МИНОБРНАУКИ РОССИИ
федеральное государственное бюджетное образовательное учреждение
высшего образования
«Нижегородский государственный технический университет
им. Р.Е. Алексеева» (НГТУ)

Институт радиоэлектроники и информационных технологий (ИРИТ)
кафедра «Информатика и системы управления» (ИСУ)

ОДОБРЕНО:
на заседании кафедры разработчика
протокол № _1__ от 02.09.2026 г.
Зав. кафедрой
_____ О.П. Тимофеева
«02» сентября 2026 г.

УТВЕРЖДЕН:
на заседании ученого совета ИРИТ
протокол № 9 от 22.09. 2026 г.
Директор института
_____ В.П. Хранилов
«02» сентября 2026 г.

Документ подписан электронной подписью в системе электронного документооборота НГТУ им. Р.Е. Алексеева
СВЕДЕНИЯ О СЕРТИФИКАТЕ ЭП
Сертификат 6e466def73b388c262d1fef93a6e3720
Владелец Ивашкин Евгений Геннадьевич
Действителен с 04.02.26 по 30.04.27

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Б1.В.ОД.16 Основы глубокого обучения

(индекс и наименование дисциплины по учебному плану)

для подготовки бакалавров

Направление подготовки:	09.03.01 Информатика и вычислительная техника
Направленность:	Интеллектуальные системы обработки информации и управления
Форма обучения:	очная, заочная
Выпускающая кафедра	ИСУ
Разработчик (и):	Шагалова П.А. доцент, к.т.н.; Савкин А.Е. ассистент
регистрационный №	09.03.01-и-56

Начальник МО _____
подпись

Заведующая отделом комплектования НТБ _____
подпись

НИЖНИЙ НОВГОРОД, 2026 год

Содержание

1. ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ (МОДУЛЮ), СООТНЕСЕННЫЕ С ПЛАНИРУЕМЫМИ РЕЗУЛЬТАТАМИ ОСВОЕНИЯ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ 3

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ 7

3. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ 8

3.1. Распределение трудоемкости дисциплины по видам работ по семестрам 8

3.2. Содержание дисциплины, структурированное по темам 12

4. УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ 22

4.1. Учебная литература, печатные издания библиотечного фонда 22

4.2. Справочно-библиографическая литература 22

4.3. Методические указания, рекомендации и другие материалы к занятиям 23

5. ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ 23

5.1. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины 23

5.1.1. Собственные образовательные ресурсы. Электронные ресурсы НТБ 23

5.1.2. Сторонние электронные образовательные и информационные ресурсы. Внешние ресурсы 23

5.1.3. Перечень современных баз данных и информационных справочных систем 24

5.2. Перечень программного обеспечения и информационных справочных систем 24

6. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ, НЕОБХОДИМОЕ ДЛЯ ОСУЩЕСТВЛЕНИЯ ОБРАЗОВАТЕЛЬНОГО ПРОЦЕССА ПО ДИСЦИПЛИНЕ 24

9. МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ ОБУЧАЮЩИМСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ 25

9.1. Общие методические рекомендации для обучающихся по освоению дисциплины, образовательные технологии 25

9.2. Методические указания для занятий лекционного типа 26

9.3. Методические указания по освоению дисциплины на лабораторных работах 26

9.5. Методические указания по самостоятельной работе обучающихся 26

10. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ДЛЯ КОНТРОЛЯ ОСВОЕНИЯ ДИСЦИПЛИНЫ И КРИТЕРИИ ОЦЕНИВАНИЯ 27

11. АДАПТАЦИЯ РАБОЧЕЙ ПРОГРАММЫ ДЛЯ ЛИЦ С ОВЗ И ИНВАЛИДОВ 27

1. ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ (МОДУЛЮ), СООТНЕСЕННЫЕ С ПЛАНИРУЕМЫМИ РЕЗУЛЬТАТАМИ ОСВОЕНИЯ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ

Настоящая рабочая программа дисциплины устанавливает планируемые результаты обучения по дисциплине, а также определяет содержание и виды учебных занятий и отчетности.

Программа разработана в соответствии с Федеральным государственным образовательным стандартом высшего образования (ФГОС ВО 3++) по направлению подготовки 09.03.01 «Информатика и вычислительная техника».

Освоение дисциплины вносит вклад в формирование профессиональной компетенции, предусмотренной образовательной программой.

Таблица 1. Компетенции, формируемые при освоении дисциплины

Код компетенции	**Формулировка компетенции**
Профессиональные (ПК)	
ПКС-3	Способен использовать формальные и интеллектуальные методы обработки и анализа информации при решении профессиональных задач

Планируемые результаты обучения по дисциплине соотнесены с индикатором достижения компетенции ИПКС-3.2 и квалификационными требованиями профессионального стандарта 06.042 «Специалист по большим данным».

