

Министерство цифрового развития, связи и массовых коммуникаций Российской Федерации  
Ордена Трудового Красного Знамени  
федеральное государственное бюджетное образовательное учреждение  
высшего образования  
«Московский технический университет связи и информатики»

Разрешаю  
допустить к защите  
Зав. кафедрой

\_\_\_\_\_ 20\_\_\_\_ г.

## ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

НА ТЕМУ

Стандартизация навыков в резке  
кардигата на основе больших языковых  
моделей

Студент: Моисеева Н.А. Моисеева  
Руководитель: Салов В.И. Салов

Москва 2024 г.

МИНИСТЕРСТВО ЦИФРОВОГО РАЗВИТИЯ, СВЯЗИ И МАССОВЫХ  
КОММУНИКАЦИЙ РОССИЙСКОЙ ФЕДЕРАЦИИ

Ордена Трудового Красного Знамени федеральное государственное  
бюджетное образовательное учреждение высшего образования  
«Московский технический университет связи и информатики»

Кафедра Математическая кибернетика и информационные технологии  
(название полностью)

«Утверждаю»

Зав. кафедрой Городничев М.Г.

«   »     20    г.

**ЗАДАНИЕ**  
на выпускную квалификационную работу

Студенту Моисеевой Надежде Александровне гр. БВТ2003

Направление (специальность) 09.03.01 Информатика и вычислительная техника

Форма выполнения выпускной квалификационной работы бакалаврская работа

(Дипломный проект, дипломная работа, магистерская диссертация, бакалаврская работа)

Тема выпускной квалификационной работы «Стандартизация навыков в резюме кандидата  
на основе больших языковых моделей»

Утверждена приказом ректора № 114-с от 24 января 2024г.

1. Исходные данные

- Высокоуровневый, интерпретируемый, объектно-ориентированный язык программирования Python
- Облачная платформа для сложных вычислений Yandex Cloud
- Виртуальная платформа по анализу данных, машинному обучению и искусственному интеллекту Kaggle.
- Библиотека моделей для обработки естественного языка Transformers

Объем работы в % и сроки  
выполнения по разделам

5% - 29.02.2024

18% - 26.03.2024

25% - 23.04.2024

46% - 17.05.2024

5% - 29.05.2024

- Библиотека для для работы с большими языковыми моделями Peft
  - Библиотека для работы с большими языковыми моделями Bitsandbytes
  - Платформа для обмена моделями машинного обучения и наборами данных HuggingFace
2. Содержание расчетно-пояснительной записки (перечень подлежащих разработке вопросов)  
Введение  
Глава 1. Теоретические основы исследования  
Глава 2. Большие языковые модели, их сравнение и выбор наиболее подходящей в рамках данного исследования.  
Глава 3. Разработка метода стандартизации навыков на основе больших языковых моделей.  
Заключение
  3. Вопросы конструктивных разработок
  4. Разработка вопросов по экологии и безопасности жизнедеятельности
  5. Техничко-экономическое обоснование (подлежащее расчету)
  6. Перечень графического материала (с точным указанием обязательных чертежей)

7. Консультанты по ВКР (с указанием относящихся к ним разделов проекта):

\_\_\_\_\_  
(подпись)

\_\_\_\_\_  
(ФИО)

\_\_\_\_\_  
(подпись)

\_\_\_\_\_  
(ФИО)

8. Срок сдачи студентом законченной ВКР:

\_\_\_\_\_

Дата выдачи задания:

\_\_\_\_\_

Руководитель \_\_\_\_\_

(подпись)

Соловьёв В.И.

(ФИО)

\_\_\_\_\_ почасовая

(штатная или почасовая)

\_\_\_\_\_ нагрузка

Задание принял к исполнению



(подпись студента)

Примечание: Настоящее задание прилагается к законченной ВКР

## ОТЗЫВ РУКОВОДИТЕЛЯ

о работе обучающегося Моисеевой Надежды Александровны в период подготовки выпускной квалификационной работы на тему «Стандартизация навыков в резюме кандидата на основе больших языковых моделей»

Рынок квалифицированных специалистов сегодня является рывком предложения: организации активно конкурируют за профессионалов с опытом работы. При этом число откликов на хорошие вакансии достигает сотен и тысяч, и их разбор является очень ресурсоемкой рутинной интеллектуальной задачей. Связано это с тем, что перечень навыков, необходимых работодателю от специалиста на данной позиции, понятен, но поскольку резюме составляется кандидатами в свободной форме на естественном языке, одни и те же навыки разными кандидатами могут называться совершенно разными словами, какие-то навыки кандидатами в явном виде не указываются, но их наличие очевидно из описания опыта работы. В связи с этим применение больших языковых моделей для стандартизации навыков соискателей, которому посвящена выпускная квалификационная работа студентки Н.А. Моисеевой, является актуальным. В более развернутом виде актуальность подтверждена и теоретическим обзором, проведенным студенткой в ходе выполнения работы в достаточном объеме.

Работа является логически завершенной, в ней проведен анализ существующих подходов к описанию навыков и компетенций в резюме соискателей, выявлены основные ограничения существующих подходов. Далее исследованы возможности применения современных больших языковых моделей к стандартизации информации о навыках. После этого была предложена методика стандартизации навыков в резюме на основе больших языковых моделей. Эта методика протестирована на реальных резюме, и результаты тестирования также представлены в работе. В завершение работы проведен анализ эффективности и результативности предложенной методики.

Значения метрик ROUGE-1 = 0.56, BLEU = 0.42, BERTScore = 0.48, полученные для дообученной студенткой модели LLaMA2 7b, говорят о достаточно высокой способности модели стандартизировать навыки соискателей.

В ходе выполнения работы студентка использовала современные большие языковые модели, в частности, была обоснованно выбрана и дообучена на собранном и предобработанном наборе резюме модель LLaMA2 7b.

Выпускная квалификационная работа Н.А. Моисеевой полностью написана самостоятельно, все ссылки на использованные источники в обзоре литературы указаны корректно. Работа не имеет существенных недостатков, все замечания студентка оперативно устраняла.

Н.А. Моисеева работала над своей выпускной квалификационной работой в полном соответствии с заданием, продемонстрировав высокий уровень сформированных компетенций и полное соответствие всем требованиям, предъявляемым к бакалаврам по направлению 09.03.01 – Информатика и вычислительная техника.

Считаю, что работа соответствует требованиям, предъявляемым к выпускным квалификационным работам бакалавров по направлению 09.03.01 – Информатика и вычислительная техника, и может быть рекомендована к защите.

Научный руководитель,  
заведующий кафедрой  
«Прикладной искусственный интеллект»,  
доктор экономических наук, профессор

В.И. Соловьев

## **Аннотация**

Тема выпускной квалификационной работы - Стандартизация навыков в резюме кандидата на основе больших языковых моделей.

Автор работы: студентка группы БВТ2003 Моисеева Надежда Александровна.

Научный руководитель: Соловьёв Владимир Игоревич.

Ключевые слова: стандартизация навыков, резюме, большие языковые модели, суммаризация, трансформеры, LLaMA 2 7b, ROUGE, BLEU, BERTScore.

Цель работы: разработка комплексного решения, направленного на повышение эффективности и объективности процессов анализа и оценки резюме кандидатов.

Содержание: введение, три главы, заключение, список использованных источников.

Объём работы - 50 страниц, рисунков - 15, таблиц - 1.

Во введении сформулированы цель и задачи работы, раскрыты научная новизна и практическая значимость. В первой главе проводится анализ существующих подходов к описанию навыков в резюме и исследование возможностей применения больших языковых моделей для извлечения и стандартизации информации о навыках из текстов резюме. Вторая глава посвящена анализу существующих БЯМ, оцениваются их возможности и ограничения в контексте решаемой задачи. Производится исследование того, на каком датасете лучше будет дообучать модель, также отбираются метрики оценки качества работы модели. В третьей главе представлена разработанная методика стандартизации навыков в резюме на основе больших языковых моделей. Описываются ключевые этапы алгоритма, включая предварительную обработку текста резюме, извлечение информации о навыках, и их стандартизацию. Приводятся результаты тестирования методики на реальных резюме кандидатов, анализируется ее эффективность и практическая применимость.

## Содержание

Содержание	6
Введение	7
Глава №1 Теоретические основы исследования	11
1.1 Анализ существующих методов оценки навыков соискателей	11
1.2 Нейронные сети	13
1.3 Трансформеры и модели на их основе	16
1.4 Квантизация	22
1.5 Задача на обобщение текста	23
Глава №2 Большие языковые модели, их сравнение и выбор наиболее подходящей в рамках данного исследования.	25
2.1 Сравнение и выбор модели	25
2.2 Метод дообучения модели	26
2.3 Выбор датасета	28
2.4 Метрики оценки качества	29
Глава №3 Разработка метода стандартизации навыков на основе больших языковых моделей.	34
3.1 Формирование датасета	34
3.2 Выбор кластера для обучения модели	37
3.3 Дообучение модели	38
ЗАКЛЮЧЕНИЕ	46
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ	48

## **Введение**

Современный рынок труда характеризуется высокой конкуренцией среди кандидатов при приеме на работу. Одним из ключевых инструментов, используемых работодателями для оценки и отбора кандидатов, является резюме - краткое изложение профессионального опыта, навыков и достижений человека. От того, насколько полно и структурировано в резюме представлена информация о навыках и компетенциях соискателя, во многом зависит его шанс успешно пройти первичный отбор и быть приглашенным на собеседование. Работодатели сталкиваются с необходимостью оценивать и сравнивать большое количество резюме, чтобы выбрать наиболее подходящих специалистов, так как существующие подходы к составлению резюме зачастую отличаются значительной вариативностью и субъективностью. Кандидаты по-разному описывают и структурируют информацию о своих навыках, используя различные формулировки и акценты. Это затрудняет объективное сравнение и оценку компетенций соискателей, создавая дополнительные сложности для рекрутеров и HR-специалистов. Традиционные подходы к анализу резюме, основанные на ручном просмотре и субъективной оценке, становятся все более трудоемкими и неэффективными в условиях растущих объемов данных.

Решение данных проблем может быть найдено в использовании современных технологий искусственного интеллекта и машинного обучения, в частности, больших языковых моделей. Эти модели, обученные на огромных объемах текстовых данных, обладают способностью понимать семантику естественного языка, извлекать смысловые связи и выявлять скрытые закономерности. Применение больших языковых моделей для анализа резюме открывает новые возможности для стандартизации и структурирования информации о навыках кандидатов.

Актуальность темы данной дипломной работы обусловлена, во-первых, растущей ролью больших языковых моделей в обработке текстовой информации, в том числе в задачах, связанных с рекрутингом и

HR-менеджментом. Во-вторых, необходимостью разработки единых подходов к описанию и оценке навыков в резюме, позволяющих повысить объективность и эффективность процессов отбора персонала. В-третьих, потребностью в автоматизации и оптимизации рутинных операций, связанных с анализом резюме, что очень актуально в условиях высокой конкуренции на рынке труда и возрастающих объемов обрабатываемых данных.

Ключевой целью данного дипломного проекта является разработка комплексного решения, направленного на повышение эффективности и объективности процессов анализа и оценки резюме кандидатов. Для достижения этой цели предлагается создание специализированного сервиса, обеспечивающего стандартизацию и структурирование информации о профессиональных навыках, представленной в резюме соискателей.

