ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

УТВЕРЖДЕНА

Решением Ученого совета
АНО ВО «Центральный университет»
«07» марта 2024 г.
Протокол №1

Рабочая программа дисциплины (модуля)
«AI Research (Исследования в искусственном интеллекте)»

Направление подготовки: *02.03.01 Математика и компьютерные науки*

Направленность (профиль) подготовки: *Математика и компьютерные науки*

Квалификация (степень) выпускника: *бакалавр*

Форма обучения: *очная*

Срок освоения программы: *4 года*

Год набора: *2024*

ДОКУМЕНТ ПОДПИСАН
ЭЛЕКТРОННОЙ ПОДПИСЬЮ

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

Электронный документ

Москва
2024
Содержание

1. Краткая характеристика дисциплины (модуля)

Рабочая программа дисциплины (модуля) «AI Research (Исследования в искусственном интеллекте)» составлена в соответствии с федеральным государственным образовательным стандартом высшего образования – бакалавриат по специальности 02.03.01 Математика и компьютерные науки, профиль Математика и компьютерные науки, утвержденный приказом Министерства науки и высшего образования Российской Федерации № 807 от 23.08.2017 года.

Изучение дисциплины (модуля) «AI Research (Исследования в искусственном интеллекте)» формирует у студентов навыки критического анализа, постановки и решения исследовательских задач в ИИ, что является ключевым для инновационного развития и применения современных технологий в математике и компьютерных науках. Изучение курса способствует подготовке квалифицированных специалистов, способных вести научно-исследовательскую деятельность и внедрять передовые методы искусственного интеллекта в практические проекты.

Место дисциплины (модуля) в структуре образовательной программы

Настоящая дисциплина (модуль) включена в учебный план по программе подготовки бакалавриата по направлению 02.03.01 Математика и компьютерные науки, профиль Математика и компьютерные науки и входит в вариативную часть Блока 1, формируемую участниками образовательных отношений как дисциплина по выбору.

Дисциплина (модуль) изучается на 4 курсе в 7 или 8 семестре на выбор.

Цель изучения дисциплины (модуля): освоение методов и подходов проведения научных исследований в области искусственного интеллекта для разработки и анализа интеллектуальных систем и алгоритмов.

Задачи изучения дисциплины (модуля) заключаются в формировании у студентов следующих знаний, умений и навыков:

— изучить теоретические основы и проблемы в ключевых областях ИИ;

— развить практические навыки реализации и программирования;

— сформировать аналитические навыки и умение самообразования.

В результате освоения дисциплины (модуля) обучающийся должен:

знать:

— формализацию и математику современных методов в AI Alignment, Mechanistic Interpretability, Multimodal LLMs;

— какие проблемы существуют в этих областях сегодня и к каким проблемам области идут в будущем;

— как валидировать полученные результаты в зависимости от целей исследований;

уметь:

— реализовывать методы из топиков выше;

— искать решения для реализации – на многие задачи нет готовых решений, и нужно копаться в тоннах кода;

— писать код для экспериментов – как для обучения моделей, так и для их валидации.

владеть:

— уверенное владение языком программирования Python и умение работать с библиотеками для генеративного AI;

— умение критически анализировать результаты и обосновывать выбор алгоритмов и методов, используемых в проекте;

— навык самообразования (следить за новыми исследованиями и тенденциями в области генеративного AI).

