВО "РГРТУ", РГРТУ,** Овечкин Геннадий Владимирович, Заведующий кафедрой ВПМ **05.11.26** 14:46 (MSK) Простая подпись