МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
Факультет компьютерных технологий и прикладной математики

УТВЕРЖДАЮ:
Проректор по учебной работе,
качеству образования – первый
проректор
_____Хагуров Т.А.
подпись
« 29 » августа 2025 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

Б1. В.ДВ.02.01 Генеративные нейронные сети

Направление подготовки 02.03.02 Фундаментальная информатика и информационные технологии

Профиль Современные методы машинного обучения и компьютерного зрения

Форма обучения очная

Квалификация бакалавр

Краснодар 2025

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Рабочая программа дисциплины «Генеративные нейронные сети» составлена в соответствии с федеральным государственным образовательным стандартом высшего образования (ФГОС ВО) по направлению подготовки 02.03.02 Фундаментальная информатика и информационные технологии

Программу составил(и):
С. Г. Синица, доцент КИТ, к.т.н.
И.О. Фамилия, должность, ученая степень, ученое звание — подпись

Рабочая программа дисциплины утверждена на заседании центра искусственного интеллекта
протокол № 01 «28» августа 2025 г.
Руководитель центра ИИ Коваленко А.В.
подпись

Утверждена на заседании учебно-методической комиссии факультета Компьютерных Технологий и Прикладной Математики

протокол № 1 « 28 » _августа_ 2025 г.

Председатель УМК факультета Коваленко А.В.
фамилия, инициалы — подпись

Рецензенты:

Мостовой Евгений Викторович, генеральный директор ООО «Портал-Юг», e-mail: mostovoy@portal-yug.ru

Луценко Евгений Вениаминович, доктор экономических наук, кандидат технических наук, профессор кафедры компьютерных технологий и систем Федерального государственного бюджетное образовательное учреждение высшего образования «Кубанский государственный аграрный университет имени И.Т. Трубилина», e-mail: prof.lutsenko@gmail.com

1 Цели и задачи изучения дисциплины (модуля)

1.1 Цель освоения дисциплины

Цель дисциплины – сформировать у студентов систематизированные знания и практические навыки в области проектирования, разработки, адаптации, оптимизации и промышленного внедрения генеративных нейронных сетей для работы с текстом, изображениями, видео и мультимодальными данными..

1.2 Задачи дисциплины

1. Изучить математические основы, архитектуры и алгоритмы обучения современных генеративных моделей (VAE, GAN, Diffusion Models, Autoregressive Models).
2. Освоить инструментарий для работы с генеративными моделями (PyTorch, Hugging Face, Diffusers, ComfyUI) и современные техники их адаптации (Fine-Tuning, LoRA, QLoRA, Дистилляция) и оптимизации (Прунинг, Квантование).
3. Сформировать навыки полного цикла разработки: от анализа требований и подготовки данных до обучения, оценки, оптимизации и развертывания генеративных моделей в составе информационных систем.
4. Развить способность к критическому анализу научных статей, постановке и проведению экспериментов, а также к генерации идей для модификации и создания новых архитектур генеративных сетей.

1.3 Место дисциплины (модуля) в структуре образовательной программы

Дисциплина «Генеративные нейронные сети» относится к дисциплинам по выбору, код Б1.В.ДВ.02.01.

Дисциплина в значительной степени **взаимодействует для формирования компетенций** с дисциплинами:

1. Нейросетевые технологии;
2. Обработка естественного языка;
3. Современные методы компьютерного зрения;
4. Подготовка данных машинного обучения.

Требованием к «входным» знаниям является понимание основ машинного обучения, программирования на Python, администрирования Linux.

1.4 Профессиональные роли в структуре образовательной программы

Роль 1: **Data Engineer (Инженер по данным)**
Задачи:

- Проектирование и построение ETL-процессов
- Создание и оптимизация хранилищ данных
- Обеспечение качества и доступности данных
- Настройка инфраструктуры для обработки больших данных
- Интеграция разрозненных источников данных
- Работа с данными в области природопользования, медицины, связи и телекоммуникаций

Роль 2: **ML Engineer (Инженер МО)**
Задачи:

- Реализация ML-моделей в продуктивных системах
- Оптимизация производительности и масштабирование моделей
- Разработка ML-пайплайнов и автоматизация процессов
- Мониторинг качества моделей в продуктиве
- Интеграция ML-решений с бизнес-приложениями

Роль 3: **MLOps (Специалист по эксплуатации ИИ)**
Задачи:

- Автоматизация процессов обучения и развертывания моделей
- Мониторинг производительности ML-систем
- Управление версиями моделей и данных
- Обеспечение CI/CD для ML-проектов
- Оптимизация вычислительных ресурсов

1.5 Перечень планируемых результатов обучения по дисциплине (модулю), соотнесенных с планируемыми результатами освоения образовательной программы

Изучение данной учебной дисциплины направлено на формирование у обучающихся следующих компетенций:

DL-1 (Э)
Способен применять и (или) разрабатывать архитектуры глубоких нейронных сетей

DL-1.6 Способен разрабатывать, адаптировать и внедрять генеративные нейронные сети для решения практических задач, включая создание новых архитектур, оптимизацию обучения и промышленное развертывание моделей	Разрабатывает новые архитектуры генеративных сетей, адаптивно применяет архитектуру VAE+GAN; разрабатывает капсульные сети
DL-1.7 Способен разрабатывать, оптимизировать и применять автоэнкодеры (AE) и вариационные автоэнкодеры (VAE) для решения задач снижения размерности, генерации данных и обнаружения аномалий, включая создание архитектур, обучение моделей и их внедрение в продуктивную среду	Применяет математические основы формирования пространства скрытых эмбеддингов; знает вероятностный характер и отличия естественного и искусственного генеративного процессов; Знает математические основы функционирования вероятностного автокодировщика; обосновывает применение дивергенции Кульбака-Лейблера через основное тождество автокодировщиков
DL-1.11 Способен применять, адаптировать и разрабатывать методы сжатия нейронных сетей для оптимизации производительности моделей, включая квантование, прунинг, дистилляцию и другие техники, с учетом требований к качеству и вычислительной эффективности.	Владеет аппаратом структурированного и неструктурированного прунинга, знает стратегии прореживания. Разрабатывает новые методы сжатия.
DL-1.12 Способен применять, адаптировать и разрабатывать методы дообучения нейронных сетей для эффективной адаптации моделей к новым задачам и доменам.	Владеет продвинутыми техниками (adapter layers, LoRA, prefix-tuning). Комбинирует различные стратегии адаптации. Работает с малыми датасетами (few-shot learning)

DL-2 (Э)

Способен применять и (или) разрабатывать современные архитектуры генеративных глубоких сетей

DL-2.1 Применяет известные архитектуры генеративных глубоких нейронных сетей для решения прикладной задачи (генерация текста, генерация изображений по тексту, синтез речи и т.д.), при необходимости проводя дообучение на наборах данных	Модифицирует архитектуры под специфические требования. Разрабатывает гибридные подходы (например, диффузионные модели + GAN). Оптимизирует архитектуры для целевых аппаратных платформ. Разрабатывает новые методы дообучения для генеративных моделей. Применяет few-shot/zero-shot learning техники. Реализует reinforcement learning для генерации.

LLM-1 (П)

Способен применять и (или) разрабатывать генеративные модели и БЯМ

LLM-1.1 Знает архитектуры генеративных моделей	Сравнивает архитектуры и выбирает подходящую под задачу
LLM-1.4 Понимает принципы генерации в мультимодальных моделях	Использует мультимодальные модели для captioning и tagging

LLM-2 (П)

Способен дообучать, адаптировать и оптимизировать генеративные модели под специфические задачи и условия применения

LLM-2.1 Понимает принципы fine-tune	Применяет fine-tune к предобученным моделям на новых датасетах

FC-1 (Б)

Способен проводить фронтирные исследования в области архитектур, алгоритмов МО, оптимизации и математики

FC-1.2 Разрабатывает новые архитектуры глубоких нейросетей	Знает передовые архитектуры в основных триадах: архитектура-данные-задача, принципы их построения, сильные и слабые стороны. Знает особенности наиболее часто встречающихся вычислителей, умеет подбирать архитектуры, адекватные особенностям вычислительных устройств.

