

Документ подписан простой электронной подписью

Информация об документе:

ФИО: Хоружий Людмила Ивановна

Должность: Директор института экономики и управления АПК

Дата подписания: 28.08.2025 15:55:55

Уникальный электронный ключ:

1e90b132d9b04dce67585160b015dddf2cb1e6a9



**МИНИСТЕРСТВО СЕЛЬСКОГО ХОЗЯЙСТВА
РОССИЙСКОЙ ФЕДЕРАЦИИ**

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ АГРАРНЫЙ УНИВЕРСИТЕТ –

МСХА имени К.А. ТИМИРЯЗЕВА»

(ФГБОУ ВО РГАУ - МСХА имени К.А. Тимирязева)

Институт экономики и управления АПК
Кафедра статистики и кибернетики

УТВЕРЖДАЮ:

Директор института
экономики и управления АПК

Л.И. Хоружий



«28» августа 2025 г.

**РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ
Б1.В.22 Большие языковые модели**

для подготовки бакалавров

ФГОС ВО

Направление: 09.03.02 Информационные системы и технологии

Направленность: Большие данные и машинное обучение,

Компьютерные науки и технологии искусственного интеллекта

Курс 4

Семестр 8

Форма обучения: очная

Год начала подготовки: 2025

Москва, 2025

Разработчики:

Демичев В.В., канд. экон. наук, доцент,
(ФИО, ученая степень, ученое звание)


(подпись)

«26» августа 2025 г.

Титов А.Д., ассистент

(ФИО, ученая степень, ученое звание)


(подпись)

«26» августа 2025 г.

Рецензент: Кийко П.В., канд.пед.наук, доцент
(ФИО, ученая степень, ученое звание)


(подпись)

«26» августа 2025 г.

Программа составлена в соответствии с требованиями ФГОС ВО, компетентно-стно-ролевых моделей в сфере искусственного интеллекта по направлению подготовки 09.03.02 Информационные системы и технологии, профессионального стандарта и учебного плана

Программа обсуждена на заседании кафедры статистики и кибернетики протокол № 11 от «26» августа 2025 г.

И.о. зав. кафедрой Уколова А.В., канд. экон. наук, доцент
(ФИО, ученая степень, ученое звание)


(подпись)

«26» августа 2025 г.

Согласовано:

Председатель учебно-методической
комиссии института экономики и управления АПК

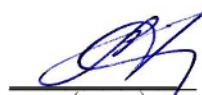
Гупалова Т.Н., канд. экон. наук, доцент
(ФИО, ученая степень, ученое звание)

протокол № 1


(подпись)

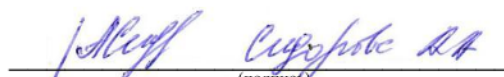
«28» августа 2025 г.

И.о. зав. выпускающей кафедрой статистики и кибернетики
Уколова А.В., канд. экон. наук, доцент
(ФИО, ученая степень, ученое звание)


(подпись)

«28» августа 2025 г.

Заведующий отделом комплектования ЦНБ


(подпись)

СОДЕРЖАНИЕ

АННОТАЦИЯ.....	5
1. ЦЕЛЬ ОСВОЕНИЯ ДИСЦИПЛИНЫ.....	6
2. МЕСТО ДИСЦИПЛИНЫ В УЧЕБНОМ ПРОЦЕССЕ	6
3. ПЕРЕЧЕНЬ ПЛАНИРУЕМЫХ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ, СООТНЕСЕННЫХ С ПЛАНИРУЕМЫМИ РЕЗУЛЬТАТАМИ ОСВОЕНИЯ ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ	7
4. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ	11
4.1 РАСПРЕДЕЛЕНИЕ ТРУДОЁМКОСТИ ДИСЦИПЛИНЫ ПО ВИДАМ РАБОТ.....	11
ПО СЕМЕСТРАМ.....	11
4.2 СОДЕРЖАНИЕ ДИСЦИПЛИНЫ.....	11
4.3 ПРАКТИЧЕСКИЕ ЗАНЯТИЯ.....	14
5. ОБРАЗОВАТЕЛЬНЫЕ ТЕХНОЛОГИИ	16
6. ТЕКУЩИЙ КОНТРОЛЬ УСПЕВАЕМОСТИ И ПРОМЕЖУТОЧНАЯ АТТЕСТАЦИЯ ПО ИТОГАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ	17
6.1. ТИПОВЫЕ КОНТРОЛЬНЫЕ ЗАДАНИЯ ИЛИ ИНЫЕ МАТЕРИАЛЫ, НЕОБХОДИМЫЕ ДЛЯ ОЦЕНКИ ЗНАНИЙ, УМЕНИЙ И НАВЫКОВ И (ИЛИ) ОПЫТА ДЕЯТЕЛЬНОСТИ.....	17
6.2. ОПИСАНИЕ ПОКАЗАТЕЛЕЙ И КРИТЕРИЕВ КОНТРОЛЯ УСПЕВАЕМОСТИ, ОПИСАНИЕ ШКАЛ ОЦЕНИВАНИЯ.....	29
7. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ	31
7.1 ОСНОВНАЯ ЛИТЕРАТУРА	31
7.2 ДОПОЛНИТЕЛЬНАЯ ЛИТЕРАТУРА.....	32
7.3 МЕТОДИЧЕСКИЕ УКАЗАНИЯ, РЕКОМЕНДАЦИИ И ДРУГИЕ МАТЕРИАЛЫ К ЗАНЯТИЯМ	32
7.4 СТАТЬИ, ОПУБЛИКОВАННЫЕ В НАУЧНЫХ ЖУРНАЛАХ 1 УРОВНЯ БЕЛОГО СПИСКА НАУЧНЫХ ЖУРНАЛОВ МИНОБРНАУКИ РОССИИ И СБОРНИКАХ НАУЧНЫХ РАБОТ КОНФЕРЕНЦИЙ УРОВНЯ А*	33
8. ПЕРЕЧЕНЬ РЕСУРСОВ ИНФОРМАЦИОННО- ТЕЛЕКОММУНИКАЦИОННОЙ СЕТИ «ИНТЕРНЕТ», НЕОБХОДИМЫХ ДЛЯ ОСВОЕНИЯ ДИСЦИПЛИНЫ.....	33
9. ПЕРЕЧЕНЬ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ И ИНФОРМАЦИОННЫХ СПРАВОЧНЫХ СИСТЕМ	33
10. ОПИСАНИЕ МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЙ БАЗЫ, НЕОБХОДИМОЙ ДЛЯ ОСУЩЕСТВЛЕНИЯ ОБРАЗОВАТЕЛЬНОГО ПРОЦЕССА ПО ДИСЦИПЛИНЕ	34
11. МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ СТУДЕНТАМ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ	37
Виды и формы отработки пропущенных занятий	37

12. МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ ПРЕПОДАВАТЕЛЯМ ПО ОРГАНИЗАЦИИ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ	37
---	----

АННОТАЦИЯ

рабочей программы учебной дисциплины Б1.В.22 Большие языковые модели для подготовки бакалавров по направлению 09.03.02 Информационные системы и технологии направленностям «Большие данные и машинное обучение», «Компьютерные науки и технологии искусственного интеллекта»

Цель освоения дисциплины: формирование у студентов профессиональных компетенций в области проектирования, реализации и адаптации больших языковых моделей (Large Language Models, LLM). Программа направлена на изучение методов тренировки, масштабирования и эффективного развертывания крупных генеративных моделей, используемых в современном искусственном интеллекте. Студенты овладеют навыками сбора и подготовки корпусов данных, тонкой настройки архитектуры моделей, оптимизации производительности и интерпретации результатов моделирования. В процессе обучения студенты научатся оценивать качество предсказательных способностей моделей с использованием специализированных инструментов и метрик. По завершении курса учащиеся будут способны разрабатывать, настраивать и внедрять большие языковые модели для широкого спектра прикладных задач, таких как обработка естественного языка, автоматическое создание контента, анализ текста и диалоговая система поддержки пользователей. В частности, особое внимание уделяется применению полученных знаний в сельскохозяйственных технологиях, таких как агроанализ текстов и документов, автоматизация обработки информации и оптимизация бизнес-процессов с использованием генеративных технологий.

Место дисциплины в учебном плане: дисциплина включена в часть учебного плана по направлению подготовки 09.03.02 Информационные системы и технологии, формируемую участниками образовательных отношений.

Требования к результатам освоения дисциплины: в результате освоения дисциплины формируются следующие компетенции (индикаторы): ПКкрмии-22 (Е-1. Базовый уровень).1; ПКкрмии-5 (FC-2. Базовый уровень).1; ПКкрмии-5 (FC-2. Базовый уровень).2; ПКкрмии-23 (LLM-1. Продвинутый уровень).1; ПКкрмии-23 (LLM-1. Продвинутый уровень).2; ПКкрмии-23 (LLM-1. Продвинутый уровень).3; ПКкрмии-23 (LLM-1. Продвинутый уровень).4; ПКкрмии-23 (LLM-1. Продвинутый уровень).5

Краткое содержание дисциплины:

Понятие большой языковой модели. Определение языковых моделей. Базовые понятия, типы моделей и роль контекста в обработке текста. Архитектура трансформеров. Модели трансформера, включая кодер и декодер, а также принципы работы механизма внутреннего внимания (self-attention). Многоуровневое многоголовое внимание. Преимущества многослойных механизмов внимания и их влияние на производительность модели. Создание больших языковых моделей. Причины большого числа параметров в трансформерах и практические последствия такого подхода. Генерация текста. Принцип работы LLM как инструмента автодополнения. Способы генерации осмысленных и грамотных текстов. Проблемы больших языковых моделей. Трудности, возникающие при обучении и эксплуатации LLM, включая потребление ресурсов и проблему

«галлюцинаций». Методы улучшения точности. Подходы для повышения эффективности и точности прогнозов, включая точную настройку (fine-tuning) и дистилляцию. Промпт инжиниринг. Сценарии использования больших языковых моделей в разных областях, таких как обработка и генерация текста, перевод и классификация данных и другие.

Общая трудоемкость дисциплины: 144 / 4 (часы/зач. ед.)