Для достижения поставленной цели в работе решаются следующие задачи:

1. Проанализировать существующие методы оценки навыков соискателей, выявить их ключевые особенности и ограничения.
2. Провести обзор и сравнительный анализ ключевых характеристик популярных больших языковых моделей, таких как GPT-3, BERT, T5 и LLaMA.
3. Найти подходящий датасет и подготовить его к работе.
4. Переобучить выбранную модель для стандартизации и категоризации навыков соискателей и протестировать её эффективность.
5. Оценить перспективы практического применения предлагаемого решения.

Научная новизна работы заключается в разработке нового подхода к стандартизации навыков в резюме, основанного на применении больших языковых моделей. В отличие от существующих решений, данный подход позволяет автоматизировать процесс извлечения, структурирования и оценки информации о компетенциях соискателей, повышая объективность и точность принимаемых решений.



Практическая значимость работы состоит в том, что предложенная методика может быть внедрена в системы автоматизированного рекрутинга, HR-технологии и платформы для поиска и подбора персонала. Это позволит оптимизировать и ускорить процессы отбора кандидатов, а также повысить качество принимаемых решений.

Теоретической и методологической основой исследования послужили работы отечественных и зарубежных ученых в области искусственного интеллекта, машинного обучения, обработки естественного языка, а также исследования, посвященные анализу и оценке профессиональных навыков.

Исходя из цели, предмета исследования и поставленных задач, в работе выделены такие разделы как: введение, первая, вторая и третья главы, заключение.

Во введении обоснована актуальность темы исследования, сформулированы цель и задачи работы, раскрыты научная новизна и практическая значимость.

В первой главе проводится анализ существующих подходов к описанию навыков и компетенций в резюме и выявляются ключевые проблемы и ограничения таких подходов.

Вторая глава посвящена исследованию возможностей применения больших языковых моделей для извлечения, структурирования и стандартизации информации о навыках из текстов резюме. Проводится анализ существующих БЯМ, оцениваются их возможности и ограничения в контексте решаемой задачи.

В третьей главе представлена разработанная методика стандартизации навыков в резюме на основе больших языковых моделей. Описываются ключевые этапы алгоритма, включая предварительную обработку текста резюме, извлечение информации о навыках, и их стандартизацию. Приводятся результаты тестирования методики на реальных резюме кандидатов, анализируется ее эффективность и практическая применимость.

В заключении подводятся итоги проведенного исследования, формулируются основные выводы и перспективы дальнейшего развития предложенной методики.

В целом, данная дипломная работа направлена на решение актуальной научно-практической задачи повышения эффективности оценки и отбора персонала на основе стандартизации информации о навыках, представленных в резюме кандидатов.

## **Глава №1 Теоретические основы исследования**

### **1.1 Анализ существующих методов оценки навыков соискателей**

Рекрутинг - это процесс поиска, привлечения и отбора персонала для замещения вакантных должностей в организации. Это одна из ключевых функций в управлении человеческими ресурсами, направленная на обеспечение компаний необходимыми квалифицированными сотрудниками.

Ключевым инструментом, используемым рекрутерами на этапе первичного отбора, является резюме кандидата. Резюме - это краткое описание профессионального опыта, навыков и достижений человека, предоставляемое работодателю для рассмотрения его кандидатуры на вакантную должность.

Традиционно оценка профессиональных навыков кандидатов осуществляется с помощью следующих методов:

1. Анализ резюме. Данный метод подразумевает ручной просмотр и интерпретацию информации, представленной в резюме соискателя. Основными ограничениями данного подхода являются:

- Субъективность оценки. Разные эксперты могут по-разному интерпретировать одни и те же формулировки в резюме.
- Неполнота информации. В резюме, как правило, приводится краткое и обобщенное описание навыков, не позволяющее в полной мере оценить их уровень.
- Отсутствие стандартизации. Не существует единых требований к структуре и содержанию резюме, что затрудняет сравнение кандидатов.

2. Собеседование. Проведение интервью с кандидатом позволяет оценить его профессиональные компетенции в процессе живого общения. Однако данный метод также имеет ряд ограничений:

- Ограниченность времени. В рамках стандартного интервью невозможно всесторонне проверить уровень владения всеми заявленными навыками.

- Субъективность оценки. Впечатление интервьюера может быть смещено личными предпочтениями или предубеждениями.

- Возможность подготовки кандидата. Соискатель может быть заранее готов к "правильным" ответам на типовые вопросы.

3. Тестирование. Данный метод предполагает оценку навыков кандидата с помощью специализированных тестов или практических заданий. Преимущества тестирования:

- Объективность оценки. Результаты тестов позволяют сравнивать кандидатов по единым критериям.

- Возможность проверки конкретных навыков. Тесты могут быть разработаны для оценки определенных компетенций.

Ограничения тестирования:

- Ограниченность охвата. Тесты, как правило, фокусируются на оценке отдельных навыков, не позволяя получить целостную картину компетенций кандидата.

- Сложность разработки. Создание качественных тестовых заданий требует значительных временных и финансовых затрат [1].

Таким образом, существующие методы оценки навыков соискателей имеют ряд существенных ограничений, связанных с субъективностью, неполнотой информации и отсутствием стандартизации. Данные недостатки создают потребность в разработке новых подходов, способных обеспечить более объективную и всестороннюю оценку профессиональных компетенций кандидатов. Это затрудняет объективное сравнение кандидатов и принятие обоснованных решений в процессе подбора персонала. Применение больших языковых моделей может стать перспективным решением для преодоления данных ограничений.

## 1.2 Нейронные сети

Нейронные сети - это передовой метод в области искусственного интеллекта, который стремится научить компьютеры обрабатывать информацию по аналогии с человеческим мозгом. Этот подход основан на концепции глубокого обучения, которая является разновидностью машинного обучения.

Основными элементами нейронной сети являются искусственные нейроны. Каждый нейрон получает входные сигналы, производит над ними некоторые вычисления и передает результат на следующие нейроны. Таким образом, информация распространяется по сети, преобразуясь на каждом слое.

Архитектура нейронной сети (Рисунок.1) обычно включает в себя следующие основные компоненты:

1. Входной слой - принимает исходные данные (например, пиксели изображения или текстовые признаки).
2. Скрытые слои - выполняют последовательные преобразования входных данных, извлекая все более абстрактные признаки и закономерности.
3. Выходной слой - генерирует результат решения задачи (например, классификацию объекта на изображении или прогнозируемое значение).

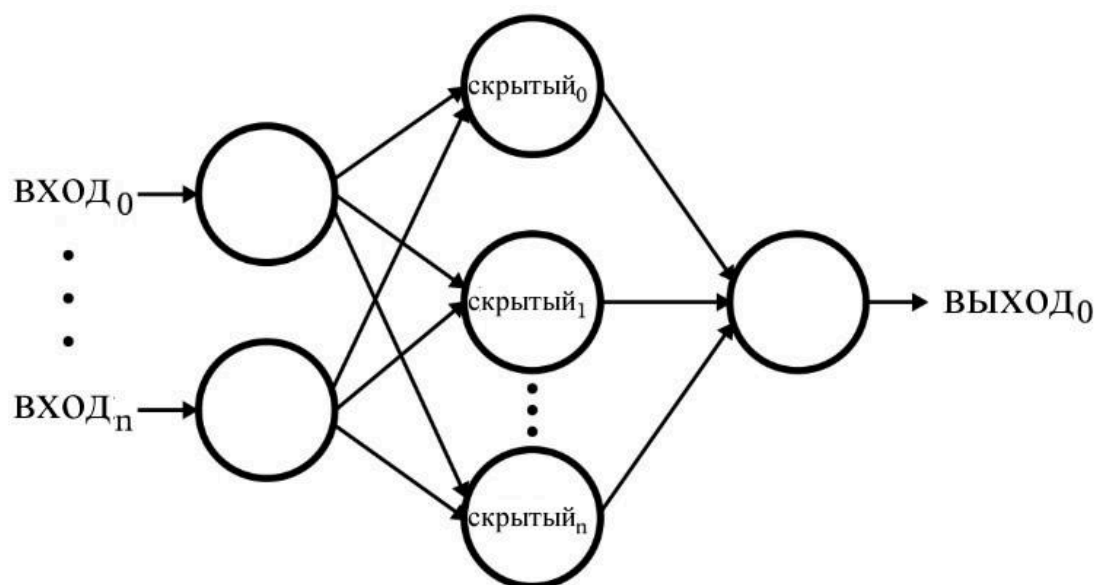


Рисунок1 - архитектура нейронных сетей

Между слоями нейроны соединены весовыми коэффициентами, которые определяют силу связи между ними. Процесс обучения нейронной сети заключается в итеративной настройке этих весовых коэффициентов на основе обучающих данных с использованием методов оптимизации, таких как градиентный спуск.

Ключевые свойства нейронных сетей:

1. Способность к обучению и адаптации. Нейронные сети могут самостоятельно обучаться на примерах, выявляя скрытые закономерности в данных.
2. Параллельная обработка информации. Нейроны работают одновременно, что позволяет эффективно обрабатывать большие объемы данных.
3. Устойчивость к шумам и ошибкам. Нейронные сети демонстрируют высокую робастность к неполным или искаженным входным данным.
4. Универсальность применения. Нейронные сети успешно применяются для решения широкого спектра задач: классификации, регрессии, распознавания образов, обработки естественного языка и многих других.

Современные нейронные сети, особенно глубокие архитектуры с множеством скрытых слоев, показывают выдающиеся результаты в решении сложных интеллектуальных задач, сопоставимые или превосходящие человеческие возможности.

Благодаря своей адаптивной природе, нейронные сети способны учиться на собственном опыте. Во время обучения сеть получает большое количество примеров и корректирует свои внутренние параметры таким образом, чтобы минимизировать ошибки и улучшить качество выходных данных. Этот процесс постоянного самосовершенствования позволяет нейронным сетям становиться все более точными и эффективными с течением времени[2-4].

Одним из ключевых преимуществ нейронных сетей является их способность решать сложные задачи, которые трудно поддаются традиционным

алгоритмическим подходам. Например, они могут использоваться для автоматического обобщения текстов, где сеть учится понимать смысл документа и генерировать краткое изложение его основных идей.

В рамках разработки методики стандартизации навыков в резюме на основе больших языковых моделей, важно рассмотреть основные архитектуры и типы нейронных сетей, которые могут быть эффективно применены для решения данной задачи.

### 1. Полносвязные нейронные сети (Feedforward Neural Networks)

1.1. Классическая архитектура, где нейроны каждого слоя связаны со всеми нейронами предыдущего слоя

1.2. Хорошо подходят для задач классификации и регрессии на структурированных данных

1.3. Могут быть использованы для классификации и кластеризации навыков, извлеченных из резюме

### 2. Сверточные нейронные сети (Convolutional Neural Networks, CNN)

2.1. Специализированы на обработке пространственно-структурированных данных, таких как изображения

2.2. Применяют операцию свертки для выявления локальных признаков

2.3. Могут быть использованы для анализа и структурирования текстовой информации в резюме

### 3. Рекуррентные нейронные сети (Recurrent Neural Networks, RNN)

3.1. Предназначены для обработки последовательных данных, таких как текст или речь

3.2. Используют внутреннее состояние для обработки элементов последовательности

3.3. Позволяют учитывать контекст при анализе навыков в резюме

### 4. Сети долго-краткосрочной памяти (Long Short-Term Memory, LSTM)

4.1.Разновидность RNN с улучшенной способностью к запоминанию длинных последовательностей

4.2.Эффективны для моделирования зависимостей в естественном языке

4.3.Могут применяться для извлечения и классификации навыков из текстов резюме

## 5.Трансформеры (Transformer Networks)

5.1.Инновационная архитектура, основанная на механизме внимания

5.2.Демонстрируют выдающуюся производительность в задачах обработки естественного языка

5.3.Являются основой для современных больших языковых моделей, таких как BERT, GPT-3, которые рассматриваются в качестве ключевого инструмента для стандартизации навыков в резюме

Среди перечисленных типов нейронных сетей, особое внимание в рамках данной дипломной работы уделяется трансформерам и большим языковым моделям на их основе. Данная архитектура показывает выдающиеся результаты в задачах понимания и обработки текстовой информации, что делает ее чрезвычайно перспективной для решения проблемы стандартизации навыков, представленных в резюме кандидатов[5].

