Приложение к ОПОП
«Информационно-управляющие системы»

СПбГЭТУ «ЛЭТИ»
ПЕРВЫЙ ЭЛЕКТРОТЕХНИЧЕСКИЙ

МИНОБРНАУКИ РОССИИ

федеральное государственное автономное образовательное учреждение высшего образования
**«Санкт-Петербургский государственный электротехнический университет
«ЛЭТИ» им. В.И.Ульянова (Ленина)»
(СПбГЭТУ «ЛЭТИ»)**

РАБОЧАЯ ПРОГРАММА

дисциплины

«ГЛУБОКОЕ ОБУЧЕНИЕ В NLP»

для подготовки бакалавров

по направлению

09.03.02 «Информационные системы и технологии»

по профилю

«Информационно-управляющие системы»

Санкт-Петербург

2025

ЛИСТ СОГЛАСОВАНИЯ

Разработчики:
ведущий инженер Кистуга А.В.

Рабочая программа рассмотрена и одобрена на заседании кафедры ИИ 28.01.2025, протокол № 1

Рабочая программа рассмотрена и одобрена учебно-методической комиссией ФКТИ, 28.01.2025, протокол № 1

Согласовано в ИС ИОТ
Начальник ОМОЛА Загороднюк О.В.

1 СТРУКТУРА ДИСЦИПЛИНЫ

Обеспечивающий факультет	ФКТИ
Обеспечивающая кафедра	ИИ
Общая трудоемкость (ЗЕТ)	3
Курс	4
Семестр	7

Виды занятий

Электронные лекции (акад. часов)	34
Электронные практические (академ. часов) (академ. часов)	34
Иная контактная работа (академ. часов)	1
Все контактные часы (академ. часов)	1
Самостоятельная работа, включая часы на контроль (академ. часов)	39
Всего (академ. часов)	108

Вид промежуточной аттестации

Дифф. зачет (курс)	4

2 АННОТАЦИЯ ДИСЦИПЛИНЫ

«ГЛУБОКОЕ ОБУЧЕНИЕ В NLP»

Курс посвящен современным методам обработки естественного языка (NLP) на основе глубокого обучения. В первой теме рассматривается архитектура Transform механизмы Self-Attention, Multi-Head Attention и позиционное кодирование, а также переход от рекуррентных сетей к моделям на основе внимания. Вторая тема охватывает дискриминативные модели (ELMo, BERT, E5), их предобучение, дообучение и применение в задачах классификации, извлечения информации и семантического поиска. Третья часть курса фокусируется на генеративных моделях (GPT, T5), языковом моделировании, настройке генерации текста и эволюции диалоговых систем, включая чат-боты. Четвертая тема посвящена ранжирующим моделям: архитектурам Cross/Bi Encoder, Poly Encoder, ColBERT, методам обучения с негативным семплированием и оценке качества ранжирования для поиска и рекомендаций. Курс сочетает теоретические основы с практическими кейсами, позволяя освоить ключевые технологии NLP для решения реальных задач.

SUBJECT SUMMARY

«DEEP LEARNING FOR NATURAL LANGUAGE PROCESSING (NLP)»

The course focuses on state-of-the-art techniques for natural language processing (NLP) based on deep learning. The first topic covers the Transformer architecture: Self-Attention, Multi-Head Attention and positional encoding mechanisms, and the transition from recurrent networks to attention-based models. The second topic covers discriminative models (ELMo, BERT, E5), their pre-training, pre-training and applications in classification, information extraction and semantic search tasks. The third part of the course focuses on generative models (GPT, T5), language modeling, text generation settings, and the evolution of dialog systems including chatbots. The

fourth topic focuses on ranking models: the Cross/Bi Encoder, Poly Encoder, ColBERT architectures, negative-sampling learning methods, and ranking quality evaluation for search and recommendation. The course combines theoretical foundations with practical cases, allowing you to master key NLP techniques to solve real-world problems.

3 ОБЩИЕ ПОЛОЖЕНИЯ

3.1 Цели и задачи дисциплины

1. Цели изучения дисциплины:

-формирование системного представления о современных архитектурах глубокого обучения для NLP, включая механизмы внимания, дискриминативные, генеративные и ранжирующие модели;

-обеспечение понимания принципов работы и областей применения ключевых моделей (Transformer, BERT, GPT, T5, ColBERT) в задачах обработки естественного языка;

-развитие умений и навыков оптимального решения практических задач: классификации текста, генерации, семантического поиска, диалоговых систем и ранжирования.

2. Задачами изучения дисциплины является:

-освоение архитектуры Transformer, механизмов Self-Attention, Multi-Head Attention и изучение их роли в замене рекуррентных сетей;

-изучение методов предобучения и дообучения дискриминативных моделей (BERT, E5), их оценки и применения в задачах извлечения информации и семантического анализа;

-изучение принципов генеративного моделирования, настройки параметров генерации текста и создания диалоговых систем;

-исследование архитектуры ранжирующих моделей, методов негативного семплирования и метрики оценки качества ранжирования;

-формирование навыков работы с современными библиотеками (Hugging Face, PyTorch) для реализации, дообучения и адаптации NLP-моделей под конкретные задачи;

-научиться анализировать ограничения моделей, интерпретировать результаты

и оптимизировать их производительность в реальных проектах.