Промежуточный контроль: экзамен

1. Цель освоения дисциплины

Целью дисциплины «Большие языковые модели» является формирование у студентов профессиональных компетенций в области проектирования, реализации и адаптации больших языковых моделей (Large Language Models, LLM). Программа направлена на изучение методов тренировки, масштабирования и эффективного развертывания крупных генеративных моделей, используемых в современном искусственном интеллекте. Студенты овладеют навыками сбора и подготовки корпусов данных, тонкой настройки архитектуры моделей, оптимизации производительности и интерпретации результатов моделирования. В процессе обучения студенты научатся оценивать качество предсказательных способностей моделей с использованием специализированных инструментов и метрик. По завершении курса учащиеся будут способны разрабатывать, настраивать и внедрять большие языковые модели для широкого спектра прикладных задач, таких как обработка естественного языка, автоматическое создание контента, анализ текста и диалоговая система поддержки пользователей. В частности, особое внимание уделяется применению полученных знаний в сельскохозяйственных технологиях, таких как агроанализ текстов и документов, автоматизация обработки информации и оптимизация бизнес-процессов с использованием генеративных технологий.

2. Место дисциплины в учебном процессе

Дисциплина «Большие языковые модели» включена в часть учебного плана, формируемую участниками образовательных отношений. Дисциплина «Большие языковые модели» реализуется в соответствии с требованиями ФГОС ВО, ОПОП ВО и Учебного плана по направлению 09.03.02 Информационные системы и технологии и компетентностно-ролевых моделей в сфере искусственного интеллекта.

Дисциплина «Большие языковые модели» изучается четвертом курсе образовательного цикла.

Предшествующими курсами, включенными в учебный план, на которых непосредственно базируются дисциплина «Большие языковые модели», являются «Алгоритмизация и программирование», «Программирование на языке Python», «Введение в компьютерные науки на иностранном языке», «Основы теории управления и цифровой обработки сигналов», «Математическая статистика», «Инжиниринг данных», «Статистика для машинного обучения», «Методы искусственного интеллекта», «Методы машинного обучения», «Генеративный искусственный интеллект».

Дисциплина «Большие языковые модели» может быть использована при написании выпускной квалификационной работы.

Особенностью дисциплины является применение программирования Python 3.

Рабочая программа дисциплины «Большие языковые модели» для инвалидов и лиц с ограниченными возможностями здоровья разрабатывается индивидуально с учетом особенностей психофизического развития, индивидуальных возможностей и состояния здоровья таких обучающихся.

3. Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения образовательной программы

Изучение дисциплины направлено на формирование у обучающихся профессиональных (инструментальных) компетенций, представленных в таблице 1.

Таблица 1

Требования к результатам освоения учебной дисциплины «Большие языковые модели».

№ п/п	Код компетенции	Содержание компетенции (или её части)	Индикаторы компетенций	В результате изучения учебной дисциплины обучающиеся должны:		
				знать	уметь	владеть
1.	ПКкрми-22 (Е-1. Базовый уровень)	Способен осуществлять трудовые функции, обусловленные профессиональной ролью, в ОПД «Экономика, финансы и управление»	ПКкрми-22 (Е-1. Базовый уровень).1 Индикатор: Применяет методы и технологии организации и управления данными и знаниями в финансовой сфере. Уровень освоения индикатора: Основы организации систем сбора и обработки финансовых данных, в том числе в формате текстовых документов	основы организации систем сбора, хранения и обработки финансовых данных, включая структурированные и неструктурированные (текстовые) источники, а также принципы применения языковых моделей для их анализа	применять методы преобразования, аннотации и интеграции финансовых текстовых данных в пайплайны работы с LLM, а также настраивать модели для решения задач классификации, извлечения информации и генерации контента в финансовой сфере	методами и технологиями организации и управления данными и знаниями в финансовой сфере.
2.	ПКкрми-5 (FC-2. Базовый уровень)	Способен проводить фронтальные исследования в области фундаментальных и генеративных моделей	ПКкрми-5 (FC-2. Базовый уровень).1 Индикатор: Исследует и разрабатывает большие языковые модели (LLM) и другие модели для символьных данных. Уровень освоения индикатора: Владеет принципами работы систем на базе символьного искусственного интеллекта. Умеет использовать готовые нейросимволических фреймворков	подходы к исследованию и разработке больших языковых моделей (LLM) и других моделей для символьных данных.	использовать готовые нейросимволических фреймворков DeepProbLog, Neurosymbolic AI Toolkit	инструментами и практиками проектирования, реализации и анализа нейросимволических моделей для решения задач, требующих как генерации, так и логического рассуждения над символьными данными.

			DeepProbLog, Neurosymbolic AI Toolkit			
			ПККрмии-5 (FC-2. Базовый уровень).2 Индикатор: Развивает методы переноса знаний с адаптацией моделей. Уровень освоения индикатора: Применяет стандартные PEFT-методы (LoRA, Adapters) для специализации LLM. Оценивает качество адаптации через домен-специфичные метрики. Работает с задачами few-shot обучения	принципы и особенности методов переноса знаний и параметрически эффективной тонкой настройки, а также подходы к оценке адаптированных моделей с использованием домен-специфичных метрик	применять PEFT-методы для специализации больших языковых моделей под задачи финансовой и других предметных областей, а также оценивать качество адаптации	инструментами и практиками тонкой настройки LLM с минимальным потреблением ресурсов, включая реализацию, тестирование и верификацию адаптированных моделей в условиях ограниченных данных и специфики предметной области
3.	ПККрмии-23 (LLM-1. Продвинутый уровень)	Способен применять и (или) разрабатывать генеративные модели и БЯМ	ПККрмии-23 (LLM-1. Продвинутый уровень).1 Индикатор: Знает архитектуры генеративных моделей. Уровень освоения индикатора: Сравнивает архитектуры и выбирает подходящую под задачу	архитектуры генеративных моделей.	проводить обоснованное сравнение генеративных архитектур по ключевым параметрам и выбирать оптимальную для решения конкретной прикладной задачи	методами аналитического и экспериментального сопоставления генеративных моделей с целью рационального выбора и адаптации архитектуры под специфику домена и требования к результату.
			ПККрмии-23 (LLM-1. Продвинутый уровень).2 Индикатор: Оценивает производительность генеративных моделей. Уровень освоения индикатора: Интерпретирует зна-	систему количественных и качественных метрик оценки производительности генеративных моделей и их интерпретацию в контексте разных задач	интерпретировать значения метрик с учётом специфики данных и целей применения, чтобы делать обоснованные выводы о пригодности модели для практического использования	методами комплексной оценки и сопоставительного анализа генеративных моделей на основе метрик, позволяющими объективно судить об их эффективности, надёжности и применимости в реальных сценариях

			чения метрик для оценки применимости модели			
			ПККрмии-23 (LLM-1. Продвинутый уровень).3 Индикатор: Понимает роль латентного пространства в генеративных моделях. Уровень освоения индикатора: Использует логит-просмотр (logit lens) для анализа вывода	роль и структуру латентного пространства в генеративных моделях, включая его влияние на процесс генерации и семантическую интерпретируемость промежуточных представлений	применять метод logit lens и сопоставимые техники для анализа активаций на промежуточных слоях модели и выявления, как формируется финальный вывод на основе латентных представлений	инструментами интерпретации внутренней динамики генеративных моделей через анализ латентного пространства.
			ПККрмии-23 (LLM-1. Продвинутый уровень).4 Индикатор: Учитывает этические аспекты генерации. Уровень освоения индикатора: Применяет фильтры и постобработку	этические аспекты генерации.	выявлять потенциально нежелательный или вредоносный контент и применять технические меры	практиками и инструментами обеспечения безопасной и ответственной генерации, включая настройку и интеграцию фильтров, а также применение методов постобработки выходов LLM в соответствии с этическими и регуляторными требованиями
			ПККрмии-23 (LLM-1. Продвинутый уровень).5 Индикатор: Проводит валидацию и тестирование генеративных моделей. Уровень освоения индикатора: Автоматизирует валидацию на основе тестов и метрик	методы валидации и тестирования генеративных моделей на основе тестов и метрик.	автоматизировать валидацию на основе тестов и метрик.	практиками непрерывной автоматизированной верификации генеративных моделей, анализа отклонений и мониторинга деградации качества в процессе эксплуатации.

4. Структура и содержание дисциплины

4.1 Распределение трудоёмкости дисциплины по видам работ по семестрам

Общая трудоёмкость дисциплины составляет 4 зач.ед. (144 часа), их распределение по видам работ и семестрам представлено в таблице 2.

Таблица 2

Распределение трудоёмкости дисциплины по видам работ в 8 семестре

Вид учебной работы	Трудоёмкость	
	час. всего/*	в т.ч. по семестрам № 8/*
Общая трудоёмкость дисциплины по учебному плану	144	144
1. Контактная работа:	38,4/4	38,4/4
Аудиторная работа	36/4	36/4
<i>в том числе:</i>		
<i>лекции (Л)</i>	12	12
<i>практические занятия (ПЗ)</i>	24/4	24/4
<i>консультации перед экзаменом</i>	2	2
<i>контактная работа на промежуточном контроле (КРА)</i>	0,4	0,4
2. Самостоятельная работа (СРС)	105,6	105,6
<i>самостоятельное изучение разделов, самоподготовка (проработка и повторение лекционного материала и материала учебников и учебных пособий, подготовка к практическим занятиям)</i>	78,6	78,6
<i>Подготовка к экзамену (контроль)</i>	27	27
Вид промежуточного контроля:	Экзамен	

*-в том числе практическая подготовка

4.2 Содержание дисциплины

Таблица 3

Тематический план учебной дисциплины

Наименование разделов и тем дисциплин (укрупнённо)	Всего	Аудиторная работа			Внеаудиторная работа СР
		Л	ПЗ всего/*	ПКР	
Тема 1 «Основные понятия больших языковых моделей»	16	2	4	-	10
Тема 2 «Архитектура трансформеров»	16	2	4	-	10
Тема 3 «Многоуровневые структуры и повышение производительности»	16	2	4	-	10
Тема 4 «Создание больших языковых моделей»	16	2	4	-	10
Тема 5 «Генерация текста и проблема «галлюцинаций»	16/2	2	4/2	-	10
Тема 6 «Улучшение качества и применение больших языковых моделей»	34,6/2	2	4/2	-	28,6
контактная работа на промежуточном контроле (КРА)	0,4	-	-	0,4	-

Наименование разделов и тем дисциплин (укрупнённо)	Всего	Аудиторная работа			Внеаудиторная работа СР
		Л	ПЗ всего/*	ПКР	
Консультации перед экзаменом	2	-	-	2	-
Подготовка к экзамену (контроль)	27			27	
Итого по дисциплине	144/4	12	24/4	29,4	78,6

Тема 1. Основные понятия больших языковых моделей.