Таблица 2. Индикаторы достижения компетенции и планируемые результаты обучения

Код и наименование компетенции	Код и наименование индикатора достижения компетенции	Код и наименование дескриптора достижения компетенции	Код ПС и ТФ	Квалификационные требования к выбранной ТФ	Оценочные материалы (ОМ)
ПКС-3. Способен использовать формальные и интеллектуальные методы обработки и анализа информации при решении профессиональных задач	ИПКС-3.2. Использует интеллектуальные методы обработки и анализа информации при решении профессиональных задач	**Знать:** – типы больших данных и особенности их обработки для задач глубокого обучения; – теоретические и прикладные основы анализа больших данных применительно к нейросетевым моделям; – архитектуры и принципы обучения больших языковых моделей (LLM): архитектура Transformer, механизм внимания, предварительное обучение, дообучение; – принципы квантизации нейросетевых моделей: числовые форматы, калибровка, низкоразрядные представления и влияние на память, качество и производительность; – принципы PEFT, LoRA и QLoRA, а также подходы к оценке качества и воспроизводимости адаптированных языковых моделей; – принципы работы и оптимизации диффузионных моделей: прямой и обратный процессы, шумовые расписания, DBCache, TaylorSeer.	06.042 A/03.6 «Подготовка данных для проведения аналитических работ по исследованию больших данных»	Трудовые действия: – определение источников больших данных для анализа; – извлечение, проверка и очистка больших объемов данных из гетерогенных источников; – агрегация и разработка представления больших объемов данных. Трудовые умения: – производить очистку данных; – проводить интеграцию и преобразование больших объемов данных; – оценивать соответствие наборов данных задачам анализа больших данных. Трудовые знания: – предметная область анализа; – теоретические и прикладные основы анализа больших данных; – современные методы и инструментальные средства	ЛР1 – энкодер-декодер; ЛР2 – квантизация и сравнительное тестирование; ЛР3 – PEFT/LoRA/QLoRA и оценка адаптации; ЛР4 – ИИ-агент (зачтено/не зачтено); практико-ориентированное задание по оптимизации диффузионных моделей (консультация в рамках КСР); зачет с оценкой.

Код и наименование компетенции	Код и наименование индикатора достижения компетенции	Код и наименование дескриптора достижения компетенции	Код ПС и ТФ	Квалификационные требования к выбранной ТФ	Оценочные материалы (ОМ)
		Уметь: – проводить интеграцию и преобразование больших объемов данных, приводить их к формату, пригодному для обучения моделей глубокого обучения; – применять методы квантизации языковых моделей и проводить сравнительное тестирование по памяти, задержке, пропускной способности и качеству; – выполнять параметрически эффективное дообучение (PEFT) LLM с использованием LoRA/QLoRA, оценивать результаты до и после адаптации и фиксировать условия эксперимента; – применять методы оптимизации диффузионных моделей: настраивать шумовое расписание, использовать кэширование промежуточных результатов и прогнозирование будущих скрытых состояний. **Владеть:** – навыками определения источников больших данных для аналитических работ с использованием методов		анализа больших данных; – типы больших данных.	

Код и наименование компетенции	Код и наименование индикатора достижения компетенции	Код и наименование дескриптора достижения компетенции	Код ПС и ТФ	Квалификационные требования к выбранной ТФ	Оценочные материалы (ОМ)
		глубокого обучения; – навыками реализации обучения моделей глубокого обучения с использованием PyTorch; – навыками постановки воспроизводимых вычислительных экспериментов, оценки и сравнения моделей с расчетом метрик качества и анализом вычислительных затрат.			

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ

Учебная дисциплина Б1.В.ОД.16 «Основы глубокого обучения» включена в перечень дисциплин части образовательной программы, формируемой участниками образовательных отношений, и определяет направленность (профиль) «Интеллектуальные системы обработки информации и управления».

Дисциплина изучается после освоения базовых дисциплин программирования, алгоритмов и математической подготовки и непосредственно опирается на результаты дисциплины Б1.В.ОД.5 «Интеллектуальный анализ данных».

К началу изучения дисциплины обучающийся должен владеть базовыми навыками программирования на Python, понимать основные задачи классификации и регрессии, принципы разделения данных на обучающую и тестовую выборки, базовые метрики качества, понятия переобучения и регуляризации, общую архитектуру искусственной нейронной сети и основы сверточных нейронных сетей.

В рамках дисциплины эти знания развиваются до уровня современных архитектур глубокого обучения, языковых и диффузионных генеративных моделей, методов эффективного инференса моделей, параметрически эффективного дообучения и агентных систем.

Результаты обучения, полученные при освоении дисциплины, необходимы при изучении дисциплины «Анализ больших данных», выполнении последующих проектных и практических работ по профилю подготовки и выпускной квалификационной работы.

Особенностью дисциплины является практико-ориентированное изучение современных методов глубокого обучения в условиях ограниченных вычислительных ресурсов. Обязательные лабораторные работы проектируются с возможностью выполнения на персональных компьютерах и/или в облачных вычислительных средах.

3. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

3.1. Распределение трудоемкости дисциплины по видам работ по семестрам

Общая трудоемкость дисциплины составляет 3 зачетные единицы, 108 часов. Распределение часов по видам работ для очной и заочной форм обучения представлено в таблице 3.

