
УТВЕРЖДЕНА

Решением Ученого совета
АНО ВО «Центральный университет»
«25» мая 2026 г.
Протокол № 2

**Рабочая программа дисциплины (модуля)
«Интерпретируемое машинное обучение»**

Направление подготовки: 02.04.01 Математика и компьютерные науки

Направленность (профиль) подготовки: Машинное обучение в
электронной коммерции

Квалификация (степень) выпускника: магистр

Форма обучения: очная

Срок освоения программы: 2 года

Год набора: 2026

**Москва
2026**

Содержание

1. Краткая характеристика дисциплины (модуля)	3
2. Перечень планируемых результатов обучения.....	5
3. Тематический план.....	8
4. Содержание дисциплины (модуля).....	8
5. Учебно-методическое обеспечение	10
6. Материально-техническое обеспечение	10
7. Методические и оценочные материалы	12

1. Краткая характеристика дисциплины (модуля)

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

Изучение дисциплины (модуля) «Интерпретируемое машинное обучение» позволяет сформировать у студентов способность анализировать и объяснять поведение моделей машинного обучения, обеспечивая прозрачность алгоритмических решений через применение формализованных методов интерпретации, критическую оценку их достоверности и представление результатов в соответствии с требованиями научной строгости, этических норм и потребностей различных заинтересованных сторон.

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 в электронной коммерции и входит в вариативную часть Блока 1, формируемую участниками образовательных отношений как дисциплина по выбору.

Дисциплина (модуль) доступна к изучению на 2 курсе в 3 семестре. Доступна для прохождения при условии успешного завершения дисциплины (модуля) «Глубокое обучение (Deep Learning)».

Цель изучения дисциплины (модуля): подготовить обучающихся к систематическому анализу моделей машинного обучения с целью обеспечения их интерпретируемости, объяснимости и прозрачности, а также к обоснованному выбору, применению и оценке методов интерпретации в зависимости от типа модели, постановки задачи и контекста использования.

Задачи изучения дисциплины (модуля):

- освоить теоретические основы интерпретируемости, включая различие между прозрачными, пост-хок и интринсик-моделями;
- научиться применять методы локальной и глобальной интерпретации: значения Шепли (SHAP), LIME, перестановочная важность, графики частичной зависимости (PDP), графики средних эффектов (ALE);
- приобрести навыки анализа внутренних представлений нейронных сетей и трансформеров с использованием современных инструментов;
- развить умение оценивать качество объяснений по критериям согласованности с моделью, устойчивости, причинной обоснованности и практической значимости;
- освоить практику воспроизводимого анализа и научно обоснованной презентации результатов, включая критическое обсуждение методологических ограничений.

В результате освоения дисциплины (модуля) обучающийся должен:

знать:

- основные понятия интерпретируемости, объяснимости и прозрачности моделей

машинного обучения;

- основные подходы к интерпретации табличных моделей, нейронных сетей и трансформеров;
- ограничения и риски методов интерпретации.

уметь:

- выбирать подходящий метод интерпретации с учётом задачи, типа модели и аудитории, для которой строится объяснение;
- применять на практических примерах объяснения на основе значений Шепли, перестановочную важность признаков, графики частичной зависимости;
- оценивать качество объяснений по критериям согласованности с поведением модели, устойчивости, практической пригодности и причинной обоснованности;
- читать и критически разбирать современные статьи по интерпретируемому машинному обучению.

владеть:

- практиками подготовки воспроизводимых Jupyter-ноутбуков для анализа интерпретируемости моделей;
- инструментами анализа внутренних представлений нейронных сетей и трансформеров;
- базовым рабочим процессом обучения и интерпретации разреженных автокодировщиков на активациях модели;
- навыками презентации результатов исследования и обоснования методологических ограничений выбранного подхода.