FC-2 (Б)

Способен проводить фронтирные исследования в области фундаментальных и генеративных моделей

FC-2.3 Исследует и создает мульти-модальные большие языковые модели (LLM)	Дообучает готовые мультимодальные модели (Flamingo, LLaVA). Строит пайплайны согласования данных разных модальностей. Владеет техниками базового выравнивания модальностей через CLIP-подобные энкодеры. Оценивает качество через стандартные метрики (cross-modal retrieval accuracy)

2. Структура и содержание дисциплины

2.1 Распределение трудоёмкости дисциплины по видам работ

Общая трудоёмкость дисциплины составляет 4 зач. ед. (144 часа), их распределение по видам работ представлено в таблице

Вид учебной работы	Всего часов	Семестры (часы)

			7					
Контактная работа, в том числе:		**72,2**	**72,2**					
Аудиторные занятия (всего):		**68**	**68**					
Занятия лекционного типа		**34**	34					
Лабораторные занятия		**34**	34					
Занятия семинарского типа (семинары, практические занятия)								
Иная контактная работа:		**4,3**	**4,2**					
Контроль самостоятельной работы (КСР)		**4**	4					
Промежуточная аттестация (ИКР)		**0,2**	0,2					
Самостоятельная работа, в том числе:		**35,8**	**35,8**					
Выполнение заданий, проект		**35,8**	**35,8**					
Контроль:								
Подготовка к зачету		**7**	7					
Общая трудоемкость	**час.**	**108**	**108**					
	в том числе контактная работа	**72,2**	**72,2**					
	зач. ед	**3**	**3**					

2.2 Структура дисциплины

Распределение видов учебной работы и их трудоемкости по разделам дисциплины.

Разделы (темы) дисциплины, изучаемые в 7 семестре

№	Наименование разделов (тем)	Количество часов				
		Всего	Аудиторная работа			Внеаудиторная работа
			Л	ПЗ	ЛР	СРС
1	2	3	4	5	6	7
1.	Введение и основы	27,8	8		8	11,8
2.	Современные архитектуры и мультимодальность	42	14		14	14
3.	Оптимизация, адаптация и инженерия	36	12		12	12
ИТОГО по разделам дисциплины		**103,8**	**34**		**34**	**35,8**
Контроль самостоятельной работы (КСР)		4				
Промежуточная аттестация (ИКР)		0,2				
Подготовка к текущему контролю						
Общая трудоемкость по дисциплине		**108**				

Примечание: Л – лекции, ПЗ – практические занятия/семинары, ЛР – лабораторные занятия, СРС – самостоятельная работа студента

2.3 Содержание разделов (тем) дисциплины

2.3.1 Занятия лекционного типа

№	Наименование раздела (темы)	Содержание раздела (темы)	Форма текущего контроля
1	2	3	4
1.	Введение и основы	Введение в генеративные модели. Постановка проблемы генерации. Обзор семейств моделей: явные и неявные модели правдоподобия. GAN, VAE,	ЛР

№	Наименование раздела (темы)	Содержание раздела (темы)	Форма текущего контроля
1	2	3	4
		Autoregressive models, Normalizing Flows, Diffusion Models. Сравнительный анализ, сильные и слабые стороны. Математические основы: распределения, правдоподобие, KL-дивергенция.	
2.	Введение и основы	Вариационные автоэнкодеры (VAE). Архитектура Encoder-Decoder. Нижняя вариационная граница (ELBO). Принцип работы и математический вывод. Условные VAE (CVAE). Проблема "размытости" и методы борьбы с ней.	ЛР
3.	Введение и основы	Состязательные сети (GAN). Архитектура генератора и дискриминатора. Функция потерь min-max. Проблемы обучения GAN (исчезающие градиенты, коллапс мод). Методы стабилизации (Wasserstein GAN, Gradient Penalty, Spectral Normalization).	ЛР
4.	Введение и основы	Авторегрессионные и трансформерные модели для генерации. Принцип цепи Маркова в генерации текста. Архитектура Transformer (Self-Attention, Positional Encoding). Эволюция GPT-семейства. In-Context Learning. Сравнение Encoder-Decoder (T5) и Decoder-only (GPT) архитектур.	ЛР
5.	Современные архитектуры и мультимодальность	Диффузионные модели. Основы. Формулировка прямого и обратного процессов. Диффузионные вероятностные модели (DDPM). Алгоритмы обучения и сэмплинга.	ЛР
6.	Современные архитектуры и мультимодальность	Диффузионные модели. Ускорение и контроль. Ускорение сэмплинга (DDIM). Classifier-free guidance. Условная и контролируемая генерация. Модели Latent Diffusion (Stable Diffusion, VQ-VAE).	ЛР
7.	Современные архитектуры и мультимодальность	Мультимодальные генеративные модели. Раннее и позднее слияние. Cross-Modal Attention. Архитектуры CLIP, BLIP. Мультимодальные большие модели: Flamingo, LLaVA, Cosmos. Задачи контрастивного обучения и выравнивания модальностей.	ЛР
8.	Современные архитектуры и мультимодальность	Генерация видео. Особенности временной согласованности. Архитектуры на основе диффузионных моделей (Video Diffusion Models).	ЛР

№	Наименование раздела (темы)	Содержание раздела (темы)	Форма текущего контроля
1	2	3	4
		Трансформеры для видео. Применение для анимации.	
9.	Современные архитектуры и мультимодальность	Модели Qwen. Multimodal Diffusion Transformer (MMDiT).	ЛР
10.	Современные архитектуры и мультимодальность	Модели Wan. Mixture of experts. Архитектура, практическое использование для генерации анимации, реставрации хроники.	ЛР
11.	Современные архитектуры и мультимодальность	Модели ESRGAN, Real-ESRGAN. Задачи восстановления и апскейла видео.	ЛР
12.	Оптимизация, адаптация и инженерия	Методы эффективного дообучения (Parameter-Efficient Fine-Tuning - PEFT). Adapter Layers, Prefix-Tuning, Prompt Tuning. LoRA (Low-Rank Adaptation) и QLoRA: математическое обоснование, преимущества, практические сценарии использования.	ЛР
13.	Оптимизация, адаптация и инженерия	Сжатие и ускорение нейронных сетей. Принципы прунинга (весовой, структурный). Квантование (пост-тренировочное, квантизация-aware training). Дистилляция знаний (Knowledge Distillation).	ЛР
14.	Оптимизация, адаптация и инженерия	Обучение с подкреплением для генерации (RLHF). Применение RL для оптимизации генерации под недифференцируемые метрики. Пайплайн RLHF для выравнивания языковых моделей (на примере ChatGPT). Проблемы и вызовы.	ЛР
15.	Оптимизация, адаптация и инженерия	Оценка качества генеративных моделей. Метрики для изображений: FID, IS, Precision/Recall. Метрики для текста: Perplexity, BLEU, ROUGE. Человеко-ориентированная оценка. Проблема оценки креативности и разнообразия.	ЛР
16.	Оптимизация, адаптация и инженерия	Этика, безопасность и легальность генеративных моделей. Проблемы bias, fairness, глубоких подделок (deepfakes). Методы обнаружения сгенерированного контента. Техники повышения безопасности и надежности моделей (согласование, красные команды).	ЛР
17.	Оптимизация, адаптация и инженерия	Промышленное внедрение и фронтир исследований. Проектирование архитектуры ИС с генеративными моделями. Системы кеширования, A/B тестирование. Обзор текущих трендов	

№	Наименование раздела (темы)	Содержание раздела (темы)	Форма текущего контроля
1	2	3	4
		(World Models, Generative AI в промышленности и науке).	

Примечание: ЛР – отчет/защита лабораторной работы, КП - выполнение курсового проекта, КР - курсовой работы, РГЗ - расчетно-графического задания, Р - написание реферата, Э - эссе, К - коллоквиум, Т – тестирование, РЗ – решение задач.