Таблица 3. Распределение трудоемкости дисциплины по видам работ для очной формы обучения

Вид учебной работы	**Трудоемкость в час**	
	Всего час.	**В т.ч. по семестрам 6 сем.**
Формат изучения дисциплины	с использованием элементов электронного обучения	с использованием элементов электронного обучения
Общая трудоемкость дисциплины по учебному плану	108	108
1. Контактная работа:	38	38
1.1 Аудиторная работа, в том числе:	34	34
лекции	17	17
лабораторные работы	17	17
практические занятия	–	–
1.2 Контрольно-самостоятельная работа	4	4
курсовая работа/курсовой проект	–	–
текущий контроль, консультации по дисциплине	4	4
контактная работа на промежуточном контроле (экзамене)	–	–
реферат, расчетно-графическая работа, контрольная работа	–	–
2. Самостоятельная работа	70	70

Вид учебной работы	Трудоемкость в час	
	Всего час.	**В т.ч. по семестрам 6 сем.**
2.1 самостоятельная работа (самостоятельное изучение разделов, самоподготовка, подготовка к лабораторным занятиям и др.)	70	70
2.2 подготовка к контролю	–	–
3. Форма контроля	зачет с оценкой	зачет с оценкой

Продолжение таблицы 3. Распределение трудоемкости дисциплины по видам работ для заочной формы обучения

Вид учебной работы	Трудоемкость в час	
	Всего час.	**В т.ч. по семестрам 8 сем.**
Формат изучения дисциплины	с использованием элементов электронного обучения	с использованием элементов электронного обучения
Общая трудоемкость дисциплины по учебному плану	108	108
1. Контактная работа:	16	16
1.1 Аудиторная работа, в том числе:	12	12
лекции	6	6
лабораторные работы	6	6
практические занятия	–	–
1.2 Контрольно-самостоятельная работа	4	4
курсовая работа/курсовой проект	–	–
текущий контроль, консультации по дисциплине	4	4
контактная работа на промежуточном контроле (экзамене)	–	–
реферат, расчетно-графическая работа, контрольная работа	–	–
2. Самостоятельная работа	92	92
2.1 самостоятельная работа (самостоятельное изучение разделов, самоподготовка, подготовка к лабораторным занятиям и др.)	88	88
2.2 подготовка к контролю	4	4
3. Форма контроля	зачет с оценкой	зачет с оценкой

3.2. Содержание дисциплины, структурированное по темам

Таблица 4. Содержание дисциплины, структурированное по темам, для очной формы обучения

Планируемые (контролируемые) результаты освоения: код УК; ОПК; ПК и индикаторы достижения компетенций	Наименование разделов, тем	Виды учебной работы				Вид СРС	Наименование используемых активных и интерактивных образовательных технологий
		Контактная работа			Самостоятельная работа студентов (СРС), час		
		Лекции, час	Лабораторные работы, час	Практические занятия, час			
6 семестр							
Раздел 1. Языковые модели и архитектура Transformer							
ПКС-3 – ИПКС-3.2	**Тема 1.1.** Данные для глубокого обучения и языковые модели: источники и типы данных, подготовка и представление данных, пакетная обработка, токенизация; PyTorch; векторные представления; RNN/LSTM/GRU; модель «последовательность-последовательность» (seq2seq); энкодер-декодер; обучение с подстановкой истинного предыдущего токена; мотивация использования механизма внимания.	2			5	Подготовка к лекции; самостоятельное изучение способов подготовки и представления данных.	Проблемная лекция; разбор примеров.
ПКС-3 – ИПКС-3.2	**Тема 1.2.** Transformer, BERT и генеративные языковые модели: механизмы самовнимание, Q/K/V, многоголовое внимание, позиционное кодирование, энкодер/декодер, каузальная маска, перекрёстное внимание, предварительное обучение, авторегрессионная генерация.	2			6	Подготовка к лекции; изучение архитектур Transformer и способов генерации.	Интерактивная лекция; разбор архитектур.
ПКС-3 – ИПКС-3.2	**Лабораторная работа №1.** Построение и исследование генеративной модели типа энкодер-декодер на основе предобученного Transformer.		5		7	Подготовка данных и окружения; выполнение вычислительного эксперимента; оформление отчета и подготовка к защите.	Вычислительный эксперимент; исследовательский метод.
Раздел 2. Квантизация и эффективное выполнение нейросетевых моделей							

Планируемые (контролируемые) результаты освоения: код УК; ОПК; ПК и индикаторы достижения компетенций	Наименование разделов, тем	Виды учебной работы				Вид СРС	Наименование используемых активных и интерактивных образовательных технологий
		Контактная работа			Самостоятельная работа студентов (СРС), час		
		Лекции, час	Лабораторные работы, час	Практические занятия, час			
ПКС-3 – ИПКС-3.2	**Тема 2.1.** Числовые форматы и базовые методы квантизации: FP32, FP16 и BF16; INT8 и FP8, форматы FP8 E4M3 и E5M2; квантизация весов, активаций и KV-кэша; масштаб и нулевая точка; симметричная и асимметричная квантизация; квантизация по тензору и по каналам; калибровка; послетренировочная квантизация (PTQ) и обучение с учётом квантизации (QAT); влияние разрядности на память, качество и производительность.	2			5	Подготовка к лекции; изучение числовых форматов, базовых схем квантизации и калибровки.	Проблемная лекция; сравнительный анализ.
ПКС-3 – ИПКС-3.2	**Тема 2.2.** Низкоразрядные форматы и продвинутые методы квантизации: INT4 и FP4; квантизация только весов и квантизация весов вместе с активациями; групповое и блочное масштабирование; микромасштабирование MXFP8 и MXFP4; проблема выбросов в распределениях весов и активаций; преобразования и повороты пространств признаков, принцип QuaRot; низкоразрядное представление KV-кэша; аппаратная поддержка низкоразрядных вычислений; объём занимаемой памяти, задержка, TTFT и пропускная способность; причины, по которым снижение разрядности не всегда обеспечивает пропорциональное ускорение.	2			5	Подготовка к лекции; изучение низкоразрядных форматов, микромасштабирования и методов работы с выбросами.	Интерактивная лекция; анализ современных методов.
ПКС-3 – ИПКС-3.2	**Лабораторная работа №2.** Исследование методов квантизации и эффективности выполнения языковых моделей.		5		7	Подготовка среды; сравнительное тестирование базовой и квантованных конфигураций; анализ памяти, производительности и качества; оформление отчета и защита.	Вычислительный эксперимент; сравнительный анализ.