## 1.3 Трансформеры и модели на их основе

В последнее время произошли значительные прорывы в области языкового моделирования, главным образом благодаря использованию трансформеров, увеличению вычислительных мощностей и доступности больших объемов обучающих данных. Эти достижения привели к революционным изменениям, позволив создавать большие языковые модели (БЯМ), которые способны достигать производительности, сопоставимой с человеческой, при решении различных задач.



Обработка естественного языка (ОЕЯ) прошла путь развития от статистического моделирования языка к нейронному, а затем от предварительно обученных языковых моделей к большим языковым моделям (БЯМ). Традиционное моделирование языка предполагает обучение моделей для конкретных задач с использованием обучения с учителем, в то время как предварительно обученные языковые модели обучаются самостоятельно на обширном массиве текстов, стремясь получить универсальное представление, применимое к различным задачам ОЕЯ. После тонкой настройки для решения прикладных задач предварительно обученные языковые модели демонстрируют более высокую производительность по сравнению с традиционным моделированием языка. Увеличение размера языковых моделей приводит к дальнейшему повышению производительности, что способствовало переходу от предварительно обученных языковых моделей к БЯМ за счет значительного увеличения параметров модели (десятки и сотни миллиардов) и объема обучающих данных (многие гигабайты и терабайты). Следуя этой тенденции, было разработано множество больших языковых моделей. Растущее количество выпущенных БЯМ в течение последних лет, представлены на графике(Рисунок.2):

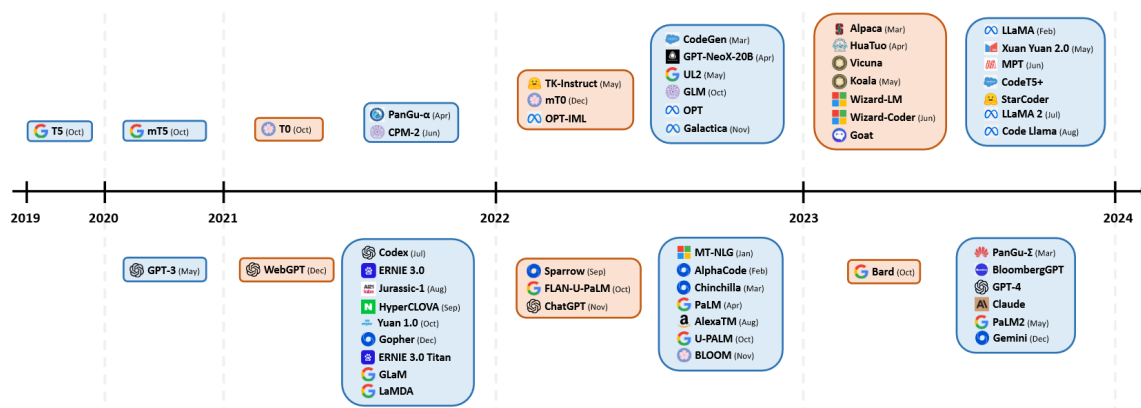


Рисунок.2 - количество выпущенных БЯМ в течение последних лет

Хронологическое отображение выпуска больших языковых моделей: синие карточки представляют «предварительно обученные» модели, а оранжевые карточки соответствуют моделям, «настроенным с помощью»

инструкций». Исходный код моделей, представленных в верхней половине - открыт, тогда как модели, находящиеся в нижней половине — с закрытым исходным кодом.

Ключевая идея трансформеров - использование механизма внимания (attention) для эффективной обработки последовательностей переменной длины.

Основные компоненты архитектуры трансформера(Рисунок.3):

1.Encoder (кодировщик): Принимает входную последовательность токенов (слов, частей слов или символов) и генерирует их векторные представления с учетом контекста. Состоит из нескольких идентичных слоев, каждый из которых содержит:

1.1.Multi-Head Self-Attention (многоголовое самовнимание): позволяет каждому токену "обращать внимание" на все остальные токены во входной последовательности, вычисляя взвешенную сумму их представлений. Это помогает уловить зависимости между удаленными токенами. "Многоголовость" означает параллельное использование нескольких механизмов внимания с разными обучаемыми весами.

1.2.Feed-Forward Network (полносвязная сеть прямого распространения): обрабатывает представления, полученные на выходе блока самовнимания, пропуская их через несколько слоев с нелинейными активациями.

1.3.Residual Connections (остаточные связи) и Layer Normalization (нормализация слоя): используются для стабилизации процесса обучения глубоких моделей.

2.Decoder (декодировщик): Принимает выходы кодировщика и генерирует выходную последовательность токенов. Также состоит из нескольких идентичных слоев, похожих на слои кодировщика, но с дополнительным блоком:

2.1.1.Masked Self-Attention (маскированное самовнимание): похоже на самовнимание в кодировщике, но каждый токен может обращать внимание

только на предшествующие ему токены, чтобы избежать утечки информации о будущих токенах при генерации текста.

2.1.2. Encoder-Decoder Attention (внимание кодировщик-декодировщик): позволяет каждому токenu в декодировщике обращать внимание на представления токенов кодировщика, что помогает генерировать выходные токены с учетом входного контекста.

3. Positional Encoding (позиционное кодирование): добавляет информацию о позиции каждого токена в последовательности, так как исходная архитектура трансформера не учитывает порядок токенов. Обычно используются синусоидальные функции или обучаемые векторы.

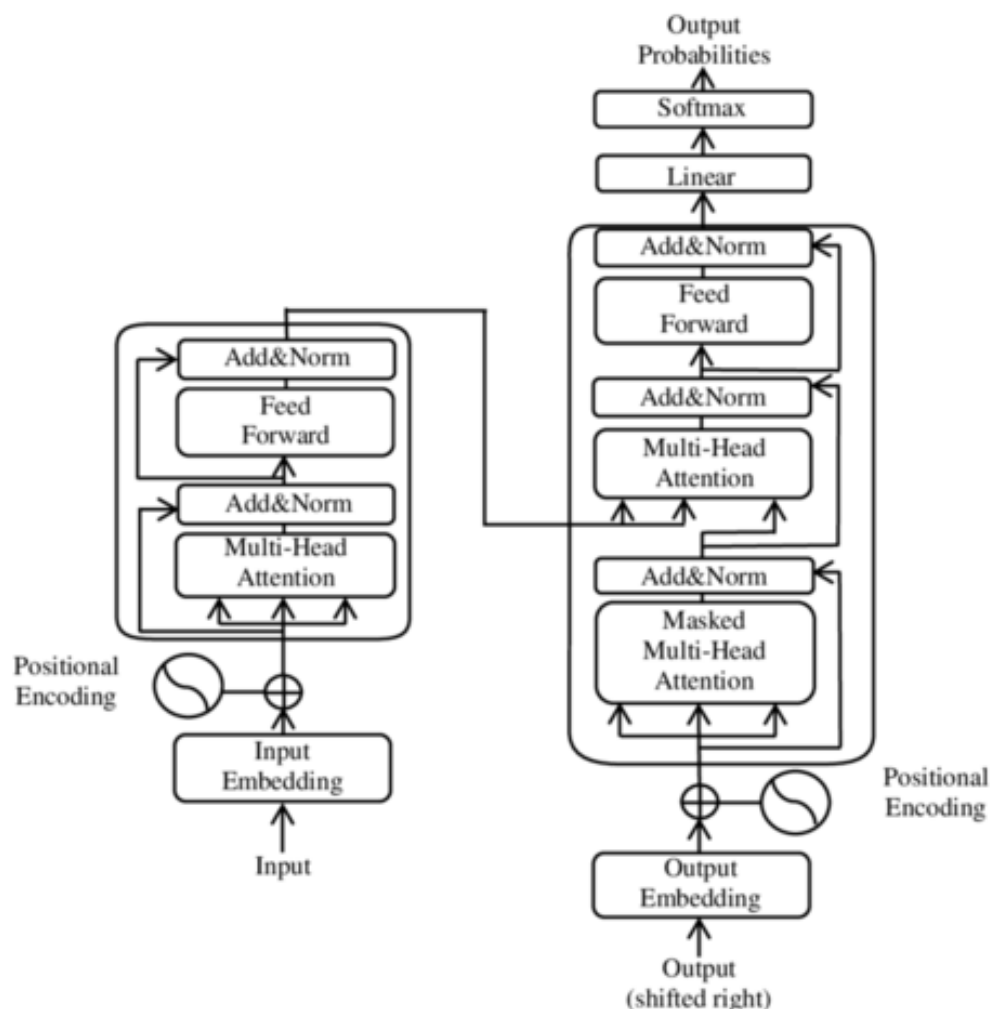


Рисунок.3 - архитектура трансформера

Процесс обучения трансформера обычно включает следующие этапы:

1. Токенизация входных текстов и их преобразование в числовые векторы (эмбеддинги).
2. Прямой проход через кодировщик для получения контекстных представлений токенов.
3. Автореферентный прямой проход через декодировщик для генерации выходных токенов один за другим.
4. Вычисление функции потерь (loss), сравнивающей сгенерированные токены с ожидаемыми.
5. Обратное распространение ошибки и обновление весов модели с помощью оптимизатора.

Благодаря механизму внимания и глубокой архитектуре трансформеры способны эффективно моделировать сложные зависимости в последовательностях и генерировать связный текст высокого качества.

Со времени своего появления в 2017 году трансформеры фактически стали стандартом для решения многих задач NLP, таких как:

1. Машинный перевод: переводы между языками.
2. Обобщение текста (summarization): генерация краткого изложения длинного текста.
3. Ответы на вопросы (question answering): поиск ответов на вопросы в большом корпусе текстов.
4. Генерация текста: создание связных текстов на заданную тему.
5. Извлечение именованных сущностей и отношений: поиск в тексте упоминаний людей, организаций, мест и т.д., а также связей между ними.

Были предложены различные модификации и усовершенствования базовой архитектуры трансформера, направленные на повышение ее эффективности, скорости обучения и работы, а также адаптацию к конкретным задачам:

1. BERT (Bidirectional Encoder Representations from Transformers): двунаправленная модель, обучаемая на задаче предсказания замаскированных токенов и определения связности предложений. Широко используется для трансферного обучения.

2. GPT (Generative Pre-trained Transformer): однонаправленная модель, обученная на задаче предсказания следующего токена в последовательности. Мощный инструмент для генерации текста.

3. T5 (Text-to-Text Transfer Transformer): унифицированная модель, которая рассматривает все задачи как преобразование текста в текст. Одна модель обучается решать широкий спектр задач.

4. ViT (Vision Transformer): адаптация трансформера для работы с изображениями. Разбивает изображение на патчи и обрабатывает их как последовательность.

