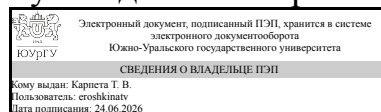


ЮЖНО-УРАЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

УТВЕРЖДАЮ:
Руководитель направления



Т. В. Карпета

РАБОЧАЯ ПРОГРАММА

дисциплины 1.О.11 Большие языковые модели (LLM)
для направления 02.04.02 Фундаментальная информатика и информационные технологии

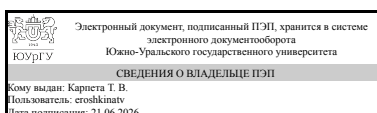
уровень Магистратура

форма обучения очная

кафедра-разработчик Прикладная математика и программирование

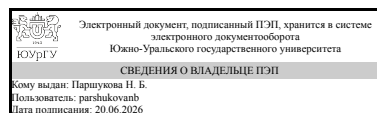
Рабочая программа составлена в соответствии с ФГОС ВО по направлению подготовки 02.04.02 Фундаментальная информатика и информационные технологии, утверждённым приказом Минобрнауки от 23.08.2017 № 811

Зав.кафедрой разработчика,
к.физ.-мат.н.



Т. В. Карпета

Разработчик программы,
к.пед.н., доцент



Н. Б. Паршукова

1. Цели и задачи дисциплины

Целью дисциплины «Большие языковые модели (LLM)» является формирование у студентов теоретических знаний и практических навыков в области разработки, обучения, дообучения и применения больших языковых моделей, включая понимание архитектурных основ современных LLM, методов эффективного обучения и развёртывания моделей в производственной среде. Задачи дисциплины:

- изучить математические и алгоритмические основы нейросетевых архитектур для обработки естественного языка;
- освоить архитектуру трансформеров и механизм внимания как фундамент современных LLM;
- изучить методы предобучения крупных языковых моделей, законы масштабирования и явления возникновения новых способностей;
- овладеть техниками промптинга, дообучения (fine-tuning), RLHF и PEFT для адаптации LLM к прикладным задачам;
- научиться применять методы квантизации, дистилляции и эффективного инференса LLM;
- освоить построение RAG-систем и LLM-агентов для практических приложений;
- получить практические навыки работы с современными фреймворками в области больших языковых моделей.

Краткое содержание дисциплины

Введение в NLP. История и современное состояние. Задачи NLP. Векторные представления слов: Count-based методы, Word2Vec, GloVe. Оценка качества эмбедингов. Методы классификации текстов: наивный Байес, логистическая регрессия, метод опорных векторов, сверточные нейросети, рекуррентные нейросети. N-граммные и нейросетевые языковые модели. Связывание весов для улучшения обучения модели. Перплексия как метрика качества языковой модели. Обучение и применение нейросетевых моделей для генерации текста. Архитектура Seq2seq, механизм внимания (Attention), алгоритм лучевого поиска (Beam Search). BLEU-метрика качества генерации текста. Архитектура трансформеров: внимание, самовнимание, заполнение пробелов во внимании. Особенности модели BERT: позиционное кодирование, кодирование пар байтов (BPE). Обучение с переносом в NLP: предобучение и дообучение модели. Контекстные векторы (CoVe) и эмбединги (ELMo) для языковых моделей. Архитектура модели BERT: масочное языковое моделирование (MLM), предсказание следующего предложения (NSP), дообучение (Fine-tuning). Улучшения модели BERT: RoBERTa, ALBERT, DistilBERT. Законы масштабирования (Scaling Laws) для качественного улучшения языковых моделей. Применение Scaling Laws для создания моделей GPT-3, LLaMA, Qwen3. Эмерджентные способности языковых моделей. Зависимость размера модели от новых способностей к рассуждению, следованию инструкциям, базовому программированию и пр. Анализ внутренних представлений больших языковых моделей (LLM). Основы промпт инжиниринга: сделай с первого раза (zero-shot), обучись на примерах (few-shot), покажи ход рассуждений (chain-of-thought). Управление креативностью и детерминированностью модели через температуру, контекст. Как заставить модель думать и проверять за себя – продвинутые техники промптинга. Рассуждение и действие (ReAct). Дерево мыслей (Tree-of-Thought) для лучшего выбора из альтернатив. Ограничения техник и интерпретация ответа модели. Обучение модели на предпочтениях людей (RLHF). Алгоритм обучения с подкреплением ближайшей оптимизации политики (PPO). Продукты моделей предпочтений: от GPT-3 к InstructGPT и ChatGPT. Прямая оптимизация