Планируемые (контролируемые) результаты освоения: код УК; ОПК; ПК и индикаторы достижения компетенций	Наименование разделов, тем	Виды учебной работы				Вид СРС	Наименование используемых активных и интерактивных образовательных технологий
		Контактная работа			Самостоятельная работа студентов (СРС), час		
		Лекции, час	Лабораторные работы, час	Практические занятия, час			
Раздел 3. Адаптация больших языковых моделей							
ПКС-3 – ИПКС-3.2	**Тема 3.1.** Дообучение больших языковых моделей и PEFT: контролируемое и инструкционное дообучение; полное дообучение и адаптерные методы; LoRA: ранг, коэффициент масштабирования, целевые модули; QLoRA: квантованная замороженная базовая модель и обучаемые LoRA-адаптеры; выбор между LoRA и QLoRA; подготовка набора данных для инструкционного дообучения.	2			5	Подготовка к лекции; изучение PEFT, LoRA/QLoRA и подготовки наборов данных.	Проблемная лекция; анализ методов.
ПКС-3 – ИПКС-3.2	**Тема 3.2.** Оценка качества и воспроизводимость адаптированных языковых моделей: базовая конфигурация; разделение данных на обучающую, валидационную и тестовую части; оценка до и после адаптации; автоматические метрики и качественный анализ результатов генерации; переобучение; фиксация параметров эксперимента и программного окружения; воспроизводимость результатов; сопоставление качества, памяти и вычислительных затрат.	2			5	Изучение методов оценки генерации, разделения данных и требований к воспроизводимости.	Интерактивная лекция; сравнительный анализ.
ПКС-3 – ИПКС-3.2	**Лабораторная работа №3.** Параметрически эффективное дообучение и оценка адаптированной языковой модели.		5		7	Подготовка набора данных; обучение LoRA/QLoRA-адаптера; оценка до/после адаптации; фиксация параметров и среды; оформление отчета и защита.	Вычислительный эксперимент; исследовательский метод.
Раздел 4. Агентные системы на базе языковых моделей							

Планируемые (контролируемые) результаты освоения: код УК; ОПК; ПК и индикаторы достижения компетенций	Наименование разделов, тем	Виды учебной работы				Вид СРС	Наименование используемых активных и интерактивных образовательных технологий
		Контактная работа			Самостоятельная работа студентов (СРС), час		
		Лекции, час	Лабораторные работы, час	Практические занятия, час			
ПКС-3 – ИПКС-3.2	**Тема 4.1.** ИИ-агенты и агентные системы: цикл работы агента, инструменты (tools) и вызов функций, навыки (skills), контекст, память, планирование, подагенты, MCP, разрешения и безопасность; агенты для работы с программным кодом; совместимый с OpenAI сервер выполнения модели.	2			4	Подготовка к лекции; изучение архитектуры агентных систем и механизмов вызова инструментов.	Кейс-метод; демонстрация процесса работы агента.
ПКС-3 – ИПКС-3.2	**Лабораторная работа №4.** Разработка и исследование агентной системы на базе языковой модели.		2		2	Настройка агента, инструмента и навыка; выполнение небольшой воспроизводимой задачи; подготовка к демонстрации.	Проектный метод; практическая демонстрация.
Раздел 5. Диффузионные генеративные модели							
ПКС-3 – ИПКС-3.2	**Тема 5.1.** Диффузионные модели и оптимизация генерации диффузионными моделями: прямой и обратный процессы диффузии, расписание шума, устранение шума, DDPM/DDIM, диффузия в латентном пространстве, U-Net/DiT, обусловливание, DBCache, TaylorSeer.	3			12	Подготовка к лекции; изучение материалов по диффузионным моделям; самостоятельное выполнение и анализ практико-ориентированного задания по расписанию шума, DBCache и TaylorSeer; подготовка к консультации в рамках КСР.	Проблемная лекция; вычислительный/аналитический эксперимент.
ПКС-3 – ИПКС-3.2	**ИТОГО**	17	17	–	70		

Продолжение таблицы 4. Содержание дисциплины, структурированное по темам, для заочной формы обучения