3. По окончании дисциплины студент должен получить знания:

-основных компонентов архитектуры Transformer;

-различных типов внимания (self-attention, multi-head attention) и их применения;

-задач, решаемых дискриминативными моделями и их практического применения;

-основных архитектур дискриминативных моделей;

-основных архитектур генеративных моделей;

-основных архитектур ранжирующих моделей.

4. По окончании дисциплины студент должен приобрести умения:

-использовать библиотеки глубокого обучения для NLP;

-анализировать и обрабатывать текстовые данные.;

-разрабатывать и обучать модели NLP для различных задач: классификация, регрессия, перевод, генерация текста;

-использовать предобученные модели NLP для решения конкретных задач.

5. По окончании дисциплины студент должен иметь навыки:

-решения различных задач обработки текстов на естественном языке;

-поиска данных и моделей для решения прикладных задач;

-визуализации результатов обучения для анализа моделей NLP;

-самостоятельного изучения и применения новых технологий в NLP;

-применения инструментов библиотек transformers и datasets;

-подготовки данных для обучения и тестирования моделей.

3.2 Место дисциплины в структуре ОПОП

Дисциплина изучается на основе ранее освоенных дисциплин учебного плана:

1. «Глубокое обучение»

и обеспечивает изучение последующих дисциплин:

1. «Генеративные нейронные сети и LLMs»

2. «Производственная практика (преддипломная практика)»

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

В результате освоения образовательной программы обучающийся должен достичь следующие результаты обучения по дисциплине:

Код компетенции/ индикатора компетенции	**Наименование компетенции/индикатора компетенции**
ПК-0	Способен разрабатывать информационные модели и применять их для решения задач профессиональной деятельности
ПК-0.1	*Знает современные виды информационных моделей, применяемых при решении задач профессиональной деятельности*
ПК-0.2	*Создает и модифицирует информационные модели для решения задач профессиональной деятельности*
ПК-0.3	*Применяет информационные модели для решения задач профессиональной деятельности*

4 СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

4.1 Содержание разделов дисциплины

4.1.1 Наименование тем и часы на все виды нагрузки

№ п/п	Наименование темы дисциплины	ЭЛек, ач	ЭПр, ач	ИКР, ач	СР, ач
1	Механизм внимания. Архитектура Transformer.	10	10		10
2	Дискриминативные модели.	8	8		10
3	Генеративные модели.	8	8	1	10
4	Ранжирующие модели.	8	8		9
	Итого, ач	34	34	1	39
	Из них ач на контроль	0	0	0	0
	Общая трудоемкость освоения, ач/зе	108/3			

4.1.2 Содержание

№ п/п	Наименование темы дисциплины	Содержание
1	Механизм внимания. Архитектура Transformer.	Преобразование последовательности в последовательность. Рекуррентные нейронные сети с вниманием. Архитектура Transformer. Механизм Self-Attention и Multi-Head Attention. Позиционное кодирование. Альтернативные реализации механизма внимания.
2	Дискриминативные модели.	Генеративные и дискриминативные модели глубокого обучения. Перенос обучения. Архитектура ELMo. Encoder-only модели. Архитектура BERT. Семейство моделей E5. Предварительное обучение и доообучение Encoder моделей. Оценка качества работы дискриминативных моделей. Задачи решаемые дискриминативными моделями и их практическое применение.
3	Генеративные модели.	Задача языкового моделирования. Decoder-only модели. Архитектура GPT. Обучение в многозадачном режиме. Encoder-Decoder модели. Архитектура T5. Настройка параметров генерации. Оценка качества работы генеративных моделей. Задачи решаемые генеративными моделями и их практическое применение. История возникновения и основные этапы развития диалоговых систем. Задачно-ориентированные диалоговые системы. Чат-боты.

№ п/п	Наименование темы дисциплины	Содержание
4	Ранжирующие модели.	Задача ранжирования. Базовые архитектуры ранжирующих моделей Cross Encoder и Bi Encoder. Мультивекторные модели. Архитектуры Poly Encoder и ColBERT. Задача реранжирования. Обучение ранжирующих моделей. Негативное семплирование. Оценка близости векторных представлений. Метрики ранжирования. Задачи, решаемые ранжирующими моделями и их практическое применение.

4.2 Перечень лабораторных работ

Лабораторные работы не предусмотрены.

4.3 Перечень практических занятий

Наименование практических занятий	Количество ауд. часов
1. Машинный перевод с русского на английский	10
2. Классификация тональности текстов	8
3. Моделирование персонифицированного диалога	8
4. Поиск текстов, содержащих ответ на вопрос	8
Итого	34

4.4 Курсовое проектирование

Курсовая работа (проект) не предусмотрены.

4.5 Реферат

Реферат не предусмотрен.

4.6 Индивидуальное домашнее задание

Задание 1. Механизм внимания. Архитектура Transformer

Цель: Освоить архитектуру Transformer, устранить ошибки в реализации классов *EncoderLayer* и *DecoderLayer*, обучить модель и провести оценку её качества с использованием метрик *loss*, *accuracy*, *precision* и *F1-score*.

Задание 2.Классификация текста. Дообучение модели ruBERT

Цель: Научиться дообучать модель *ruBERT* для задач текстовой классификации с применением стратегий *mean_pooling* и *cls_pooling*, сравнить их эффективность и оптимизировать гиперпараметры для достижения показателей качества выше 0.85.

