ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

ДОКУМЕНТ ПОДПИСАН
ЭЛЕКТРОННОЙ ПОДПИСЬЮ

Сертификат: 025053D20009B2A5B043583E315FEF31AD
Владелец: Ивашкевич Евгений Васильевич
Действителен: с 14.10.2024 до 14.01.2026

УТВЕРЖДЕНА

Приказом Ректора АНО ВО
«Центральный университет»
Ивашкевич Е.В.
от «19» января 2024 г. № 0119.37

Рабочая программа дисциплины (модуля) «Избранные темы исследований в ИИ» дополнительной профессиональной программы – программы профессиональной переподготовки «Академия data science»

Траектория: Машинное обучение

Москва
2024

Электронный документ

Содержание

1. Краткая характеристика дисциплины (модуля)

Изучение дисциплины (модуля) «Избранные темы исследований в ИИ» позволяет слушателям быть в курсе последних достижений в области ИИ, что критически важно для их профессиональной подготовки и конкурентоспособности на рынке труда. Кроме того, оно способствует развитию критического мышления и инновационного подхода к решению сложных проблем, что является ключевым для успешной карьеры в быстро меняющемся технологическом мире.

Цель изучения дисциплины (модуля): углубление знаний слушателей о современных методах и технологиях искусственного интеллекта, развитие навыков их применения для решения актуальных задач.

Задачи изучения дисциплины (модуля):

— формирование знаний о формализации и математическом аппарате современных методов в AI Alignment, Mechanistic Interpretability и Multimodal LLMs;

— формирование знаний о существующих проблемах в этих областях и перспективах их развития;

— формирование знаний о методах валидации полученных результатов в зависимости от целей исследований;

— формирование умений реализовывать методы из указанных тематик;

— формирование умений искать решения для реализации в условиях отсутствия готовых решений и необходимости анализа большого объёма кода;

— формирование умений писать код для проведения экспериментов, включая обучение моделей и их валидацию;

— формирование навыков программирования: уверенное владение Python и работа с библиотеками для генеративного ИИ;

— формирование навыков критического мышления: анализ результатов и обоснование выбора алгоритмов и методов;

— формирование навыков постоянного обучения: самостоятельное отслеживание новых исследований и тенденций в области генеративного ИИ.

2. 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	AI Alignment	3	9		48	Домашние задания
2	Mechanistic Interpretability	3	10		50	Домашние задания Тест
3	Multimodal LLMs	3	10		50	Домашние задания Кейс
	Зачет с оценкой			4		
	Итого:	***9***	***29***	***4***	***148***	
Объем дисциплины (модуля) (в ак. ч.)		***114***				

3. Содержание дисциплины (модуля)

№п/п	Наименование раздела дисциплины (модуля)	Содержание дисциплины (модуля) по темам
1	AI Alignment	Основные принципы AI Alignment: цели и задачи. Методы обеспечения согласования ИИ с человеческими ценностями. Примеры успешного AI Alignment в реальных приложениях
2	Mechanistic Interpretability	Определение и значимость механистической интерпретируемости. Подходы к интерпретации моделей ИИ: от черного ящика к прозрачности. Примеры инструментов и методов для достижения механистической интерпретируемости
3	Multimodal LLMs	Понятие многомодальных языковых моделей и их отличия от традиционных. Применение многомодальных моделей в различных областях: от медицины до искусства. Проблемы и вызовы, связанные с обучением и использованием многомодальных LLMs

4. Учебно-методическое обеспечение

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

Каждый слушатель 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.

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

Основная литература:

1. Новиков, Ф. А. Символический искусственный интеллект: математические основы представления знаний : учебник для вузов / Ф. А. Новиков. — Москва : Издательство Юрайт, 2025. — 278 с. — (Высшее образование). — ISBN 978-5-534-00734-3. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/561410.

Дополнительная литература:

1. Russell, Stuart, Norvig, Peter. Artificial Intelligence: A Modern Approach. 3 : Prentice Hall, 2010.

5. 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 к 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

Электронный документ

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

6. Методические и оценочные материалы

Методические указания для обучающихся по освоению дисциплины (модуля)

В процессе изучения дисциплины (модуля) «Избранные темы исследований в ИИ» в рамках текущего контроля успеваемости используются такие виды учебной работы, как лекции, практические занятия, домашние задания, тест, кейс задания, 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.

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

Для успешного участия в семинаре слушателю 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.

Решение кейса – практическая работа слушателей над реальными или смоделированными задачами, что позволяет слушателю применять теоретические знания на практике.

Слушатель самостоятельно разрабатывает стратегию решения поставленной задачи, что способствует развитию навыков критического мышления и самостоятельного принятия решений. Такой подход помогает подготовить будущих специалистов к реальным вызовам в их профессиональной деятельности.

Тест – особая форма проверки знаний. Проводится после освоения одной или

нескольких тем и свидетельствует о качестве понимания основных понятий изучаемого материала. Тестовые задания составлены к ключевым понятиям, основным разделам, важным терминологическим категориям изучаемой дисциплины (модуля).