Планируемые (контролируемые) результаты освоения: код УК; ОПК; ПК и индикаторы достижения компетенций	Наименование разделов, тем	Виды учебной работы				Вид СРС	Наименование используемых активных и интерактивных образовательных технологий
		Контактная работа			Самостоятельная работа студентов (СРС), час		
		Лекции, час	Лабораторные работы, час	Практические занятия, час			
8 семестр							
Раздел 1. Языковые модели и архитектура Transformer							
ПКС-3 – ИПКС-3.2	**Тема 1.1.** Данные для глубокого обучения и языковые модели: подготовка и представление данных, PyTorch, векторные представления, RNN/LSTM/GRU, модель «последовательность-последовательность» (seq2seq), энкодер-декодер, механизм внимания.	1			7	Самостоятельная проработка материалов; подготовка данных и окружения.	Обзорная проблемная лекция.
ПКС-3 – ИПКС-3.2	**Тема 1.2.** Transformer, BERT и генеративные языковые модели.	1			9	Самостоятельное изучение самовнимания, архитектур с энкодером/декодером, BERT/GPT и стратегий генерации.	Интерактивная лекция; разбор архитектуры.
ПКС-3 – ИПКС-3.2	**Лабораторная работа №1.** Построение модели типа энкодер-декодер на основе предобученного Transformer (сокращенный вариант).		2		10	Подготовка блокнота Jupyter Notebook; выполнение основной части эксперимента; отчет.	Вычислительный эксперимент.
Раздел 2. Квантизация и эффективное выполнение нейросетевых моделей							
ПКС-3 – ИПКС-3.2	**Тема 2.1.** Числовые форматы и базовые методы квантизации: FP32/FP16/BF16, INT8 и FP8 (E4M3/E5M2), масштаб и нулевая точка, симметричная/асимметричная квантизация, квантизация по тензору/каналам, калибровка, PTQ/QAT.	1			8	Самостоятельное изучение числовых форматов, базовых методов квантизации и калибровки.	Обзорная лекция; сравнительный анализ.

Планируемые (контролируемые) результаты освоения: код УК; ОПК; ПК и индикаторы достижения компетенций	Наименование разделов, тем	Виды учебной работы				Вид СРС	Наименование используемых активных и интерактивных образовательных технологий
		Контактная работа			Самостоятельная работа студентов (СРС), час		
		Лекции, час	Лабораторные работы, час	Практические занятия, час			
ПКС-3 – ИПКС-3.2	**Тема 2.2.** Низкоразрядные форматы и продвинутые методы квантизации: INT4/FP4, групповое и блочное масштабирование, MXFP8/MXFP6/MXFP4/MXINT8, выбросы, QuaRot/SpinQuant, низкоразрядный KV-кэш, аппаратные ограничения и показатели эффективности.				7	Самостоятельное изучение низкоразрядных форматов, микромасштабирования, поворотов и сравнительных показателей.	Сравнительный анализ.
ПКС-3 – ИПКС-3.2	**Лабораторная работа №2.** Исследование методов квантизации и эффективности выполнения языковых моделей (сокращенный вариант).		1		3	Сравнительное тестирование базовой и квантованной конфигураций; анализ памяти, производительности и качества; отчет.	Вычислительный эксперимент; сравнительный анализ.
Раздел 3. Адаптация больших языковых моделей							
ПКС-3 – ИПКС-3.2	**Тема 3.1.** Дообучение больших языковых моделей и PEFT: контролируемое и инструкционное дообучение, полное дообучение и адаптерные методы, LoRA, QLoRA, выбор метода и подготовка данных.	1			9	Самостоятельное изучение PEFT, LoRA/QLoRA и подготовки наборов данных для инструкционного дообучения.	Обзорная лекция; анализ методов.
ПКС-3 – ИПКС-3.2	**Тема 3.2.** Оценка качества и воспроизводимость адаптированных языковых моделей: базовая конфигурация, разделение данных, оценка до/после, метрики и качественный анализ, переобучение и утечка данных, фиксация параметров и среды, сопоставление качества и затрат.				7	Изучение методов оценки адаптированных моделей и требований к воспроизводимости.	Сравнительный анализ.
ПКС-3 – ИПКС-3.2	**Лабораторная работа №3.** Параметрически эффективное дообучение и оценка адаптированной языковой модели (сокращенный вариант).		2		8	Подготовка набора данных; настройка LoRA/QLoRA-адаптера; сравнение генерации до/после; фиксация параметров и среды; отчет.	Вычислительный эксперимент; исследовательский метод.

Планируемые (контролируемые) результаты освоения: код УК; ОПК; ПК и индикаторы достижения компетенций	Наименование разделов, тем	Виды учебной работы				Вид СРС	Наименование используемых активных и интерактивных образовательных технологий
		Контактная работа			Самостоятельная работа студентов (СРС), час		
		Лекции, час	Лабораторные работы, час	Практические занятия, час			
Раздел 4. Агентные системы на базе языковых моделей							
ПКС-3 – ИПКС-3.2	**Тема 4.1.** ИИ-агенты и агентные системы: цикл работы агента, инструменты, навыки, контекст, MCP, разрешения и безопасность.	1			5	Самостоятельное изучение процесса работы агента и вызова инструментов.	Кейс-метод; демонстрация.
ПКС-3 – ИПКС-3.2	**Лабораторная работа №4.** Настройка и демонстрация агентной системы на базе языковой модели.		1		2	Настройка агента, инструмента и навыка; подготовка краткой демонстрации.	Проектный метод; практическая демонстрация.
Раздел 5. Диффузионные генеративные модели							
ПКС-3 – ИПКС-3.2	**Тема 5.1.** Диффузионные модели и оптимизация генерации диффузионными моделями: DDPM/DDIM, диффузия в латентном пространстве, U-Net/DiT, расписание шума, DBCache, TaylorSeer.	1			13	Самостоятельное изучение диффузионных моделей; самостоятельная подготовка и анализ задания по оптимизации генерации диффузионными моделями; консультация в рамках КСР.	Проблемная лекция; аналитический эксперимент.
ПКС-3 – ИПКС-3.2	**ИТОГО**	6	6	–	88		

4. УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

4.1. Учебная литература, печатные издания библиотечного фонда

1. Goodfellow, I. Deep Learning / I. Goodfellow, Y. Bengio, A. Courville. – Cambridge, MA: MIT Press, 2016. – 800 p. – Электронная версия: https://www.deeplearningbook.org/.