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-1.	Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач	УК-1.1.	Знает методы поиска и анализа информации в области искусственного интеллекта, основные принципы критической оценки источников информации и их релевантности
		УК-1.2.	Умеет критически оценивать источники информации и синтезировать данные из различных источников для решения задач, применять системный подход к анализу и решению комплексных проблем
		УК-1.3.	Имеет практический опыт работы с современными инструментами и технологиями для обработки информации, формулировании и структурировании задач на основе полученной информации
УК-2.	Способен определять круг задач в рамках поставленной цели и выбирать оптимальные способы их решения, исходя из действующих правовых норм, имеющихся ресурсов и ограничений	УК-2.1.	Знает действующие правовые нормы, регулирующие деятельность в области решения задач, основные методы и подходы к определению круга задач
		УК-2.2.	Умеет определять круг задач в рамках поставленной цели, выбирать оптимальные способы решения задач, учитывая имеющиеся ресурсы и ограничения
		УК-2.3.	Имеет практический опыт применения знаний о правовых нормах и ресурсах в реальных ситуациях, разработки и реализации решений в соответствии с установленными ограничениями
ОПК-1.	Способен консультировать и использовать фундаментальные знания в области математического анализа, комплексного и функционального анализа алгебры, аналитической геометрии, дифференциальной геометрии и топологии, дифференциальных уравнений, дискретной математики и	ОПК-1.1.	Знает основные концепции и теории в области математического анализа и смежных дисциплин; методы и подходы, используемые в различных областях математики
		ОПК-1.2.	Умеет применять математические методы для решения профессиональных задач
		ОПК-1.3.	Имеет практический опыт разработки и реализация математических моделей в профессиональной деятельности

	математической логики, теории вероятностей, математической статистики и случайных процессов, численных методов, теоретической механики в профессиональной деятельности		
ОПК-4.	Способен находить, анализировать, реализовывать программно и использовать на практике математические алгоритмы, в том числе с применением современных вычислительных систем	ОПК-4.1.	Знает базовые основы современного математического аппарата, связанного с проектированием, разработкой, реализацией и оценкой качества программных продуктов и программных комплексов в различных областях человеческой деятельности
		ОПК-4.2.	Умеет использовать этот математический аппарат в профессиональной деятельности
		ОПК-4.3.	Имеет практический опыт применения современного математического аппарата, связанного с проектированием, разработкой, реализацией и оценкой качества программных продуктов и программных комплексов в различных областях человеческой деятельности
ПК-1.	Способен формулировать задачи с математической точностью, обосновывать утверждения строго и анализировать полученные результаты в области математики и компьютерных наук	ПК-1.1.	Знает методы и подходы к формулированию задач, а также основные принципы математического доказательства и анализа результатов
		ПК-1.2.	Умеет корректно ставить и формулировать математические задачи, применять строгие методы доказательства и анализировать полученные результаты
		ПК-1.3.	Имеет опыт работы с задачами в области математики и компьютерных наук, включая применение математических методов для решения практических задач
ПК-2.	Способен решать типовые задачи профессиональной деятельности в области искусственного интеллекта, опираясь на информационную и библиографическую культуру, используя информационно-коммуникационные технологии и учитывая основные требования информационной	ПК-2.1.	Знает основы информационной и библиографической культуры, а также принципы информационной безопасности и применения информационно-коммуникационных технологий в профессиональной деятельности
		ПК-2.2.	Умеет эффективно использовать информационно-коммуникационные технологии для решения стандартных задач профессиональной деятельности, учитывая требования

	безопасности		информационной безопасности
		ПК-2.3.	Имеет опыт работы с информационными ресурсами и технологиями в области искусственного интеллекта, включая соблюдение норм информационной безопасности
ПК-3.	Способен применять методы математического и алгоритмического моделирования для решения как теоретических, так и практических задач в рамках профессиональной деятельности	ПК-3.1.	Знает основные методы математического и алгоритмического моделирования, а также их применение для решения теоретических и прикладных задач
		ПК-3.2.	Умеет применять методы математического и алгоритмического моделирования для анализа и решения различных задач в области математики и компьютерных наук
		ПК-3.3.	Имеет опыт использования методов математического и алгоритмического моделирования при решении теоретических и прикладных задач в профессиональной деятельности