5. LLaMA (Large Language Model Meta AI): семейство языковых моделей, разработанных компанией Meta AI (ранее Facebook AI Research). Эти модели обучены на огромных объемах текстовых данных и демонстрируют впечатляющие возможности генерации и понимания текста. Отличительные особенности LLaMA:

5.1. Эффективное обучение на большом количестве не размеченных текстовых данных.

5.2. Возможность генерации длинных последовательностей текста без потери связности и релевантности.

5.3. Высокая скорость работы за счет оптимизации архитектуры и процесса вывода.

5.4. Способность к few-shot обучению - адаптации к новым задачам по нескольким примерам.

5.5. Открытость исследовательского сообщества - модели LLaMA доступны для некоммерческого использования.

## 1.4 Квантизация

Квантизация - это процесс сжатия и уменьшения разрядности весов и активаций нейронной сети, что позволяет значительно сократить размер модели и снизить требования к вычислительным ресурсам без существенной потери в качестве.

Основные цели применения квантизации:

### 1. Уменьшение размера модели:

Большие языковые модели, могут занимать десятки и сотни гигабайт памяти, что затрудняет их развертывание и использование, особенно в средах с ограниченными ресурсами. Квантизация позволяет сжать модель, сократив размер весов и активаций, что делает ее более компактной и эффективной в использовании памяти.

### 2. Повышение производительности:

Операции с числами меньшей разрядности (например, 8-битными вместо 32-битных) выполняются значительно быстрее на большинстве аппаратных платформ. Применение квантизации позволяет ускорить вывод прогнозов модели, что важно для реального времени или приложений с высокой нагрузкой, это позволит использовать менее мощное оборудование.

Таким образом, квантизация в сочетании с другими методами, такими как файнтюнинг, позволяет значительно улучшить эффективность больших языковых моделей, что особенно важно для таких ресурсоемких задач, как стандартизация навыков в резюме кандидата. Без квантизации обучение и вывод такой модели были бы гораздо более затратными и долгими, а возможно и вовсе невозможными на имеющемся оборудовании.

## 1.5 Задача на обобщение текста

Обобщение текста это процесс автоматического создания краткого содержательного саммари (summary) из более объемного текстового документа или набора документов. Цель суммаризации - выделить ключевую информацию из первоисточника, отбросив второстепенные детали.

В сфере обработки естественного языка (ОЕЯ) выделяют два основных вида суммаризации:

### 1. Экстрактивная суммаризация (Extractive Summarization):

При этом подходе алгоритм анализирует текст, оценивает важность каждого предложения на основе различных признаков (частота ключевых слов, позиция в тексте и др.) и выбирает наиболее значимые предложения для включения в итоговое саммари. То есть в итоге выводится подмножество предложений из оригинального документа.

### 2. Абстрактивная суммаризация (Abstractive Summarization):

В отличие от экстрактивного подхода, при абстрактивной суммаризации алгоритм генерирует совершенно новый текст, который по смыслу передает основное содержание первоисточника. Абстрактивное саммари может включать фразы и предложения, которых не было в исходном тексте. Это более сложная задача, требующая глубокого понимания текста и способности к генерации осмысленного контента.

Для реализации абстрактивной суммаризации обычно используют нейросетевые модели типа Sequence-to-Sequence (например, архитектуры на базе Transformer), обученные на больших парах текст-саммари.

Для моего дипломного проекта по стандартизации навыков в резюме я выбрала абстрактивную суммаризацию по причине того, что описание навыков в резюме - это, как правило, несколько предложений в свободной форме. Чтобы выделить из них стандартизированный список навыков, нужно не просто выбрать ключевые предложения (как в экстрактивной суммаризации), а

обобщить и сформулировать их в виде стандартизированных формулировок навыков.



## **Глава №2 Большие языковые модели, их сравнение и выбор наиболее подходящей в рамках данного исследования.**

### **2.1 Сравнение и выбор модели**

В рамках данного дипломного проекта я рассматривала различные варианты моделей, включая BERT, T5, ViT и LLaMa 2. В итоге, для решения поставленной задачи я выбрала LLaMa 2 7b по следующим причинам:

#### **1. Оптимальный баланс производительности и доступности:**

- BERT, T5: Хотя эти модели отлично подходят для задач NLP, они требуют значительных вычислительных ресурсов, что ограничивает их доступность для индивидуальной разработки и экспериментов.
- ViT: Будучи моделью компьютерного зрения, ViT не подходит для обработки текстовой информации, содержащейся в резюме.
- LLaMa 2 7b: Эта модель предлагает хороший баланс между производительностью и эффективностью. Она достаточно мощная для обработки естественного языка и в то же время доступна для использования на стандартном оборудовании.

#### **2. Открытый исходный код и лицензия:**

- LLaMa 2: Является моделью с открытым исходным кодом, доступной для коммерческого использования. Это даёт мне больше свободы в использовании, модификации и интеграции модели в разрабатываемое приложение.

#### **3. Возможность дообучения и специализации:**

LLaMa 2: Может быть легко дообучена на специализированном наборе данных резюме, что позволит улучшить её точность и релевантность результатов.

LLaMa 2 7b предлагает оптимальное сочетание производительности, доступности, открытости и возможностей для дообучения, что делает её идеальным выбором для моего дипломного проекта по стандартизации навыков в резюме с использованием больших языковых моделей.

## **2.2 Метод дообучения модели**

Fine-tuning (дообучение) - это процесс адаптации предобученной языковой модели под конкретную задачу путем дополнительного обучения на специализированном наборе данных. Большие языковые модели, такие как GPT, BERT, T5 и другие, предварительно обучаются на огромных массивах текстовых данных, что позволяет им приобрести общие знания о языке, контексте и различных предметных областях. Однако для эффективного решения конкретных задач, например, стандартизации навыков в резюме, одного предобучения недостаточно.

Основные причины, по которым необходимо проводить fine-tuning:

1. Адаптация к специфике задачи: Дообучение позволяет настроить модель на особенности конкретной задачи, в данном случае - на анализ и стандартизацию навыков в резюме. Модель учится понимать структуру резюме, распознавать релевантные навыки и приводить их к единому стандарту.
2. Повышение точности: Fine-tuning на специализированном наборе данных (например, на коллекции реальных резюме с размеченными навыками) позволяет значительно повысить точность модели в решении поставленной задачи по сравнению с использованием только предобученной модели.
3. Учет особенностей предметной области: Каждая предметная область имеет свою специфическую терминологию и контекст. Дообучение позволяет модели лучше понимать и интерпретировать термины и концепции, характерные для сферы HR и рекрутинга.

4. Снижение требований к объему данных: Благодаря трансферу знаний от предобученной модели, fine-tuning можно проводить на относительно небольших наборах данных, что особенно важно, когда доступ к большим объемам размеченных данных ограничен

Для адаптации большой языковой модели под задачу стандартизации навыков в резюме кандидата необходимо провести дообучение (файн-тюнинг) на специализированном датасете. Процесс адаптации можно разбить на следующие шаги:

1. Подготовка датасета:

- Собрать большое количество реальных резюме кандидатов из различных сфер деятельности.

- Разметить навыки в каждом резюме, приведя их к стандартизированным формулировкам. Это можно сделать вручную с привлечением экспертов в HR или попытаться автоматизировать с помощью готовых словарей навыков.

- Разбить размеченные резюме на обучающую и тестовую выборки.

2. Выбор архитектуры и претренированной языковой модели:

- Подойдут модели, обученные на большом объеме текстовых данных, например BERT, GPT, T5, LLaMA.

- Размер модели зависит от объема данных для обучения и вычислительных ресурсов. Можно начать с небольшой модели и при необходимости увеличить.

3. Файн-тюнинг модели:

- Добавить к языковой модели слои для решения задачи извлечения именованных сущностей (навыков). Обычно это линейный слой и функция активации.

- Инициализировать веса добавленных слоев, а веса языковой модели использовать из претренированной модели.

- Обучать модель на размеченных резюме из обучающей выборки, используя кросс-энтропию как функцию потерь. Варьировать гиперпараметры обучения.

4. Оценка качества:

- Проверить модель на тестовой выборке, используя метрики.
- Проанализировать ошибки модели, при необходимости дообучить на дополнительных данных.

Процесс адаптации требует значительных ресурсов для сбора и разметки данных, вычислений и тестирования. Но это позволяет получить модель, которая будет выдавать релевантные результаты для конкретной бизнес-задачи. Использование трансферного обучения на базе больших языковых моделей ускорит достижение приемлемого качества по сравнению с обучением модели с нуля.

## **2.3 Выбор датасета**

При выборе датасета для переобучения модели суммаризации текста нужно учитывать несколько важных факторов:

1. Соответствие предметной области: Датасет должен содержать тексты, относящиеся к той же предметной области, что и тексты, которые модель будет обрабатывать в реальном применении. Это может быть новостные статьи, научные публикации, технические руководства и т.д. Несоответствие предметной области может привести к ухудшению качества суммаризации.

2. Размер и репрезентативность: Датасет должен быть достаточно большим, чтобы модель могла эффективно обучиться на нем. Кроме того, он должен быть репрезентативным, то есть охватывать разнообразные типы текстов, стилей, тематик в пределах выбранной предметной области.

3. Наличие эталонных суммариев: Для использования методов оценки, основанных на сравнении с человеческими суммариями (ROUGE, BLEU и др.), необходимо, чтобы в датасете присутствовали не только исходные тексты, но и их ручные суммарии, написанные экспертами.

4. Качество аннотаций: Эталонные суммарии в датасете должны быть высокого качества, чтобы служить надежным ориентиром для оценки модели. Плохое качество аннотаций исказит результаты оценки.

5. Доступность и лицензирование: Важно, чтобы датасет был свободно доступен для использования в исследовательских и коммерческих целях. Лицензионные ограничения могут помешать применению датасета.

6. Актуальность: Для обучения модели, которая будет работать с современными текстами, желательно использовать относительно свежие данные, отражающие актуальные тенденции в языке и тематике.

## **2.4 Метрики оценки качества**

Оценка качества модели суммаризации текста является крайне важной задачей по нескольким ключевым причинам:

1. Объективность и сравнимость: Без использования формальных метрик и методов оценки, качество модели можно оценить только субъективно. Это затрудняет сравнение различных моделей между собой и отслеживание прогресса в улучшении модели. Метрики позволяют получить объективные и сопоставимые показатели производительности.

2. Выявление сильных и слабых сторон: Анализ результатов оценки по различным метрикам помогает понять, в каких аспектах модель работает хорошо, а в каких - требует улучшения. Это дает ценные insights для дальнейшей оптимизации и доработки модели.

3. Соответствие требованиям: Модели суммаризации должны удовлетворять определенным требованиям, чтобы быть полезными в реальных приложениях. Метрики позволяют проверить, насколько хорошо модель справляется с задачами, важными для конечных пользователей, таких как сохранение ключевой информации, связность текста, лаконичность.

4. Воспроизводимость и прозрачность: Использование стандартных метрик делает процесс оценки моделей прозрачным и воспроизводимым. Это важно для научной работы, публикаций и обмена результатами в исследовательском сообществе.

Таким образом, комплексная оценка качества модели суммаризации с использованием разнообразных метрик является ключом к созданию эффективных и практически полезных решений в этой области. Она позволяет обеспечить высокое качество, соответствие требованиям и возможность дальнейшего совершенствования модели.