Задание 3. Генерация текста. Дообучение модели T5

Цель: Освоить методы дообучения и настройки генеративных моделей (*ruGPT3-small*, *T5*), а также провести оценку качества генерации текста с использованием метрик *BLEU* и *METEOR*, сравнив результаты базовых и дообученных моделей.

Задание 4. Асимметричный семантический поиск. Дообучение модели ruBERT

Цель: Реализовать и дообучить Bi-encoder модели (*Multilingual E5*, *Sentence-transformers*) для задач асимметричного семантического поиска и оценить качество поиска по метрикам *Recall@5*, *MRR*, *MAP* и *NDCG@10*, проведя сравнительный анализ различных подходов.

Форма сдачи отчета: распечатанный на листах формата А4 отчет с типовым титульным листом, отвечающий требованиям, принятым в СПбГЭТУ ”ЛЭТИ”.

Объем отчета: от 5 до 25 стр. (указывать использованные источники не требуется).

Отчет загружается на проверку в электронном виде на платформе онлайн-обучения LETIteach.

Отчет о выполненной работе должен содержать:

- Тему работы
- Наименование дисциплины
- Ф.И.О. и номер группы исполнителя (-лей)
- Постановку задания и использованные исходные данные
- Последовательное выполнение пунктов работы с приведением полученных результатов и пояснений, каким образом они были получены

- При необходимости и сообразности – выводы по отдельным пунктам работы и по работе в целом.

Требования к оформлению и содержанию отчета о выполненной работе:

- Задача отчета – полноценно раскрыть проделанную работу (от постановки задачи до получения результатов и формулировки выводов), по возможности раскрыть ее актуальность.
- Схемы должны быть читаемыми как в смысле их масштаба, так и в смысле используемой системы обозначений.
- Обязательна нумерация/поименование страниц, пунктов работы, рисунков (снизу) и таблиц (сверху).
- Удобочитаемость текста, в частности форматирование основного текста: основной шрифт - TNR, 14 кегль, межстрочный интервал 1.2 п.
- Фрагменты программного кода должны быть либо содержательно интегрированы в текст отчета, либо включать комментарий, раскрывающие их назначение.

4.7 Доклад

Доклад не предусмотрен.

4.8 Кейс

Кейс не предусмотрен.

4.9 Организация и учебно-методическое обеспечение самостоятельной работы

Изучение дисциплины сопровождается самостоятельной работой студентов с рекомендованными преподавателем литературными источниками и информационными ресурсами сети Интернет.

Планирование времени для изучения дисциплины осуществляется на весь период обучения, предусматривая при этом регулярное повторение пройденного материала. Обучающимся, в рамках внеаудиторной самостоятельной работы, необходимо регулярно дополнять сведениями из литературных источников материал, законспектированный на лекциях. При этом на основе изучения рекомендованной литературы целесообразно составить конспект основных положений, терминов и определений, необходимых для освоения разделов учебной дисциплины.

Особое место уделяется консультированию, как одной из форм обучения и контроля самостоятельной работы. Консультирование предполагает особым образом организованное взаимодействие между преподавателем и студентами, при этом предполагается, что консультант либо знает готовое решение, которое он может предписать консультируемому, либо он владеет способами деятельности, которые указывают путь решения проблемы.

Самостоятельное изучение студентами теоретических основ дисциплины обеспечено необходимыми учебно-методическими материалами (учебники, учебные пособия, конспект лекций и т.п.), выполненными в печатном или электронном виде.

По каждой теме содержания рабочей программы могут быть предусмотрены индивидуальные домашние задания (расчетно-графические работы, рефераты, конспекты изученного материала, доклады и т.п.).

Изучение студентами дисциплины сопровождается проведением регулярных консультаций преподавателей, обеспечивающих практические занятия по дисциплине, за счет бюджета времени, отводимого на консультации (внеаудиторные занятия, относящиеся к разделу «Самостоятельные часы для изучения дисциплины»).

В случае применения ДОТ с заменой аудиторных занятий:

Самостоятельной записи на курс нет. Студент заходит на курс, используя логин/пароль от единой учетной записи университета (единый логин и пароль). Каждую неделю будет доступна новая тема курса: видеолекции, кратко раскрывающие содержание каждой темы, презентации и конспекты, с которыми обучающиеся смогут ознакомиться в любое удобное время. Все темы включают практические занятия, которые предусматривают самостоятельное выполнение заданий, а также задания с автоматической проверкой, результаты которых учитываются при общей аттестации полученных знаний. В конце каждой лекции необходимо пройти небольшой контрольный тест, который покажет насколько усвоен предложенный материал. Рекомендуем изучать материал последовательно, что существенно облегчит работу. У каждого контрольного задания имеется своя форма (тест или практическое задание) есть срок выполнения (окончательный срок), по истечении которого даже правильные ответы система принимать не будет! В расписании курса указан окончательный срок каждого задания, который варьируется от двух до четырех недель в зависимости от его сложности. Весь учебный курс рассчитан на 16 недель. Его итоги будут подведены в течение нескольких недель после его окончания.