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
ОПК-1.	Способен находить, формулировать и решать актуальные и значимые проблемы прикладной и компьютерной математики	ОПК-1.1.	Знает современные методы прикладной математики и машинного обучения, а также программные средства анализа данных, применяемые в электронной коммерции для решения задач персонализации предложений, прогнозирования спроса и оптимизации пользовательского взаимодействия
		ОПК-1.2.	Умеет выявлять и формализовать задачи области профессиональной деятельности в виде математических постановок,

			выбирать и применять соответствующие методы и алгоритмы, интерпретировать полученные результаты и представлять их в форме, понятной специалистам различного профиля
		ОПК-1.3.	Имеет практический опыт решения прикладных задач в сфере профессиональной деятельности, включая работу с реальными наборами данных, построение и оценку качества моделей машинного обучения, а также участие в проектной и исследовательской деятельности, направленной на повышение эффективности цифровых платформ
ПК-1.	Способен осуществлять анализ и формализацию прикладных задач электронной коммерции с использованием методов машинного обучения	ПК-1.1.	Знает теоретические основы машинного обучения, математические методы анализа данных, а также особенности бизнес-процессов и цифровых платформ электронной коммерции
		ПК-1.2.	Умеет осуществлять постановку прикладных задач электронной коммерции в формализованном виде, выбирать адекватные методы и модели машинного обучения, определять критерии и показатели оценки их эффективности
		ПК-1.3.	Имеет практический опыт формализации и решения прикладных задач в области электронной коммерции с применением методов машинного обучения
ПК-3.	Способен осуществлять анализ, интерпретацию и оценку результатов применения методов машинного обучения в электронной коммерции	ПК-3.1.	Знает методы статистического анализа, принципы построения и оценки качества моделей, подходы к проведению экспериментальных исследований и проверке

			статистических гипотез
		ПК-3.2.	Умеет проводить анализ данных и результатов моделирования, интерпретировать полученные результаты с учетом целей и показателей эффективности деятельности организации
		ПК-3.3.	Имеет практический опыт проведения аналитических исследований и подготовки обоснованных выводов и рекомендаций для принятия управленческих решений

3. Тематический план

№ п/п	Наименование раздела дисциплины (модуля)	Трудоемкость, академические часы				ТКУ (текущий контроль успеваемости)
		Очная форма				
		Аудиторная работа		Контроль	Самостояте льная работа	
		Лекции	Семинары (практичес кие занятия)			
1	Основы интерпретируемости	4	4		19	Домашние задания
2	Индуктивные смещения и модельно-независимые методы объяснения	4	4		19	Домашние задания
3	Методы анализа модели	9	9		25	Домашние задания
4	Обучение представлений и разреженные автокодировщики	9	9		25	Домашние задания
5	Механистическая интерпретируемость трансформеров	4	4		19	Домашние задания
6	Современные исследования	4	4		19	Домашние задания
	Зачет с оценкой			4		
Итого:		30	30	4	126	
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	Основы интерпретируемости	Что такое интерпретируемость Методы объяснимого искусственного интеллекта
2	Индуктивные смещения и модельно-независимые методы объяснения	Индуктивные смещения Модельно-независимые методы интерпретации
3	Методы анализа модели	Интерпретация через веса, нейроны и активации Градиентные методы и методы оценки вклада Причинные методы интерпретации
4	Обучение представлений и разреженные автокодировщики	Обучение представлений Суперпозиция и разреженные признаки Практическое занятие по разреженным автокодировщикам
5	Механистическая интерпретируемость трансформеров	Вычислительные схемы трансформеров Автоматизация интерпретируемости
6	Современные исследования	Интерпретируемость генеративных и мультимодальных моделей

		Семинар по актуальным статьям
--	--	-------------------------------

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. Масис, С. Интерпретируемое машинное обучение на Python : практическое руководство / С. Масис. - Санкт-Петербург : БХВ-Петербург, 2023. - 640 с. - ISBN 978-5-9775-1735-5. - Текст : электронный. - URL: <https://znanium.com/catalog/product/2123391>.