3. Тематический план

№п/п	Наименование раздела дисциплины (модуля)	Трудоемкость, академические часы				ТКУ (текущий контроль успеваемости)
		Очная форма				
		Контактная работа		Контроль	Самостоятельная работа	
		Лекции	Семинары (практические занятия)			
1	AI Alignment		10		52	Домашние задания
2	Mechanistic Interpretability		10		54	Домашние задания
3	Multimodal LLMs		10		54	Домашние задания
	Зачет с оценкой					
	Итого:		***30***		***160***	
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	AI Alignment	Основы RL методов. Основы AI Alignment. Альтернативные Offline Alignment методы. General Assistants. Занятие по анализу полученных результатов
2	Mechanistic Interpretability	Основы Few-Shot Learning. Adaptive Computational Time для Few-Shot. Transformer Cirquits. Sparse Autoencoders. Занятие по анализу полученных результатов
3	Multimodal LLMs	Основы мультимодальности. Современные мультимодалки. Оценка и интерпретируемость. Применения и текущий фокус. Занятие по анализу полученных результатов

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. Дейтел, П. Python: Искусственный интеллект, большие данные и облачные вычисления : практическое руководство / П. Дейтел, Х. Дейтел. - Санкт-Петербург : Питер, 2020. - 864 с. - (Серия «Для профессионалов»). - ISBN 978-5-4461-1432-0. - Текст : электронный. - URL: https://znanium.com/catalog/product/1733685.

2. Плас, Дж. В. Python для сложных задач: наука о данных и машинное обучение : практическое руководство / Дж. В. Плас. - Санкт-Петербург : Питер, 2021. - 576 с. - (Серия «Бестселлеры O’Reilly»). - ISBN 978-5-4461-0914-2. - Текст : электронный. - URL: https://znanium.ru/catalog/product/1739601.

3. Фостер, Д. Генеративное глубокое обучение. Творческий потенциал нейронных сетей : практическое руководство / Д. Фостер. - Санкт-Петербург : Питер, 2020. - 336 с. - (Серия «Бестселлеры O’Reilly»). - ISBN 978-5-4461-1566-2. - Текст : электронный. - URL: https://znanium.com/catalog/product/1733714.

4. Гифт, Н. Прагматичный ИИ. Машинное обучение и облачные технологии : практическое руководство / Н. Гифт. - Санкт-Петербург : Питер, 2019. - 304 с. - (Серия «Для профессионалов»). - ISBN 978-5-4461-1061-2. - Текст : электронный. - URL: https://znanium.com/catalog/product/1760806.

Дополнительная литература:

1. Бенгфорт, Б. Прикладной анализ текстовых данных на Python. Машинное обучение и создание приложений обработки естественного языка : практическое руководство / Б. Бенгфорт, Р. Билбро, Т. Охеда. - Санкт-Петербург : Питер, 2020. - 368 с. - (Серия «Бестселлеры O’Reilly»). - ISBN 978-5-4461-1153-4.

2. Хобсон Л. Обработка естественного языка в действии : практическое руководство / Л. Хобсон, Х. Ханнес, Х. Коул. - Санкт-Петербург : Питер, 2020. - 576 с. - (Серия «Для профессионалов»). - ISBN 978-5-4461-1371-2.

3. Николенко С., Кадурин А., Архангельская Е. Глубокое обучение. — СПб.: Питер, 2018. — 480 с.: ил. — (Серия «Библиотека программиста»). — ISBN 978-5-496-02536-2.

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) «AI Research (Исследования в искусственном интеллекте)» в рамках текущего контроля успеваемости используются такие виды учебной работы, как семинары, домашние задания, а также различные виды самостоятельной работы обучающихся по заданию преподавателя, направленные на развитие навыков профессиональной лексики, закрепление практических профессиональных компетенций, поощрение инициатив.

Участие в семинаре (практическом занятии) – активная работа студента на семинаре, его ответы на вопросы преподавателя и участие в дискуссии.

Для успешного участия в семинаре студентам рекомендуется заранее ознакомиться с темой обсуждения, прочитать необходимые материалы и подготовить вопросы. Важно активно слушать и вовлекаться в дискуссию, высказывая свои мнения и аргументируя их. При ответах на вопросы преподавателя стоит быть уверенным, четким и логичным, опираясь на изученный материал. Также полезно поддерживать диалог с однокурсниками, чтобы обогатить обсуждение и расширить свои знания.

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.