Текущая СРС	**Примерная трудоемкость, ач**
Работа с лекционным материалом, с учебной литературой	6
Опережающая самостоятельная работа (изучение нового материала до его изложения на занятиях)	0
Самостоятельное изучение разделов дисциплины	8
Выполнение домашних заданий, домашних контрольных работ	12
Подготовка к лабораторным работам, к практическим и семинарским занятиям	9
Подготовка к контрольным работам, коллоквиумам	0
Выполнение расчетно-графических работ	0
Выполнение курсового проекта или курсовой работы	0
Поиск, изучение и презентация информации по заданной проблеме, анализ научных публикаций по заданной теме	0
Работа над междисциплинарным проектом	0
Анализ данных по заданной теме, выполнение расчетов, составление схем и моделей, на основе собранных данных	0
Подготовка к зачету, дифференцированному зачету, экзамену	4
ИТОГО СРС	39

5 Учебно-методическое обеспечение дисциплины

5.1 Перечень основной и дополнительной литературы, необходимой для освоения дисциплины

№ п/п	Название, библиографическое описание	К-во экз. в библ.
Основная литература		
1	Брайан Макмахан Знакомство с PyTorch : глубокое обучение при обработке естественного языка / Макмахан Брайан, Рао Делип, 2021. -256 с. -Текст : непосредственный.	неогр.
2	Лейн Хобсон Обработка естественного языка в действии / Хобсон Лейн, Ханнес Хапке, Коул Ховард, 2021. -576 с. -Текст : непосредственный.	неогр.
Дополнительная литература		
1	Бенджамин Бенгфорт Прикладной анализ текстовых данных на Python : Машинное обучение и создание приложений обработки естественного языка / Бенгфорт Бенджамин, Билбро Ребекка, Охеда Тони, 2021. -368 с. -Текст : непосредственный.	неогр.

5.2 Перечень ресурсов информационно-телекоммуникационной сети «Интернет», используемых при освоении дисциплины

№ п/п	Электронный адрес
1	NLP: что это такое и как она работаетhttps://skillbox.ru/media/code/nlp-chto-eto-takoe-i-kak-ona-rabotaet/
2	Методы обработки естественного языкаhttps://developers.sber.ru/help/gigachat-api/natural-language-processing-techniques
3	11 примеров применений NLP в бизнесеhttps://allsee.team/natural-language-processing-applications

5.3 Адрес сайта курса

Адрес сайта курса: https://vec.etu.ru/moodle/course/view.php?id=25050

6 Критерии оценивания и оценочные материалы

6.1 Критерии оценивания

Для дисциплины «Глубокое обучение в NLP» предусмотрены следующие формы промежуточной аттестации: зачет с оценкой.

Зачет с оценкой

Оценка	Описание
Неудовлетворительно	теоретическое содержание курса не освоено, необходимые практически навыки и умения не сформированы, выполненные учебные задания содержат грубые ошибки, дополнительная самостоятельная работа над курсом не приведет к существенному повышению качества выполнения учебных заданий
Удовлетворительно	теоретическое содержание курса освоено частично, но пробелы не носят существенного характера, необходимые практические навыки и умения работы с освоенным материалом в основном сформированы, большинство предусмотренных программой обучения учебных заданий выполнено, некоторые из выполненных заданий содержат ошибки
Хорошо	теоретическое содержание курса освоено полностью, без пробелов, некоторые практические навыки и умения сформированы недостаточно, все предусмотренные программой обучения учебные задания выполнены, качество выполнения ни одного из них не оценено минимальным числом баллов, некоторые виды заданий выполнены с ошибками
Отлично	теоретическое содержание курса освоено полностью, без пробелов, необходимые практические навыки и умения сформированы, все предусмотренные программой обучения учебные задания выполнены, качество их выполнения оценено количеством баллов, близким к максимальному

Особенности допуска

К зачету с оценкой допускаются студенты, получившие оценку ”удовлетворительно” и выше по результатам текущего контроля (т.е. не ниже, чем ”удовлетворительно” за каждый тест, практическую работу и ИДЗ). Итоговая оценка по дисциплине выставляется на основе оценки студента за зачёт в форме теста.

6.2 Оценочные материалы для проведения текущего контроля и промежуточной аттестации обучающихся по дисциплине

Вопросы к дифф.зачету

№ п/п	Описание
1	Опишите ключевые недостатки рекуррентных нейронных сетей (RNN) и объясните, как архитектура Transformer их устраняет.
2	Раскройте принцип работы механизма Self-Attention, включая вычисление матриц Query, Key и Value.
3	Обсудите, зачем в архитектуре Transformer используется Multi-Head Attention и как это повышает эффективность модели.
4	Объясните важность позиционного кодирования в Transformer и сравните методы его реализации (например, синусоидальное vs. обучаемое).
5	Сравните подходы «последовательность-в-последовательность» (seq2seq) в Transformer и классических RNN.
6	Приведите примеры альтернативных механизмов внимания (например, sparse attention) и опишите их преимущества.
7	Обоснуйте, почему архитектура Transformer лучше справляется с длинными текстовыми последовательностями, чем RNN.
8	Сравните генеративные и дискриминативные модели, указав их различия и типичные задачи для каждого класса.
9	Опишите, как механизм Masked Language Modeling (MLM) используется в предобучении модели BERT.
10	Объясните, почему ELMo считается контекстуализированной моделью, в отличие от статических эмбеддингов word2vec.
11	Охарактеризуйте назначение моделей семейства E5 и их ключевые отличия от BERT.
12	Распишите шаги дообучения (fine-tuning) модели BERT для задачи классификации текста.
13	Перечислите метрики для оценки качества дискриминативных моделей в семантическом поиске и обоснуйте их выбор.
14	Объясните, почему Encoder-only модели (например, BERT) непригодны для генерации текста.