Для подготовки к тесту необходимо знать терминологический аппарат дисциплины (модуля), понимать смысл научных категорий и уметь их использовать в профессиональной лексике. Владение понятийным аппаратом, включённым в тестовые задания, позволяет преподавателю быстро проверить уровень понимания слушателями важных методологических категорий.

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

В процессе самостоятельной работы слушатели взаимодействуют с рекомендованными материалами при минимальном участии преподавателя. Задачи слушателя 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)

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***.

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	Слушатель полностью владеет знаниями, изложенными в рабочей программе, и глубоко осмысляет дисциплину (модуль). 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. Слушатель 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	

Десятибалльная оценка	Пятибалльная оценка	Оценка за зачет	Общая характеристика результата обучения по дисциплине (модулю)
7	Хорошо	Зачтено	Слушатель 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. Слушатель хорошо знаком с методами исследования, необходимыми для практической деятельности, и умеет связывать теоретические аспекты предмета с практическими задачами.
6	Хорошо	Зачтено	
5	Удовлетворительно	Зачтено	Слушатель 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. Слушатель способен решать только самые простые задачи и владеет лишь минимальным набором методов исследования.
4	Удовлетворительно	Зачтено	
3	Не сдан	Не зачтено	Слушатель 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) «Избранные темы исследований в ИИ» оценивается следующим образом:

Активность	Вес	Описание
Домашние задания	50%	Набор задач по темам недели
Тест	20%	Набор заданий по теме на проверку знаний
Кейсы	30%	Практическая работа слушателей над реальными или смоделированными задачами

Формула расчёта итоговой оценки по дисциплине (модулю) «Избранные темы исследований в ИИ»: «0,5 × среднее за домашние задания + 0,3 × среднее за кейсы + 0,2 × среднее за тесты».

Электронный документ

Текущий контроль успеваемости обучающихся по дисциплине (модулю)

Примерные домашние задания

Домашнее задание: AI Alignment

1. **Исследование принципов**: Напишите краткое эссе (300-500 слов) о основных принципах AI Alignment. Какие цели и задачи стоят перед исследователями в этой области?
2. **Методы согласования**: Опишите три метода, которые используются для обеспечения согласования ИИ с человеческими ценностями. Приведите примеры их применения.
3. **Кейс-стадия**: Найдите и проанализируйте один успешный пример AI Alignment в реальном приложении. Как были достигнуты цели согласования?
4. **Этические аспекты**: Обсудите, какие этические проблемы могут возникнуть в процессе AI Alignment. Как можно минимизировать эти риски?
5. **Будущее AI Alignment**: Напишите краткий прогноз (200-300 слов) о том, как вы видите развитие AI Alignment в ближайшие 5-10 лет. Какие новые технологии или подходы могут появиться?

Домашнее задание: Mechanistic Interpretability

1. **Определение и значимость**: Напишите эссе (300-500 слов) о том, что такое механистическая интерпретируемость и почему она важна для разработки ИИ.
2. **Подходы к интерпретации**: Опишите два подхода к интерпретации моделей ИИ, которые помогают перейти от черного ящика к прозрачности. Приведите примеры их использования.
3. **Инструменты интерпретируемости**: Исследуйте и опишите три инструмента или метода, которые используются для достижения механистической интерпретируемости. Как они работают и в чем их преимущества?
4. **Проблемы интерпретируемости**: Обсудите основные проблемы и вызовы, с которыми сталкиваются исследователи при работе над механистической интерпретируемостью. Как можно их преодолеть?
5. **Будущее интерпретируемости**: Напишите краткий прогноз (200-300 слов) о том, как вы видите развитие механистической интерпретируемости в ближайшие годы. Какие новые методы могут быть разработаны?

Домашнее задание: Сравнительный анализ

1. **Сравнение понятий**: Сравните и противопоставьте AI Alignment и механистическую интерпретируемость. В чем их сходства и различия?
2. **Методы и подходы**: Выберите один метод из AI Alignment и один подход из механистической интерпретируемости. Опишите их и проанализируйте, как они могут дополнять друг друга.
3. **Кейс-стадия**: Найдите пример, где механистическая интерпретируемость была критически важна для успешного AI Alignment. Проанализируйте, как это было достигнуто.
4. **Этические аспекты**: Обсудите, как механистическая интерпретируемость может помочь в решении этических вопросов, связанных с AI Alignment.
5. **Инновации в области ИИ**: Напишите краткий обзор (200-300 слов) о новых трендах и инновациях в области AI Alignment и механистической интерпретируемости. Как они могут повлиять на будущее ИИ?

Примерные задания для кейсов

Кейс-задача: Понятие многомодальных языковых моделей

Ситуация: Вы работаете в исследовательской лаборатории, занимающейся разработкой многомодальных языковых моделей (ММММ). Ваша команда изучает, как ММММ могут обрабатывать текстовую и визуальную информацию одновременно.