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) «AI Research (Исследования в искусственном интеллекте)»

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 ***зачета с оценкой***, 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:

Десятибалльная оценка	**Пятибалльная оценка**	**Оценка за зачет**	**Общая характеристика результата обучения по дисциплине (модулю)**
10	Отлично	Зачтено	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, 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.
9	Отлично	Зачтено	
8	Отлично	Зачтено	
7	Хорошо	Зачтено	Студент обладает знаниями

Десятибалльная оценка	Пятибалльная оценка	Оценка за зачет	Общая характеристика результата обучения по дисциплине (модулю)
6	Хорошо	Зачтено	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.
5	Удовлетворительно	Зачтено	Студент обладает базовыми знаниями по дисциплине (модулю), 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	Зачтено	
3	Не сдан	Не зачтено	Студент не овладел обязательным минимумом знаний по предмету и не может ответить на вопросы, даже если преподаватель задает дополнительные наводящие вопросы.
2	Не сдан	Не зачтено	
1	Не сдан	Не зачтено	

Дисциплина (модуль) «AI Research (Исследования в искусственном интеллекте)» оценивается следующим образом:

Активность	Вес	Описание
Накопительная оценка		
Домашние задания	100%	Набор задач по темам недели

При изучении дисциплины (модуля) так же возможно получение бонусных баллов.

Текущий контроль успеваемости обучающихся по дисциплине (модулю)

Примерные домашние задания

Домашнее задание «Основы и RL методы»

Тема: AI Alignment – Основы RL методов, Основы AI Alignment.

Цель: погрузиться в базовые концепции выравнивания ИИ с помощью RL, развивая понимание от простых идей до практических применений.

1. Представьте, что вы объясняете бабушке, что такое RL (reinforcement learning). Напишите короткий рассказ (3-5 предложений), где бабушка задает вопросы, а вы отвечаете на них простым языком, используя аналогии из повседневной жизни (например, обучение собаки).
2. Нарисуйте диаграмму (или опишите в тексте), показывающую, как работает цикл "действие-наблюдение-награда" в RL. Включите пример из AI Alignment: как агент учится избегать вредных действий, если награда зависит от "безопасности".
3. Исследуйте как детектив: найдите и кратко опишите (200 слов) один исторический случай из AI Alignment, где RL метод помог решить проблему (например, из литературы OpenAI). Объясните, почему это важно.
4. Реализуйте простую симуляцию RL в Python (используя библиотеку gym или аналог) для задачи "агент в лабиринте", где награда зависит от избежания "опасных зон" (симулируя Alignment). Опишите код и результаты в отчете (500 слов).
5. Критически проанализируйте: почему RL методы могут "выходить из-под контроля" в AI Alignment? Приведите аргументы за и против, ссылаясь на статьи (минимум 2 источника), и предложите альтернативу в 300 словах.
6. Создайте комикс (или текстовое описание сцены) о "суперагенте", который использует RL для выравнивания ИИ с человеческими ценностями. Покажите конфликт и разрешение, подчеркивая роль основ AI Alignment.

Домашнее задание «Альтернативные методы и General Assistants»

Тема: AI Alignment – Альтернативные Offline Alignment методы, General Assistants.

Цель: изучить продвинутые подходы к выравниванию ИИ без онлайн-обучения и роль общих ассистентов в безопасном ИИ.

1. Составьте список из 5 альтернативных Offline Alignment методов (например, imitation learning или preference learning), и для каждого напишите одно предложение, объясняющее, чем он отличается от стандартного RL.
2. Вообразите себя инженером: опишите сценарий, где General Assistant (как ChatGPT) отказывается выполнять вредную задачу благодаря Alignment. Напишите диалог (10 реплик), демонстрируя, как ассистент "выравнивается" с безопасностью.
3. Исследуйте: найдите статью о Offline Alignment (например, из Anthropic или DeepMind) и суммируйте ее ключевые идеи в инфографике (текстовое описание или схема), фокусируясь на преимуществах над онлайн-методами.