15	Раскройте суть языкового моделирования (language modeling) в decoder-only архитектурах на примере GPT.
16	Сравните обучение моделей T5 и GPT, выделив особенности подхода «текст-в-текст» в T5.
17	Объясните, как параметры генерации (temperature, top-k, top-p) влияют на качество и разнообразие выходного текста.
18	Сравните задачно-ориентированные диалоговые системы и чат-боты общего назначения, приведя примеры использования.
19	Обсудите причины возникновения «галлюцинаций» в генеративных моделях и методы их минимизации.
20	Опишите метрики оценки качества генерации текста (BLEU, ROUGE, Perplexity) и их ограничения.
21	Сравните архитектуры Encoder-Decoder (T5) и decoder-only (GPT), указав их сильные стороны.
22	Сравните архитектуры Cross Encoder и Bi Encoder, указав сценарии их применения.
23	Объясните роль негативного семплирования в обучении ранжирующих моделей и его влияние на качество.
24	Опишите, как ColBERT использует раздельное кодирование запроса и документа для улучшения ранжирования.
25	Перечислите метрики оценки ранжирования (MRR, NDCG) и объясните, в каких задачах они применяются.
26	Раскройте концепцию реранжирования (re-ranking) и опишите, как комбинируются Bi Encoder и Cross Encoder.
27	Объясните, чем Poly Encoder улучшает подход Bi Encoder, фокусируясь на взаимодействии контекстов.
28	Обоснуйте выбор между BERT и GPT для разработки чат-бота, учитывая их архитектурные особенности.
29	Опишите этапы создания системы семантического поиска с использованием моделей E5 и ColBERT.
30	Проанализируйте этические и технические ограничения современных NLP-моделей (например, BERT, GPT) при их внедрении.

Вариант теста

1. Какой ключевой недостаток рекуррентных нейронных сетей (RNN) устранила архитектура Transformer?

а) Неспособность работать с изображениями.

б) Вычислительная неэффективность из-за последовательной обработки данных и проблема исчезающих градиентов.

в) Слишком малое количество обучаемых параметров.

г) Отсутствие механизма нелинейной активации.

2. Что НЕ является частью вычисления механизма Self-Attention?

а) Умножение матрицы Query на матрицу Key для получения матрицы внимания.

б) Применение функции softmax к матрице внимания для получения весов.

в) Суммирование всех векторов входной последовательности.

г) Умножение весов на матрицу Value для получения итогового контекстуализированного вектора.

3. Зачем в архитектуре Transformer используется механизм Multi-Head Attention?

а) Чтобы уменьшить общее количество параметров модели.

б) Чтобы модель могла параллельно обращать внимание на информацию из разных подпространств представлений (например, на синтаксис и семантику).

в) Чтобы сделать модель обязательно глубокой.

г) Это необходимо для работы механизма Dropout.

4. Почему в Transformer важно позиционное кодирование?

а) Для увеличения скорости обучения.

б) Поскольку Self-Attention не учитывает порядок элементов последовательности по своей природе, позиционное кодирование добавляет информацию об их расположении.

в) Чтобы уменьшить требования к памяти.

г) Для того чтобы модель могла работать с изображениями.

5. Какое ключевое архитектурное отличие в подходе «последовательность-в-последовательность» (seq2seq) у Transformer по сравнению с классиче-

скими RNN?

а) Transformer использует только один слой LSTM.

б) Transformer заменяет RNN-энкодер и RNN-декодер на стеки слоев Self-Attention и Feed-Forward, что позволяет обрабатывать последовательность целиком параллельно.

в) Transformer не может использоваться для задач seq2seq.

г) В Transformer декодер должен работать строго последовательно, а энкодер — параллельно.

6. Какой из перечисленных механизмов внимания позволяет модели обрабатывать чрезвычайно длинные последовательности, игнорируя большинство связей между токенами?

а) Полное внимание (Full Attention)

б) Симметричное внимание (Symmetric Attention)

в) Разреженное внимание (Sparse Attention)

г) Линейное внимание (Linear Attention)

7. Почему архитектура Transformer лучше справляется с длинными текстовыми последовательностями, чем RNN?

а) RNN по своей природе не могут обрабатывать текст.

б) Transformer обрабатывает все токены последовательности параллельно, а длина пути для зависимости между любыми двумя токенами постоянна и мала, что решает проблему исчезающих градиентов.

в) Transformer имеет меньшую вычислительную сложность по сравнению с RNN для любых задач.

г) Transformer не требует позиционного кодирования.

8. Какая из перечисленных задач является типичной для генератив-

ной модели?

а) Классификация отзывов на положительные и отрицательные.

б) Определение тональности текста.

в) Создание нового текста, продолжающего заданный промпт.

г) Определение, является ли электронное письмо спамом.

9. Какой из следующих шагов НЕ входит в механизм Masked Language Modeling (MLM) в BERT?

а) Замена случайно выбранных токенов в последовательности на специальный токен [MASK].

б) Предсказание исходного токена по его контексту (окружающим словам).

в) Авторегрессивное предсказание следующего токена в последовательности.

г) Обучение модели бидиректив но, учитывая как левый, так и правый контекст.