2.3.2 Занятия семинарского типа

Не предусмотрены

2.3.3 Лабораторные занятия

№	Наименование раздела (темы)	Наименование лабораторных работ	Форма текущего контроля
1	2	3	4
1.	Введение и основы	Настройка окружения conda и основы PyTorch для генеративных задач. Работа с Hugging Face.	ЛР
2.	Введение и основы	Реализация и обучение VAE на датасете MNIST/FashionMNIST.	ЛР
3.	Введение и основы	Реализация и обучение DCGAN на датасете CIFAR-10.	ЛР
4.	Введение и основы	Интеграция генеративных моделей в веб-сервис с использованием FastAPI.	ЛР
5.	Современные архитектуры и мультимодальность	Визуальное программирование пайплайнов генерации в ComfyUI: основы.	ЛР
6.	Современные архитектуры и мультимодальность	Создание сложных workflow в ComfyUI: использование ControlNet, работа с нодами. Ablation studies.	ЛР
7.	Современные архитектуры и мультимодальность	Fine-tuning предобученной GPT-модели для генерации текста.	ЛР
8.	Современные архитектуры и мультимодальность	Применение LoRA для дообучения Stable Diffusion на собственном датасете.	ЛР
9.	Современные архитектуры и мультимодальность	Работа с диффузионными моделями через Hugging Face Diffusers: безусловная и условная генерация.	ЛР
10.	Современные архитектуры и мультимодальность	Мультимодальность: использование CLIP для семантического поиска и улучшения генерации.	ЛР
11.	Современные архитектуры и мультимодальность	Работа с моделями повышения качества изображений: Real-ESRGAN.	ЛР
12.	Оптимизация, адаптация и инженерия	Практика прунинга и квантования.	ЛР
13.	Оптимизация, адаптация и инженерия	Практика дистилляции: сжатие языковой модели.	ЛР
14.	Оптимизация, адаптация и инженерия	Комплексная оценка качества генеративной модели: расчет FID, субъективная оценка.	ЛР
15.	Оптимизация, адаптация и инженерия	Дообучение мультимодальной модели LLaVA на собственном датасете.	ЛР

№	Наименование раздела (темы)	Наименование лабораторных работ	Форма текущего контроля
1	2	3	4
16.	Оптимизация, адаптация и инженерия	RLHF в RL4LMs.	ЛР
17.	Оптимизация, адаптация и инженерия	Fine-tuning модели Qwen с использованием LoRA на собственном датасете в ComfyUI. Logit lens.	ЛР

Примечание: ЛР – отчет/защита лабораторной работы, КП - выполнение курсового проекта, КР - курсовой работы, РГЗ - расчетно-графического задания, Р - написание реферата, Э - эссе, К - коллоквиум, Т – тестирование, РЗ – решение задач.

2.3.4 Примерная тематика курсовых работ (проектов)

Курсовая работа не предусмотрена. В качестве курсового проекта студенты защищают инфраструктуру проекта веб-приложения с использованием ML по заданию от индустриального партнера.

2.4 Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине (модулю)

№	Вид СРС	Перечень учебно-методического обеспечения дисциплины по выполнению самостоятельной работы
1	2	3
1	Изучение теоретического материала	Методические указания по организации самостоятельной работы студентов, утвержденные кафедрой информационных технологий, протокол №1 от 30.08.2019
2	Решение задач	Методические указания по организации самостоятельной работы студентов, утвержденные кафедрой информационных технологий, протокол №1 от 30.08.2019

Учебно-методические материалы для самостоятельной работы обучающихся из числа инвалидов и лиц с ограниченными возможностями здоровья (ОВЗ) предоставляются в формах, адаптированных к ограничениям их здоровья и восприятия информации:

Для лиц с нарушениями зрения:

– в печатной форме увеличенным шрифтом,

– в форме электронного документа,

– в форме аудиофайла,

– в печатной форме на языке Брайля.

Для лиц с нарушениями слуха:

– в печатной форме,

– в форме электронного документа.

Для лиц с нарушениями опорно-двигательного аппарата:

– в печатной форме,

– в форме электронного документа,

– в форме аудиофайла.

Данный перечень может быть конкретизирован в зависимости от контингента обучающихся.

3. Образовательные технологии

В соответствии с требованиями ФГОС в программа дисциплины предусматривает использование в учебном процессе следующих образовательные технологии: чтение

лекций с использованием мультимедийных технологий; метод малых групп, разбор практических задач и кейсов.

При обучении используются следующие образовательные технологии:

- Технология коммуникативного обучения – направлена на формирование коммуникативной компетентности студентов, которая является базовой, необходимой для адаптации к современным условиям межкультурной коммуникации.
- Технология разноуровневого (дифференцированного) обучения – предполагает осуществление познавательной деятельности студентов с учётом их индивидуальных способностей, возможностей и интересов, поощряя их реализовывать свой творческий потенциал. Создание и использование диагностических тестов является неотъемлемой частью данной технологии.
- Технология модульного обучения – предусматривает деление содержания дисциплины на достаточно автономные разделы (модули), интегрированные в общий курс.
- Информационно-коммуникационные технологии (ИКТ) - расширяют рамки образовательного процесса, повышая его практическую направленность, способствуют интенсификации самостоятельной работы учащихся и повышению познавательной активности. В рамках ИКТ выделяются 2 вида технологий:
- Технология использования компьютерных программ – позволяет эффективно дополнить процесс обучения языку на всех уровнях.
- Интернет-технологии – предоставляют широкие возможности для поиска информации, разработки научных проектов, ведения научных исследований.
- Технология индивидуализации обучения – помогает реализовывать личностно-ориентированный подход, учитывая индивидуальные особенности и потребности учащихся.
- Проектная технология – ориентирована на моделирование социального взаимодействия учащихся с целью решения задачи, которая определяется в рамках профессиональной подготовки, выделяя ту или иную предметную область.
- Технология обучения в сотрудничестве – реализует идею взаимного обучения, осуществляя как индивидуальную, так и коллективную ответственность за решение учебных задач.
- Игровая технология – позволяет развивать навыки рассмотрения ряда возможных способов решения проблем, активизируя мышление студентов и раскрывая личностный потенциал каждого учащегося.
- Технология развития критического мышления – способствует формированию разносторонней личности, способной критически относиться к информации, умению отбирать информацию для решения поставленной задачи.

Комплексное использование в учебном процессе всех вышеназванных технологий стимулируют личностную, интеллектуальную активность, развивают познавательные процессы, способствуют формированию компетенций, которыми должен обладать будущий специалист.

Основные виды интерактивных образовательных технологий включают в себя:

- работа в малых группах (команде) - совместная деятельность студентов в группе под руководством лидера, направленная на решение общей задачи путём творческого сложения результатов индивидуальной работы членов команды с делением полномочий и ответственности;
- проектная технология - индивидуальная или коллективная деятельность по отбору, распределению и систематизации материала по определенной теме, в результате которой составляется проект;
- анализ конкретных ситуаций - анализ реальных проблемных ситуаций, имевших место в соответствующей области профессиональной деятельности, и поиск вариантов лучших решений;

– развитие критического мышления – образовательная деятельность, направленная на развитие у студентов разумного, рефлексивного мышления, способного выдвинуть новые идеи и увидеть новые возможности.

Подход разбора конкретных задач и ситуаций широко используется как преподавателем, так и студентами во время лекций, лабораторных занятий и анализа результатов самостоятельной работы. Это обусловлено тем, что при исследовании и решении каждой конкретной задачи имеется, как правило, несколько методов, а это требует разбора и оценки целой совокупности конкретных ситуаций.

Семестр	Вид занятия	Используемые интерактивные образовательные технологии	количество интерактивных часов
7	ЛР	Практические занятия в режимах взаимодействия «преподаватель – студент» и «студент – студент»	46
Итого			**46**

Примечание: Л – лекции, ПЗ – практические занятия/семинары, ЛР – лабораторные занятия, СРС – самостоятельная работа студента

Темы, задания и вопросы для самостоятельной работы призваны сформировать навыки поиска информации, умения самостоятельно расширять и углублять знания, полученные в ходе лекционных и практических занятий.

Подход разбора конкретных ситуаций широко используется как преподавателем, так и студентами при проведении анализа результатов самостоятельной работы.

Для лиц с ограниченными возможностями здоровья предусмотрена организация консультаций с использованием электронной почты.

Для лиц с нарушениями зрения:

– в печатной форме увеличенным шрифтом,

– в форме электронного документа.

Для лиц с нарушениями слуха:

– в печатной форме,

– в форме электронного документа.

Для лиц с нарушениями опорно-двигательного аппарата:

– в печатной форме,

– в форме электронного документа.

Для лиц с ограниченными возможностями здоровья предусмотрена организация консультаций с использованием электронной почты.

Данный перечень может быть конкретизирован в зависимости от контингента обучающихся.

4. Оценочные и методические материалы

4.1 Оценочные средства для текущего контроля успеваемости и промежуточной аттестации

Оценочные средства предназначены для контроля и оценки образовательных достижений обучающихся, освоивших программу учебной дисциплины «DataOps & ML Ops».

Оценочные средства включает контрольные материалы для проведения **текущего контроля** в форме оценки лабораторных работ к проекта к **экзамену**.