Метрики оценки качества для моделей выполняющих задачу по обобщению текста:

1. ROUGE (Recall-Oriented Understudy for Gisting Evaluation):

$$ROUGE - N(A_i) = \frac{\sum_{M_y} count(Ngram(A_i) \cap Ngram(M_{ij}))}{\sum_{M_y} count(Ngram(M_{ij}))} \quad (1)$$

Где:

$A_i$  – оцениваемая обзорная аннотация i-того кластера.

$M_{ij}$  – ручные аннотации i-того кластера.

Ngram (D) – множество всех n-грамм из лемм соответствующего документа D.

1.1. ROUGE-N (N=1,2,3,4)(1): Эта метрика оценивает n-граммное перекрытие между сгенерированным резюме и эталонными резюме. Она рассчитывается как доля n-грамм в сгенерированном резюме, которые

присутствуют в эталонных резюме. Более высокие значения ROUGE-N указывают на большее лексическое сходство.

1.2. ROUGE-L: Эта метрика оценивает наибольшую общую последовательность (LCS) между сгенерированным и эталонным резюме. Она отражает насколько хорошо сгенерированное резюме может воспроизвести последовательность слов из эталонного.

1.3. ROUGE-SU4: Эта метрика оценивает перекрытие пар слов с учетом промежуточных слов. Она более гибкая, чем ROUGE-N, и позволяет учитывать синонимы и парафразы.

2. BLEU (Bilingual Evaluation Understudy)(2):

$$BLEU = brevity\ penalty * \left( \prod_{i=1}^n precision_i \right)^{\frac{1}{n}} * 100\% \quad (2)$$

$$\text{Где } brevity\ penalty = \min\left(1, \frac{output\ length}{reference\ length}\right) \quad (3)$$

BLEU оценивает n-граммное перекрытие между сгенерированным резюме и эталонными, но с учетом как точности (precision), так и полноты (recall). Она рассчитывается как взвешенная геометрическая средняя прецизионности n-грамм, с штрафом за краткость сгенерированного текста. Более высокие значения BLEU указывают на большее сходство сгенерированного и эталонного текстов.

3. METEOR (Metric for Evaluation of Translation with Explicit Ordering)(4):

$$METEOR = F_{mean} * (1 - p) \quad (4)$$

Где:

$$\text{Гармоническое среднее точности и полноты} - F_{mean} = \frac{\alpha * \beta * P * R}{\alpha * R + \beta * P}; \quad (5)$$

$$\text{Точность} - P = \frac{m}{w_t}, \quad (6)$$

$m$  - число N-грамм, которые присутствуют как в идеальном резюме, так и в сгенерированном,

$w_t$  - общее число N-грамм в сгенерированном резюме,

$$\text{Полнота} - R = \frac{m}{w_r}, \quad (7)$$

$w_r$  - общее число N-грамм в идеальном резюме;

$$\text{Оценка штрафа} - p = 0,5 * \left(\frac{c}{u_m}\right)^3, \quad (8)$$

Где:

$c$  - несколько юни-грамм, которые соответствуют друг другу в произведенном и идеальном текстах (чем больше соответствий в текстах, тем меньше будет количество таких отрезков, а следовательно, и значение  $c$ );

$u_m$  - количество юни-грамм, которым нашлось соответствие;

METEOR учитывает синонимию, парафразы и стемминг при сравнении сгенерированного и эталонного резюме. Она основана на гибком сопоставлении лексем и учитывает как точность, так и полноту. Более высокие значения METEOR указывают на большее семантическое сходство.

4. BERTScore(9) использует контекстуализированные эмбединги токенов предобученной модели BERT. Она вычисляет семантическую близость двух предложений, суммируя косинусную близость между эмбедингами их токенов. Далее вычисляется F1 мера по следующим формулам:

$$F_{BERT} = 2 \frac{P_{BERT} * R_{BERT}}{P_{BERT} + R_{BERT}}, \quad (9)$$

$$R_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^T \hat{x}_j, \quad (10)$$

$$P_{BERT} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^T \hat{x}_j, \quad (11)$$

Где  $x$  - это эмбединги предсказания, а  $\hat{x}$  - эмбединги эталонного перефразирования. Чем больше метрика, тем лучше качество



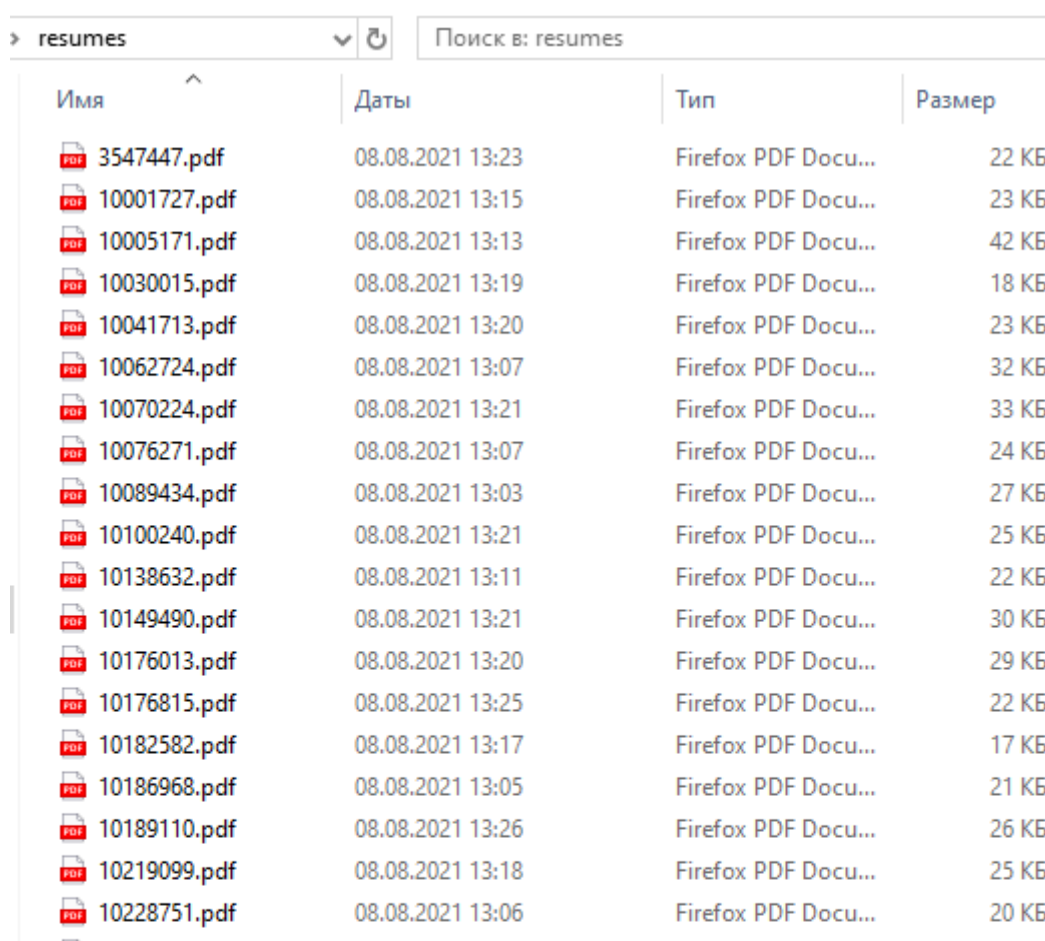
BERTScore использует предобученную модель BERT для вычисления семантического сходства между сгенерированным и эталонным резюме на уровне токенов. Она учитывает контекстуальные значения слов, что позволяет лучше оценить смысловое сходство текстов.

Среди всех рассмотренных метрик оценки качества генерируемого текста, метрика BERTScore демонстрирует наиболее тесную корреляцию с экспертной оценкой, выполненной человеком.

## Глава №3 Разработка метода стандартизации навыков на основе больших языковых моделей.

### 3.1 Формирование датасета

Для выполнения работы мне было необходимо собрать подходящий набор данных. Я обратилась к платформе Kaggle, где нашла готовый набор резюме в формате .docx и .pdf(Рисунок.4). Чтобы использовать эти данные в своем исследовании, я преобразовала их в более удобный формат .csv. Это позволило мне сформировать требуемый датасет и приступить к дообучению и тестированию своей модели по стандартизации навыков, извлекаемых из резюме кандидатов.



Имя	Даты	Тип	Размер
PDF 3547447.pdf	08.08.2021 13:23	Firefox PDF Docu...	22 КБ
PDF 10001727.pdf	08.08.2021 13:15	Firefox PDF Docu...	23 КБ
PDF 10005171.pdf	08.08.2021 13:13	Firefox PDF Docu...	42 КБ
PDF 10030015.pdf	08.08.2021 13:19	Firefox PDF Docu...	18 КБ
PDF 10041713.pdf	08.08.2021 13:20	Firefox PDF Docu...	23 КБ
PDF 10062724.pdf	08.08.2021 13:07	Firefox PDF Docu...	32 КБ
PDF 10070224.pdf	08.08.2021 13:21	Firefox PDF Docu...	33 КБ
PDF 10076271.pdf	08.08.2021 13:07	Firefox PDF Docu...	24 КБ
PDF 10089434.pdf	08.08.2021 13:03	Firefox PDF Docu...	27 КБ
PDF 10100240.pdf	08.08.2021 13:21	Firefox PDF Docu...	25 КБ
PDF 10138632.pdf	08.08.2021 13:11	Firefox PDF Docu...	22 КБ
PDF 10149490.pdf	08.08.2021 13:21	Firefox PDF Docu...	30 КБ
PDF 10176013.pdf	08.08.2021 13:20	Firefox PDF Docu...	29 КБ
PDF 10176815.pdf	08.08.2021 13:25	Firefox PDF Docu...	22 КБ
PDF 10182582.pdf	08.08.2021 13:17	Firefox PDF Docu...	17 КБ
PDF 10186968.pdf	08.08.2021 13:05	Firefox PDF Docu...	21 КБ
PDF 10189110.pdf	08.08.2021 13:26	Firefox PDF Docu...	26 КБ
PDF 10219099.pdf	08.08.2021 13:18	Firefox PDF Docu...	25 КБ
PDF 10228751.pdf	08.08.2021 13:06	Firefox PDF Docu...	20 КБ
PDF .....	-----	-----	-----

Рисунок.4 - Найденный на Kaggle датасет

Чтобы преобразовать данные в формат, удобный для дообучения модели, нужно было сделать следующее:

Выделить важную информацию, на основе которой будет происходить обучение, разбив данные на группы:

- ключевая часть резюме,
- информация об образовании,
- ключевые навыки соискателя, с помощью которых в рамках обучения с учителем будет происходить оценка правильного ответа модели. Эта колонка со скиллами кандидатов станет основой для дальнейшего обучения и тестирования модели стандартизации навыков резюме.

Для выделения важной части текста резюме на которой будет происходить обучение я использовала токенайзер из библиотеки `nlk`. С помощью неё получим представление лейблов по тексту, а именно лейблы “Organisation” и “Person”. Данные, размеченные данными лейблами относятся к составителю резюме и, соответственно, будут включать в себя такие сведения о человеке как: личные данные, хобби, навыки, образование и прочее. Из них модель должна будет научиться формировать навыки, относящиеся к работе.

Сами навыки для верификации выбрала следующим образом:

С платформы Kaggle я забрала файл со списком ключевых навыков, относящихся к различным профессиям. Текущие данные будут служить для выделения столбца с ключевыми навыками из резюме.

Проанализировав вручную данные из скачанных резюме, я заметила, что люди зачастую указывали ключевые навыки в блоках с такими ключевыми словами: 'highlights', 'skills', 'languages', 'environment'.