10. В чем ключевое отличие эмбеддингов ELMo от статических эмбеддингов word2vec?

а) ELMo создает одно статическое векторное представление для каждого слова.

б) ELMo генерирует контекстуализированные представления слов, которые зависят от всего предложения, в котором это слово находится.

в) word2vec учитывает морфологию слова.

г) ELMo не может быть использован для решения практических задач.

11. Каково основное назначение моделей семейства E5?

а) Генерация поэзии.

б) Создание универсальных векторных представлений текста (эмбеддингов) для задач поиска и семантического сравнения.

в) Классификация изображений.

г) Машинный перевод.

12. Какой из этих шагов является заключительным при дообучении (fine-tuning) BERT для классификации текста?

а) Добавление классификационного слоя (головы) поверх выходного представления [CLS]-токена и обучение всей модели на целевом наборе данных.

б) Замена всех слоев энкодера BERT.

в) Использование только выходных представлений всех токенов без [CLS].

г) Обучение модели только на задаче MLM.

13. Какая метрика наиболее подходит для оценки качества семантического поиска, когда важно, чтобы релевантный документ был на первом месте?

а) Precision@1

б) Recall@10

в) Mean Reciprocal Rank (MRR)

г) Точность (Accuracy)

14. Почему Encoder-only модели (например, BERT) непригодны для генерации связного текста?

а) Они не умеют работать с текстом.

б) Их архитектура и предобучение (например, MLM) не предназначены для авторегрессивного предсказания следующего токена, которое лежит в основе генерации.

в) Они слишком медленные.

г) Они не имеют механизма внимания.

15. В чем суть языкового моделирования в decoder-only архитектурах, таких как GPT?

а) Предсказание случайного слова в середине предложения.

б) Авторегрессивное предсказание следующего токена в последовательности на основе предыдущих токенов.

в) Определение тональности всего текста.

г) Классификация текста по категориям.

16. Какой подход лежит в основе модели T5?

а) Все задачи, включая классификацию, перевод и суммаризацию, преобразуются в формат «текст-в-текст».

б) Модель обучается только на задаче генерации диалогов.

в) Модель использует только энкодер, как BERT.

г) Модель не использует механизм внимания.

17. Как параметр temperature влияет на генерацию текста?

а) Высокое значение temperature делает распределение вероятностей более пиковым, увеличивая уверенность модели и снижая разнообразие.

б) Низкое значение temperature делает распределение вероятностей более сглаженным, увеличивая случайность и разнообразие выходного текста.

в) Высокое значение temperature делает распределение вероятностей более сглаженным, увеличивая случайность и разнообразие выходного текста.

г) Параметр temperature не влияет на распределение вероятностей.

18. Что характерно для задачно-ориентированных диалоговых систем?

а) Они предназначены для ведения светской беседы на любые темы.

б) Они сфокусированы на выполнении конкретной задачи в узкой предметной области (например, бронирование столика, служба поддержки).

в) Они всегда используют генеративные модели.

г) Они не требуют интеграции с внешними базами данных.

19. Что такое «галлюцинация» в контексте генеративных моделей?

а) Слишком высокая скорость генерации текста.

б) Генерация моделью информации, которая является неправдоподобной или не содержится в исходных данных.

в) Использование слишком низкого значения temperature.

г) Отсутствие нелинейных активаций в модели.

20. Какая метрика оценивает качество генерации текста путем сравнения с эталонными текстами на основе совпадения n-грамм?

а) BLEU

б) Perplexity

в) Accuracy

г) F1-Score

21. Какая архитектура обычно более эффективна для задач чистой генерации текста, таких как написание статей или диалоги?

а) Encoder-Decoder (T5)

б) Decoder-only (GPT)

в) Encoder-only (BERT)

г) Симметричная (Autoencoder)

22. В каком сценарии Bi-Encoder предпочтительнее Cross-Encoder?

а) Когда требуется максимально точное попарное сравнение двух текстов,

и время выполнения не критично.

б) Когда необходимо предварительно отобрать кандидатов из большого набора данных, так как эмбеддинги от Bi-Encoder можно предвычислить и использовать для быстрого косинусного сходства.

в) Когда оба текста очень короткие.

г) Когда модель должна быть обязательно генеративной.

23. Какова основная роль негативного семплирования при обучении ранжирующих моделей?

а) Увеличение скорости обучения модели.

б) Предоставление модели примеров нерелевантных документов, чтобы научить ее лучше различать релевантные и нерелевантные результаты.

в) Уменьшение размера модели.

г) Генерация позитивных примеров.

24. В чем ключевая особенность подхода ColBERT?

а) Он не использует механизм внимания.

б) Он кодирует запрос и документ независимо, но вычисляет внимание на позднем этапе, используя тщательно спроектированные взаимодействия между их токенами, что обеспечивает хороший баланс между скоростью и качеством.

в) Он требует полного попарного взаимодействия запроса и документа на каждом шаге инференса, как Cross-Encoder.

г) Он генерирует текст на основе запроса.

25. Какая метрика для оценки ранжирования учитывает позицию первого релевантного документа и хорошо подходит для задач, где пользователь ожидает найти ответ сразу?

а) NDCG@10

б) MRR (Mean Reciprocal Rank)

в) Precision@5

г) Recall@100

26. Что такое реранжирование (re-ranking) в конвейере семантического поиска?

а) Удаление нерелевантных документов из индекса.