предпочтений (DPO) как технически более сложная альтернатива RLHF. Семейство методов дообучения модели с изменением малого количества параметров: Adapters, LoRA, Prefix Tuning. Сравнение методов и практические рекомендации по применению. Уменьшение размеров LLM с использованием квантования весов модели (PTQ). Алгоритмы 4-х битного квантования весов: GPTQ, AWQ. Методы уменьшения размера сети (дистилляция), удаление избыточных компонентов (прунинг), прореживание сети (разреживание). RAG (Retrieval-Augmented Generation): архитектура, векторные базы данных. LLM-агенты. Обзор фреймворков

2. Компетенции обучающегося, формируемые в результате освоения дисциплины

Планируемые результаты освоения ОП ВО (компетенции)	Планируемые результаты обучения по дисциплине
ПК-2 Способен исследовать и разрабатывать архитектуры интеллектуальных информационных систем для различных предметных областей на основе комплексов методов и инструментальных средств систем искусственного интеллекта	Знает: принципы архитектуры и работы больших языковых моделей, основные типы LLM, методы их обучения и дообучения, подходы к оценке качества генерации текста, ограничения, риски и области ответственного применения LLM в различных предметных областях Умеет: формулировать прикладные задачи, пригодные для решения с использованием LLM, подбирать и адаптировать конкретные модели под требования проекта, разрабатывать эффективные промпты, комбинировать LLM с внешними инструментами и источниками знаний, анализировать результаты генерации и вносить корректировки в стратегию использования модели Имеет практический опыт: применения LLM-платформ и библиотек для разработки прототипов интеллектуальных приложений обработки естественного языка, настройки параметров моделей и промптов под конкретные сценарии, интеграции LLM в программные системы, тестирования и документирования полученных решений

3. Место дисциплины в структуре ОП ВО

Перечень предшествующих дисциплин, видов работ учебного плана	Перечень последующих дисциплин, видов работ
Нет	Не предусмотрены

Требования к «входным» знаниям, умениям, навыкам студента, необходимым при освоении данной дисциплины и приобретенным в результате освоения предшествующих дисциплин:

Нет

4. Объём и виды учебной работы

Общая трудоемкость дисциплины составляет 5 з.е., 180 ч., 76,5 ч. контактной работы

Вид учебной работы	Всего часов	Распределение по семестрам
		в часах
		Номер семестра
		3
Общая трудоёмкость дисциплины	180	180
<i>Аудиторные занятия:</i>	64	64
Лекции (Л)	32	32
Практические занятия, семинары и (или) другие виды аудиторных занятий (ПЗ)	0	0
Лабораторные работы (ЛР)	32	32
<i>Самостоятельная работа (СРС)</i>	103,5	103,5
Выполнение индивидуальных заданий	30	30
Выполнение лабораторных работ	43,5	43,5
Подготовка к экзамену	30	30
Консультации и промежуточная аттестация	12,5	12,5
Вид контроля (зачет, диф.зачет, экзамен)	-	экзамен

5. Содержание дисциплины

№ раздела	Наименование разделов дисциплины	Объем аудиторных занятий по видам в часах			
		Всего	Л	ПЗ	ЛР
1	Введение в NLP. Векторные представления слов	8	4	0	4
2	Классификация текстов и языковые модели	8	4	0	4
3	Seq2seq, механизм внимания. Архитектура Transformer	8	4	0	4
4	Transfer Learning. Предобученные модели BERT и GPT	8	4	0	4
5	Большие языковые модели: масштабирование, эмерджентные свойства	8	4	0	4
6	Промптинг и продвинутое техники промптинга	8	4	0	4
7	Дообучение больших языковых моделей	8	4	0	4
8	Эффективный инференс, RAG и практическое применение больших языковых моделей	8	4	0	4

5.1. Лекции

№ лекции	№ раздела	Наименование или краткое содержание лекционного занятия	Кол-во часов
1	1	Введение в NLP. История и современное состояние. Задачи NLP	2
2	1	Векторные представления слов: Count-based методы, Word2Vec, GloVe. Оценка качества эмбедингов	2
3	2	Методы классификации текстов: наивный Байес, логистическая регрессия, метод опорных векторов, сверточные нейросети, рекуррентные нейросети.	2
4	2	N-граммные и нейросетевые языковые модели. Связывание весов для улучшения обучения модели. Перплексия как метрика качества языковой модели.	2

