Федеральное государственное образовательное бюджетное

учреждение высшего образования

**«Финансовый университет при Правительстве Российской Федерации»**

**(Финансовый университет)**

Факультет «Информационных технологий и анализа больших данных»

Выпускная квалификационная работа

на тему «Разработка приложения для взаимодействия с GPT-моделью»

(наименование темы выпускной квалификационной работы)

Направление подготовки 09.03.03 «Прикладная информатика»

(код и наименование направления подготовки)

Профиль «ИТ-сервисы и технологии обработки данных в экономике и финансах»

(наименование направленности)

Выполнил студент учебной группы

\_\_\_\_\_\_\_\_\_\_\_\_\_\_ПИ21-1\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(номер учебной группы)

\_\_\_\_\_\_Ишмаев Руслан Ильдарович\_\_\_\_

(фамилия, имя, отчество полностью)

Руководитель к.т.н., доцент \_

(ученая степень и/или звание)

\_\_\_\_Алюнов Александр Николаевич\_\_\_\_

(фамилия, имя, отчество полностью)

**ВКР соответствует предъявляемым**

**требованиям**

должность

ученая степень

Фамилия

(И.О. Фамилия)

«\_\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_ 202\_ г.

Москва – 202\_г.

# Введение

Современные технологии искусственного интеллекта стремительно развиваются, предлагая новые возможности для решения задач, которые еще недавно или были полностью неразрешимы методами, предлагаемыми машинным обучением, или решались неэффективно. Одним из наиболее значимых достижений в этой области стало появление GPT-моделей, которые способны обрабатывать и генерировать текст на естественном языке с высокой степенью осмысленности. Такие модели теперь находят применение в самых различных областях: от автоматизации бизнес-процессов до создания адаптируемых под пользователя приложений и цифровых ассистентов.

В последние несколько лет наблюдается значительный рост не только интереса к приложениям, использующих возможности крупных языковых моделей, но и количества реализации самих моделей. Это связано не только с их высокой точностью и адаптивностью, но и с их потенциалом для упрощения различных аспектов жизни человека. Приложения, основанные на GPT, могут выполнять широкий спектр функций: от написания текстов до сложного анализа данных.

Целью данной работы является разработка приложения, которое позволило бы более эффективно взаимодействовать с различными GPT-моделями, представляя пользователю понятный интерфейс и широкий функционал. В рамках работы будут рассмотрены как теоретические аспекты работы языковых моделей, так и современные подходы к реализации языковых моделей, будут рассмотрены современные реализации, пользующиеся наибольшей популярностью. На основе проведенного анализа будет спроектировано и реализовано приложение, которое позволит пользователю выбирать модель из набора доступных для дальнейшего взаимодействия с ней, что продемонстрирует практическую ценность и возможность использования GPT-моделей в различных сценариях.

Для достижения цели были поставлены следующие задачи:

* Изучение теоретических основ языковых моделей и технологии GPT;
* Изучение наиболее популярных реализаций языковых моделей, построенных на технологии GPT;
* Реализация программного интерфейса для взаимодействия с различными GPT-моделями;
* Разработка графического интерфейса;
* Проведение тестирования и апробация готового приложения.

Объектом дипломного исследования является семейство языковых моделей, построенных на GPT-технологии. Предметом – возможность создания информационной системы для взаимодействия с различными генеративными моделями.

Актуальность выбранной темы обусловлена кратным ростом использования технологий искусственного интеллекта как в повседневной жизни, так и в рамках различных бизнес-процессов. Согласно статистике, инвестиции по миру в развитие искусственного интеллекта значительно выросли за последнее десятилетие. Инвестиции же в генеративный искусственный интеллект на конец 2023 года составили 25.2 миллиардов долларов США, что примерно в 8 раз больше, чем на конец 2022 года. Данная статистика подтверждает цифрами заинтересованность рынка в искусственном интеллекте, в частности GPT-моделях.

Полученные наработки могут быть полезны для компаний, создающих решения по взаимодействию с искусственными интеллектом, а также людям, часто использующим искусственный интеллект в своей повседневной жизни.

# Теоретические основы языковых моделей

## История и эволюция языковых моделей

История языковых моделей тесно связана с развитием вычислительных технологий, лингвистки и машинного обучения. На протяжении нескольких десятилетий ученые, инженеры и исследователи разрабатывают методы для обработки естественного языка (Natural Language Processing, NLP), которые позволяют создавать современные модели, такие как GPT. Рассмотрим основные этапы эволюции языковых моделей.

### Ранние этапы: статистические языковые модели

В начале XX века зародились первые идеи о моделировании естественного языка при помощи математики, однако только с развитием компьютеров в середине XX века стало возможно реализовать эти идеи на практике.

#### N-граммные модели (1950-1970-е годы)

Одной из первых попыток моделировать языки стали N-граммные модели, также известные как статистически модели. Такие модели анализируют последовательность слов и рассчитывают вероятность появления каждого слова на основе предыдущих. Например, вероятность слова в триграммной модели рассчитывается на основе двух предыдущих слов.