б) Двухэтапный процесс: быстрый поиск кандидатов (например, с помощью Bi-Encoder), а затем их точное ранжирование с помощью более медленной, но точной модели (например, Cross-Encoder).

в) Генерация новых документов по запросу.

г) Использование только одной модели для всего процесса.

27. Чем механизм Poly-Encoder улучшает подход Bi-Encoder?

а) Полностью отказывается от использования эмбеддингов.

б) Вводит промежуточный шаг, где вычисляются внимания между запросом и несколькими кодами-кандидатами из документа, что позволяет уловить более сложные взаимодействия, оставаясь при этом быстрее Cross-Encoder.

в) Использует только один глобальный вектор представления для документа.

г) Является чисто генеративной архитектурой.

28. Какую модель следует выбрать для разработки чат-бота, способного генерировать креативные и развернутые ответы?

а) BERT, потому что она лучше понимает контекст.

б) GPT, так как ее decoder-only архитектура предобучена именно для авторегрессивной генерации текста.

в) E5, потому что она создает качественные эмбеддинги.

г) T5, но только для задач классификации.

29. Какой из этапов НЕ является частью создания системы семантического поиска с использованием E5 и ColBERT?

а) Создание векторного индекса для коллекции документов с помощью модели E5.

б) Генерация поэтических текстов на каждый запрос пользователя.

в) Быстрый поиск кандидатов по векторному индексу (с использованием E5).

г) Точное реранжирование топ-кандидатов с помощью модели ColBERT.

30. Какое из перечисленных утверждений НЕ является типичным ограничением современных больших языковых моделей (LLM), таких как BERT или GPT?

а) Модели могут воспроизводить и усиливать социальные предубеждения, присутствующие в данных для обучения.

б) Модели требуют значительных вычислительных ресурсов для обучения и развертывания, что создает экологические и экономические барьеры.

в) Модели обладают истинным пониманием смысла и причинно-следственных связей, как человек.

г) Модели склонны к ”галлюцинациям” — генерации правдоподобной, но фактически неверной информации.

Образцы задач (заданий) для контрольных (проверочных) работ

Вопросы в тестах формируются аналогично приведенным ниже примерам.

1. Какой ключевой недостаток RNN устраняет архитектура Transformer?

a) Низкая скорость обучения

b) Проблема исчезающего градиента

c) Неспособность работать с изображениями

d) Ограниченная длина входных данных

2. Какие матрицы используются в механизме Self-Attention?

a) Query

b) Gradient

c) Key

d) Value

3. Зачем в Transformer применяется Multi-Head Attention?

a) Для уменьшения вычислительной сложности

b) Чтобы модель учитывала разные аспекты контекста

c) Для ускорения сходимости обучения

d) Чтобы избежать переобучения

4. Почему в Transformer важно позиционное кодирование?

a) Для учета порядка слов, так как Transformer не имеет рекуррентности

b) Чтобы уменьшить размерность эмбеддингов

c) Для стабилизации градиентов

d) Чтобы избежать внимания к padding-токенам

5. Чем подход seq2seq в Transformer отличается от классического RNN?

a) Использует механизм внимания вместо скрытых состояний

b) Обрабатывает последовательность целиком, а не пошагово

c) Требует меньше памяти

d) Работает только с фиксированной длиной последовательности

6. Какое преимущество у Sparse Attention?

a) Уменьшает вычислительные затраты

b) Улучшает интерпретируемость модели

c) Позволяет учитывать только локальные зависимости

d) Все варианты верны

7. Почему Transformer лучше RNN для длинных текстов?

a) Параллельная обработка и отсутствие проблем с долгосрочными зависимостями

b) Меньше параметров

c) Не требует позиционного кодирования

d) Использует только локальное внимание

8. Какие задачи решают генеративные модели?

a) Классификация текста

b) Генерация текста

c) Перевод

d) Семантический поиск

9. Как работает Masked Language Modeling?

a) Предсказывает следующее слово в последовательности

b) Восстанавливает замаскированные токены на основе контекста

c) Классифицирует предложения

d) Оптимизирует perplexity

10. Почему ELMo считается контекстуализированной моделью?

a) Использует разные эмбеддинги для одного слова в разных контекстах

b) Обучается на меньших данных

c) Не требует предобучения

d) Работает только с частотными словами

11. Почему BERT (Encoder-only) непригоден для генерации текста?

a) Не имеет механизма авторегрессии

b) Использует только Masked Language Modeling

c) Не учитывает позиционные эмбеддинги

d) Обучается только на классификации

12. Как GPT генерирует текст?

a) Предсказывает следующее слово на основе предыдущих

b) Восстанавливает замаскированные токены

c) Использует bidirectional контекст

d) Классифицирует токены

13. Чем подход T5 («текст-в-текст») отличается от GPT?

a) Все задачи преобразуются в текстовый формат

b) T5 использует только encoder

c) GPT не поддерживает transfer learning

d) T5 объединяет encoder и decoder

14. Как параметр temperature влияет на генерацию?

a) Увеличивает разнообразие, но снижает точность

b) Уменьшает длину текста

c) Фильтрует редкие слова

d) Ускоряет инференс

15. Какие особенности у чат-ботов общего назначения?

a) Ориентированы на узкую предметную область

b) Используют генеративные модели (например, GPT)

c) Требуют жестких правил

d) Поддерживают открытые диалоги

16. Как можно уменьшить галлюцинации?