Определение языковой модели как системы, способной оценивать вероятность возникновения определенного фрагмента текста в заданном контексте. Изучение основных функций языковых моделей, таких как оценка вероятности предложений, завершение предложений и помощь в создании качественного контента. Характеристика общих черт языковых моделей, различающихся методами представления текста (слова, символы, подслова) и способами обработки контекста. Исторический экскурс от простых вероятностных моделей до сложных глубоких нейросетей. Важность учета контекста при построении лингвистически значимых решений. Использование механизмов, позволяющих эффективно обрабатывать широкий контекст (например, внутреннее внимание) для лучшего понимания смысла и смысловых оттенков текста. Классификация языковых моделей по типу обработки данных (последовательные, двунаправленные, трансформерные), масштабу (малые, средние, большие) и назначению (общего назначения, специализированные). Примеры популярных архитектур и областей их применения. Детализация особенностей больших языковых моделей (Big Language Model, BLM), их отличительные черты (количество параметров, объем данных для обучения, универсальность). Влияние размера модели на ее гибкость и производительность в решении сложных лингвистических задач.

Тема 2. Архитектура трансформеров.

Структура трансформеров, их отличия от предыдущих поколений моделей обработки естественного языка. Основные идеи, заложенные в архитектуру трансформеров, позволяющие эффективно решать задачи перевода, классификации и генерации текста. Рассмотрение устройства кодера и декодера в трансформере. Функция кодера заключается в преобразовании входного текста в компактное представление (вектор признаков), тогда как задача декодера – восстановление выходного текста на основе полученного векторного представления. Центральная концепция трансформеров – Self-Attention. Внутреннее устройство слоев внимания, структура вычислений и визуализации активации внимания.

Тема 3. Многоуровневые структуры и повышение производительности.

Концепции многоглавого внимания (Multi-Head Attention), его роли в увеличении выразительности и информативности модели. Особенности многоуровневой организации механизмов внимания, включающей большое число параллельно работающих блоков внимания. Анализ преимуществ использования многослойных схем внимания, таких как улучшенная обработка иерархической структуры текста, лучшая работа с длинными последовательностями и высокая эффективность выделения скрытых паттернов. Оценка влияния увеличения

глубины модели на ее производительность. Практические рекомендации по выбору оптимального количества слоев и головы внимания для достижения баланса между точностью и скоростью вычислений. Экспериментальные данные и примеры успешных реализаций многослойных моделей в приложениях обработки естественного языка.

Тема 4. Создание больших языковых моделей.

Факторы, определяющих необходимость увеличения количества параметров в больших языковых моделях. Преимущества глубокой архитектуры и ее связь с производительностью моделей на различных задачах. Последствия, возникающие вследствие увеличения размеров моделей, таких как повышенные требования к ресурсам хранения и вычислительным мощностям, замедление скорости вывода и увеличение стоимости обучения. Этапы обучения больших языковых моделей, включая сбор и подготовку датасетов, выбор гиперпараметров, технику предварительной и последующей настройки, оценку качества и мониторинг прогресса. Практические рекомендации по эффективному обучению больших моделей.

Тема 5. Генерация текста и проблема «галлюцинаций».

Понимание основ функционирования больших языковых моделей, таких как механизмы принятия решений, правила распределения вероятностей для продолжения текста и роль предварительного обучения на обширных датасетах. Факторы, влияющие на качество сгенерированного текста, включая связность, когерентность и правильность с точки зрения грамматики и стилистики. Критерии оценки качества сгенерированных текстов и техники улучшения их читабельности. Явление «галлюцинаций» – ситуации, когда модель генерирует недостоверную или несуществующую информацию. Причина возникновения таких артефактов, пути минимизации риска ошибочных выводов и стратегия проверки достоверности создаваемых данных.

Тема 6. Улучшение качества и применение больших языковых моделей.

Техники тонкой настройки (fine-tuning) и дистилляции (distillation) для улучшения производительности больших языковых моделей. Особенности выбора подходящей стратегии для конкретных задач и характеристик данных. Введение в понятие prompt engineering, техники формирования эффективных инструкций (промптов) для больших языковых моделей. Правила формулировки запросов и инструкций для получения качественных и точных ответов от модели. Обзор применений больших языковых моделей в задачах обработки естественного языка, таких как перевод, генерация текста, классификация документов и других направлениях. Примеры успешного внедрения больших языковых моделей в различных предметных областях.

4.3 Практические занятия

Таблица 4

Содержание практических занятий и контрольные мероприятия

Название раздела, темы	№ и название лекций/ практических занятий	Формируемые компетенции (индикаторы)	Вид контрольного мероприятия	Кол-во часов
Тема 1 «Основные понятия больших языковых моделей»	Лекция 1 «Основные понятия больших языковых моделей»	ПКкрмии-22 (Е-1. Базовый уровень).1, ПКкрмии-5 (ФС-2. Базовый уровень).1	-	2
	Практическая работа 1 «Реализация простой языковой модели на основе n-грамм»	ПКкрмии-22 (Е-1. Базовый уровень).1, ПКкрмии-5 (ФС-2. Базовый уровень).1	Устный опрос	4
Тема 2 «Архитектура трансформеров»	Лекция 2 «Архитектура трансформеров»	ПКкрмии-23 (LLM-1. Продвинутый уровень).1, ПКкрмии-23 (LLM-1. Продвинутый уровень).3	-	2
	Практическая работа 2 «Создание простого трансформера для задач перевода».	ПКкрмии-23 (LLM-1. Продвинутый уровень).1, ПКкрмии-23 (LLM-1. Продвинутый уровень).3	Устный опрос	4
Тема 3 «Многоуровневые структуры и повышение производительности»	Лекция 3 «Многоуровневые структуры и повышение производительности»	ПКкрмии-23 (LLM-1. Продвинутый уровень).1, ПКкрмии-23 (LLM-1. Продвинутый уровень).2	-	2
	Практическая работа 3 «Работа с механизмом многоголового внимания».	ПКкрмии-23 (LLM-1. Продвинутый уровень).1, ПКкрмии-23	Устный опрос	4

Название раздела, темы	№ и название лекций/ практических занятий	Формируемые компетенции (индикаторы)	Вид контрольного мероприятия	Кол-во часов
		(LLM-1. Продвинутый уровень).2		
Тема 4 «Создание больших языковых моделей»	Лекция 4 «Создание больших языковых моделей»	ПКкрмии-23 (LLM-1. Продвинутый уровень).1, ПКкрмии-23 (LLM-1. Продвинутый уровень).4	-	2
	Практическая работа 4 «Настройка и обучение небольшой языковой модели»	ПКкрмии-5 (FC-2. Базовый уровень).2, ПКкрмии-23 (LLM-1. Продвинутый уровень).5	Устный опрос	4
Тема 5 «Генерация текста и проблема «галлюцинаций»	Лекция 5 «Генерация текста и проблема «галлюцинаций»	ПКкрмии-23 (LLM-1. Продвинутый уровень).2, ПКкрмии-23 (LLM-1. Продвинутый уровень).4	-	2
	Практическая работа 5 «Fine-tuning предварительно обученной модели для задачи генерации текста»	ПКкрмии-5 (FC-2. Базовый уровень).2, ПКкрмии-23 (LLM-1. Продвинутый уровень).5	Устный опрос	4
Тема 6 «Улучшение качества и применение больших языковых моделей»	Лекция 6 «Улучшение качества и применение больших языковых моделей»	ПКкрмии-23 (LLM-1. Продвинутый уровень).4, ПКкрмии-23 (LLM-1. Продвинутый уровень).5	-	2
	Практическая работа 6 «Prompt Engineering и тестирование качества генерации»	ПКкрмии-5 (FC-2. Базовый уровень).2	Устный опрос	4

Таблица 5

Перечень вопросов для самостоятельного изучения дисциплины

№ п/п	Название раздела, темы	Перечень рассматриваемых вопросов для самостоятельного изучения
1.	Тема 1 «Основные понятия больших языковых моделей»	Развитие языков программирования и парадигмы искусственного интеллекта: связь с появлением и развитием больших языковых моделей. (ПККрмии-5 (FC-2. Базовый уровень).1, ПККрмии-22 (Е-1. Базовый уровень).1)
2.	Тема 2 «Архитектура трансформеров»	Реализация трансформера на PyTorch. (ПККрмии-23 (LLM-1. Продвинутый уровень).1, ПККрмии-5 (FC-2. Базовый уровень).1)
3.	Тема 3 «Многоуровневые структуры и повышение производительности»	Эксперимент с влиянием числа слоев и размерности head attention на качество и скорость обучения трансформера. (ПККрмии-23 (LLM-1. Продвинутый уровень).1, ПККрмии-23 (LLM-1. Продвинутый уровень).2, ПККрмии-23 (LLM-1. Продвинутый уровень).3)
4.	Тема 4 «Создание больших языковых моделей»	Оптимизация обучения больших языковых моделей: исследование эффективности Mixed Precision Training (FP16) и распределенного обучения. (ПККрмии-23 (LLM-1. Продвинутый уровень).2, ПККрмии-23 (LLM-1. Продвинутый уровень).5)
5.	Тема 5 «Генерация текста и проблема «галлюцинаций»	Исследование способов снижения эффекта «галлюцинаций»: введение фильтров и ограничений на выходные данные модели. (ПККрмии-23 (LLM-1. Продвинутый уровень).4, ПККрмии-23 (LLM-1. Продвинутый уровень).5)
6.	Тема 6 «Улучшение качества и применение больших языковых моделей»	Автоматизированная генерация тестовых заданий на основе large language models. (ПККрмии-23 (LLM-1. Продвинутый уровень).5, ПККрмии-5 (FC-2. Базовый уровень).2)

5. Образовательные технологии

Таблица 6

Применение активных и интерактивных образовательных технологий

№ п/п	Тема и форма занятия		Наименование используемых активных и интерактивных образовательных технологий (форм обучения)
1.	Тема 1 «Основные понятия больших языковых моделей»	ПЗ	Разбор конкретных ситуаций
2.	Тема 2 «Архитектура трансформеров»	ПЗ	Мозговой штурм
3.	Тема 3 «Многоуровневые структуры и повышение производительности»	ПЗ	Разбор конкретных ситуаций
4.	Тема 4 «Создание больших языковых моделей»	ПЗ	Разбор конкретных ситуаций