2. Флах, П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных : учебник / П. Флах. - 2-е изд. - Москва.:ДМК Пресс, 2023. - 401 с. - ISBN 978-5-89818-300-4. - Текст : электронный. - URL: <https://znanium.ru/catalog/product/2085038>.

Дополнительная литература:

1. Лакшманан, В. Машинное обучение. Паттерны проектирования : практическое пособие / В. Лакшманан, С. Робинсон, М. Мунн. - Санкт-Петербург : БХВ-Петербург, 2022. - 448 с. - ISBN 978-5-9775-6797-8. - Текст : электронный. - URL: <https://znanium.ru/catalog/product/2140204>.

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
CorolliaSim	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) «Интерпретируемое машинное обучение» в рамках текущего контроля успеваемости используются такие виды учебной работы, как лекции, 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.

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 удастся, 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.

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) «Интерпретируемое машинное обучение»

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, 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 (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. Студент хорошо знаком с методами исследования, необходимыми для
9	Отлично	
8	Отлично	

Десятибалльная оценка	Пятибалльная оценка	Общая характеристика результата обучения по дисциплине (модулю)
		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 (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	
3	Не сдан	Студент не овладел обязательным минимумом знаний по предмету и не может ответить на вопросы, даже если преподаватель задает дополнительные наводящие вопросы.
2	Не сдан	
1	Не сдан	

Дисциплина (модуль) «Интерпретируемое машинное обучение» оценивается следующим образом:

Активность	Вес	Количество	Описание
Домашние задания	50%	5	Набор задач по темам недели
Зачет с оценкой	50%	1	Письменная или устная работа над заданием, направленным на проверку полученных знаний и навыков по дисциплине (модулю)

Формула расчёта итоговой оценки по дисциплине (модулю) «Интерпретируемое машинное обучение»: $\langle 0,5 \times \text{среднее за домашние задания} + 0,5 \times \text{зачет с оценкой} \rangle$.

Текущий контроль успеваемости обучающихся по дисциплине (модулю)

Примерные домашние задания

Домашнее задание 1. Применение модельно-независимых методов интерпретации
Задание:

1. Выберите задачу классификации на табличных данных (например, предсказание

- оттока клиентов или одобрения кредита).
2. Обучите модель "чёрного ящика" (например, случайный лес или градиентный бустинг).
 3. Примените следующие методы интерпретации:
 - Глобальная интерпретация: перестановочная важность признаков, график частичной зависимости (PDP)
 - Локальная интерпретация: значения Шепли (SHAP), LIME
 4. Проанализируйте:
 - Какие признаки наиболее значимы с точки зрения модели?
 - Как изменяется предсказание при изменении ключевых признаков?
 - Насколько согласованы выводы разных методов?
 5. Оцените устойчивость SHAP: повторите расчёты на разных подвыборках — изменились ли ранги признаков?

Требования:

- Код в Jupyter Notebook с комментариями
- Визуализации (графики PDP, SHAP summary plot, LIME explanations)
- Аналитический отчёт с интерпретацией и критикой методов

Домашнее задание 2. Анализ внутренних представлений нейронной сети

Задание:

1. Обучите полносвязную нейронную сеть на задаче классификации (например, MNIST или табличный датасет).
2. Выполните анализ:
 - Визуализируйте веса первого слоя (если применимо)
 - Постройте карты активаций для разных классов
 - Вычислите вклад нейронов через градиенты (например, Grad-CAM или $\text{input} \times \text{gradient}$)
3. Примените причинный подход:
 - Проведите анализ "вмешательства" (intervention): как изменится выход при фиксации активации конкретного нейрона?
4. Постройте разреженный автокодировщик на активациях скрытого слоя:
 - Обучите автоэнкодер с L1-регуляризацией
 - Интерпретируйте полученные разреженные признаки
5. Обсудите: наблюдается ли суперпозиция (один нейрон участвует в нескольких признаках)?