5	3	Обучение и применение нейросетевых моделей для генерации текста. Архитектура Seq2seq, механизм внимания (Attention), алгоритм лучевого поиска (Beam Search). BLEU-метрика качества генерации текста.	2
6	3	Архитектура трансформеров: внимание, самовнимание, заполнение пробелов во внимании. Особенности модели BERT: позиционное кодирование, кодирование пар байтов (BPE).	2
7	4	Обучение с переносом в NLP: предобучение и дообучение модели. Контекстные векторы (CoVe) и эмбединги (ELMo) для языковых моделей.	2
8	4	Архитектура модели BERT: масочное языковое моделирование (MLM), предсказание следующего предложения (NSP), дообучение (Fine-tuning). Улучшения модели BERT: RoBERTa, ALBERT, DistilBERT.	2
9	5	Законы масштабирования (Scaling Laws) для качественного улучшения языковых моделей. Применение Scaling Laws для создания моделей GPT-3, LLaMA, Qwen3	2
10	5	Эмерджентные способности языковых моделей. Зависимость размера модели от новых способностей к рассуждению, следованию инструкциям, базовому программированию и пр. Анализ внутренних представлений больших языковых моделей (LLM).	2
11	6	Основы промпт инжиниринга: сделай с первого раза (zero-shot), обучись на примерах (few-shot), покажи ход рассуждений (chain-of-thought). Управление креативностью и детерминированностью модели через температуру, контекст.	2
12	6	Как заставить модель думать и проверять за себя – продвинутые техники промптинга. Рассуждение и действие (ReAct). Дерево мыслей (Tree-of-Thought) для лучшего выбора из альтернатив. Ограничения техник и интерпретация ответа модели.	2
13	7	Обучение модели на предпочтениях людей (RLHF). Алгоритм обучения с подкреплением ближайшей оптимизации политики (PPO). Продукты моделей предпочтений: от GPT-3 к InstructGPT и ChatGPT. Прямая оптимизация предпочтений (DPO) как технически более сложная альтернатива RLHF.	2
14	7	Семейство методов дообучения модели с изменением малого количества параметров: Adapters, LoRA, Prefix Tuning. Сравнение методов и практические рекомендации по применению.	2
15	8	Уменьшение размеров LLM с использованием квантования весов модели (PTQ). Алгоритмы 4-х битного квантования весов: GPTQ, AWQ. Методы уменьшения размера сети (дистилляция), удаление избыточных компонентов (прунинг), прореживание сети (разреживание).	2
16	8	RAG (Retrieval-Augmented Generation): архитектура, векторные базы данных. LLM-агенты. Обзор фреймворков	2

5.2. Практические занятия, семинары

Не предусмотрены

5.3. Лабораторные работы

№ занятия	№ раздела	Наименование или краткое содержание лабораторной работы	Кол-во часов
1	1	Работа с векторными представлениями слов: Word2Vec, FastText, визуализация семантического пространства	2
2	1	Построение системы семантического поиска на основе эмбедингов. Оценка качества эмбедингов	2
3	2	Классификация текстов с использованием CNN и RNN. Сравнение	2

		классических и нейросетевых методов	
4	2	N-граммная языковая модель. Нейросетевые языковые модели. Сглаживание и перплексия	2
5	3	Реализация базовой seq2seq модели для машинного перевода. Применение механизма внимания. BLEU-метрика	2
6	3	Использование предобученных трансформеров для решения реальных NLP-задач – от простого применения готовых моделей до тонкой настройки (fine-tuning) под свою задачу.	2
7	4	Дообучение предобученной BERT-модели для классификации текстов и NER	2
8	4	Сравнение семейств предобученных моделей (BERT, RoBERTa, GPT-2) на прикладных задачах.	2
9	5	Исследование масштабирующих законов: эксперименты с моделями разных размеров	2
10	5	Анализ emergent abilities: тестирование LLM (Qwen3, LLaMA) на задачах рассуждения	2
11	6	Промптинг базовых моделей. Шаблон чата. Zero-shot, Few-shot, Chain-of-Thought. Оценка результатов	2
12	6	Автоматическая оптимизация промптов. ReAct-агенты на базе LLM. DSPy-фреймворк	2
13	7	Ресурсно-эффективное дообучение LLM с использованием библиотек PEFT и trl от HuggingFace.	2
14	7	Обучение модели на предпочтениях людей (RLHF). Оценка качества выровненных моделей.	2
15	8	Реализация методов пост-тренировочной квантизации (PTQ) для больших языковых моделей.	2
16	8	Практическая реализация RAG-системы «с нуля» – от обучения эмбеддингов до генерации ответов LLM	2