Оценочные средства для инвалидов и лиц с ограниченными возможностями здоровья выбираются с учетом их индивидуальных психофизических особенностей.

– при необходимости инвалидам и лицам с ограниченными возможностями здоровья предоставляется дополнительное время для подготовки ответа на экзамене;

– при проведении процедуры оценивания результатов обучения инвалидов и лиц с ограниченными возможностями здоровья предусматривается использование технических средств, необходимых им в связи с их индивидуальными особенностями;

– при необходимости для обучающихся с ограниченными возможностями здоровья и инвалидов процедура оценивания результатов обучения по дисциплине может проводиться в несколько этапов.

Процедура оценивания результатов обучения инвалидов и лиц с ограниченными возможностями здоровья по дисциплине (модулю) предусматривает предоставление информации в формах, адаптированных к ограничениям их здоровья и восприятия информации:

Для лиц с нарушениями зрения:

– в печатной форме увеличенным шрифтом,

– в форме электронного документа.

Для лиц с нарушениями слуха:

– в печатной форме,

– в форме электронного документа.

Для лиц с нарушениями опорно-двигательного аппарата:

– в печатной форме,

– в форме электронного документа.

Данный перечень может быть конкретизирован в зависимости от контингента обучающихся.

Структура оценочных средств для текущей и промежуточной аттестации

№ п/п	Контролируемые разделы (темы) дисциплины*	Код контролируемой компетенции (или ее части)	Наименование оценочного средства	
			Текущий контроль	Промежуточная аттестация
1	Введение и основы	LLM-1	*ЛР 1-4*	*Проект с кейсом от индустриального партнера*
2	Современные архитектуры и мультимодальность	LLM-1, LLM-2	*ЛР 5-11*	*Проект с кейсом от индустриального партнера*
3	Оптимизация, адаптация и инженерия	DL-1, DL-2, FC-1, FC-2	*ЛР 12-17*	*Проект с кейсом от индустриального партнера*

Показатели, критерии и шкала оценки сформированных компетенций

Соответствие **<u>пороговому уровню</u>** освоения компетенций планируемым результатам обучения и критериям их оценивания (оценка: **зачтено**):

DL-1 (Э)

Способен применять и (или) разрабатывать архитектуры глубоких нейронных сетей

DL-1.6 Способен разрабатывать, адаптировать и внедрять генеративные нейронные сети для решения практических задач, включая создание новых архитектур, оптимизацию	Разрабатывает новые архитектуры генеративных сетей, адаптивно применяет архитектуру VAE+GAN; разрабатывает капсульные сети

обучения и промышленное развертывание моделей	
DL-1.7 Способен разрабатывать, оптимизировать и применять автоэнкодеры (AE) и вариационные автоэнкодеры (VAE) для решения задач снижения размерности, генерации данных и обнаружения аномалий, включая создание архитектур, обучение моделей и их внедрение в продуктивную среду	Применяет математические основы формирования пространства скрытых эмбеддингов; знает вероятностный характер и отличия естественного и искусственного генеративного процессов; Знает математические основы функционирования вероятностного автокодировщика; обосновывает применение дивергенции Кульбака-Лейблера через основное тождество автокодировщиков
DL-1.11 Способен применять, адаптировать и разрабатывать методы сжатия нейронных сетей для оптимизации производительности моделей, включая квантование, прунинг, дистилляцию и другие техники, с учетом требований к качеству и вычислительной эффективности.	Владеет аппаратом структурированного и неструктурированного прунинга, знает стратегии прореживания. Разрабатывает новые методы сжатия.
DL-1.12 Способен применять, адаптировать и разрабатывать методы дообучения нейронных сетей для эффективной адаптации моделей к новым задачам и доменам.	Владеет продвинутыми техниками (adapter layers, LoRA, prefix-tuning). Комбинирует различные стратегии адаптации. Работает с малыми датасетами (few-shot learning)
DL-2 (Э)	
Способен применять и (или) разрабатывать современные архитектуры генеративных глубоких сетей	
DL-2.1 Применяет известные архитектуры генеративных глубоких нейронных сетей для решения прикладной задачи (генерация текста, генерация изображений по тексту, синтез речи и т.д.), при необходимости проводя дообучение на наборах данных	Модифицирует архитектуры под специфические требования. Разрабатывает гибридные подходы (например, диффузионные модели + GAN). Оптимизирует архитектуры для целевых аппаратных платформ. Разрабатывает новые методы дообучения для генеративных моделей. Применяет few-shot/zero-shot learning техники. Реализует reinforcement learning для генерации.
LLM-1 (П)	
Способен применять и (или) разрабатывать генеративные модели и БЯМ	
LLM-1.1 Знает архитектуры генеративных моделей	Сравнивает архитектуры и выбирает подходящую под задачу
LLM-1.4 Понимает принципы генерации в мультимодальных моделях	Использует мультимодальные модели для captioning и tagging
LLM-2 (П)	
Способен дообучать, адаптировать и оптимизировать генеративные модели под специфические задачи и условия применения	
LLM-2.1 Понимает принципы fine-tune	Применяет fine-tune к предобученным моделям на новых датасетах
FC-1 (Б)	
Способен проводить фронтирные исследования в области архитектур, алгоритмов МО, оптимизации и математики	

FC-1.2 Разрабатывает новые архитектуры глубоких нейросетей	Знает передовые архитектуры в основных триадах: архитектура-данные-задача, принципы их построения, сильные и слабые стороны. Знает особенности наиболее часто встречающихся вычислителей, умеет подбирать архитектуры, адекватные особенностям вычислительных устройств.

FC-2 (Б)
Способен проводить фронтирные исследования в области фундаментальных и генеративных моделей

FC-2.3 Исследует и создает мульти-модальные большие языковые модели (LLM)	Дообучает готовые мультимодальные модели (Flamingo, LLaVA). Строит пайплайны согласования данных разных модальностей. Владеет техниками базового выравнивания модальностей через CLIP-подобные энкодеры. Оценивает качество через стандартные метрики (cross-modal retrieval accuracy)

Типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций в процессе освоения образовательной программы

Пример лабораторной работы

Лабораторная работа №1: Настройка окружения и введение в PyTorch для генеративных задач

Цель работы:

Сформировать базовые практические навыки работы с инструментарием для разработки генеративных нейронных сетей, включая создание изолированных окружений, освоение основных возможностей PyTorch и знакомство с экосистемой Hugging Face как фундамента для последующей работы с современными генеративными моделями.

Задачи:

Изучить принципы работы с системами управления окружениями (Conda)
Настроить изолированное Python-окружение с помощью Conda
Ознакомиться с организацией репозитория моделей Hugging Face Hub
Научиться работать с предобученными моделями через Hugging Face Hub с помощью PyTorch

Ожидаемые результаты:
Генератор датасета .

Ход работы

1. Установите conda, создайте виртуальное окружение и установите необходимые пакеты:

```
# Установите conda
wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh -O ~/miniconda3/miniconda.sh

bash ~/miniconda3/miniconda.sh -b -u -p ~/miniconda3

rm ~/miniconda3/miniconda.sh
```

source ~/miniconda3/bin/activate

conda init --all

```
# Проверить установку Conda
conda --version

# Обновить Conda при необходимости
conda update -n base -c defaults conda

# Создать новое окружение с Python 3.12

conda create -n generative-ai python=3.12

# Активировать окружение
conda activate generative-ai

# Установить основные пакеты
conda install pytorch torchvision torchaudio pytorch-cuda=12.1 -c pytorch -c
nvidia

conda install matplotlib numpy pandas transformers datasets huggingface_hub

# Фиксим баг libtorch_cpu.so: undefined symbol: iJIT_NotifyEvent на AMD CPU
conda install mkl=2023.1.0


# test_environment.py
import torch
import transformers
print(f"PyTorch version: {torch.__version__}")
print(f"CUDA available: {torch.cuda.is_available()}")
print(f"Transformers version: {transformers.__version__}")
```