6. Текущий контроль успеваемости и промежуточная аттестация по итогам освоения дисциплины

6.1. Типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений и навыков и (или) опыта деятельности

Вопросы для подготовки к устным опросам

Тема 1. Основные понятия больших языковых моделей

1. Понятие большой языковой модели
2. Определение языковых моделей
3. Роль контекста в обработке текста
4. Долгосрочные и краткосрочные зависимости в обработке языка
5. Типы языковых моделей (n-граммные, рекуррентные, трансформеры)
6. Латентные переменные и эмбединги
7. Задача автокомплита и её значимость
8. Оценка вероятности появления токенов

Тема 2. Архитектура трансформеров

1. Устройство модели трансформера
2. Кодер и декодер в трансформере
3. Формирование входных и выходных векторов
4. Масштабируемость архитектуры трансформера
5. Компоненты трансформера: положение вложений
6. Нормализация и регуляризация в трансформерах
7. Внутренняя структура компонента Self-Attention
8. Параллельная обработка данных в трансформерах
9. Матрица внимания и её свойства
10. Многоголовое внимание и его роль в трансформерах

Тема 3. Многоуровневые структуры и повышение производительности

1. Многоуровневое многоголовое внимание
2. Связь количества уровней и точности модели
3. Переносимость знаний между уровнями
4. Ограничения использования глубокого многоуровневого внимания
5. Преимущества многослойных механизмов внимания
6. Скалируемость моделей с глубокими слоями
7. Автоматическое выделение важных участков текста
8. Практическое влияние на точность и скорость работы модели
9. Алгоритмы вычисления весов внимания
10. Оптимизация глубоких моделей для производственных задач

Тема 4. Создание больших языковых моделей

1. Причины большого числа параметров в трансформерах
2. Эволюция размеров моделей: BERT → GPT → LLAMA
3. Выбор оптимальной архитектуры и объёма параметров
4. Последствия высокого объёма параметров
5. Проблемы масштабирования и ресурсозатраты
6. Оптимизация расходов на хранение и обработку данных

7. Процесс обучения больших языковых моделей
8. Предварительное обучение vs тонкая настройка
9. Переиспользование ранее обученных моделей
10. Специфичные техники обучения больших моделей

Тема 5. Генерация текста и проблема «галлюцинаций»

1. Принцип работы больших языковых моделей (LLM)
2. Человеческий фактор в оценке генерации текста
3. Генерируемые тексты: осмысленность и грамматичность
4. Частота возникновения феномена «галлюцинаций»
5. Методы предотвращения ложных утверждений
6. Стратегии исправления и модификации текстов
7. Недостатки и ограничения больших языковых моделей («галлюцинации»)
8. Качество выхода моделей в сравнении с человеческим текстом
9. Испытания на стрессоустойчивость генеративных моделей
10. Возможности тестирования и оценки адекватности генерации

Тема 6. Улучшение качества и применение больших языковых моделей

1. Повышение точности прогнозов: fine-tuning и дистилляция
2. Методология передачи знаний при дистилляции
3. Применение transfer learning для больших языковых моделей
4. Методы оперативной инженерии (prompt engineering)
5. Формируемые запросами инструкции (Prompt templates)
6. Запросы на условную генерацию текста
7. Минимизация когнитивной нагрузки при взаимодействии с моделью
8. Применения больших языковых моделей в текстовой обработке
9. Решение мультязычных задач с помощью больших моделей
10. Фильтрация нежелательной информации и этические аспекты

Кейс-задачи от индустриального партнера «Россельхозбанка»

Кейс-задача 1

Тема 3 «Многоуровневые структуры и повышение производительности»

Кейс-задача от индустриального партнера: «Многоагентная LLM-система для комплексной оценки кредитоспособности агропредприятий»

«АгроФинанс Банк» - один из крупнейших кредиторов предприятий агропромышленного комплекса. Ежегодно мы обрабатываем тысячи кредитных заявок от сельхозпроизводителей. Каждое решение о кредитовании требует комплексного анализа:

1. Финансовых отчетов (структурированные данные)
2. Бизнес-планов и технологических карт (неструктурированные тексты)
3. Внешних факторов: прогнозы погоды, рыночные цены, геоданные полей
4. Качественных показателей: репутация заемщика, опыт управления, отраслевые экспертизы

Проблема: Существующие скоринговые модели, основанные на табличных данных, не учитывают более 40% релевантной информации, содержащейся в текстовых документах и внешних источниках. Это приводит к:

- Риску одобрения ненадежных заемщиков (когда в бизнес-плане есть неочевидные противоречия)
- Упущенным возможностям (когда перспективный стартап не проходит по формальным критериям, но имеет инновационную технологию)
- Высоким временным затратам аналитиков (до 8 часов на комплексную проверку одного предприятия)

Цель: разработать прототип многоуровневой LLM-системы «AgroRisk Analyzer», которая автоматизирует сбор, анализ и консолидацию разнородных данных о сельхозпредприятии-заемщике, формируя структурированное заключение для риск-менеджера.

Задание

Этап 1. Проектирование архитектуры многоагентной системы

Задача: спроектировать модульную систему, где разные LLM-агенты специализируются на конкретных типах данных и задач.

Требования:

1. Анализ документов (Document Agent):

Должен извлекать ключевые показатели из PDF-файлов: выручка, себестоимость, планы по посевным площадям, наименования культур.

Должен определять противоречия: например, если в тексте бизнес-плана заявлена урожайность 50 ц/га пшеницы, а в приложенном техкарте используются семена с потенциалом 30 ц/га.

2. Анализ внешних данных (External Data Agent):

Должен интегрироваться с внешними API (погода, цены на сырье).

Должен оценивать реалистичность планов заемщика: например, насколько заявленная цена реализации соответствует рыночным прогнозам на следующий сезон.

3. Агрегатор и генератор отчета (Orchestrator & Report Agent):

Оркестратор координирует работу всех агентов, распределяет задачи и управляет потоком данных.

Агент-отчет на основе выводов всех специалистов формирует итоговый отчет по структурированному шаблону.

Выход этапа: Схема архитектуры (Miro, draw.io) с описанием ролей, потоков данных и API-интерфейсов между агентами.

Этап 2: Разработка прототипа ключевых агентов

Реализовать двух наиболее критичных агентов на Python.

Требования к прототипу:

1. Document Agent (на основе RAG):

Реализовать загрузку и чанкование PDF-документов.

Настроить векторное хранилище (Chroma/FAISS) для семантического поиска.

Создать промты для извлечения финансовых и производственных метрик, а также для выявления логических несоответствий.

Тестовые данные: 50 сгенерированных бизнес-планов АПК (часть - с преднамеренными ошибками).

2. Orchestrator (Интеллектуальный маршрутизатор):

Реализовать классификатор входящих запросов (например, «Проанализировать бизнес-план», «Оценить риски по региону»).

Разработать простой workflow engine, который на основе запроса запускает нужную последовательность агентов.

Реализовать контекстный менеджер, который передает результаты работы одного агента другому.

Выход этапа: Рабочий код двух агентов на GitHub с примером запуска пайплайна на тестовом документе.

Этап 3. Оценка качества и создание MVP

Протестировать систему на реалистичных данных и измерить её эффективность.

Действия:

1. Создание тестового набора данных:

Подготовить 20 аннотированных кейсов (бизнес-план + внешние данные + вердикт эксперта-аналитика банка).

2. Оценка метрик:

Точность извлечения фактов (Fact Accuracy): % корректно извлеченных числовых показателей.

Полнота выявления рисков (Risk Recall): % обнаруженных системой противоречий и рисков из общего числа, отмеченных экспертом.

Время обработки: сравнение со средним временем аналитика.

3. Генерация отчета:

Настроить финального агента на формирование отчета в формате JSON, включающего: ключевые показатели, выявленные риски с уровнем критичности, общую рекомендацию (низкий/средний/высокий риск).

Выход этапа: Отчет с метриками, примерами успешных и неуспешных кейсов, демо-интерфейс (Streamlit/Gradio) для загрузки документа и просмотра отчета.

Этап 4. Подготовка стратегии внедрения и презентация

Подготовить материалы для принятия решения о пилотировании системы в банке.

Требования к финальной презентации:

1. Бизнес-ценность:

Оценка потенциального снижения кредитных потерь (%).

Расчет экономии времени аналитиков (часов в год).

Сценарии использования: скрининг заявок, мониторинг действующих заемщиков, стресс-тестирование портфеля.

2. Техническая дорожная карта:

Этапы промышленного внедрения (позапное подключение источников данных, обучение сотрудников).

Инфраструктурные требования (GPU, облако, безопасность).

План по дальнейшему улучшению системы (дообучение на реальных данных, добавление новых агентов).

3. Архитектурная схема итоговой системы:

Как система будет интегрирована с внутренними IT-системами банка (CRM, скоринговая платформа).

Кейс-задача 2

Тема 4 «Создание больших языковых моделей»

Кейс-задача от индустриального партнера: «Разработка доменно-специализированной LLM для консультирования в органическом земледелии»

Бизнес-контекст и проблема

Холдинг «БиоАгро» - лидер на рынке органической сельхозпродукции в России. Мы столкнулись с ключевой проблемой:

Острый дефицит экспертов-агрономов по органическому земледелию. Консультации требуют глубоких знаний:

1. Биологических методов защиты растений (биопрепараты, энтомофаги)
2. Международных стандартов сертификации (EU 848/2018, NOP, ГОСТ 33980-2016)
3. Локальных почвенно-климатических условий
4. Экономики органического производства

Существующие LLM дают опасно неточные рекомендации:

Советуют запрещенные в органике вещества

Не знают российских нормативов

Не учитывают специфику регионов

Допускают критические «галлюцинации» в дозировках

Бизнес-последствия: ошибки в консультациях приводят к потере сертификации, убыткам фермеров и репутационным рискам для холдинга.

Цель: разработать с нуля или дообучить специализированную LLM «BioAgroExpert» с контролируемой точностью $\geq 95\%$ в предметной области.

Задание

Этап 1. Планирование цикла создания модели и сбор данных

Сформировать стратегию и корпус данных для обучения.