5.4. Самостоятельная работа студента

Выполнение СРС			
Подвид СРС	Список литературы (с указанием разделов, глав, страниц) / ссылка на ресурс	Семестр	Кол-во часов
Выполнение индивидуальных заданий	https://github.com/yandexdataschool/nlp_course Выполнение домашних работ	3	30
Выполнение лабораторных работ	https://github.com/yandexdataschool/nlp_course Выполнение заданий лабораторных работ	3	43,5
Подготовка к экзамену	https://github.com/yandexdataschool/nlp_course Гольдберг, Й. Нейросетевые методы в обработке естественного языка : руководство / Й. Гольдберг ; перевод с английского А. А. Слинкина. — Москва : ДМК Пресс, 2019. — 282 с. — ISBN 978-5-97060-754-1. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/131704 (дата обращения: 15.06.2026). — Режим доступа: для авториз. пользователей. Гудфеллоу, Я. Глубокое обучение / Я. Гудфеллоу, И. Бенджио, А. Курвилль ; перевод с английского А. А. Слинкина. — 2-е изд. — Москва : ДМК Пресс, 2018. — 652 с. —	3	30

	ISBN 978-5-97060-618-6. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/107901 (дата обращения: 15.06.2026). — Режим доступа: для авториз. пользователей.		
--	--	--	--

6. Фонд оценочных средств для проведения текущего контроля успеваемости, промежуточной аттестации

Контроль качества освоения образовательной программы осуществляется в соответствии с Положением о балльно-рейтинговой системе оценивания результатов учебной деятельности обучающихся.

6.1. Контрольные мероприятия (КМ)

№ КМ	Семестр	Вид контроля	Название контрольного мероприятия	Вес	Макс. балл	Порядок начисления баллов	Учитывается в ПА
1	3	Текущий контроль	Контрольная точка 1. Работа с векторными представлениями слов. Семантический поиск на основе эмбедингов.	1	5	5 - выполнены все задания; 4 - выполнены 4 из 5 заданий; 3 - выполнены 3 из 5 заданий; 2 - выполнены 2 из 5 заданий; 1 - выполнено 1 из 5 заданий; 0 - задание не выполнено.	экзамен
2	3	Текущий контроль	Контрольная точка 2. N-граммная языковая модель. Нейросетевые языковые модели.	1	5	5 - выполнены все задания; 4 - выполнены 4 из 5 заданий; 3 - выполнены 3 из 5 заданий; 2 - выполнены 2 из 5 заданий; 1 - выполнено 1 из 5 заданий; 0 - задание не выполнено.	экзамен
3	3	Текущий контроль	Контрольная точка 3. Использование предобученных трансформеров для решения реальных NLP-задач. Дообучение предобученной BERT-модели для классификации текстов и NER	1	5	Всего 5 заданий. За каждое задание 2 балла. 0 - задание не выполнено; 1 - задание выполнено с существенными недостатками; 2 - задание выполнено.	экзамен
4	3	Текущий контроль	Контрольная точка 4. Промптинг базовых моделей	1	5	5 - выполнены все задания; 4 - выполнены 4 из 5 заданий; 3 - выполнены 3 из 5 заданий; 2 - выполнены 2 из 5 заданий; 1 - выполнено 1 из 5 заданий; 0 - задание не выполнено.	экзамен
5	3	Текущий контроль	Контрольная точка 5. Квантизация больших языковых моделей	1	5	5 - выполнены все задания; 4 - выполнены 4 из 5 заданий; 3 - выполнены 3 из 5 заданий; 2 - выполнены 2 из 5 заданий;	экзамен

						1 - выполнено 1 из 5 заданий; 0 - задание не выполнено.	
6	3	Текущий контроль	Контрольная точка 6. Практическая реализация RAG-системы	1	5	5 - выполнены все задания; 4 - выполнены 4 из 5 заданий; 3 - выполнены 3 из 5 заданий; 2 - выполнены 2 из 5 заданий; 1 - выполнено 1 из 5 заданий; 0 - задание не выполнено.	экзамен
7	3	Промежуточная аттестация	Экзамен	-	5	5 - получены исчерпывающие ответы на все вопросы билета; 4 - получены ответы на оба вопроса, но допущены небольшие ошибки; 3 - получены ответа на оба вопроса, но допущены существенные ошибки; 2 - получен полноценный ответ только на один вопрос билета; 1 - получен полноценный ответ только на один вопрос билета, при этом допущены неточности; 0 - не получено ответа ни на один вопрос.	экзамен