2. Протестируйте скачивание и работу LLM с Hugging Face:

```
# hf.py
import torch
from huggingface_hub import list_models, list_datasets
from transformers import pipeline

# Поиск моделей для генеративных задач
generative_models = list_models(filter="text-generation", limit=5)
print("Популярные модели для генерации текста:")
for model in generative_models:
    print(f"- {model.modelId}")

# Загрузка и использование готовой модели
generator = pipeline('text-generation',
                     model='distilgpt2',
                     device=0 if torch.cuda.is_available() else -1)

# Генерация текста
result = generator("The future of artificial intelligence",
                   max_length=50,
                   num_return_sequences=1)
print("Сгенерированный текст:")
print(result[0]['generated_text'])

# datasets.py
from datasets import load_dataset
```

```
# Просмотр доступных датасетов
text_datasets = list_datasets(filter="task_categories:text-generation",
limit=3)
print("Датасеты для генерации текста:")
for dataset in text_datasets:
    print(f"- {dataset.id}")

# Загрузка небольшого датасета для экспериментов
try:
    dataset = load_dataset("imdb", split="train[:100]")  # первые 100
примеров
    print(f"Размер датасета: {len(dataset)}")
    print(f"Пример текста: {dataset[0]['text'][:200]}...")
except Exception as e:
    print(f"Ошибка загрузки датасета: {e}")

#generator.py

def simple_text_generator(model_name, prompt, max_length=100):
    """Простая функция для генерации текста"""
    try:
        # Создание пайплайна
        generator = pipeline('text-generation',
                             model=model_name,
                             device=0 if torch.cuda.is_available() else -1)

        # Генерация
        results = generator(prompt,
                            max_length=max_length,
                            num_return_sequences=1,
                            temperature=0.7,
                            do_sample=True)

        return results[0]['generated_text']

    except Exception as e:
        return f"Ошибка: {e}"

# Тестирование на разных промптах
prompts = [
    "In a distant future where AI",
    "The secret to understanding neural networks is",
    "When I started learning deep learning"
]

for prompt in prompts:
    generated = simple_text_generator('distilgpt2', prompt)
    print(f"Промпт: {prompt}")
    print(f"Результат: {generated}")
    print("-" * 50)
```

3. Составьте отчет и загрузите в Moodle.

Требования к отчету

1. Титульный лист

Название работы, ФИО студента, группа, дата.

2. Введение

Цель и задачи работы.

Описание приложения.

3. Теоретическая часть
Кратко описать суть работы.

4. Реализация
Код на Python.

5. Результаты
Скриншоты работы .

6. Выводы
Проблемы, возникшие при выполнении задания.

Критерии оценки
Зачтено: Окружение настроено, протестирована работа модели с HuggingFace.
Не зачтено: Окружение не настроено или модель не работает.

Рекомендуемые инструменты
- Python.

Зачетно-экзаменационные материалы для промежуточной аттестации (зачет)

Описание командного проекта

Название: "Разработка и развертывание системы с использованием генеративного ИИ"

Цель: Спроектировать, реализовать и развернуть веб-приложение с использованием генеративного ИИ.

Задачи команды (3-4 человека):

1. Анализ предметной области и проектирование архитектуры:
- Выбрать предметную область (см. задачи индустриальных партнеров).
- Сформулировать требования к системе: функциональные (точность, latency) и нефункциональные (воспроизводимость, масштабируемость).
- Спроектировать архитектуру системы с использованием диаграмм (UML), выделив компоненты для сбора данных, их проверки, обучения, сервинга и мониторинга.
- Обосновать выбор стека технологий для каждого компонента.

2. Разработка и настройка инфраструктуры:
- Создать GitLab-репозиторий с правильной структурой папок (data, models, src, pipelines).
- Настроить DVC для версионирования данных и моделей, интегрировав его с удаленным хранилищем (Yandex Cloud).
- Реализовать конвейер CI/CD в GitLab, который:
- Запускает тесты кода и данных при каждом коммите.
- Запускает тренировочный пайплайн при изменении данных или кода модели (Continuous Training).

3. Реализация ML-пайплайна и мониторинг:

- Реализовать скрипты для сбора и симуляции потока новых данных, fine-тюнинга модели.
- Реализовать скрипт проверки качества новых данных (Data Validation) и детектирования дрейфа концепций (Concept Drift).

4. Разработка веб-приложения:
- Реализовать веб-приложение, использующее разработанную генеративную модель.

Результат: Команда представляет работающий прототип веб-приложения с использованием генеративного ИИ, документацию по архитектуре, репозиторий с кодом и пайплайнами, а также отчет по результатам работы системы в режиме Continuous Training.

Перечень компетенций (части компетенции), проверяемых оценочным средством
LLM-1, LLM-2, DL-1, DL-2

4.2 Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций

Методические рекомендации, определяющие процедуры оценивания на экзамене:

Процедура промежуточной аттестации проходит в соответствии с Положением о текущем контроле и промежуточной аттестации обучающихся ФГБОУ ВО «КубГУ».

Итоговой формой контроля сформированности компетенций у обучающихся по дисциплине является зачет. Студенты обязаны сдать зачет в соответствии с расписанием и учебным планом.

ФОС промежуточной аттестации состоит из заданий и результатов текущего контроля.

Форма проведения экзамена: устно (защита проекта).

Преподавателю предоставляется право задавать студентам дополнительные вопросы по всей учебной программе дисциплины.

Результат сдачи зачета заносится преподавателем в экзаменационную ведомость и зачетную книжку.

Оценивание уровня освоения дисциплины основывается на качестве выполнения студентом заданий текущего контроля и проекта.

Критерии оценки:

Зачтено – выполнено 60% лабораторных работ и выполнен проект.

Не удовлетворительно – выполнено менее 60% лабораторных работ или не выполнен проект.

Методические рекомендации, определяющие процедуры оценивания лабораторных работ:

Процедура оценивания лабораторных работ проходит в соответствии с Положением о текущем контроле и промежуточной аттестации обучающихся ФГБОУ ВО «КубГУ».

По каждой лабораторной работе оформляется отчет. Отчеты сдаются на проверку руководителю в течение курса по мере их выполнения, и защищаются студентами в установленном порядке.

При защите отчета студенту могут быть заданы вопросы и дополнительные задания по сути лабораторной работы, в том числе из списка контрольных вопросов к данной лабораторной работе. При неудовлетворительной оценке знаний студента по теме данного отчета, студент возвращается к повторному изучению соответствующих материалов, после

чего допускается к повторной защите. Неудовлетворительно выполненный отчет также возвращается на доработку.

Отчет должен содержать заголовок, тему лабораторной работы, цель, задание, индивидуальную тему, описание хода выполнения работы, необходимые прикладные материалы (схемы, макеты документов и т.п.), в соответствии с требованиями к содержанию, и выводы по работе.

Оценочные средства для инвалидов и лиц с ограниченными возможностями здоровья выбираются с учетом их индивидуальных психофизических особенностей.

– при необходимости инвалидам и лицам с ограниченными возможностями здоровья предоставляется дополнительное время для подготовки ответа на экзамене;

– при проведении процедуры оценивания результатов обучения инвалидов и лиц с ограниченными возможностями здоровья предусматривается использование технических средств, необходимых им в связи с их индивидуальными особенностями;

– при необходимости для обучающихся с ограниченными возможностями здоровья и инвалидов процедура оценивания результатов обучения по дисциплине может проводиться в несколько этапов.

Процедура оценивания результатов обучения инвалидов и лиц с ограниченными возможностями здоровья по дисциплине предусматривает предоставление информации в формах, адаптированных к ограничениям их здоровья и восприятия информации:

Для лиц с нарушениями зрения:

– в печатной форме увеличенным шрифтом,

– в форме электронного документа.

Для лиц с нарушениями слуха:

– в печатной форме,

– в форме электронного документа.

Для лиц с нарушениями опорно-двигательного аппарата:

– в печатной форме,

– в форме электронного документа.

Данный перечень может быть конкретизирован в зависимости от контингента обучающихся.

4.3 Методические указания по организации лабораторных работ по дисциплине "Генеративный ИИ"

1. Общие сведения

Образовательная программа: «Современные методы машинного обучения и компьютерное зрение».
Дисциплина: "Генеративный искусственный интеллект".
Вид обеспечения: Проведение лабораторных работ.
Условия применения:
Для успешного выполнения лабораторных работ требуется:

Программное обеспечение:

- **Язык программирования и окружение:** Python, Jupyter Notebook/Lab, Google Colab Pro / Kaggle Notebooks.
- **Основные фреймворки глубокого обучения:** PyTorch (предпочтительно для исследований), TensorFlow/Keras.
- **Специализированные библиотеки для генеративных моделей:**

- **Diffusion-модели:** `diffusers` (Hugging Face), `denoising-diffusion-pytorch`.
- **GANs:** PyTorch GAN Zoo, `pytorch-gan`.
- **VAE и авторегрессивные модели:** Стандартные реализации в PyTorch/TensorFlow.
- **Трансформеры для генерации:** Hugging Face `transformers` (GPT, T5, CLIP).