Требования:

1. Выбор базовой архитектуры и обоснование:

Провести сравнительный анализ: обучение с нуля Russian GPT vs. дообучение открытой модели (Llama 3, Qwen, GigaChat).

Рассчитать требования к ресурсам (VRAM, время обучения) для каждого варианта.

Рекомендация: выбрать оптимальный вариант с учетом бюджета 5 млн руб. и срока 6 месяцев.

2. Формирование доменного корпуса:

Источники данных:

- Научные статьи и диссертации по органическому земледелию (PDF, 10 000+ документов)
- Технологические регламенты и стандарты (EU, ГОСТ, внутренние стандарты «БиоАгро»)
- Протоколы полевых экспериментов холдинга за 10 лет
- Вопрос-ответные пары из чатов агрономов с фермерами (анонимизированные)

Объем целевого корпуса: не менее 50 млн токенов качественного текста.

3. Дизайн пайплайна предобработки:

Разработать ETL-пайплайн для очистки и нормализации данных.

Особое внимание: извлечение текста из сканов, обработка таблиц с нормами внесения, удаление дубликатов.

Выход этапа: технический отчет с архитектурной схемой, планом сбора данных и оценкой ресурсов.

Этап 2. Проектирование и реализация процесса обучения

Создать полный тренировочный пайплайн.

Требования:

1. Проектирование многоэтапного обучения:

Этап А (Pre-training / Continued Pre-training): Адаптация модели на общем агротексте.

Этап В (Supervised Fine-Tuning): Обучение на инструкциях (50 000+ пар "вопрос-развернутый ответ").

Этап С (Alignment): Выравнивание под предпочтения экспертов (безопасность, точность, стиль).

2. Разработка датасетов:

Создать инструмент для полуавтоматической генерации SFT-данных на основе корпуса.

Разработать методику сбора human feedback для этапа RLHF/DPO:

- Привлечь 20 экспертов-агрономов холдинга
- Создать интерфейс для сравнения ответов модели
- Определить оценочную рубрику (критерии: точность, безопасность, полнота)

3. Выбор методов оптимизации:

Обосновать использование LoRA/QLoRA для эффективного дообучения.

Спланировать Distributed Training стратегию для ускорения.

Выход этапа: Ноутбуки с кодом предобработки, генерации данных и тренировочные скрипты. Прототип интерфейса для сбора обратной связи.

Этап 3. Валидация, оценка и создание MVP

Провести комплексное тестирование модели.

Требования:

1. Разработка системы оценки:

Стандартные метрики: Perplexity, BLEU, ROUGE.

Доменные метрики:

- Точность рекомендаций: % ответов, соответствующих стандартам органики.
- Коэффициент безопасности: отсутствие запрещенных веществ в советах.
- Географическая релевантность: учет региональных условий.

2. Создание тестового набора (Test Suite):

1000+ валидированных вопросов разного типа:

- Фактологические ("Норма внесения триходермы?")
- Диагностические ("Желтеют листья томата в теплице")
- Планирование ("Севооборот после картофеля")
- Критические ("Можно ли использовать глифосат?")

3. Сравнение с baseline:

Тестирование против ChatGPT-4, Claude 3, GigaChat.

Слепое оценивание экспертами.

Выход этапа: детальный отчет с метриками, сравнительной таблицей, примерами ответов. Рабочий API прототипа модели.

Этап 4. План промышленного внедрения и бизнес-презентация

Подготовить решение к использованию в холдинге.

Требования к финальной презентации:

1. Архитектура конечного продукта:

Схема развертывания (облако/он-премис)

Интеграция с мобильным приложением для фермеров

Система мониторинга и переобучения модели

2. Бизнес-модель и ROI:

Расчет экономии на консультациях (эксперты → модель)

Оценка снижения рисков потери сертификации

Потенциал монетизации (SaaS для других хозяйств)

3. Дорожная карта развития:

Этапы коммерциализации

План расширения функционала (мультимодальность, прогнозы урожайности)

Стратегия защиты IP (патенты на датасеты и методы обучения)

4. Этический протокол:

Система ответственности за рекомендации

Механизм эскалации сложных кейсов к человеку

Политика обновления знаний (новые стандарты, исследования)

Кейс-задача 3

Тема 5 «Генерация текста и проблема «галлюцинаций»

Кейс-задача от индустриального партнера: «Разработка системы генерации производственных отчетов для ГК "АгроТех" с нулевым уровнем галлюцинаций»

Проблема: каждое из 50+ предприятий холдинга обязано предоставлять:

1. Ежедневные операционные отчеты (что сделано, какие проблемы)
2. Еженедельные производственные сводки (выполнение плана, анализ отклонений)
3. Аварийные отчеты (при нештатных ситуациях)

Текущий процесс:

- Полевые работники → диктуют информацию по телефону/в мессенджере → менеджер → вручную структурирует в отчет → отправляет в головной офис.
- Время на один отчет: 1-3 часа
- Потери от задержек: до 15% операционных решений принимаются на устаревших данных
- Качественные ошибки: человеческий фактор, разночтения в терминологии

Попытка автоматизации: Пилотное внедрение стандартной LLM (GPT-4) для генерации отчетов из сырых сообщений дало катастрофические результаты:

- Галлюцинации по техникам: "Трактор John Deere 8310" → "Комбайн CASE IH 9250"
- Вымышленные показатели: Приписывание несуществующих объемов работ
- Опасные рекомендации: "Увеличить дозу препарата X в 2 раза" (при отсутствии такой информации в исходных данных)
- Юридические риски: Расхождения между сгенерированным отчетом и фактическими работами

Цель: разработать систему генерации отчетов с гарантированной фактологической точностью $\geq 99.9\%$ и полным исключением вредоносных галлюцинаций.

Задание

Этап 1. Анализ проблемы и проектирование архитектуры "Генерация + Верификация"

Спроектировать многоуровневую систему, где каждый этап контролирует достоверность.

Требования:

1. Анализ источников галлюцинаций в контексте АПК:
Составить таксономию типов ошибок: вымысел фактов, искажение значений, добавление несуществующих связей.
Проанализировать 100 реальных примеров сгенерированных отчетов и классифицировать ошибки.
2. Архитектура с тройным контролем:

Модуль извлечения фактов (Fact Extractor): Выделяет из сырого текста утверждения (субъект-действие-объект-параметр).

Модуль генерации отчета (Constrained Generator): Генерирует текст СТРОГО на основе извлеченных фактов.

Модуль верификации (Fact Checker): Сравнивает сгенерированный отчет с исходными фактами.

3. Дизайн промтов и ограничений:

Разработать систему промтов с жесткими шаблонами и ограничениями генерации.

Предложить метод контекстуального grounding (привязки каждого утверждения к источнику).

Выход этапа: Архитектурная схема системы, таксономия ошибок, спецификация промтов и протоколов контроля.

Этап 2. Разработка и обучение модуля "Constrained Generator"

Создать специализированную модель для генерации с нулевым вымыслом.

Требования:

1. Создание датасета без галлюцинаций:

Собрать и разметить 10 000+ пар "сырые данные → эталонный отчет"

Внести контролируемые искажения для обучения модели распознаванию галлюцинаций

Особое внимание: числовые данные, наименования техники, химикатов, полей

2. Выбор и обоснование подхода:

Вариант А: Fine-tuning существующей LLM с строгим контролем на уровне лосс-функции

Вариант В: Разработка rule-based + нейросетевого гибрида

Вариант С: Применение контролируемой генерации с графом знаний

3. Метрики обучения:

Основная: Fact F1-score (точность воспроизведения фактов)

Дополнительная: Hallucination Rate (доля вымышленных утверждений)

Критический порог: Hallucination Rate < 0.1%

Выход этапа: Датасет, тренировочные скрипты, обученная модель с метриками качества.

Этап 3. Создание системы верификации и конвейера "человек в петле"

Разработать механизмы обнаружения и исправления ошибок.

Требования:

1. Многоуровневая верификация:

Уровень 1: Автоматическая проверка числовых диапазонов (урожайность не может быть 1000 ц/га)

Уровень 2: Сверка с базами знаний (существует ли такое поле/техника у предприятия)

Уровень 3: LLM-based fact checking (модель проверяет модель)

2. Система эскалации:

Разработать алгоритм оценки уверенности системы в корректности отчета

Создать протокол передачи отчетов с низкой уверенностью менеджеру-человеку

Интерфейс для быстрой правки и обратной связи

3. Контрольный датасет "ловушек":

Создать 500+ специальных тестовых примеров, провоцирующих галлюцинации

Измерять устойчивость системы к преднамеренным искажениям

Выход этапа: Код системы верификации, интерфейс для эскалации, тестовый набор "ловушек" с результатами.

Этап 4. Промышленные испытания и план внедрения

Подготовить систему к использованию на 5 пилотных предприятиях.

Требования к финальной презентации:

1. Результаты тестирования:

Сравнение с baseline (GPT-4, Claude) по критериям:

- Фактическая точность
- Скорость генерации
- Стоимость обработки

Результаты на "ловушках"

2. Архитектура промышленного решения:

Схема интеграции с ERP-системой холдинга (1C, SAP)

План развертывания на edge-устройствах (для работы без интернета)

Система мониторинга качества в реальном времени

3. Экономическое обоснование:

Расчет ROI: экономия времени менеджеров vs. стоимость системы

Оценка предотвращенных потерь от ошибочных решений

Модель масштабирования на все 50+ предприятий

4. Юридический и этический протокол:

Механизм цифровой подписи и аудита сгенерированных отчетов

Распределение ответственности между системой и оператором

Протоколы реагирования на обнаруженные ошибки

Кейс-задача 4

Тема 6 «Улучшение качества и применение больших языковых моделей»

Кейс-задача от индустриального партнера: «Оптимизация системы чат-поддержки для сети агромагазинов «АгроМаркет»

Бизнес-контекст и проблема

«АгроМаркет» обслуживает 50 000+ клиентов ежемесячно через:

1. Онлайн-чат на сайте и в приложении
2. Телефонную поддержку
3. Мессенджеры (WhatsApp, Telegram)

Проблемы существующей системы:

1. Низкая точность базовой LLM:
 - 40% ответов требуют эскалации к оператору
 - Частые ошибки в рекомендациях препаратов
 - Незнание специфики региональных ассортиментов
2. Высокая нагрузка на экспертов:
 - 15 штатных агрономов-консультантов
 - Среднее время ответа: 15 минут
 - Пиковые нагрузки (сезон посевной/уборки) — до 2 часов ожидания
3. Потеря продаж:
 - 25% клиентов не дожидаются ответа
 - Конверсия из чата в покупку: всего 12%

Финансовые потери: Оценка упущенной выручки — 120+ млн руб. в год.