Столбец с компетенциями кандидата заполнялся следующим образом:

Был написан код, который анализировал следующие двадцать слов в каждом из упомянутых выше блоков.

Каждое слово сверялось с файлом, в котором находился готовый список навыков по разным профессиям, в случае совпадения - навык добавлялся в колонку 'Skills' в .csv файле.

Если упомянутых блоков в резюме было несколько - навыки добавлялись из каждого, так как человек мог иметь компетенции в разных сферах.

Постобработка файла, создаваемого для дообучения модели происходила следующим образом:

Для того чтобы убрать все символы, которые не могут быть прочитаны через библиотеку Pandas строки были закодированы по стандарту “ascii”, а потом декодированы обратно. Данная процедура позволяет избавиться от нечитаемых спец. символов.

Пройдя по всем строкам, выкинула данные, в которых ключевые поля 'Skills' или 'Context' были пустыми, так как это было бы неэффективно при дообучении модели. Потери данных для обучения были незначительными.

Последний спец. столбец 'Text', был собран из текущих столбцов 'Skills' и 'Context' с добавлением промта(‘###Instruction’, ‘###Skills’, ‘###Resume’).

Для более удобного использования собранного датасета, я использовала платформу HuggingFace(Рисунок.5), а именно - загрузила туда файл и далее по необходимости подгружала его с помощью библиотеки “datasets”.

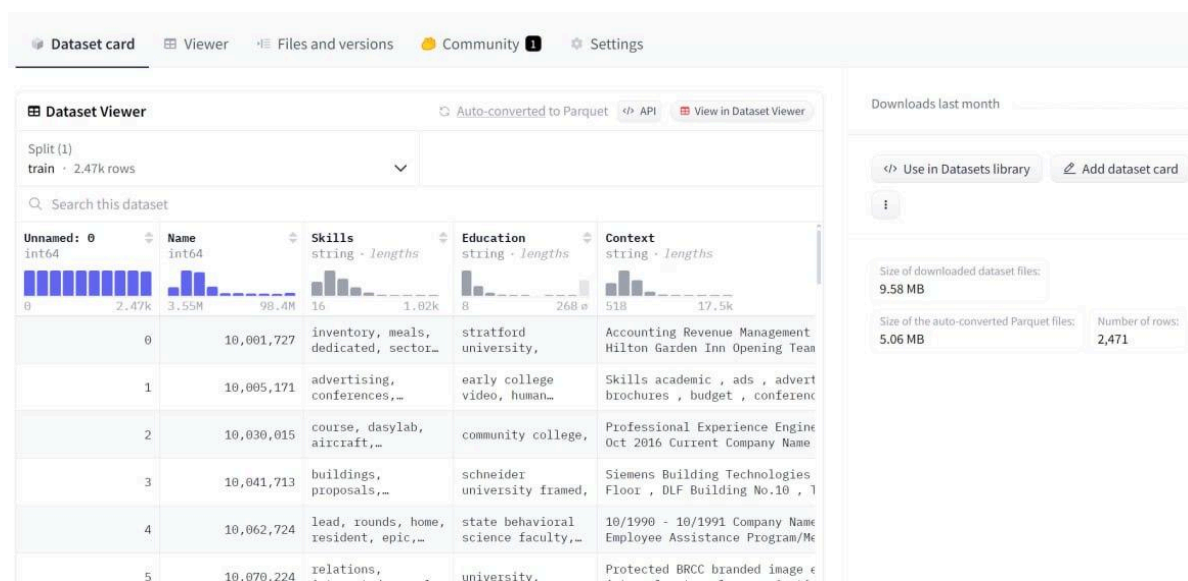


Рисунок.5 - Датасет, выложенный на HuggingFace.

Процесс сбора и подготовки данных был важным первым шагом в моей дипломной работе. Имея этот набор резюме в структурированном формате, я смогла приступить к основной части исследования, связанной с разработкой и оценкой методов стандартизации навыков на основе больших языковых моделей.

### **3.2 Выбор кластера для обучения модели**

В начале исследования был использован GoogleColab так как он ежедневно предоставляет ограниченное количество ресурсов для работы. При работе с ним, была использована среда Jupyter notebook. При использовании бесплатного аккаунта в день можно было использовать только 15 гб GPU RAM, которых хватало для обучения только с использованием одной эпохи, после этого ресурсов не хватало для использования обученной модели.

Данная проблема была решена с помощью смены кластера, я перешла на Yandex Cloud. По причине того, что тут другой подход к использованию ресурсов - они оплачиваются по часам, в отличие от Google Cloud, здесь нет бесплатного аккаунта, зато предоставляется грант на 3000 рублей, который можно использовать на любые ресурсы. Для исследования была арендована виртуальная машина, память которой я расширила до 40 гб, так как было необходимо пространство для хранения языковых моделей, их токенизаторов, а так же датасета, подгружаемого с HuggingFace. Средой разработки в данном случае выступала Jupyter Lab(Рисунок.6) с 1 GPU V100 8vCPUs от 48 до 96 гб RAM.

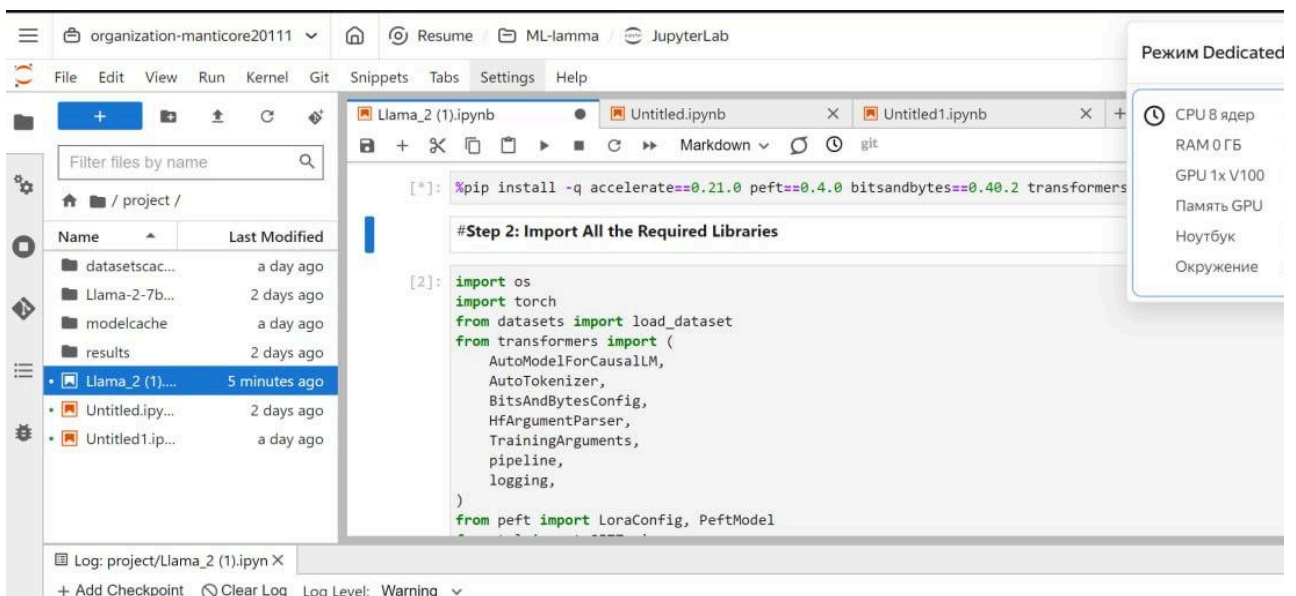


Рисунок.6 - Среда разработки.

### 3.3 Дообучение модели

Был написан скрипт для обучения языковой модели на основе архитектуры трансформера с использованием библиотек Hugging Face Transformers, Peft и Bitsandbytes.

В начале, устанавливаются необходимые пакеты с помощью `pip`. Затем импортируются необходимые библиотеки и модули. Для `datasets` версия не указывается потому что таким образом будет использована последняя не конфликтующая с другими библиотеками версия.

1. Импортируется функция `load_dataset` из библиотеки `datasets`. Эта функция позволяет загрузить готовый датасет из репозитория Hugging Face.
2. Импортируется класс `Dataset` из библиотеки `datasets`. Этот класс представляет собой абстрактный класс для работы с датасетами.
3. Загружается датасет "Resume-labels" из репозитория на Hugging Face.

4. Разделяется набор данных на три поднабора: тренировочный (train\_dataset) и валидационный (validation\_dataset) и тестовый (test\_dataset). В процентном соотношении - 70/20/10.

5. Создается словарь DatasetDict, который хранит три датасета: тренировочный, валидационный и тестовый.

В целом, этот код загружает датасет "Resume-labels" из репозитория на Hugging Face, разделяет его на тренировочный, валидационный и тестовый наборы данных, и создает словарь DatasetDict для хранения этих датасетов.

Далее задаются параметры для обучения модели.

1. Определяется имя модели, которую необходимо обучить. В этом случае это модель LLaMA-2-7b-chat-hf из репозитория NousResearch на Hugging Face.

Параметры QLoRA

2. lora\_r = 64: Определяется размерность внимания LoRA.  
3. lora\_alpha = 16: Определяется параметр масштабирования LoRA.  
4. lora\_dropout = 0.1: Определяется вероятность выброса для слоев LoRA.

Вероятность выброса (dropout probability) в слоях LoRA используется для регуляризации и предотвращения переобучения модели при тонкой настройке.

Во время обучения с LoRA существует риск переобучения, когда модель слишком сильно подстраивается под обучающие данные и теряет способность обобщать на новых данных. Для борьбы с этим явлением используется регуляризация, одним из методов которой является dropout.

Dropout случайным образом "выбрасывает" (отключает) некоторую долю нейронов во время обучения. Это вынуждает остальные нейроны более сильно адаптироваться, что препятствует соадаптации нейронов и переобучению. Параметр вероятности выброса (dropout probability) определяет, какая доля нейронов будет отключена на каждой итерации обучения.

Таким образом, использование dropout с подходящей вероятностью выброса в слоях LoRA помогает предотвратить переобучение и улучшить обобщающую способность модели на новых данных после тонкой настройки.

#### Параметры bitsandbytes

5. `use_4bit = True`: Активируется загрузка базовой модели с 4-битной точностью.
6. `bnb_4bit_compute_dtype = "float16"`: Определяется тип вычислений для 4-битных базовых моделей.
7. `bnb_4bit_quant_type = "nf4"`: Определяется тип квантизации для 4-битных базовых моделей.
8. `use_nested_quant = False`: Не активируется вложенная квантизация для 4-битных базовых моделей.

#### Параметры TrainingArguments

9. `output_dir = "./results"`: Определяется директория для хранения предсказаний модели и контрольных точек.
10. `num_train_epochs = 5`: Определяется количество эпох для обучения модели.
11. `fp16 = False` и `bf16 = False`: Не активируется обучение с 16-битной или 32-битной точностью.
12. `per_device_train_batch_size = 4` и `per_device_eval_batch_size = 4`: Определяется размер пакета для обучения и оценки на каждом устройстве.
13. `gradient_accumulation_steps = 1`: Определяется количество шагов для накопления градиентов.
14. `gradient_checkpointing = True`: Активируются контрольные точки градиентов.
15. `max_grad_norm = 0.3`: Определяется максимальная норма градиента.
16. `learning_rate = 2e-4`: Определяется начальная скорость обучения.
17. `weight_decay = 0.001`: Определяется коэффициент регуляризации весов.