2. Zhang, A. Dive into Deep Learning / A. Zhang, Z. C. Lipton, M. Li, A. J. Smola. – Cambridge: Cambridge University Press, 2023. – Интерактивное электронное издание: https://d2l.ai/.

3. Jurafsky, D. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models / D. Jurafsky, J. H. Martin. – 3rd ed. – Online manuscript, 2026. – https://web.stanford.edu/~jurafsky/slp3/.

4.2. Справочно-библиографическая литература

1. Vaswani, A. Attention Is All You Need / A. Vaswani, N. Shazeer, N. Parmar [и др.] // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – arXiv:1706.03762. – https://arxiv.org/abs/1706.03762.

2. Devlin, J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M.-W. Chang, K. Lee, K. Toutanova. – 2018. – arXiv:1810.04805. – https://arxiv.org/abs/1810.04805.

3. Hu, E. J. LoRA: Low-Rank Adaptation of Large Language Models / E. J. Hu, Y. Shen, P. Wallis [и др.]. – 2021. – arXiv:2106.09685. – https://arxiv.org/abs/2106.09685.

4. Dettmers, T. QLoRA: Efficient Finetuning of Quantized LLMs / T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer. – 2023. – arXiv:2305.14314. – https://arxiv.org/abs/2305.14314.

5. Micikevicius, P. FP8 Formats for Deep Learning / P. Micikevicius, D. Stosic, N. Burgess [и др.]. – 2022. – arXiv:2209.05433. – https://arxiv.org/abs/2209.05433.

6. Rouhani, B. D. Microscaling Data Formats for Deep Learning / B. D. Rouhani, R. Zhao, A. More [и др.]. – 2023. – arXiv:2310.10537. – https://arxiv.org/abs/2310.10537.

7. Ashkboos, S. QuaRot: Outlier-Free 4-Bit Inference in Rotated LLMs / S. Ashkboos, A. Mohtashami, M. L. Croci [и др.]. – 2024. – arXiv:2404.00456. – https://arxiv.org/abs/2404.00456.

8. Liu, Z. SpinQuant: LLM Quantization with Learned Rotations / Z. Liu, C. Zhao, I. Fedorov [и др.]. – 2024. – arXiv:2405.16406. – https://arxiv.org/abs/2405.16406.

9. Ho, J. Denoising Diffusion Probabilistic Models / J. Ho, A. Jain, P. Abbeel. – 2020. – arXiv:2006.11239. – https://arxiv.org/abs/2006.11239.

10. Rombach, R. High-Resolution Image Synthesis with Latent Diffusion Models / R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer. – 2021. – arXiv:2112.10752. – https://arxiv.org/abs/2112.10752.

11. Liu, J. From Reusing to Forecasting: Accelerating Diffusion Models with TaylorSeers / J. Liu, C. Zou, Y. Lyu, J. Chen, L. Zhang // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). – 2025. – P. 15853–15863. – https://arxiv.org/abs/2503.06923.

Периодические и научные издания по профилю дисциплины:

1. Journal of Machine Learning Research (JMLR) – https://www.jmlr.org/.

2. Transactions on Machine Learning Research (TMLR) – https://jmlr.org/tmlr/.

4.3. Методические указания, рекомендации и другие материалы к занятиям

1. Основы глубинного обучения: материалы курса НИУ ВШЭ, 2023. – http://wiki.cs.hse.ru/Основы_глубинного_обучения/2023.

2. Основы глубинного обучения: репозиторий учебных материалов НИУ ВШЭ. – https://github.com/hse-ds/iad-deep-learning.

3. PyTorch Tutorials: официальные учебные материалы по работе с PyTorch. – https://docs.pytorch.org/tutorials/.

4. Комплект авторских презентаций по дисциплине «Основы глубокого обучения» (архитектуры нейронных сетей, LLM, квантизация, эффективное выполнение моделей, диффузионные модели, агентные системы). – Электронные материалы кафедры ИСУ, 2026.

5. ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

Учебный процесс по дисциплине обеспечивается свободно распространяемым программным обеспечением и электронными образовательными ресурсами. Состав программного обеспечения может обновляться с сохранением функциональной эквивалентности, необходимой для достижения планируемых результатов обучения.