Цель: повысить качество автоматических ответов с 60% до 90% точности с помощью комбинации fine-tuning и продвинутого prompt engineering.

Задание

Этап 1. Анализ и проектирование стратегии улучшения

Проанализировать текущие диалоги и разработать двухуровневую стратегию улучшения.

Требования:

1. Анализ 5000+ реальных диалогов:
 - Классификация типов запросов (справочные, рекомендательные, проблемные)
 - Выявление паттернов успешных/неуспешных ответов
 - Анализ самых частых и самых дорогостоящих ошибок
2. Сравнение подходов:
 - Prompt engineering: Создание библиотеки промтов для разных категорий запросов
 - Fine-tuning: Дообучение модели на исторических диалогах экспертов
 - Гибридный подход: Когда применять каждый метод
3. Выбор модели и обоснование:
 - Сравнение GPT-4 Turbo, Claude 3, отечественных моделей (GigaChat, YandexGPT)
 - Критерии: стоимость API, скорость, поддержка русского языка, возможность fine-tuning

Выход этапа: Детальный отчет с классификацией запросов, рекомендациями по архитектуре решения и экономическим обоснованием выбранного подхода.

Этап 2. Prompt Engineering для повышения точности

Разработать и протестировать систему умных промтов.

Требования:

1. Создание библиотеки контекстных промтов:
 - Системные промты для разных ролей (консультант по СЗР, эксперт по технике, агроном)
 - Few-shot примеры для 20+ наиболее частых сценариев
 - Цепочки промтов (Chain-of-Thought) для сложных диагностических вопросов
2. Динамический выбор промта:
 - Разработать классификатор запросов (на основе эмбедингов)
 - Система автоматического выбора оптимального промта
 - Механизм А/В тестирования разных версий промтов
3. Интеграция с внешними данными:
 - Промты с RAG (база препаратов, каталог техники, региональные рекомендации)
 - Ограничение ответов только проверенной информацией

Выход этапа: Библиотека промтов (50+ вариантов), классификатор запросов, результаты А/В тестирования на 1000 диалогах.

Этап 3. Fine-tuning доменной модели

Создать специализированную модель для агроконсультаций.

Требования:

1. Подготовка датасета для fine-tuning:
 - 10 000+ пар "вопрос-экспертный ответ" из исторических диалогов
 - Разметка по категориям сложности
 - Балансировка датасета (редкие, но важные сценарии)
2. Выбор стратегии fine-tuning:
 - Full fine-tuning vs LoRA/QLoRA
 - Многоэтапное обучение: сначала на справочных данных, потом на диалогах
 - Контрольные точки для разных уровней специализации
3. Метрики качества:
 - Точность рекомендаций (экспертная оценка)
 - Коэффициент эскалации (сколько диалогов передано человеку)
 - Конверсия в продажу (интеграция с CRM)

Выход этапа: Обученная модель, скрипты fine-tuning, сравнительный анализ качества до/после обучения.

Этап 4. Интеграция и промышленные испытания

Внедрить систему на реальном трафике.

Требования:

1. Архитектура системы:
 - Многоуровневая система ответов (быстрые ответы → сложные консультации)
 - Механизм плавной передачи диалога человеку

Система обучения на новых диалогах

2. Пилотное внедрение:

3 региона, 5000+ клиентов ежемесячно

Сравнение метрик с контрольной группой (старая система)

Мониторинг в реальном времени

3. Критерии успеха пилота:

Увеличение конверсии в покупку на 40%

Снижение времени ответа до 1 минуты

Удовлетворенность клиентов (NPS) ≥ 8.5

Выход этапа: Рабочая система в production, дашборд с метриками, отчет об эффективности пилота.

Этап 5. Презентация результатов и план масштабирования

Требования к финальной презентации:

1. Результаты улучшения качества:

Сравнение точности ответов: базовая LLM vs промты vs fine-tuned модель

Экономия на операционных расходах

Рост выручки от чат-продаж

2. Архитектура итогового решения:

Схема работы системы end-to-end

Интеграция с CRM, ERP, системой управления знаниями

Механизмы постоянного улучшения (обратная связь от операторов)

3. Бизнес-кейс для масштабирования:

ROI по месяцам

План внедрения на все 250+ магазинов

Дополнительные возможности (прогноз спроса, персонализированные предложения)

4. Риски и их минимизация:

Юридические аспекты рекомендаций препаратов

Защита от злоупотреблений

План аварийного перехода на операторов

6.2. Описание показателей и критериев контроля успеваемости, описание шкал оценивания

В восьмом семестре для оценки знаний ведется на основе рейтинговой оценки студента, которая формируется как сумма баллов за участие в устном опросе на практических занятиях. Студент допускается к сдаче экзамена с оценкой при достижении рейтинга 60%.

Максимальная оценка за участие в опросе также 10 баллов.

9 баллов - ставится при наличии незначительных неточностей в ответе.

8 баллов - при наличии негрубых ошибок в ответе, которые не привели к ложным выводам и неверному пониманию сути вопроса.

7 баллов - сделаны неверные выводы по применяемым методам, при этом общее понимание применяемых методов не искажено.

6-5 баллов - нарушена логика в понимании применяемых методов.

Количество баллов складывается следующим образом: 6 устных вопросов * 10 (максимальное количество) баллов = 60 баллов (максимально возможное количество набранных баллов). В процентах (количество набранных баллов / максимально возможная сумма баллов) * 100.

Участие в интерактивных занятиях может быть зачтено активным студентам как участие в опросе по теме, на котором применялись интерактивные технологии.

На экзамене студент может получить максимальное количество баллов равное 100. Далее итоговая оценка определяется следующим образом. Если текущий рейтинг студента составляет 51 балл и выше, а на зачете с оценкой студент получил от 85 баллов и выше («отлично»), то итоговая оценка 51 балл + 85 баллов («отлично»).

Промежуточный контроль – экзамен.

Таблица 7

Шкала оценивания (средний балл)	Экзамен
> 136	Отлично
120-135	Хорошо
104-119	Удовлетворительно
0-103	Неудовлетворительно

Положительными оценками, при получении которых дисциплина засчитывается в качестве пройденной, являются оценки «удовлетворительно», «хорошо» и «отлично».

Если получена оценка «неудовлетворительно» по дисциплине, то необходимо, после консультации с преподавателем, в течение 10 календарных дней следующего семестра подготовить ответы на ряд вопросов, предусмотренных программой обучения, и представить результаты этих ответов преподавателю.

Критерии оценивания результатов обучения

Таблица 8

Оценка	Критерии оценивания
Высокий уровень «5» (отлично)	оценку «отлично» заслуживает студент, освоивший знания, умения, компетенции и теоретический материал без пробелов; выполнивший все задания, предусмотренные учебным планом на высоком качественном уровне; практические навыки профессионального применения освоенных знаний сформированы. Компетенции, закреплённые за дисциплиной, сформированы на уровне – высокий.
Средний уровень «4» (хорошо)	оценку «хорошо» заслуживает студент, практически полностью освоивший знания, умения, компетенции и теоретический материал, учебные задания не оценены максимальным числом баллов, в основном сформировал практические навыки. Компетенции, закреплённые за дисциплиной, сформированы на уровне – хороший (средний).
Пороговый уровень «3» (удовлетворительно)	оценку «удовлетворительно» заслуживает студент, частично с пробелами освоивший знания, умения, компетенции и теоретический материал, многие учебные задания либо не выполнил, либо они оценены числом баллов близким к минимальному, некоторые практические навыки не сформированы. Компетенции, закреплённые за дисциплиной, сформированы на уровне – достаточный.
Минимальный уровень «2» (неудовлетворительно)	оценку «неудовлетворительно» заслуживает студент, не освоивший знания, умения, компетенции и теоретический материал, учебные задания не выполнил, практические навыки не сформированы. Компетенции, закреплённые за дисциплиной, не сформированы.

7. Учебно-методическое и информационное обеспечение дисциплины

7.1 Основная литература

1. Нугуманова, А. Б. Автоматизированная обработка текстовых массивов : учебник и практикум для вузов / А. Б. Нугуманова. — 2-е изд. — Моск-

ва : Издательство Юрайт, 2024. — 82 с. — (Высшее образование). — ISBN 978-5-534-20738-5. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: <https://urait.ru/bcode/558668> (дата обращения: 30.07.2025).

2. Станкевич, Л. А. Интеллектуальные системы и технологии: учебник и практикум для вузов / Л. А. Станкевич. — 2-е изд., перераб. и доп. — Москва : Издательство Юрайт, 2025. — 478 с. — (Высшее образование). — ISBN 978-5-534-20363-9. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: <https://urait.ru/bcode/560754> (дата обращения: 21.08.2025).

3. Шунейко, А. А. Квантитативная лингвистика и новые информационные технологии : учебник для вузов / А. А. Шунейко, И. А. Авдеенко. — Москва : Издательство Юрайт, 2025. — 347 с. — (Высшее образование). — ISBN 978-5-534-15446-7. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: <https://urait.ru/bcode/567628> (дата обращения: 30.07.2025).

7.2 Дополнительная литература

1. Layer by Layer: Uncovering Hidden Representations in Language Models [Электронный ресурс] / Oscar Skea, Md Rifat Arefin, Dan Zhao, Niket Patel, Jalal Naghiyev, Yann LeCun, Ravid Shwartz-Ziv // Proceedings of the 40th International Conference on Machine Learning (ICML 2025). — Режим доступа: <https://icml.cc/virtual/2025/oral/47214> (дата обращения: 01.07.2025).

2. When Do LLMs Help With Node Classification? A Comprehensive Analysis [Электронный ресурс] / Xixi Wu, Yifei Shen, Fangzhou Ge, Caihua Shan, Yizhu Jiao, Xiangguo Sun, Hong Cheng // Proceedings of the 40th International Conference on Machine Learning (ICML 2025). — Режим доступа: <https://icml.cc/virtual/2025/poster/45461> (дата обращения: 01.07.2025).