- 18. `optim = "paged_adamw_32bit"`: Определяется оптимизатор для обучения.
- 19. `lr_scheduler_type = "cosine"`: Определяется тип графика изменения скорости обучения.
- 20. `max_steps = -1`: Не ограничивается количество шагов обучения.
- 21. `warmup_ratio = 0.03`: Определяется соотношение шагов для линейного нагрева.
- 22. `group_by_length = True`: Активируется группировка последовательностей по длине для экономии памяти и ускорения обучения.
- 23. `save_steps = 0` и `logging_steps = 25`: Определяются шаги для сохранения контрольных точек и логирования.

#### Параметры SFT

- 24. `max_seq_length = None`: Не ограничивается максимальная длина последовательности.
- 25. `packing = False`: Не активируется упаковка нескольких коротких примеров в одну входную последовательность для увеличения эффективности.
- 26. `device_map = {"": 0}`: Определяется карта устройств для загрузки модели (`device map`). В этом случае модель загружается на устройство GPU 0.

Далее загружается токенайзер и базовая модель с использованием конфигурации QLoRA. Если использовать 4-битную точность, то задается тип вычислений и конфигурация квантизации. Затем проверяю, поддерживает ли GPU формат `bfloat16`, и вывожу соответствующее сообщение.

Затем загружается конфигурация LoRA и задаются параметры обучения. Затем создается объект `SFTTrainer` для обучения модели.

Далее происходит обучение в пять эпох с 2156 итерациями(Рисунок.7 - Рисунок.8).

```
# Set supervised fine-tuning parameters
trainer = SFTTrainer(
    model=model,
    train_dataset=dataset["train"],
    eval_dataset=dataset["val"],
    peft_config=peft_config,
    dataset_text_field="Text",
    max_seq_length=max_seq_length,
    tokenizer=tokenizer,
    args=training_arguments,
    packing=packing,
)

# Train model
trainer.train()

Last executed at 2024-06-10 02:52:46 in 2h 8m 8s
```

```
/home/jupyter/.local/lib/python3.10/site-packages/peft/utils/other.py:102: FutureWarning:
    warnings.warn(
/home/jupyter/.local/lib/python3.10/site-packages/trl/trainer/sft_trainer.py:159: UserWarning:
    warnings.warn(
Map: 100%|██████████| 1729/1729 [00:01<00:00, 884.71 examples/s]
Map: 100%|██████████| 520/520 [00:00<00:00, 925.65 examples/s]
 0%|          | 0/2165 [00:00<?, ?it/s]You're using a LlamaTokenizerFast tokenizer.
call to the `pad` method to get a padded encoding.
 1%|          | 25/2165 [01:49<2:20:58, 3.95s/it]
{'loss': 2.9761, 'learning_rate': 7.692307692307693e-05, 'epoch': 0.06}
 2%|          | 50/2165 [02:54<1:00:24, 1.71s/it]
{'loss': 2.6452, 'learning_rate': 0.00015384615384615385, 'epoch': 0.12}
 3%|          | 75/2165 [04:44<2:19:25, 4.00s/it]
{'loss': 2.6221, 'learning_rate': 0.00019998881018102737, 'epoch': 0.17}
 5%|          | 100/2165 [05:50<57:49, 1.68s/it]
{'loss': 2.2813, 'learning_rate': 0.0001998629534754574, 'epoch': 0.23}
 6%|          | 125/2165 [07:40<2:14:37, 3.96s/it]
{'loss': 2.5515, 'learning_rate': 0.00019959742939952392, 'epoch': 0.29}
 7%|          | 150/2165 [08:44<51:59, 1.55s/it]
{'loss': 2.2369, 'learning_rate': 0.00019919260931265664, 'epoch': 0.35}
 8%|          | 175/2165 [10:37<2:23:32, 4.33s/it]
{'loss': 2.4297, 'learning_rate': 0.00019864905939235214, 'epoch': 0.4}
 9%|          | 200/2165 [11:47<58:21, 1.78s/it]
```

Рисунок.7 - обучение модели

```
95%|██████████| 2050/2165 [2:01:22<08:41, 4.54s/it]
{'loss': 2.1193, 'learning_rate': 1.4762346775940793e-06, 'epoch': 4.73}
96%|██████████| 2075/2165 [2:02:47<03:45, 2.51s/it]
{'loss': 1.9842, 'learning_rate': 9.0502382320653e-07, 'epoch': 4.79}
97%|██████████| 2100/2165 [2:04:20<04:55, 4.54s/it]
{'loss': 2.1627, 'learning_rate': 4.7240625348735633e-07, 'epoch': 4.85}
98%|██████████| 2125/2165 [2:05:45<01:39, 2.48s/it]
{'loss': 2.0308, 'learning_rate': 1.7898702322648453e-07, 'epoch': 4.91}
99%|██████████| 2150/2165 [2:07:14<00:56, 3.79s/it]
{'loss': 2.0602, 'learning_rate': 2.5176505749346936e-08, 'epoch': 4.97}
100%|██████████| 2165/2165 [2:07:51<00:00, 3.54s/it]
{'train_runtime': 7671.0706, 'train_samples_per_second': 1.127, 'train_steps_per_second': 0.282, 'train_loss': 2.1932825128145637, 'epoch': 5.0}

TrainOutput(global_step=2165, training_loss=2.1932825128145637, metrics={'train_runtime': 7671.0706, 'train_samples_per_second': 1.127, 'train_steps_per_second': 0.282})

# Save trained model
trainer.model.save_pretrained(new_model)

Last executed at 2024-06-10 02:52:47 in 1.03s
```

Рисунок.8 - обучение модели

После обучения модели она сохраняется в формате FP16, а затем загружается в объект PeftModel и объединяется с базовой моделью. Затем токенайзер сохраняется отдельно.

Сохраняется модель на Hugging Face(Рисунок.9) для того, чтобы иметь возможность использовать модель с других устройств в дальнейшем.

```
[15]: model.push_to_hub("Llama-2-7b-chat-resume-epoches", check_pr=True, use_auth_token="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")

tokenizer.push_to_hub("Llama-2-7b-chat-resume-epoches", check_pr=True, use_auth_token="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")

Last executed at 2024-06-10 03:07:33 in 1m 22.75s

tokenizer.model: 100%|██████████| 500k/500k [00:00<00:00, 672kB/s]

[15]: CommitInfo(commit_url='https://huggingface.co/Tiger20111/Llama-2-7b-chat-resume-epoches/commit/20410af1d9c31cd47dff227985936cc00e59474c', commit_message='Upload tokenizer', commit_description='', oid='20410af1d9c31cd47dff227985936cc00e59474c', pr_url=None, pr_revision=None, pr_num=None)
```

Рисунок.9 - сохранение модели на Hugging Face

Затем модель и токенайзер загружаются в память GPU, и с помощью метода generate генерируется текст на основе входного текста(Рисунок.10 - Рисунок.15).

```
### Instruction: Resume with skills

### Resume:
Manage process agency staff invoices Pharmacy Rehabilitation department .. Serve head Requisitioner various departments provide technical support staff coordinators Procurement Suite excel data base Risk Management incidents fall analysis .. Analyzed budgets sub-grantees communicated results program directors CEOs .. Education Training Bachelor Arts : Communication f New Rochelle City , State Communications 8/08-11/08 Dale Carnegie Course - Public Speaking , Effective Communication Human Relation May 2004 Activities Honors NYS Notary Public , n , Basic Life Support , PROFESSIONAL MEMBERSHIPS : Urban League Young Professionals Skills administrative , administrative support , agency , budgets , conferences , contracts , collection , data management , data base , database , delivery , Department Health , DOH , dialysis , staff training , expense reports , filing , grant applications , grant proposals Human Resource , Lexis Nexis , logistics , MAC , Director , managing , meetings , access , Excel , Outlook , PowerPoint , MS Windows , Word , policies , presentations , Procurement public Speaking , Quality Assurance , Rehabilitation , reporting , Risk Management , sales , spreadsheets , technical support , phone , trade shows , travel arrangements. Oversee Co department ensure annual staff training .. Skills MS Windows , Word , Excel , PowerPoint , Outlook MAC , Lexis Nexis , Soarian MediNotes EMR Experience Coordinating Manager 09/2013 , State Management annual updates procurement policies , standards procedures guidelines reflect changes operations including regulations , risks best practices .. COORDINATING MANAGER erations : Special events , meeting , travel logistics ; correspondence , file , records , database management ; project administration executive-level management * Sales Support : management , problem trouble-shooting resolution ; contract administration , order review , shipping management ; sales tracking reporting * Communications : Business writer , professional medical terminology , policies standards HIPAA * Financial/Budget Administration : Budget oversight , invoice verification , requisitions , expense tracking ; purchasing , support .. Serve member Culture Change Committee .. Program Associate 12/2005 07/2009 Company Name City , State Housing Community Development .. Managed , grantee level , Housing & Urban rehensive housing counseling program , including annual grant proposal , affiliate contractual process , data collection entry .. Developed standardized work procedures improve work database .. Manage Department Health ( DOH ) Joint Commission ( JCAHO ) annual survey files well upload Plan Corrections ( POC ) Health Commerce System ( HCS ) .. Provide Executive Director Deputy Executive Director .. Assist managing Quality Assurance / Risk Management Department Board Reports quarterly Performance Improvement reports .. Executive Assistant @ e City , State Provided administrative services Executive Director .. Updated maintained calendar ; acted as right arm as gatekeeper Senior VP Housing .. Sales Executive 03/1 ity , State Maintained relationships existing customers regular review visits .. Assist managing Human Resource annual evaluation audits ..
```

Рисунок.10 - generate

```
### Skills:
support, powerpoint, conferences, word, outlook, agency, excel, windows, ms, administrative, budgets,
```

Рисунок.11 - первая генерация

```
### Instruction: Resume with skills

### Resume:
Education Associate Arts , Business Management 2010 University Phoenix City , State , USA 3.69 GPA Skills Account Management , Accounts Payable/Receivable , Adobe , Bookkeeping , Client management , Expense Reports , Hiring Human Resources , Inventory Management Control , Marketing Strategies , Meeting Planning , New product development , Office Management , Online Accounting , Policies Procedures , Property Management , Purchasing , Quickbooks Pro , Real Estate , Production Scheduling. Albuquerque West , Santa Fe/Southern Colorado , Central Utah .. Assistant Community Sales Associate May 1998 Sep 2001 Company Name City , State Communicated recommendations builder home phase construction .. Experience Secretary II Dec 2001 City , State Served primary support channel five sales executive Albuquerque East , .. Associate Head Ericsson Electronics relocation contract team using high multitasking ensured completed 100 % satisfaction .. Associate Relocation Consultant/Property Management Coordinator Jan 2001 Jan 2002 Company Name City , State Maintained portfolio 75 residential properties management employees working abroad Ericsson Electronics acted sole point contact property management clients .. Marketing Director/Executive Assistant Jan 2003 Oct 2007 Company Name implementing online marketing strategy helped increase property listings 42 % .. Direct manager staff 3-15 employees Supervised aspects day-to-day business Dallas , TX office employing reports , budgets , information satellite Houston , TX office .. Telco Engineer/Upper Tier Escalations Jan 2002 Jan 2003 Company Name City , State Consistently maintained 92 % ticket install escalations .. Director Operations Oct 2007 Aug 2014 Company Name City , State Implemented complete accounting system transition Peachtree Manufacturing Quickbooks Pro increased better tracking expenses income ..
```