6.2. Процедура проведения, критерии оценивания

Вид промежуточной аттестации	Процедура проведения	Критерии оценивания
экзамен	Промежуточная аттестация курса представляет собой экзамен в устной форме. Прохождение промежуточной аттестации не обязательно (но студент может изъявить желание ее прохождения). Прохождение устного экзамена оценивается от 0 до 5 баллов. Билет состоит из 2-х теоретических вопросов. Оценивание экзамена проводится по следующей шкале: 5 - получены исчерпывающие ответы на все вопросы билета; 4 - получены ответы на оба вопроса, но допущены небольшие ошибки; 3 - получены ответа на оба вопроса, но допущены существенные ошибки; 2 - получен полноценный ответ только на один вопрос билета; 1 - получен полноценный ответ только на один вопрос билета, при этом допущены неточности; 0 - не получено ответа ни на один вопрос. В случае, если студент желает пройти промежуточную аттестацию, его рейтинг по курсу вычисляется как $R = 0.6 \cdot R_t + 0.4 \cdot R_e$, где R_t – текущий рейтинг по курсу, R_e – экзаменационный рейтинг.	В соответствии с пп. 2.5, 2.6 Положения

6.3. Паспорт фонда оценочных средств

Компетенции	Результаты обучения	№ KM						
		1	2	3	4	5	6	7
ПК-2	Знает: принципы архитектуры и работы больших языковых моделей, основные типы LLM, методы их обучения и дообучения, подходы к оценке качества генерации текста, ограничения, риски и области ответственного применения LLM в различных предметных областях	+	+	+			+	+
ПК-2	Умеет: формулировать прикладные задачи, пригодные для решения с	+	+	+	+	+	+	+

	использованием LLM, подбирать и адаптировать конкретные модели под требования проекта, разрабатывать эффективные промпты, комбинировать LLM с внешними инструментами и источниками знаний, анализировать результаты генерации и вносить корректировки в стратегию использования модели								
ПК-2	Имеет практический опыт: применения LLM-платформ и библиотек для разработки прототипов интеллектуальных приложений обработки естественного языка, настройки параметров моделей и промптов под конкретные сценарии, интеграции LLM в программные системы, тестирования и документирования полученных решений								

Типовые контрольные задания по каждому мероприятию находятся в приложениях.

7. Учебно-методическое и информационное обеспечение дисциплины

Печатная учебно-методическая документация

а) *основная литература:*

Не предусмотрена

б) *дополнительная литература:*

Не предусмотрена

в) *отечественные и зарубежные журналы по дисциплине, имеющиеся в библиотеке:*

Не предусмотрены

г) *методические указания для студентов по освоению дисциплины:*

1. Языковое моделирование

из них: *учебно-методическое обеспечение самостоятельной работы студента:*

1. Языковое моделирование

Электронная учебно-методическая документация

№	Вид литературы	Наименование ресурса в электронной форме	Библиографическое описание
1	Основная литература	ЭБС издательства Лань	Гудфеллоу, Я. Глубокое обучение / Я. Гудфеллоу, И. Бенджио, А. Курвилль ; перевод с английского А. А. Слинкина. — 2-е изд. — Москва : ДМК Пресс, 2018. — 652 с. — ISBN 978-5-97060-618-6. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/107901 (дата обращения: 15.06.2026). — Режим доступа: для авториз. пользователей.
2	Основная литература	ЭБС издательства Лань	Гольдберг, Й. Нейросетевые методы в обработке естественного языка : руководство / Й. Гольдберг ; перевод с английского А. А. Слинкина. — Москва : ДМК Пресс, 2019. — 282 с. — ISBN 978-5-97060-754-1. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/131704 (дата обращения: 15.06.2026). — Режим доступа: для авториз. пользователей.

Перечень используемого программного обеспечения:

Нет

Перечень используемых профессиональных баз данных и информационных справочных систем:

Нет

8. Материально-техническое обеспечение дисциплины

Не предусмотрено