3. Synthesizing Privacy-Preserving Text Data via Finetuning without Finetuning Billion-Scale LLMs [Электронный ресурс] / Bowen Tan, Zheng Xu, Eric Xing, Zhiting Hu, Shanshan Wu // Proceedings of the 40th International Conference on Machine Learning (ICML 2025). — Режим доступа: <https://icml.cc/virtual/2025/poster/45901> (дата обращения: 01.07.2025)

7.3 Методические указания, рекомендации и другие материалы к занятиям

1. Большие данные: Учебное пособие / В. В. Демичев, Д. В. Быков, А. С. Невзоров, В. С. Токарев. — Москва : Российский государственный аграрный университет, 2024. — 86 с.

2. Алгоритмизация и программирование: Учебное пособие / В. В. Демичев, Д. В. Быков, Д. Э. Храмов [и др.]. – Москва: Российский государственный аграрный университет, 2024. – 248 с.

7.4 Статьи, опубликованные в научных журналах 1 уровня Белого списка научных журналов Минобрнауки России и сборниках научных работ конференций уровня А*

1. Ashwani, S., Hegde, K., Reddy Mannuru, N., Singh Sengar, D., Jindal, M., Chaitanya Rao Kathala, K., Banga, D., Jain, V., & Chadha, A. (2024). Cause and Effect: Can Large Language Models Truly Understand Causality?. Proceedings of the AAAI Symposium Series, 4(1), 2-9. <https://doi.org/10.1609/aaais.v4i1.31764>

2. Ding, N., Qin, Y., Yang, G. et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. Nat Mach Intell 5, 220–235 (2023). <https://doi.org/10.1038/s42256-023-00626-4>

3. McDermott, M.B.A., Yap, B., Szolovits, P. et al. Structure-inducing pre-training. Nat Mach Intell 5, 612–621 (2023). <https://doi.org/10.1038/s42256-023-00647-z>

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

1. Официальный сайт Python. URL: <https://www.python.org/> (открытый доступ)

2. Официальный сайт дистрибутива языков программирования Python и R Anaconda. URL: <https://www.anaconda.com/> (открытый доступ)

3. Kaggle. URL: <https://www.kaggle.com/> (открытый доступ).

9. Перечень программного обеспечения и информационных справочных систем

Таблица 9

Перечень программного обеспечения

№ п/п	Наименование раздела учебной дисциплины	Наименование программы	Тип программы	Автор	Год разработки
1	Тема 1 «Основные понятия больших языковых моделей»	Jupyter Notebook/nltk/natasha	Интерактивная веб-среда для разработки/ Библиотека (пакет) для обработки естественного языка)	Project Jupyter/Стивен Бёрд, Эдвард Лопер и др./ Александр Кошелев и сообщество.	2015/2001/2016
2	Тема 2 «Архитектура трансформеров»	Transformers/torch	Библиотека (пакет) для Python, предоставляющая API и инструменты для работы с архитектурами трансформе-	Hugging Face, Inc./ Адам Паске, Сэм Гросс, Сумит Чинтала	2019/2016

			ров/Фреймворк для машинного обучения с открытым исходным кодом		
3	Тема 3 «Многоуровневые структуры и повышение производительности»	Библиотека transformers	Библиотека (пакет) для Python, предоставляющая API и инструменты для работы с архитектурами трансформеров	Hugging Face, Inc.	2019
4	Тема 4 «Создание больших языковых моделей»	Библиотеки transformers, datasets, GPU-ускорители (CUDA)	Библиотека (пакет) для Python, предоставляющая API и инструменты для работы с архитектурами трансформеров/ Библиотека для лёгкой и эффективной загрузки и обработки наборов данных/ Параллельная вычислительная платформа и API модель программирования	Hugging Face, Inc./ Hugging Face, Inc./ NVIDIA Corporation	2022/2020/ 2006
5.	Тема 5 «Генерация текста и проблема «галлюцинаций»	Pytorch-lightning	Обёртка (high-level interface) над PyTorch для упрощения процесса исследования	William Falcon и сообщество.	2019
6.	Тема 6 «Улучшение качества и применение больших языковых моделей»	Платформа Hugging Face Hub	Онлайн-платформа для совместной работы в области машинного обучения	Hugging Face, Inc.	2020

10. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине

Таблица 10

**Сведения об обеспеченности специализированными аудиториями,
кабинетами, лабораториями**

Наименование специальных помещений и помещений для самостоятельной работы (№ учебного корпуса, № аудитории)	Оснащенность специальных помещений и помещений для самостоятельной работы
1	2
<i>учебная аудитория для проведения занятий лекционного типа, учебная аудитория для проведения занятий семинарского типа, учебная аудитория для проведения курсового проектирования (выполнения курсовых работ), учебная аудитория для групповых и индивидуальных консультаций, учебная аудитория для текущего контроля и промежуточной аттестации (2й учебный корпус, 102 ауд.)</i>	1. Компьютер – 29 шт.; 2. Стенд «Сергеев Сергей Степанович 1910-1999» (Инв.№591013/25) – 1 шт.; 3. Огнетушитель порошковый (Инв. №559527) – 1 шт.; 4. Подвесное крепление к огнетушителю (Инв. № 559528) – 1 шт.; 5. Жалюзи (Инв. №1107-221225, Инв. №1107-221225) – 2шт.; 6. Стул – 29 шт.; 7. Стол компьютерный – 28 шт.; 8. Стол для преподавателя – 1 шт.; 9. Доска маркерная (Инв. № 558762/5) – 1 шт.; 10. Трибуна напольная (без инв. №) – 1 шт.
<i>учебная аудитория для проведения занятий семинарского типа, учебная аудитория для проведения курсового проектирования (выполнения курсовых работ), учебная аудитория для групповых и индивидуальных консультаций, учебная аудитория для текущего контроля и промежуточной аттестации, помещение для самостоятельной работы (2й учебный корпус, 106 ауд.)</i>	1. Рабочая станция FORSITE THI516G512G, Российская Федерация A4Tech Fstyler F1512 – 16 шт.; 2. Стол наборный (Инв. №410136000010828) – 1 шт. 3. Стол компьютерный (Инв. № 410136000010813-410136000010827) – 15 шт.; 4. Стул (Инв. № 410136000010829-410136000010853) – 25 шт.; 5. Интерактивная панель (Инв. № 410124000603715) – 1 шт.
<i>учебная аудитория для проведения занятий семинарского типа, учебная аудитория для проведения курсового проектирования (выполнения курсовых работ), учебная аудитория для групповых и индивидуальных консультаций, учебная аудитория для текущего контроля и промежуточной аттестации, помещение для самостоятельной работы (2й учебный корпус, 302 ауд.)</i>	1. Компьютер – 16 шт. 2. Телевизор – 1 шт. 3. Стол для преподавателя – 1 шт. 4. Стол компьютерный – 16 шт. 5. Стул офисный – 17 шт. 6. Компьютер: PRO-3159209 Intel Core i5-10400 2900МГц, Intel B460, 16Гб DDR4, Intel UHD Graphics 630 (встроенная), SSD 240Гб, 500Вт, Mini-Tower – 1 шт. 7. Кондиционер HAIER HSU -24HPL03/R3 (Инв. № 210134000062198) – 1 шт. 8. Вешалка напольная (Инв.№1107-333144, Инв.№1107-333144) – 2 шт.
<i>учебная аудитория для проведения занятий лекционного типа, учебная аудитория для проведения занятий семинарского типа, учебная аудитория для проведения курсового проектирования (выполнения курсовых работ), учебная аудитория для групповых и индивидуальных консультаций, учебная аудитория</i>	1. Трибуна напольная (Инв.№ 599206) – 1 шт.; 2. Жалюзи (Инв.№591110) – 1 шт.; 3. Доска маркетинговая (Инв.№ 35643/4) – 1 шт.; 4. Стол – 15 шт.; 5. Скамейка – 14 шт.; 6. Стол эрго – 1 шт.; 7. Стул – 16 шт.

Наименование специальных помещений и помещений для самостоятельной работы (№ учебного корпуса, № аудитории)	Оснащенность специальных помещений и помещений для самостоятельной работы
1	2
для текущего контроля и промежуточной аттестации, помещение для самостоятельной работы (2й учебный корпус, 303 ауд.)	
учебная аудитория для проведения занятий лекционного типа, учебная аудитория для проведения занятий семинарского типа, учебная аудитория для проведения курсового проектирования (выполнения курсовых работ), учебная аудитория для групповых и индивидуальных консультаций, учебная аудитория для текущего контроля и промежуточной аттестации (1й учебный корпус, 212 ауд.)	Количество рабочих мест: 24 Встроенные сетевые адаптеры (Intel I219-V или Realtek RTL8111H), интерфейс RJ-45, скорость 10/100/1000 Мбит/с. Точки доступа: Ubiquiti UniFi AP AC Pro, стандарты IEEE 802.11a/b/g/n/ac, частоты 2.4 ГГц (450 Мбит/с) и 5 ГГц (1300 Мбит/с), поддержка MU-MIMO, питание PoE. Структурное подразделение: Кафедра Цифровая кафедра
учебная аудитория для проведения занятий лекционного типа, учебная аудитория для проведения занятий семинарского типа, учебная аудитория для проведения курсового проектирования (выполнения курсовых работ), учебная аудитория для групповых и индивидуальных консультаций, учебная аудитория для текущего контроля и промежуточной аттестации (1й учебный корпус, 214 ауд.)	Количество рабочих мест: 24 Встроенные сетевые адаптеры (Intel I219-V или Realtek RTL8111H), интерфейс RJ-45, скорость 10/100/1000 Мбит/с. Точки доступа: Ubiquiti UniFi AP AC Pro, стандарты IEEE 802.11a/b/g/n/ac, частоты 2.4 ГГц (450 Мбит/с) и 5 ГГц (1300 Мбит/с), поддержка MU-MIMO, питание PoE. Структурное подразделение: Кафедра Цифровая кафедра
Центр «Проектный институт цифровой трансформации АПК»	Рабочие станции глубокого обучения – 3 шт. Графическая станция 1008100Intel i9-10940X 3.3 ГГц, X299, 128Гб 2666 МГц, SSD 1Тб, HDD 4Тб, RTX 8000 48Гб (NVIDIA QUADRO), 750Вт, Midi-Tower (Серия WS-ADVANCED) Ноутбук – 4 шт. Процессор 10-ядерный Intel Core i9 13980HX 3.70GHz, Оперативная память64GB (2x32Gb) DDR4 4800Mhz - x1, Видеокарта NVIDIA GeForce RTX 4090 32GB - Single GPU вычислительный кластер: на базе NVIDIA H100. (7168 ГБ ОЗУ, 110 производительных ядер и 220 высокоэффективных потоков уровня XEON GOLD и PLATINUM, 400 ГБ VRAM и 84480 cuda ядер для задач машинного обучения, 72000 ГБ высокоскоростное хранилище enterprise - уровня, 10 Гбит сеть с функцией резервирования и система питания с горячей заменой). и другое оборудование Центра
Центральная научная библиотека имени Н.И. Железнова	Читальные залы библиотеки

Наименование специальных помещений и помещений для самостоятельной работы (№ учебного корпуса, № аудитории)	Оснащенность специальных помещений и помещений для самостоятельной работы
1	2
Студенческое общежитие	Комната для самоподготовки

11. Методические рекомендации студентам по освоению дисциплины

Приступая к изучению дисциплины «Большие языковые модели», студенты должны ознакомиться с учебной программой, учебной, научной и методической литературой, имеющейся в библиотеке РГАУ-МСХА им. К.А. Тимирязева, получить в библиотеке рекомендованные учебники и учебно-методические пособия, завести новую тетрадь для работы с первоисточниками.