a) Увеличить температуру генерации

b) Добавить проверку фактов через поиск

c) Игнорировать контекст

d) Обучать только на коротких текстах

17. Какие метрики оценивают качество генерации текста?

a) BLEU

b) Accuracy

c) ROUGE

d) Perplexity

18. В чем сила архитектуры T5?

a) Только генерация текста

b) Универсальность для задач «текст-в-текст»

c) Отсутствие позиционного кодирования

d) Использование только autoregressive подхода

19. Когда используют Cross Encoder?

a) Для быстрого поиска кандидатов

b) Для точного реранжирования

c) Для генерации текста

d) Для обучения без учителя

20. Зачем нужно негативное семплирование в ранжировании?

a) Чтобы увеличить скорость инференса

b) Для улучшения различимости релевантных/нерелевантных пар

c) Чтобы уменьшить размер модели

d) Для генерации негативных примеров

Весь комплект контрольно-измерительных материалов для проверки сформированности компетенции (индикатора компетенции) размещен в закрытой части по адресу, указанному в п. 5.3

6.3 График текущего контроля успеваемости

Неделя	Темы занятий	Вид контроля
1	Механизм внимания. Архитектура Transformer.	Практическая работа
2	Механизм внимания. Архитектура Transformer.	
3		ИДЗ / ИДРГЗ / ИДРЗ
4	Дискриминативные модели.	Практическая работа
5	Дискриминативные модели.	
6		ИДЗ / ИДРГЗ / ИДРЗ
7	Механизм внимания. Архитектура Transformer. Дискриминативные модели.	
8		Тест
9	Генеративные модели.	Практическая работа
10	Генеративные модели.	
11		ИДЗ / ИДРГЗ / ИДРЗ
12	Ранжирующие модели.	Практическая работа
13	Ранжирующие модели.	
14		ИДЗ / ИДРГЗ / ИДРЗ
15	Генеративные модели. Ранжирующие модели.	
16		
17		Тест

6.4 Методика текущего контроля

Методика текущего контроля на лекционных занятиях

Текущий контроль усвоения материала лекционных занятий проводится в форме тестирования на платформе онлайн-обучения LETIteach.

Каждый тест включает в себя 20 тестовых заданий.

«неудовлетворительно» <=9 правильных ответов;

«удовлетворительно» 10-14 правильных ответов;

«хорошо» 15-17 правильных ответов;

«отлично» 18-20 правильных ответов;

Тест считается успешно пройденным, если решено не менее 10 тестовых заданий.

Принципы и критерии оценки практических работ:

Практические занятия направлены на подготовку студентов к выполнению и сдаче ИДЗ.

Оценка выставляется по 4-балльной шкале по итогам выполнения работы:

«отлично» – работа выполнена полностью без ошибок, студент уверенно владеет материалом.

«хорошо» – работа выполнена с 1-2 незначительными ошибками, исправленными с помощью преподавателя.

«удовлетворительно» – работа выполнена частично или с существенными ошибками, потребовалась значительная помощь.

«неудовлетворительно» – работа не выполнена или выполнена неверно, студент не владеет материалом.

Принципы и критерии оценки ИДЗ:

ИДЗ должно быть выполнено на основе выданного преподавателем набора входных данных и соответствовать закрепленному за студентом варианту задания.

Выполнение ИДЗ оценивается по следующим критериям:

- вовремя сданное задание 0-5;
- полнота и корректность выполнения задания 0-5;
- качество защиты задания 0-5.

Средний балл за ИДЗ = (баллы за вовремя сданное задание + баллы за полноту икорректность выполнения задания + баллы за качество защиты задания) / 3.

Максимальная оценка одно ИДЗ - 5 баллов (оценка ”отлично”).

Для допуска к дифф. зачету студенту необходимо получить оценку не ниже, чем ”удовлетворительно” за каждое ИДЗ, практическую работу и тест.

Методика текущего контроля самостоятельной работы студентов.

Контроль самостоятельной работы студентов осуществляется на лекционных и практических занятиях студентов по методикам, описанным выше.

7 Описание информационных технологий и материально-технической базы

Тип занятий	Тип помещения	Требования к помещению	Требования к программному обеспечению
Лекция	Лекционная аудитория	Количество посадочных мест в соответствии с контингентом. Рабочее место преподавателя, ноутбук, проектор, экран, маркерная доска.	Альт Рабочая станция
Практические занятия	Аудитория	Количество посадочных мест в соответствии с контингентом. Рабочее место преподавателя, ноутбук, проектор, экран, маркерная доска. Наличие ПК на рабочих местах студентов.	Альт Рабочая станция
Самостоятельная работа	Помещение для самостоятельной работа	Оснащено компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду университета.	Альт Рабочая станция

8 Адаптация рабочей программы для лиц с ОВЗ

Адаптированная программа разрабатывается при наличии заявления со стороны обучающегося (родителей, законных представителей) и медицинских показаний (рекомендациями психолого-медико-педагогической комиссии). Для инвалидов адаптированная образовательная программа разрабатывается в соответствии с индивидуальной программой реабилитации.

ЛИСТ РЕГИСТРАЦИИ ИЗМЕНЕНИЙ

№ п/п	Дата	Изменение	Дата и номер протокола заседания УМК	Автор	Начальник ОМОЛА