- Библиотеки для работы с данными и метрик: Pillow, OpenCV, `torchvision`, `lpips` (метрика схожести), `clean-fid` (FID метрика).
- Инструменты для экспериментов: Weights & Biases (W&B), MLflow.
- Оптимизация и инференс: ONNX Runtime, TensorRT (для продвинутого инференса).

Аппаратное обеспечение:

- Обязательно: Доступ к GPU с достаточным объемом памяти (не менее 8 ГБ, рекомендуется 16+ ГБ). Для обучения базовых диффузионных моделей и больших GAN/VAE требуется значительная память.
- Рекомендуется: Высокопроизводительные виртуальные машины с GPU (NVIDIA A100/V100) через облачные сервисы.

Облачная инфраструктура и данные:

- Платформы для вычислений: **Yandex DataSphere (с GPU), Google Colab Pro, Lambda Labs, RunPod.**
- Хранилища данных: **Yandex Object Storage / Amazon S3 для хранения датасетов и чекпоинтов моделей.**
- Датасеты: **Доступ к публичным датасетам (CelebA-HQ, FFHQ, LAION, COCO, WikiArt) и возможность загрузки пользовательских данных.**
- Платформы для развертывания: **Hugging Face Spaces, Replicate (для демонстрации моделей).**

2. Цели, задачи и ожидаемые результаты

Цель:
Сформировать у студентов глубокое практическое понимание архитектур, принципов обучения, применения и оценки современных генеративных моделей. Научить создавать системы для синтеза, трансформации и управления контентом (изображения, текст, мультимодальные данные) с контролируемыми свойствами.

Задачи:

1. **Работа с базовыми архитектурами:**

- Реализация и обучение моделей Generative Adversarial Networks (GAN): DCGAN, StyleGAN (упрощенная версия). Анализ режима коллапса и методов его стабилизации.
- Реализация и обучение Variational Autoencoders (VAE) для генерации и латентных представлений. Исследование латентного пространства.

2. **Диффузионные модели:**

- Реализация базового пайплайна диффузии (прямой/обратный процесс) для генерации изображений.
- Практическое применение современных диффузионных фреймворков (diffusers): текстово-изображенческая генерация (Stable Diffusion), Inpainting, контроль генезера через ControlNet, fine-tuning (LoRA, Dreambooth).

3. **Трансформеры в генеративных задачах:**

- Использование предобученных языковых моделей (GPT-2, LLaMA) для генерации и продолжения текста.
- Работа с мультимодальными моделями (CLIP): использование для управления диффузионными моделями (текст/изображение -> изображение), оценка схожести.

4. **Оценка, управление и оптимизация:**

- Расчет и интерпретация метрик качества генерации: FID (Frechet Inception Distance), IS (Inception Score), Precision/Recall для распределений, CLIP Score.
- Техники управления атрибутами в латентном пространстве GAN/VAE (latent space arithmetic, StyleGAN mixing). Интерполяция в латентном пространстве.
- Оптимизация моделей для эффективного инференса: квантизация, pruning, использование специализированных движков.

5. **Прикладные и интеграционные задачи:**

- **Создание end-to-end пайплайна для решения прикладной задачи: например, генерация дизайна по описанию, аугментация датасетов, реставрация изображений (super-resolution, inpainting).**
- **Разработка простого веб-интерфейса (Gradio, Streamlit) для интерактивного взаимодействия с обученной генеративной моделью.**

Ожидаемые результаты:
После выполнения лабораторных работ студенты смогут:

- Изучить математические основы, архитектуры и алгоритмы обучения современных генеративных моделей (VAE, GAN, Diffusion Models, Autoregressive Models).
- Освоить инструментарий для работы с генеративными моделями (PyTorch, Hugging Face, Diffusers, ComfyUI) и современные техники их адаптации (Fine-Tuning, LoRA, QLoRA, Дистилляция) и оптимизации (Прунинг, Квантование).
- Сформировать навыки полного цикла разработки: от анализа требований и подготовки данных до обучения, оценки, оптимизации и развертывания генеративных моделей в составе информационных систем.
- Развить способность к критическому анализу научных статей, постановке и проведению экспериментов, а также к генерации идей для модификации и создания новых архитектур генеративных сетей.
- Проводить комплексную оценку качества генеративных моделей, используя как стандартные метрики, так и качественный анализ.
- Управлять выходом генеративных моделей через текстовые промпты, латентные векторы и контрольные сигналы (ControlNet).
- Создавать прототипы приложений на основе генеративного ИИ с интерактивным интерфейсом.

– Понимать и учитывать ключевые практические аспекты: вычислительные затраты, этические риски (deepfakes, bias) и ограничения современных генеративных моделей.

5. Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины (модуля)

5.1 Основная литература:

1 Генеративное глубокое обучение. Как не мы рисуем картины, пишем романы и музыку. 2-е межд изд. Фостер Д. O'Relly, 2024.

2 OpenAI API Documentation– URL: https://platform.openai.com/docs

3 Speech and Language Processing – 3rd ed. – Pearson, 2024. – 1024 p.

4 Кревецкий, А. В. Основы технологий искусственного интеллекта : учебное пособие : [16+] / А. В. Кревецкий, Ю. А. Ипатов, Н. И. Роженцова ; под общ. ред. А. В. Кревецкого ; Поволжский государственный технологический университет. – Йошкар-Ола : Поволжский государственный технологический университет, 2023. – 272 с. : ил., табл., схем. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=714624 (дата обращения: 13.12.2025). – Библиогр.: с. 264-267. – ISBN 978-5-8158-2358-7. – Текст : электронный.

5 Обухов, А. Д. Системный анализ и обработка информации в интеллектуальных системах : учебное пособие / А. Д. Обухов, И. Л. Коробова ; Тамбовский государственный технический университет. – Тамбов : Тамбовский государственный технический университет (ТГТУ), 2020. – 81 с. : ил. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=720763 (дата обращения: 13.12.2025). – Библиогр. в кн. – ISBN 978-5-8265-2217-2. – Текст : электронный.

6 Generative Adversarial Neural Architecture Search. Seyed Saeed Changiz Rezaei, Fred X. Han, Di Niu, Mohammad Salameh, Keith Mills, Shuo Lian, Wei Lu, Shangling Jui// Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence Main Track. Pages 2227-2234. https://doi.org/10.24963/ijcai.2021/307

5.2 Дополнительная литература:

1 Generative AI for Cloud Solutions. Paul Singh, Anurag Karuparti. Packt Publishing, 2024.

2 Deep Learning – MIT Press, 2016.

3 Radford, A., Kim, J.W., Hallacy, C., et al. Learning Transferable Visual Models From Natural Language Supervision // ICML. – 2021.

4 Документация PyTorch.

5.3. Периодические издания:

1 Базы данных компании «Ист Вью» http://dlib.eastview.com

2 Электронная библиотека GREBENNIKON.RU https://grebennikon.ru/

5.4. Интернет-ресурсы, в том числе современные профессиональные базы данных и информационные справочные системы

Электронно-библиотечные системы (ЭБС):

1 ЭБС «ЮРАЙТ» https://urait.ru/

2 ЭБС «УНИВЕРСИТЕТСКАЯ БИБЛИОТЕКА ОНЛАЙН» http://www.biblioclub.ru/

3 ЭБС «BOOK.ru» https://www.book.ru

4 ЭБС «ZNANIUM.COM» www.znanium.com

5 ЭБС «ЛАНЬ» https://e.lanbook.com

Профессиональные базы данных

1 Scopus http://www.scopus.com/

2 ScienceDirect https://www.sciencedirect.com/
3 Журналы издательства Wiley https://onlinelibrary.wiley.com/
4 Научная электронная библиотека (НЭБ) http://www.elibrary.ru/
5 Полнотекстовые архивы ведущих западных научных журналов на Российской платформе научных журналов НЭИКОН http://archive.neicon.ru
6 Национальная электронная библиотека (доступ к Электронной библиотеке диссертаций Российской государственной библиотеки (РГБ) https://rusneb.ru/
7 Президентская библиотека им. Б.Н. Ельцина https://www.prlib.ru/
8 База данных CSD Кембриджского центра кристаллографических данных (CCDC) https://www.ccdc.cam.ac.uk/structures/
9 Springer Journals: https://link.springer.com/
10 Springer Journals Archive: https://link.springer.com/
11 Nature Journals: https://www.nature.com/
12 Springer Nature Protocols and Methods: https://experiments.springernature.com/sources/springer-protocols
13 Springer Materials: http://materials.springer.com/
14 Nano Database: https://nano.nature.com/
15 Springer eBooks (i.e. 2020 eBook collections): https://link.springer.com/
16 "Лекториум ТВ" http://www.lektorium.tv/
17 Университетская информационная система РОССИЯ http://uisrussia.msu.ru