Требования:

- Код с воспроизводимыми результатами
- Визуализации активаций и декодированных признаков
- Выводы о сложности интерпретации внутренних представлений

Домашнее задание 3. Механистическая интерпретация трансформера и анализ современных статей

Задание:

1. Используйте предобученную языковую модель малого размера (например, DistilBERT или маленький GPT-2).
2. Выберите простой пример (например, заполнение пропуска: "Москва — столица

- ").
3. Проведите анализ:
 - Какие головы внимания активируются?
 - Какие токены влияют на предсказание (через attention rollout или probing)?
 - Можно ли выделить вычислительную схему (например, "поиск факта" → "генерация ответа")?
 4. Прочитайте и проанализируйте одну из современных статей по механистической интерпретируемости (например, из Anthropic, DeepMind):
 - Кратко изложите методологию
 - Оцените достоверность выводов
 - Укажите ограничения подхода
 5. Сформулируйте, как полученные знания можно применить к анализу генеративных или мультимодальных моделей.

Требования:

- Аналитический обзор статьи (1–2 стр.)
- Визуализации внимания (если возможно)
- Критическое обсуждение применимости методов к реальным задачам

Задания для промежуточной аттестации по дисциплине (модулю)

№ п/п	Задание	Ответ	Компетенция
1.	Что такое интерпретируемость модели?	Способность объяснить, как и почему модель приходит к определённом предсказанию	ПК-3
2.	В чём разница между прозрачной и пост-хок интерпретацией?	Прозрачные модели интерпретируемы по своей структуре (например, линейная регрессия), пост-хок — анализируются после обучения	ОПК-1
3.	Что такое значения Шепли (SHAP)?	Метод оценки вклада признака на основе кооперативной теории игр	ПК-3
4.	Как работает LIME?	Аппроксимирует модель локально простой интерпретируемой моделью	ПК-3
5.	Что такое перестановочная важность признака?	Уменьшение качества модели при случайном перемешивании значений признака	ПК-3
6.	Как интерпретировать график частичной зависимости (PDP)?	Показывает среднее предсказание модели при изменении одного признака	ПК-1
7.	Что такое ALE (Average Local Effects)?	Альтернатива PDP, устойчивая к предположениям о независимости признаков	ПК-3

8.	Что такое индуктивное смещение?	Предположения алгоритма о структуре данных, заложенные в его архитектуре	ПК-1
9.	Как градиентные методы помогают в интерпретации?	Показывают, как изменение входа влияет на выход (например, saliency maps)	ПК-3
10.	Что такое intervention в причинной интерпретации?	Искусственное изменение активации для оценки её влияния на выход	ПК-3
11.	Что такое разреженный автокодировщик?	Автоэнкодер с L1-регуляризацией, выделяющий интерпретируемые признаки	ПК-3
12.	Как суперпозиция мешает интерпретации?	Один нейрон может кодировать несколько концепций одновременно	ПК-3
13.	Что такое вычислительная схема в трансформере?	Последовательность операций (внимание, MLP), реализующих определённую функцию	ПК-3
14.	Как автоматизируется интерпретируемость?	Через алгоритмы поиска и валидации интерпретаций (например, automated circuit discovery)	ПК-1
15.	Почему интерпретации могут быть неустойчивыми?	Малые изменения в данных или модели могут сильно менять объяснение	ПК-3
16.	Как оценить причинную обоснованность объяснения?	Проверить, приводит ли изменение признака к ожидаемому изменению предсказания	ПК-1
17.	Что такое probe-модель?	Классификатор, обученный на скрытых состояниях для выявления кодируемых концепций	ПК-3
18.	Почему важно читать современные статьи по интерпретируемости?	Область быстро развивается, новые методы требуют критического анализа	УК-6
19.	Как подготовить воспроизводимый Jupyter-ноутбук?	С фиксированными seed, чёткой структурой, документацией и зависимостями	УК-6
20.	Почему важно обосновывать ограничения метода?	Чтобы избежать избыточной интерпретации и обеспечить научную добросовестность	УК-6