5.1. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины (модуля)

5.1.1. Собственные образовательные ресурсы. Электронные ресурсы НТБ

– Научно-техническая библиотека НГТУ – https://www.nntu.ru/structure/view/podrazdeleniya/nauchno-tehnicheskaya-biblioteka/resursy
– Библиотека электронных учебников НГТУ – https://fdp.nntu.ru/knizhnaya-polka/
– Электронный каталог книг и периодических изданий (АИБС «МегаПро») – https://library.nntu.ru/megapro/web

5.1.2. Сторонние электронные образовательные и информационные ресурсы. Внешние ресурсы

– PyTorch Documentation и PyTorch Tutorials – https://docs.pytorch.org/
– Hugging Face Transformers – https://huggingface.co/docs/transformers/
– Hugging Face PEFT – https://huggingface.co/docs/peft/
– Hugging Face Datasets – https://huggingface.co/docs/datasets/
– Hugging Face Diffusers – https://huggingface.co/docs/diffusers/
– SGLang Documentation – https://docs.sglang.ai/
– OpenCode Documentation – https://opencode.ai/docs/
– Google Colaboratory – https://colab.research.google.com/

5.1.3. Перечень современных баз данных и информационных справочных систем

Таблица 5. Перечень современных профессиональных баз данных и информационно-справочных систем

№	Наименование профессиональной базы данных, информационно-справочной системы	Доступ к ресурсу
1	Научная электронная библиотека eLIBRARY.RU	https://www.elibrary.ru/
2	Научная электронная библиотека открытого доступа «КиберЛенинка»	https://cyberleninka.ru/
3	arXiv – открытый архив научных публикаций	https://arxiv.org/
4	DBLP Computer Science Bibliography	https://dblp.org/
5	Semantic Scholar	https://www.semanticscholar.org/

5.2. Перечень программного обеспечения и информационных справочных систем

Таблица 6. Перечень программного обеспечения

Программное обеспечение, используемое в университете на договорной основе	Программное обеспечение свободного распространения и веб-сервисы
Специализированное проприетарное программное обеспечение для освоения дисциплины не требуется. При проведении занятий в компьютерных аудиториях используется базовое системное и офисное программное обеспечение, установленное в НГТУ.	Python 3.x; Jupyter Notebook/JupyterLab; PyTorch; torchvision; Hugging Face Transformers; Datasets; PEFT; Accelerate; bitsandbytes; Diffusers; SGLang; Git; OpenCode; веб-браузер. Для ресурсоемких вычислительных экспериментов допускается использование Google Colab или функционально эквивалентной облачной среды. Версии библиотек выбираются совместимыми с утвержденными учебными материалами в формате Jupyter Notebook.

6. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ, НЕОБХОДИМОЕ ДЛЯ ОСУЩЕСТВЛЕНИЯ ОБРАЗОВАТЕЛЬНОГО ПРОЦЕССА ПО ДИСЦИПЛИНЕ

Для проведения лекционных занятий используются учебные аудитории, оснащенные рабочим местом преподавателя, мультимедийным проектором или интерактивной панелью, экраном и доступом к информационно-телекоммуникационной сети «Интернет».

Лабораторные работы проводятся в компьютерной аудитории либо с использованием персональных компьютеров обучающихся, обеспечивающих работу Python/Jupyter-окружения и доступ к информационно-телекоммуникационной сети «Интернет».

Наличие локального дискретного GPU у каждого обучающегося не является обязательным требованием. Для лабораторных работ по квантизации, PEFT/LoRA/QLoRA и практико-ориентированного задания по диффузионным моделям предусматривается возможность использования облачной вычислительной среды с GPU либо предоставленного преподавателем удаленного вычислительного ресурса.

Для лабораторной работы по агентным системам требуется доступ к языковой модели через OpenAI-совместимый или функционально эквивалентный программный

интерфейс. Локальное размещение LLM на компьютере обучающегося не является обязательным.

Помещения для самостоятельной работы должны быть оснащены компьютерной техникой с доступом к сети «Интернет», электронной информационно-образовательной среде НГТУ и электронным библиотечным ресурсам. Сведения о конкретных аудиториях и их оснащенности определяются расписанием и актуальным перечнем материально-технических ресурсов университета.

9. МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ ОБУЧАЮЩИМСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ

9.1. Общие методические рекомендации для обучающихся по освоению дисциплины, образовательные технологии

Дисциплина реализуется посредством контактной работы с обучающимися, самостоятельной работы и промежуточной аттестации. Контактная работа проводится в аудиториях и, при необходимости, с использованием электронной информационно-образовательной среды университета.

Изучение дисциплины строится по принципу последовательного перехода от архитектур глубокого обучения к современным генеративным моделям, методам оптимизации и прикладным системам ИИ. Основными образовательными технологиями являются проблемное изложение материала, разбор архитектур и программных реализаций, вычислительный эксперимент, сравнительное тестирование, кейс-метод и элементы проектного обучения.

При освоении дисциплины рекомендуется уделять основное внимание не воспроизведению готового кода, а постановке вычислительного эксперимента, выбору контролируемых параметров, воспроизводимости результатов, интерпретации метрик и анализу компромиссов между качеством модели, скоростью, потреблением памяти и стоимостью обучения или выполнения модели.

9.2. Методические указания для занятий лекционного типа

Лекционный курс предполагает систематизированное изложение основных вопросов тематического плана. Перед занятием обучающемуся рекомендуется ознакомиться с материалами предыдущей темы и указанными преподавателем первичными источниками. В ходе лекции необходимо фиксировать основные определения, схемы архитектур, алгоритмические этапы и связи между рассматриваемыми методами.