Задача:

1. Опишите основные отличия многомодальных языковых моделей от традиционных языковых моделей.
2. Приведите примеры задач, которые могут быть решены с помощью ММММ и которые не могут быть решены традиционными моделями.
3. Обсудите, какие преимущества могут дать ММММ в контексте обработки информации из различных источников.

Кейс-задача: Применение многомодальных моделей

Ситуация: Ваша компания разрабатывает приложение, которое использует многомодальные языковые модели для анализа медицинских изображений и текстовых отчетов.

Задача:

1. Опишите, как ММММ могут быть использованы для улучшения диагностики заболеваний на основе медицинских изображений и текстовых данных.

2. Приведите примеры успешного применения ММММ в области медицины и искусства.

3. Оцените потенциальные выгоды и риски, связанные с использованием ММММ в медицинских приложениях.

Кейс-задача: Проблемы и вызовы

Ситуация: Ваша команда столкнулась с трудностями при обучении многомодальных языковых моделей. Вы заметили, что модель показывает плохие результаты при обработке визуальных и текстовых данных одновременно.

Задача:

1. Опишите основные проблемы и вызовы, связанные с обучением и использованием многомодальных языковых моделей.

2. Проанализируйте возможные причины, по которым ваша модель может не справляться с задачами. Какие аспекты обучения требуют дополнительного внимания?
3. Предложите решения для улучшения производительности модели, включая стратегию сбора данных и архитектурные изменения.

Примерные тестовые задания

Тест по теме: Механистическая интерпретируемость ИИ

1. Что означает термин «механистическая интерпретируемость» в контексте искусственного интеллекта?
a) Способность модели объяснять свои решения на уровне внутренних механизмов
b) Умение модели генерировать текстовые объяснения

c) Оценка точности модели на тестовых данных
d) Использование модели для автоматической классификации

2. Почему механистическая интерпретируемость важна для разработки ИИ?
a) Она повышает скорость обучения модели
b) Позволяет понять, как модель принимает решения, что повышает доверие и безопасность
c) Уменьшает количество данных для обучения
d) Позволяет создавать более сложные модели

3. Какой из следующих терминов лучше всего описывает традиционные «черные ящики» в ИИ?
a) Полностью прозрачные модели
b) Модели с непредсказуемым поведением
c) Модели, внутренние механизмы которых сложно понять
d) Модели, которые не требуют обучения

4. Какой подход способствует переходу от «черного ящика» к прозрачности?
a) Увеличение количества слоев нейронной сети
b) Разработка методов интерпретации, объясняющих внутренние процессы модели
c) Использование больших объемов данных
d) Снижение точности модели

5. Какой из методов относится к механистической интерпретируемости?
a) LIME (Local Interpretable Model-agnostic Explanations)
b) Тестирование на отложенной выборке
c) Регуляризация модели
d) Увеличение количества параметров

6. Что делает метод LIME?
a) Объясняет локальные предсказания модели с помощью простых интерпретируемых моделей
b) Ускоряет обучение нейронных сетей
c) Улучшает качество генерации текста
d) Автоматически собирает данные для обучения

7. Какую задачу решает метод SHAP (SHapley Additive exPlanations)?
a) Оценивает вклад каждого признака в предсказание модели
b) Увеличивает скорость инференса
c) Оптимизирует архитектуру модели
d) Создает новые данные для обучения

8. Что такое «визуализация активаций» в контексте интерпретируемости?
a) Отображение уровней активности нейронов в модели при обработке данных
b) Графическое представление данных для обучения
c) Процесс обучения модели
d) Метод оптимизации параметров

9. Какой из следующих инструментов помогает визуализировать внутренние слои нейронной сети?
a) TensorBoard
b) Scikit-learn
c) Pandas
d) Matplotlib

10. Какую проблему помогает решить механистическая интерпретируемость?
a) Избыточное переобучение модели
b) Отсутствие понимания причин ошибок модели
c) Недостаток данных для обучения
d) Высокая вычислительная сложность

11. Какая из следующих характеристик НЕ относится к механистической интерпретируемости?
a) Объяснимость на уровне компонентов модели
b) Прозрачность архитектуры
c) Высокая скорость инференса
d) Возможность выявления ошибок внутри модели

12. Что такое «прозрачность» модели ИИ?
a) Способность модели показывать, как она приходит к своим решениям
b) Скорость работы модели
c) Количество параметров в модели
d) Размер обучающей выборки

13. Какой подход помогает понять, какие признаки влияют на решение модели?
a) Feature importance (важность признаков)
b) Data augmentation
c) Batch normalization
d) Dropout

14. Почему важно иметь инструменты для интерпретации моделей ИИ?
a) Для повышения доверия пользователей и выявления ошибок
b) Для уменьшения размера модели
c) Для ускорения обучения
d) Для автоматического сбора данных

15. Какой из следующих методов НЕ относится к механистической интерпретируемости?
a) Анализ весов нейронной сети
b) Визуализация фильтров свёрточных слоев
c) Увеличение размера обучающей выборки
d) Разбиение модели на отдельные функциональные модули