Бесплатные образовательные ресурсы

1 Jupyter Notebook – интерактивные вычисления
2 Visual Studio Code – редактор кода с поддержкой Python
3 Google Scholar/arXiv – доступ к научным публикациям

Ресурсы свободного доступа

1 КиберЛенинка http://cyberleninka.ru/;
2 Американская патентная база данных http://www.uspto.gov/patft/
3 Министерство науки и высшего образования Российской Федерации https://www.minobrnauki.gov.ru/;
4 Федеральный портал "Российское образование" http://www.edu.ru/;
5 Информационная система "Единое окно доступа к образовательным ресурсам" http://window.edu.ru/;
6 Единая коллекция цифровых образовательных ресурсов http://school-collection.edu.ru/ .
7 Проект Государственного института русского языка имени А.С. Пушкина "Образование на русском" https://pushkininstitute.ru/;
8 Справочно-информационный портал "Русский язык" http://gramota.ru/;
9 Служба тематических толковых словарей http://www.glossary.ru/;
10 Словари и энциклопедии http://dic.academic.ru/;
11 Образовательный портал "Учеба" http://www.ucheba.com/;
12 Законопроект "Об образовании в Российской Федерации". Вопросы и ответы http://xn--273--84d1f.xn--p1ai/voprosy_i_otvety

Собственные электронные образовательные и информационные ресурсы КубГУ

1 Электронный каталог Научной библиотеки КубГУ http://megapro.kubsu.ru/MegaPro/Web
2 Электронная библиотека трудов ученых КубГУ http://megapro.kubsu.ru/MegaPro/UserEntry?Action=ToDb&idb=6
3 Среда модульного динамического обучения http://moodle.kubsu.ru

4 База учебных планов, учебно-методических комплексов, публикаций и конференций http://infoneeds.kubsu.ru/

5 Библиотека информационных ресурсов кафедры информационных образовательных технологий http://mschool.kubsu.ru;

6 Электронный архив документов КубГУ http://docspace.kubsu.ru/

7 Электронные образовательные ресурсы кафедры информационных систем и технологий в образовании КубГУ и научно-методического журнала "ШКОЛЬНЫЕ ГОДЫ" http://icdau.kubsu.ru/

5.5. Публикации конференций A*

1. Ian J. Goodfellow at al. Generative Adversarial Nets. NEURIPS. 2014. https://proceedings.neurips.cc/paper_files/paper/2014/file/f033ed80deb0234979a61f95710dbe25-Paper.pdf
2. f -GAN: Training Generative Neural Samplers using Variational Divergence Minimization. Sebastian Nowozin et al, NEURIPS. 2016. https://proceedings.neurips.cc/paper_files/paper/2016/file/cedebb6e872f539bef8c3f919874e9d7-Paper.pdf
3. NVAE: A Deep Hierarchical Variational Autoencoder. Arash Vahda et al. NEURIPS. 2020. https://proceedings.neurips.cc/paper/2020/file/e3b21256183cf7c2c7a66be163579d37-Paper.pdf
4. D-VAE: A Variational Autoencoder for Directed Acyclic Graphs. Muhan Zhang et al. NEURIPS. 2019. https://proceedings.neurips.cc/paper_files/paper/2019/file/e205ee2a5de471a70c1fd1b46033a75f-Paper.pdf
5. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. Chunyuan Li et al. NEURIPS. 2023. https://papers.neurips.cc/paper_files/paper/2023/file/5abcdf8ecdcacba028c6662789194572-Paper-Datasets_and_Benchmarks.pdf

6. Методические указания для обучающихся по освоению дисциплины (модуля)

По курсу предусмотрено проведение лекционных занятий, на которых дается систематизированный материал по генеративным сетям. В ходе лекций рассматриваются ключевые концепции. После каждой лекции рекомендуется выполнение практических заданий для закрепления ключевых понятий и методов.

Лабораторные занятия курса посвящены практическому освоению работы с генеративными сетями. На занятиях студенты используют, дообучают и адаптируют готовые мультимодальные модели, разрабатывают приложение с их использованием.

При самостоятельной работе студентам необходимо изучать рекомендованную литературу в виде официальной документации к используемым открытым программным продуктам, облачным платформам.

Важнейшим компонентом курса является самостоятельная проектная работа, в ходе которой студент разрабатывает законченное решение с уровнем технологической готовности (УТГ) 5-7 с применением CI/CD/CT для решения задач (кейсов) индустриальных партнеров. Допускается выполнение проектов в командах 3-4 человека.

Кейсы ПАО «Сбербанк»

1. Генеративный ИИ для автоматического составления инвестиционных обзоров
Описание:
Аналитики Сбера ежедневно составляют десятки аналитических и инвестиционных обзоров по рынкам, компаниям, макроэкономике. Задача — исследовать применение LLM для генерации кратких сводок и аналитических отчетов на основе входных данных: биржевые котировки, макроэкономические показатели, рыночные события.
Цель:
Разработать инструмент, способный по структурированным данным и краткому описанию формировать инвестиционный обзор в деловом стиле.
Ожидаемый результат:
Модель, генерирующая аналитические тексты длиной 500–1000 слов с разделами «обзор событий», «рекомендации», «прогнозы», оформленные в формате банка.
2. НЛП-анализ жалоб клиентов в свободной форме
Описание:
В рамках клиентского сервиса Сбербанк обрабатывает обращения из чатов, мобильного приложения и жалобной формы. Требуется построить модель семантического анализа, выделяющую суть обращения, определяющую тональность и потенциальную серьёзность инцидента.
Цель:
Автоматизировать классификацию обращений для ускорения маршрутизации и выявления повторяющихся болевых точек в продуктах и процессах.
Ожидаемый результат:
Прототип модели, автоматически выделяющей темы жалоб (например, «ошибка в приложении», «двойное списание»), их эмоциональную окраску и критичность.
3. Генерация сценариев фишинговых писем для обучения сотрудников
Описание:
Банк проводит киберучения, включая рассылку тестовых фишинговых писем сотрудникам для повышения их устойчивости к социальным атакам. Проект предполагает использование генеративной модели для создания реалистичных фишинговых писем различных типов (поддельные счета, HR-запросы, ИТ-поддержка).
Цель:
Создать генератор, способный на основе заданных параметров (тема, стиль, уровень угрозы) создавать тексты фишинга для тренировок.
Ожидаемый результат:
Набор разнообразных примеров фишинга и оценка их эффективности по реакции сотрудников, а также классификация моделей угроз.
4. Мультимодальный ассистент для банковских отделений
Описание:
Физические отделения Сбербанка внедряют интерактивных консультантов. Предполагается создание мультимодального ИИ-ассистента, который воспринимает речь и визуально ориентируется в пространстве (распознаёт клиента, документы, банкоматы), а также отвечает голосом.
Цель:
Разработать базовый прототип, имитирующий функциональность помощника: ответы на типовые запросы, визуальные подсказки, навигация по отделению.
Ожидаемый результат:
Интерактивная модель, объединяющая голосовой ввод, зрительное восприятие (например, QR-код паспорта), текстовый вывод и жестовую реакцию.
5. Объяснимость и контроль генеративных моделей в банковском ИИ
Описание:
Банк активно использует LLM и NLP-сервисы (в чат-ботах, генерации шаблонов ответов,

автоответах на e-mail), однако встает вопрос: как объяснять и контролировать поведение таких моделей, особенно в юридически значимых коммуникациях?

Цель:

Исследовать подходы к трассировке решений LLM (например, через логирование reasoning chain, пост-фильтрацию ответов, встроенные правила).

Ожидаемый результат:

Концепция системы explainability + compliance-модуля, обеспечивающего соответствие генерации стандартам банка и регулятора.

6. Генерация пользовательских сценариев работы в мобильном приложении

Описание:

Банк хочет использовать генеративный ИИ для быстрой симуляции пользовательских сценариев — например, как клиент оформляет вклад, переводит средства, получает уведомление о риске мошенничества.