В ходе занятий вести конспектирование учебного материала. Обращать внимание на категории, формулировки, раскрывающие содержание тех или иных явлений и процессов, научные выводы и практические рекомендации. Задавать преподавателю уточняющие вопросы с целью уяснения теоретических положений, разрешения спорных ситуаций.

В ходе подготовки к практическим занятиям изучить основную литературу, ознакомиться с дополнительной литературой в соответствии с поставленной задачей. При этом учесть рекомендации преподавателя и требования учебной программы. Необходимо дорабатывать свой конспект, делая в нем соответствующие записи из литературы, рекомендованной преподавателем и предусмотренной учебной программой.

При подготовке к экзамену (в конце семестра) повторять пройденный материал в строгом соответствии с учебной программой. Использовать конспекты и литературу, рекомендованную преподавателем. Обратит особое внимание на темы учебных занятий, пропущенных студентом по разным причинам. При необходимости обратиться за консультацией и методической помощью к преподавателю.

Виды и формы отработки пропущенных занятий

Студент, пропустивший занятия обязан самостоятельно подготовиться к теме устного опроса, которые состоялись на практическом занятии. В рамках часов консультаций студент может ответить на вопросы пропущенного устного опроса, которые были пропущены.

12. Методические рекомендации преподавателям по организации обучения по дисциплине

Курс «Большие языковые модели» должен давать не абстрактно-формальные, а прикладные знания. Данная цель может быть реализована только при условии соблюдения в учебных планах преемственности учебных дисциплин. Базовые знания для изучения «Большие языковые модели» дают такие дисциплины, как «Алгоритмизация и программирование», «Программирование на языке Python», «Введение в компьютерные науки на иностранном языке», «Основы теории управления и цифровой обработки сигналов», «Математиче-

ская статистика», «Инжиниринг данных», «Статистика для машинного обучения», «Методы искусственного интеллекта», «Методы машинного обучения». Освоение основных тем данной дисциплины позволит студентам сформировать представление о таком сложном предмете как «Большие языковые модели», понять всю ширину науки и получить необходимые знания для последующего профессионального развития в этой области.

Студент может подготовить доклад по теме, представляющей его научный интерес, представить результаты в виде презентации. В случае надлежащего качества, его работа может быть заслушана на научном кружке кафедры или на студенческой научной конференции. По решению кафедры, студенты, занявшие призовые места на научных студенческих конференциях, могут освобождаться от сдачи зачета с оценкой по этой дисциплине.

Преподаватель должен указывать, в какой последовательности следует изучать материал дисциплины, обращать внимание на особенности изучения отдельных тем и разделов, помогать отбирать наиболее важные и необходимые сведения из учебных пособий, а также давать объяснения вопросам программы курса, которые обычно вызывают затруднения. При этом преподавателю необходимо учитывать следующие моменты:

1. Не следует перегружать студентов творческими заданиями.
2. Чередовать творческую работу на занятиях с заданиями во внеаудиторное время.
3. Давать студентам четкий инструктаж по выполнению самостоятельных заданий: цель задания; условия выполнения; объем; сроки; требования к оформлению.
4. Осуществлять текущий учет и контроль за самостоятельной работой.
5. Давать оценку и обобщать уровень усвоения навыков самостоятельной, творческой работы.

Программу разработали:

Демичев В.В., кандидат экономических наук, доцент,



(подпись)

Титов А.Д., ассистент



(подпись)

РЕЦЕНЗИЯ

на рабочую программу дисциплины Б1.В.22 «Большие языковые модели» ОПОП ВО по направлению 09.03.02 Информационные системы и технологии, направленностям «Большие данные и машинное обучение», «Компьютерные науки и технологии искусственного интеллекта» (квалификация выпускника – бакалавр)

Кийко Павлом Владимировичем, доцентом кафедры высшей математики, кандидатом педагогических наук (далее по тексту рецензент), проведено рецензирование рабочей программы дисциплины «Большие языковые модели» ОПОП ВО по направлению 09.03.02 Информационные системы и технологии, направленностям «Большие данные и машинное обучение», «Компьютерные науки и технологии искусственного интеллекта» (бакалавриат) разработанной в ФГБОУ ВО «Российский государственный аграрный университет – МСХА имени К.А. Тимирязева», на кафедре статистики и кибернетики (разработчики – Демичев Вадим Владимирович, доцент, кандидат экономических наук, доцент кафедры статистики и кибернетики, Титов Артем Денисович, ассистент кафедры статистики и кибернетики).

Рассмотрев представленные на рецензирование материалы, рецензент пришел к следующим выводам:

1. Предъявленная рабочая программа дисциплины «Большие языковые модели» (далее по тексту Программа) соответствует требованиям ФГОС ВО и компетентностно-ролевых моделей в сфере искусственного интеллекта по направлению 09.03.02 Информационные системы и технологии. Программа содержит все основные разделы, соответствует требованиям к нормативно-методическим документам.

2. Представленная в Программе **актуальность** учебной дисциплины в рамках реализации ОПОП ВО не подлежит сомнению – дисциплина относится ко части дисциплин, формируемых участниками образовательных отношений – Б1.В.

3. Представленные в Программе **цели** дисциплины соответствуют требованиям ФГОС ВО направления 09.03.02 Информационные системы и технологии.

4. В соответствии с Программой за дисциплиной «Большие языковые модели» закреплены **8 профессиональных компетенций, соответствующих компетентностно-ролевой модели по искусственному интеллекту**. Дисциплина «Большие языковые модели» и представленная Программа способна реализовать их в объявленных требованиях.

5. **Результаты обучения**, представленные в Программе в категориях знать, уметь, владеть и в уровне освоений компетенций соответствуют специфике и содержанию дисциплины и демонстрируют возможность получения заявленных результатов.

6. Общая трудоёмкость дисциплины «Большие языковые модели» составляет 4,0 зачётные единицы (144 часа).

7. Информация о взаимосвязи изучаемых дисциплин и вопросам исключения дублирования в содержании дисциплин соответствует действительно-

сти. Дисциплина «Большие языковые модели» взаимосвязана с другими дисциплинами ОПОП ВО и Учебного плана по направлению *09.03.02 Информационные системы и технологии* и возможность дублирования в содержании отсутствует.

8. Представленная Программа предполагает использование современных образовательных технологий, используемые при реализации различных видов учебной работы. Формы образовательных технологий соответствуют специфике дисциплины.

9. Программа дисциплины «Большие языковые модели» предполагает проведение занятий в интерактивной форме.

10. Виды, содержание и трудоёмкость самостоятельной работы студентов, представленные в Программе, соответствуют требованиям к подготовке выпускников, содержащимся во ФГОС ВО направления 09.03.02 Информационные системы и технологии.

11. Представленные и описанные в Программе формы *текущей* оценки знаний (устный опрос, коллоквиум), соответствуют специфике дисциплины и требованиям к выпускникам.

Форма промежуточного контроля знаний студентов, предусмотренная Программой, осуществляется в форме зачета с оценкой в седьмом семестре, что соответствует статусу дисциплины, формируемой участниками образовательных отношений – Б1.В. ФГОС ВО направления 09.03.02. Информационные системы и технологии.

12. Формы оценки знаний, представленные в Программе, соответствуют специфике дисциплины и требованиям к выпускникам.

13. Учебно-методическое обеспечение дисциплины представлено: основной литературой – 3 источника (базовые учебники), дополнительной литературой – 3 наименования, статьи, опубликованные в научных журналах 1 уровня Белого списка научных журналов Минобрнауки России и сборниках научных работ конференций уровня А* – 3 источника, Интернет-ресурсы – 4 источника и соответствует требованиям ФГОС ВО направления *09.03.02 Информационные системы и технологии*.

14. Материально-техническое обеспечение дисциплины соответствует специфике дисциплины «Большие языковые модели» и обеспечивает использование современных образовательных, в том числе интерактивных методов обучения.

15. Методические рекомендации студентам и методические рекомендации преподавателям по организации обучения по дисциплине дают представление о специфике обучения по дисциплине «Большие языковые модели».

ОБЩИЕ ВЫВОДЫ

На основании проведенного рецензирования можно сделать заключение, что характер, структура и содержание рабочей программы дисциплины «Большие языковые модели» ОПОП ВО по направлению 09.03.02 Информационные системы и технологии, направленностям «Большие данные и машинное обучение», «Компьютерные науки и технологии искусственного интеллекта» (квалификация выпускника – бакалавр), разработанная Демичевым Вадимом Влади-

мировичем, доцентом, кандидатом экономических наук, доцентом кафедры статистики и кибернетики и Титовым Артемом Денисовичем, ассистентом кафедры статистики и кибернетики соответствует требованиям ФГОС ВО, компетентностно-ролевых моделей в сфере искусственного интеллекта, современным требованиям экономики, рынка труда и позволит при её реализации успешно обеспечить формирование заявленных компетенций.

Рецензент:

Кийко Павел Владимирович, доцент кафедры высшей математики, кандидат педагогических наук


(подпись)

«26» августа 2025 г.