Рисунок.12 - generate

### Skills:  
adobe, ratio, maintain, accounts, background, aid, scenarios, management, field, completion, close, bookkeeping, account, pay

Рисунок.13 - вторая генерация

### Instruction: Resume with skills

### Resume:

Upgrade hardware software , removed viruses-spyware , sold computers accessories , setup new computers BestBuy specifications quick sales ... Created modified publication databases ' Research Information Tracking System ( RITS ) .. Company Name City , State Computer Department Intern 09/2008 06/2009 Designed proto-typed web-based , IT assets inventory , Apache , MySQL , PHP , HTML , CSS .. Coordinated building use Homelink River charter school , provide safe functional classrooms facilities .. Assist Citrix software , MS Office Ubuntu applications .. Company Name City , State Computer Assistant , GS-5 03/2010 08/2013 Managed internet intranet sites .. Education Training Master Science : Management Or 15 Warner Pacific University , City , State , USA Management Organizational Leadership Certification : ITIL Foundation 2016 New Horizons , City , State , USA Bachelor Science : shington University , City , State , USA Technology Site Builder Theming : Drupal , Content Management 08/2012 OpenSourcery , City , State , USA Additional Skills Web , Content laborator/facilitator , innovator inspection , audits , presentations , researcher , self-starter , T1 , problem solver. Coordinated testing improvement Chief Information Office er training guide , use Forest Service IT specialist , transform web services Content Management System , using Drupal Linux platform .. Instrumental researching , planning coll e increase knowledge base Drupal , SharePoint eBooks PIMRS , R & D members CIO/NO levels .. Awarded Forest Service , Pacific Northwest Research Station , delivering superb compu Application Program , demonstrated outstanding support maintenance stations ' websites , demonstrating extra effort updating research related databases , researching new media s ications .. Administered SharePoint sites .. Company Name City , State Office/Building Manager 01/2005 10/2005 32-hour work week Managed church office , created correspondence r tellite seminars media presentations .. INFORMATION TECHNOLOGY SPECIALIST ( WEB ) , GS-11 Career Overview Objective IT Specialist , GS-2210-9 ( CUSTSPT ) NOC Merit-2016-0031 Ex erse industry experience government , maritime , forestry , research development .. Key developer supporter new Regional Examination Center ( REC ) Merchant Mariner database , w across United States , U.S. Coast Guard .. Converted print publications eBook format eReaders , including : embedding video audio media clips .. Company Name City , State Databa ate patient tracking system using MS Access relationship database help market naturopathic clinic .. Establish fleet Linux OS laptops , saved school district \$ 250,000 , repurpc led .. Ensure material presented compliance copyright requirements section 508 Rehabilitation Act .. Work Experience Company Name City , State Information Technology Specialist 16 Over two years planning , coordinating , identifying business research functions , resources services working Forest Services ' Climate Change Resource Center ( CCRC ) .. Pla chers students Camas School District , including software migration , computer server upgrades .. Qualifications Excellent communicator Adopts technology business needs Stays cu terpersonal skills MS SharePoint , MS Access MS Office , Adobe Suite OS ( ) Windows , Linux , Mac Skype , WebEx , Adobe Connect , MS Lync Technical Skills Skills Level I , II II talls upgrades agency software System Admin Accomplishments Experience Regional System Manager 5 5 Total Years Last Used September 2014 September 2015 Awarded Forest Service , F tion , developing publishing innovations , multiple website support , championing SharePoint .. Company Name City , State Marine Science Technician , E-6 11/1984 05/2004 Provide rt Regional System Manager 400 workstations , across multiple Coast Guard campus , remote field stations , facilities ships .. Company Name City , State Geek Squad Tech 10/2004 t customers troubleshooting computer problems , technical questions ..

Рисунок.14 - generate

### Skills:  
system, office, sharepoint, management, ii, ms, support, content, innovator, built, relationships, team, suite, successful, systems, maintained, installs, adobe, i, interactions, level, providers, software

Рисунок.15 - третья генерация.

Таблица 1. - Итоговая оценка модели по метрикам:

	ROUGE-1	BLEU	BERTScore
LLaMA2 7b.	0.56086119462	0.42419378673	0.4848532874

Значение метрики ROUGE-1 указывает на достаточно высокий уровень совпадения одиночных слов между сгенерированным и эталонным текстами, что говорит о том, что модель в целом хорошо справляется с воспроизведением отдельных ключевых слов, это безусловно важно для задачи стандартизации навыков в резюме.

Значение метрики BLEU указывает на умеренно высокое сходство сгенерированного текста с эталонным на уровне более длинных последовательностей слов. Это означает, что модель способна воспроизводить слова в правильной последовательности.

BERTScore - это метрика, которая оценивает семантическое сходство между сгенерированным и эталонным текстами, используя предварительно обученную языковую модель BERT. Её значение свидетельствует о том, что модель способна генерировать текст, который сохраняет семантику эталонного текста, что важно для стандартизации навыков.

## ЗАКЛЮЧЕНИЕ

В рамках данной работы были рассмотрены традиционные способы представления информации о навыках в резюме. Выявлены основные проблемы и ограничения существующих подходов, такие как неструктурированность, разнородность формулировок, субъективность. Обоснована необходимость применения новых технологий для стандартизации навыков.

Для решения задачи суммаризации текста и извлечения навыков из резюме была выбрана языковая модель LLaMA2 с 7 миллиардами параметров. После дообучения на специально сформированном датасете, модель продемонстрировала достаточно высокую эффективность в решении поставленной задачи.

Получившийся суммаризатор, успешно справляется с извлечением ключевых навыков из резюме кандидатов. Полученные результаты показывают, что разработанная программа имеет большой потенциал для применения в области управления человеческими ресурсами (HR). Автоматизация процесса анализа резюме и извлечения ключевых навыков поможет значительно облегчить работу HR-специалистов, сократить время на обработку большого количества резюме и повысить эффективность процесса подбора персонала.

Для улучшения производительности модели в задаче стандартизации навыков в резюме можно рассмотреть следующие методы:

1. Увеличение объема и разнообразия обучающих данных: Модель может быть обучена на более обширном наборе резюме из различных отраслей и областей деятельности. Это позволит модели лучше распознавать и стандартизировать навыки.

2. Улучшение методов оценки и валидации: Необходимо тщательно проанализировать метрики оценки и методы валидации, используемые для оценки производительности модели. Это должно помочь выявить области, требующие дальнейшего улучшения, и направить усилия на устранение конкретных недостатков.

3. Экспериментирование с архитектурой модели: Можно исследовать возможность модификации архитектуры модели LLaMA2 7b или использования альтернативных архитектур, которые могут быть лучше приспособлены для задачи стандартизации навыков в резюме.

Практическое применение предлагаемого решения, основанного на использовании дообученной языковой модели, позволит значительно повысить качество и объективность процессов оценки резюме кандидатов. Автоматизированное извлечение, стандартизация и структурирование информации о профессиональных навыках соискателей обеспечит более полный учет их компетенций и сократит временные и трудовые затраты рекрутеров и специалистов по подбору персонала.

## СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Егоршин А.П. Основы управления персоналом: учебное пособие / А.П Егоршин. 2015. 352с
2. Васильев А.Н. Принципы и техника нейросетевого моделирования / А.Н. Васильев, Д.А. Тархов. Москва: Высшая школа, 2014.-218 с.
3. Галушкин А.И. Нейронные сети: основы теории. / А.И. Галушкин. М.: РиС, 2015. 496 с
4. Редько В.Г. Эволюция, нейронные сети, интеллект: Модели и концепции эволюционной кибернетики / В.Г. Редько. М.: Ленанд, 2019. 224 с.
5. Москалев, Н. С. Виды архитектур нейронных сетей / Н. С. Москалев. - Текст : непосредственный // Молодой ученый. - 2016. с. 30-34. - URL: <https://moluch.ru/archive/133/37121/>
6. Фрэнк Розенблатт. Принципы нейродинамики. Перцептроны и теория механизмов мозга. Издательство «Мир», 1965г — стр. 82
7. Федотов С., Синицин Ф. Учебник по машинному обучению ШАД URL: <https://education.yandex.ru/handbook/ml>
8. Anna Rogers. How the Transformers Broke NLP Leaderboards. [Электронный ресурс] URL: <https://hackingsemantics.xyz/2019/leaderboards, 2019>
9. Макмахан Б., Рао Д.: Знакомство с PyTorch: глубокое обучение при обработке естественного языка. Питер, 2020 г. 256 стр
10. Барский, А. Б. Нейросетевые методы оптимизации решений : учебное пособие / Барский А. Б. - Санкт-петербург : ИЦ Интермедия, 2017. - 312 с. // ЭБС "Консультант студента" : [сайт]. -URL :<https://www.studentlibrary.ru/book/ISBN9785438301349.html>
11. Лю, Ю. (Х. ) Обучение с подкреплением на PyTorch : сборник рецептов / Лю Ю. (Х. ), пер. с англ. А. А. Слинкина. - Москва : ДМК Пресс, 2020. - 282 с. // ЭБС "Консультант студента" : [сайт]. -URL: <https://www.studentlibrary.ru/book/ISBN9785970608531.html>



12. Журавлева, Л. В. Исследования особенностей развития нейронных сетей в современном мире / Л. В. Журавлева, К. А. Стригулин. // Технические науки: проблемы и перспективы : материалы IV Междунар. науч. конф. (г. Санкт-Петербург, июль 2016 г.). - Санкт-Петербург : Свое издательство, 2016. - С. 9-11. - URL: <https://moluch.ru/conf/tech/archive/166/10748/>
13. С.Д. Тарасов “Исследование и оптимизация параметров алгоритма Manifold Ranking на основе метрики автоматической оценки качества обзорного реферирования ROUGE-RUS” / учеб. пособие Балтийский Государственный Технический Университет им. Д.Ф.Устинова «ВОЕНМЕХ», URL: [http://rcdl.ru/doc/2009/086\\_093\\_DIIS-seminar-1-2009-3.pdf](http://rcdl.ru/doc/2009/086_093_DIIS-seminar-1-2009-3.pdf)
14. Галушкин А. И. Нейронные сети: основы теории. - Изд-во: Горячая линия - Телеком, 2012. - 496 с
15. Тархов Д. А. Нейросетевые модели и алгоритмы / Справочник. - Изд-во: Радиотехника, 2014. - 352 с
16. Новиков И. Нужен ли вам fine-tuning моделей и что это такое / [Электронный ресурс]: <https://vc.ru/ml/984133-nuzhen-li-vam-fine-tuning-modelei-i-chto-eto-takoe>
17. Usman Malik Fine Tuning Llama-2 for Question Answering Tasks in Python
18. Шумский С.А - Машинный интеллект. Очерки по теории машинного обучения и искусственного интеллекта / 2020. - 340 с.
19. Hatice Özolat “Text Summarization: How to Calculate BertScore” 2023 [Электронный ресурс]: <https://haticeozolat17.medium.com/text-summarization-how-to-calculate-bertscore-771a51022964>
20. Priyanka “Evaluation Metrics in Natural Language Processing — BLEU”, 2022 [Электронный ресурс]: <https://medium.com/@priyankads/evaluation-metrics-in-natural-language-processing-bleu-dc3cfa8faaa5>

21. Lavie, A., Sagae, K. and Jayaraman, S. (2004) "The Significance of Recall in Automatic Metrics for MT Evaluation" URL: [https://link.springer.com/chapter/10.1007/978-3-540-30194-3\\_16](https://link.springer.com/chapter/10.1007/978-3-540-30194-3_16)