Цель:

Разработать генератор пошаговых сценариев пользовательского поведения с вариативностью (молодой клиент, пенсионер, ИП).

Ожидаемый результат:

Набор автоматически сгенерированных UX-сценариев, оформленных в виде сценариев для QA или UX-исследований, с логикой действий и типичными ошибками пользователя.

7. Генерация synthetic data для банковских моделей

Описание:

Модели в Сбере требуют большого объёма транзакционных и клиентских данных, которые нельзя использовать напрямую из-за требований ЦБ и ФЗ-152. Задача — разработать метод генерации синтетических банковских данных, максимально близких к реальным по распределениям и поведению.

Цель:

Создать безопасный pipeline генерации данных (например, транзакций, профилей клиентов, шаблонов расходов) для обучения моделей.

Ожидаемый результат:

Синтетический датасет и отчет о метриках приближённости к реальному (TSNE, K-L divergence и др.), с оценкой пригодности для обучения скоринговых или антифрод-моделей.

8. Сравнение text2video / text2img моделей

Описание:

Сбербанк заинтересован в сравнении text2video / text2img моделей (открытые модели, особенно китайские). Задача требует применения облачных ресурсов партнера для машинного обучения. От студентов требуется навык запуска открытых моделей, планирования, структурирования и логирования экспериментов, совместной работы. Задача может быть распараллелена для сравнения множества моделей независимо в группе студентов.

Цель:

Провести сравнение работы актуальных открытых моделей text2video / text2img.

Ожидаемый результат:

Таблица с результатами экспериментов модель / репозиторий / функционал / требования / оценка производительности / X примеров генераций (было/стало), human_eval по принципу арены (какая лучше)

Кейсы от «АВАЛАБ»

1. LLM и RAG для BI-системы Fastboard

Описание:

Для разрабатываемой компанией BI-системы Fastboard требуется разработать интерфейс на естественном языке для построения отчетов на больших массивах данных в ClickHouse.

С помощью LLM необходимо классифицировать запросы пользователей на естественном языке и извлекать фактические параметры для дальнейшего вызова веб-сервиса отчетов.

Цель:

Разработать промпты для классификации и обработки запросов пользователей LLM и преобразования их к вызовам типовых отчетов с фактическими параметрами, извлекаемыми из запроса.

Ожидаемый результат:

Инструмент на основе LLM, позволяющий запрашивать данные о продажах.

2. Генеративный ИИ для создания проектной документации по ТЗ

Описание:

В рамках проектирования объектов девелоперской компании архитекторы и инженеры тратят значительное время на подготовку текстовой проектной документации (обоснование решений, пояснительные записки, описания инженерных систем). Задача — исследовать возможность использования LLM для генерации черновиков проектной документации на основе исходных данных: этажность, материалы, климат, назначение, нормы.

Цель:

Разработать прототип текстового генератора, который помогает специалистам быстрее формировать документацию в соответствии с шаблонами и нормативами.

Ожидаемый результат:

Инструмент на основе LLM, создающий логически стройный и нормативно грамотный текст, поддающийся быстрой редактуре инженером.

3. Мультимодальный агент для анализа строительных площадок

Описание:

ООО «АВА ЛАБ» разрабатывает систему для мониторинга строительных объектов. Требуется создать прототип мультимодального ИИ-агента, способного анализировать изображения со стройплощадки (видео/фото), а также принимать голосовые и текстовые запросы (например, «проверь монтаж перекрытия на 5 этаже»).

Цель:

Объединить возможности компьютерного зрения (распознавание стадии строительства, техники, нарушений) и НЛП (понимание запросов, отчётов).

Ожидаемый результат:

Интерактивный агент, который на запрос специалиста может показать нужный участок, прокомментировать прогресс, зафиксировать нарушения.

4. Генерация рекламного контента для жилых комплексов

Описание:

«АВА ГРУПП» регулярно запускает маркетинговые кампании для жилых комплексов. Необходимо исследовать использование диффузионных моделей для генерации изображений (визуализации интерьеров, окрестностей, видов из окон) и LLM — для описаний квартир, преимуществ района, инфраструктуры.

Цель:

Создать инструменты для быстрой генерации продающих материалов без привлечения дизайнеров и копирайтеров на первых этапах.

Ожидаемый результат:

Набор сгенерированных карточек объектов с текстом, изображением и логикой «живого» рекламного сообщения.

5. Генерация документации и шаблонов договоров

Описание:

Юридический департамент регулярно работает с договорами долевого участия, актами приёма-передачи и другими документами. Использование LLM может значительно сократить время на подготовку черновиков — достаточно ввести параметры сделки.

Цель:
Создать систему, которая генерирует адаптированные тексты документов по вводным данным (тип объекта, этаж, площадь, ФИО, сроки и пр.).
Ожидаемый результат:
Генератор документов в формате Word или PDF с автоматической подстановкой параметров и соблюдением юридического стиля.
6. ИИ-помощник для риэлторов
Описание:
Агенты по недвижимости готовят рекламные материалы в соц. сети для еще не построенных объектов. Требуется согдать генератор видео роликов, использующих планировку квартиры, изображения риэлтора и текст, который произносит риэлтор в кадре, показывая квартиру.
Цель:
Упростить подготовку рекламного контента для риэлторов.
Ожидаемый результат:
Приложение, способное генерировать реалистичные видео ролики с показом виртуальных квартир риэлторами.

Для студентов с ограниченными возможностями здоровья предусмотрены дополнительные индивидуальные консультации, на которых преподаватель подробно разъясняет сложные аспекты дисциплины, помогает адаптировать практические задания и обеспечивает специальные условия для освоения методов работы с системами искусственного интеллекта. Индивидуальный подход позволяет таким студентам полноценно участвовать в учебном процессе и достигать требуемых результатов обучения.

7. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю)

7.1 Перечень информационно-коммуникационных технологий

1. Облачные платформы и сервисы
cloud.ru,YandexCloud,AWS/GCP/Azure– облачные вычисления

2. Системы управления версиями и коллаборации
Git/GitHub/GitLab – контроль версий кода и совместная разработка

4. Система управления обучением
Moodle – сдача работ

7.2 Перечень лицензионного и свободно распространяемого программного обеспечения

1. Свободное ПО (Open Source)
GitLab, GIT, Python, PyTorch.

8. Материально-техническое обеспечение по дисциплине (модулю)

Виртуальные машины, кластер Managed Kubernetes и ресурсы GPU в облаке предоставляется индустриальным партнером ПАО «Сбербанк»:

№	Продукт	Параметры продукта	Кол-во	Кол-во конфигураций	Ед. изм.
1	Виртуальная машина с GPU	Виртуальная машина с GPU NVIDIA® Tesla® V100 2 GPU 8 vCPU 128 ГБ RAM	1	1	Шт
		ОС Ubuntu_24.04	1		Шт
		Системный диск SSD	1		Шт
			2000		Гб
		Диск SSD	2		Шт
			4096		Гб
		Аренда публичного IP	1		Шт
2	ML Inference Instance Type GPU	Время работы в месяц	40	1	Ч
		Инстанс 8 x NVIDIA® H100 NVLink PCIe 160 vCPU 1520 GB RAM	1		Шт
		Количество запросов к ML-моделям	1		Млн. Шт
		Кэш ML-моделей	160		Гб

Дополнительные облачные ресурсы предоставляются технологическим партнером Yandex Cloud.

№	Вид работ	Наименование учебной аудитории, ее оснащенность оборудованием и техническими средствами обучения
1	Лекционные занятия	Аудитория, укомплектованная специализированной мебелью и техническими средствами обучения
2	Лабораторные занятия	Аудитория, укомплектованная специализированной мебелью и техническими средствами обучения, компьютерами, проектором, программным обеспечением
3	Практические занятия	Аудитория, укомплектованная специализированной мебелью и техническими средствами обучения
4	Групповые (индивидуальные) консультации	Аудитория, укомплектованная специализированной мебелью и техническими средствами обучения, компьютерами, программным обеспечением
5	Текущий контроль, промежуточная аттестация	Аудитория, укомплектованная специализированной мебелью и техническими средствами обучения, компьютерами, программным обеспечением
6	Самостоятельная работа	Кабинет для самостоятельной работы, оснащенный компьютерной техникой с возможностью подключения к сети «Интернет», программой экранного увеличения и обеспеченный доступом в электронную информационно-образовательную среду университета.

Примечание: Конкретизация аудиторий и их оснащение определяется ОПОП.