При изучении тем, связанных с Transformer, квантизацией, PEFT/LoRA/QLoRA, оценкой адаптированных моделей, диффузионными моделями и агентными системами, рекомендуется сопоставлять теоретические положения с документацией используемых библиотек и результатами последующих лабораторных экспериментов.

9.3. Методические указания по освоению дисциплины на лабораторных работах

Лабораторные работы выполняются преимущественно в формате Jupyter Notebook или совместимом с ним формате. До начала работы обучающийся должен ознакомиться с постановкой задачи, подготовить программное окружение и исходные данные. Ресурсоемкие вычисления допускается выполнять в облачной среде или на удаленном ресурсе, предоставленном преподавателем.

Отчет по лабораторной работе должен содержать постановку задачи, краткое описание данных и модели, параметры эксперимента, основные фрагменты программной реализации, результаты измерений, графики или сравнительные таблицы и содержательные выводы. При наличии стохастических процедур следует обеспечивать воспроизводимость результатов путем фиксации начального состояния генератора псевдослучайных чисел и документирования существенных параметров среды.

ЛР №1 посвящена построению и исследованию модели типа энкодер-декодер; ЛР №2 – квантизации и эффективности выполнения языковых моделей; ЛР №3 – параметрически эффективному дообучению с PEFT/LoRA/QLoRA и оценке адаптированной модели; ЛР №4 – настройке и исследованию агентной системы. Для ЛР №4 применяется бинарный текущий контроль «зачтено/не зачтено»: оцениваются корректность конфигурации, использование инструментов и навыков, демонстрация хода выполнения и понимание архитектуры агентной системы, а не качество ответа языковой модели само по себе.

При работе с агентными системами запрещается включать в отчеты и общедоступные репозитории реальные API-ключи, пароли и иные секреты. Инструменты агента должны запускаться с минимально необходимыми разрешениями, а эксперимент – выполняться в тестовой директории или учебном репозитории.

9.5. Методические указания по самостоятельной работе обучающихся

Самостоятельная работа обеспечивает подготовку к лекциям, лабораторным работам, текущему контролю и промежуточной аттестации. Она включает изучение рекомендуемой литературы и официальной технической документации, подготовку данных и программной среды, выполнение части вычислительных экспериментов, оформление отчетов и подготовку к защите лабораторных работ.

Практико-ориентированное задание по оптимизации диффузионных моделей выполняется самостоятельно с консультационной поддержкой в рамках КСР. Обучающийся исследует влияние параметров процесса диффузии и/или методов сокращения избыточных вычислений (DBCache, TaylorSeer), анализирует изменение времени выполнения и качества результата и формулирует выводы о применимости метода.

Для обучающихся заочной формы самостоятельная работа является основным видом учебной деятельности; основные вычислительные материалы предоставляются в форме воспроизводимых заданий в формате Jupyter Notebook и ссылок на первичные источники.

10. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ДЛЯ КОНТРОЛЯ ОСВОЕНИЯ ДИСЦИПЛИНЫ И КРИТЕРИИ ОЦЕНИВАНИЯ

Оценочные материалы текущей и промежуточной аттестации по дисциплине оформляются отдельным комплектом в соответствии с макетом оценочных материалов НГТУ. Комплект обеспечивает контроль достижения планируемых результатов обучения, указанных в разделе 1 настоящей РПД.

В состав текущего контроля по дисциплине входят:

– выполнение и защита лабораторной работы №1 по построению и исследованию модели типа энкодер-декодер;

– выполнение и защита лабораторной работы №2 по квантизации и сравнительному тестированию языковых моделей;
– выполнение и защита лабораторной работы №3 по PEFT/LoRA/QLoRA и оценке адаптированной языковой модели;
– выполнение лабораторной работы №4 по агентным системам с бинарной оценкой «зачтено/не зачтено»;
– выполнение практико-ориентированного задания по оптимизации генерации диффузионными моделями;
– устный и/или письменный контроль теоретического материала по темам дисциплины.

Промежуточная аттестация проводится в форме зачета с оценкой. К промежуточной аттестации допускаются обучающиеся, выполнившие обязательные лабораторные работы и обязательные задания текущего контроля. Конкретные задания, показатели, критерии и шкалы оценивания приведены в комплекте оценочных материалов текущей и промежуточной аттестации.

11. АДАПТАЦИЯ РАБОЧЕЙ ПРОГРАММЫ ДЛЯ ЛИЦ С ОВЗ И ИНВАЛИДОВ

В соответствии с Федеральным законом от 29.12.2012 № 273-ФЗ «Об образовании в Российской Федерации» профессиональное обучение и профессиональное образование обучающихся с ограниченными возможностями здоровья осуществляются на основе образовательных программ, адаптированных при необходимости для обучения указанных обучающихся.

Адаптация рабочей программы осуществляется при наличии заявления обучающегося и документов, определяющих необходимые специальные условия обучения. Для инвалидов специальные условия определяются также с учетом индивидуальной программы реабилитации или абилитации.

При освоении дисциплины могут применяться электронные учебные материалы в доступных форматах, увеличенное время на выполнение отдельных видов работ, альтернативные способы представления результатов и консультации с использованием ЭИОС. Конкретные формы адаптации выбираются с учетом индивидуальных потребностей обучающегося и технических возможностей университета без снижения требований к планируемым результатам обучения.