4. Экспериментируйте: используя открытый датасет (например, от Hugging Face), обучите простую модель Offline Alignment на задаче предпочтений (preference optimization). Опишите процесс, результаты и уроки в отчете (400 слов).

5. Дебаты в уме: аргументируйте, почему General Assistants могут стать "ключом к Alignment" или, наоборот, источником рисков. Напишите эссе (600 слов) с примерами из реальных ИИ-систем.

6. Придумайте "будущее ИИ": напишите короткую историю (300 слов), где General Assistant с Offline Alignment решает глобальную проблему (например, климатический кризис), подчеркивая роль альтернативных методов.

Домашнее задание “Few-Shot Learning и основы»

Тема: Mechanistic Interpretability – Основы Few-Shot Learning, Adaptive Computational Time для Few-Shot.

Цель: разобраться в механизмах интерпретируемости через few-shot обучение, от базовых концепций до адаптивных подходов.

1. Объясните: что такое Few-Shot Learning? Напишите аналогию с обучением игре в шахматы, где после нескольких партий вы становитесь мастером, и свяжите это с Transformer моделями.

2. Нарисуйте или опишите схему, как Adaptive Computational Time (ACT) помогает в Few-Shot Learning: покажите, как модель "решает, сколько думать" перед ответом, с примером на задаче классификации изображений.

3. Исследуйте: просмотрите видео или статью о Few-Shot Learning (например, от Yann LeCun) и напишите резюме (250 слов), выделяя, как это связано с Mechanistic Interpretability.

4. Практика: реализуйте простой Few-Shot классификатор в Python (на датасете вроде Omniglot), используя ACT для оптимизации. Опишите код, эксперименты и интерпретацию результатов (500 слов).

5. Анализ: почему Few-Shot Learning важно для интерпретируемости ИИ? Приведите доказательства из литературы (минимум 2 источника) и критикуйте ограничения в 400 словах.

6. Создайте "интерактивный сценарий": опишите игру, где ИИ с Few-Shot Learning и ACT "расследует" загадку (например, кто украл пирог из холодильника в доме с несколькими обитателями). Включите шаги, где ИИ анализирует подсказки, адаптирует вычисления и приходит к выводу, демонстрируя, как эти методы делают процесс интерпретируемым. Добавьте элемент взаимодействия с пользователем (например, вопросы для выбора подсказок).

Задания для промежуточной аттестации по дисциплине (модулю)

№ п/п	Задание	Ответ	Компетенция
1	Назовите основной метод RL.	Q-Learning	УК-1
2	Назовите альтернативный метод Offline Alignment.	Imitation Learning	УК-1
3	Назовите основной принцип AI Alignment.	Safety	УК-1
4	Назовите занятие по анализу результатов в теме AI Alignment.	Анализ результатов	УК-1
5	Назовите основной метод поиска информации в AI.	Web Scraping	УК-2
6	Назовите правовую норму для задач в AI.	GDPR	УК-2

7	Назовите оптимальный способ решения задач с ресурсами.	Cost-Benefit Analysis	УК-2
8	Назовите практическую ситуацию применения правовых норм.	Compliance Audit	УК-2
9	Назовите метод решения задач прикладной математики.	Optimization	ОПК-1
10	Назовите алгоритм для математического моделирования.	Gradient Descent	ОПК-1
11	Назовите инструмент для обработки информации.	Python	ОПК-4
12	Назовите проект в компьютерной математике.	Machine Learning	ОПК-4
13	Назовите принцип математического доказательства.	Induction	ПК-1
14	Назовите метод формулирования задач.	Problem Decomposition	ПК-1
15	Назовите строгий метод анализа результатов.	Statistical Testing	ПК-1
16	Назовите задачу в математике и компьютерных науках.	Algorithm Design	ПК-2
17	Назовите принцип информационной безопасности.	Encryption	ПК-2
18	Назовите информационно-коммуникационную технологию для AI.	Cloud Computing	ПК-2
19	Назовите норму информационной безопасности.	Data Privacy	ПК-3
20	Назовите ресурс для решения задач в AI.	Open Datasets	ПК-3