Основные преимущества и недостатки данной модели представлены в таблице 1. Несмотря на то, что модель действительно простая в реализации и дешевая по ресурсам, недостатки, не позволяющие воспринимать контекст больше, чем N слов, а также использование только самых часто-встречаемых слов, оказались главными ограничениями. Из-за этих факторов N-граммные модели использовались, в основном, только в авто-дополнении текста, также известное как Т9.

|  |  |
| --- | --- |
| Преимущества | Недостатки |
| Простота реализации | Ограниченность контекста |
| Скорость работы | Генерация редких последовательностей слов |

Таблица 1

#### Модели на основе скрытых марковских процессов

В 1970–1980-х годах в обработке текста начали использовать скрытые марковские модели, которые рассматривали текст как последовательность состояний. HMM оказались полезными для задач, таких как распознавание речи и разметка текста.

Скрытые марковские модели (Hidden Markov Models, HMM) основаны на статистическом подходе к моделированию последовательностей данных, где наблюдаемые события зависят от скрытых (латентных) состояний, которые эволюционируют согласно марковскому процессу первого порядка. Принцип действия HMM заключается в следующем:

1. Марковское свойство скрытых состояний: Последовательность скрытых состояний . Т.е. удовлетворяет марковскому свойству, где вероятность перехода в следующее состояние зависит только от текущего состояния.
2. Наблюдения зависят от скрытых состояний: Наблюдаемая последовательность генерируется из скрытых состояний с вероятностями, заданными функцией эмиссии. То есть, каждое наблюдение зависит только от текущего скрытого состояния .
3. Параметры модели:
   1. Начальные вероятности определяют распределение первого скрытого состояния.
   2. Матрица переходов задаёт вероятности переходов между скрытыми состояниями.
   3. Матрица эмиссий определяет вероятность наблюдения при нахождении в состоянии .
4. Основные задачи:
   1. Оценка вероятности последовательности: Вычисление , где , с помощью алгоритма вперёд-назад (Forward-Backward).
   2. Декодирование: Поиск наиболее вероятной последовательности скрытых состояний , породившей наблюдения , с использованием алгоритма Витерби.
   3. Обучение модели: Оптимизация параметров по наблюдаемым данным с использованием алгоритма Баум-Велша.

Основные преимущества и недостатки данной модели представлены в таблице 2.

|  |  |
| --- | --- |
| Преимущества | Недостатки |
| Простота и интуитивность | Ограниченность в представлении сложных зависимостей |
| Доказанная эффективность для задач с последовательными данными | Не всегда хорошо работает с большими объемами данных или задачах с высокой сложностью |
| Хорошо изученные алгоритмы |  |

Таблица 2

### Переход к нейронным сетям (1990-2010-е годы)

С развитием вычислительных мощностей и нейросетевых подходов началась новая эра языковых моделей.

#### Рекуррентные нейронные сети (RNN) (1990-е годы)

Рекуррентные нейронные сети (Recurrent Neural Networks, RNN) стали важным этапом в развитии технологий обработки последовательных данных, таких как текст, аудио и временные ряды. Они были разработаны для решения задачи моделирования последовательностей, где текущий элемент данных зависит от предыдущих. Это делает их особенно полезными для задач обработки естественного языка, где порядок слов играет ключевую роль.

Ключевая особенность RNN заключается в том, что они имеют рекуррентные связи, позволяющие сохранять "память" о предыдущих шагах. Это достигается за счет введения скрытого состояния (hidden state), которое обновляется на каждом временном шаге и хранит информацию о предыдущих элементах последовательности.

Преимущества и недостатки RNN-моделей представлены в таблице 3.

|  |  |
| --- | --- |
| Преимущества | Недостатки |
| Обработка последовательностей произвольной длины | Затухающие и взрывающиеся градиенты |
| Память о предыдущих шагах | Краткосрочная память |
| Гибкость – возможность использование в задачах разного типа (классификация, генерация, регрессия, и т.д.) | Медленная скорость обучения |

Таблица 3

#### LSTM и GRU (1997-2010-е годы)

Рекуррентные нейронные сети (RNN) имеют ограничения, такие как проблема затухающих и взрывающихся градиентов, что делает их малопригодными для работы с длинными последовательностями. Для решения этих проблем были разработаны две ключевые модификации RNN — LSTM (Long Short-Term Memory) и GRU (Gated Recurrent Unit).

LSTM была предложена в 1997 году Шмидхубером и Хохрайтером (Hochreiter Schmidhuber). Основная цель LSTM — эффективно сохранять информацию на длительных временных интервалах, предотвращая затухание или взрыв градиентов. Для этого LSTM использует специальные механизмы управления потоком информации через "ячейки памяти" (memory cells) и "гейты" (gates).

Преимущества и недостатки LSTM-моделей представлены в таблице 4.

|  |  |
| --- | --- |
| Преимущества | Недостатки |
| Долгосрочная память | Медленное обучение |
| Гибкость – благодаря гейтам модель может динамически решать, что запоминать, а что забывать. | Высокие вычислительные затраты |
| Широкое применение |  |

Таблица 4

GRU была предложена в 2014 году Чо и его коллегами (Cho et al.) как упрощенная альтернатива LSTM. GRU сохраняет ключевые идеи LSTM (гейты для управления потоком информации), но устраняет некоторые элементы, такие как отдельная ячейка памяти. Это делает GRU проще и быстрее в обучении, при этом она достигает схожих результатов.

Преимущества и недостатки GRU-моделей представлены в таблице 5.

|  |  |
| --- | --- |
| Преимущества | Недостатки |
| Легковесность | Ограниченная память |
| Эффективность | Меньшая гибкость |
| Универсальность |  |

Таблица 5

#### Word2Vec и GloVe (2013-2015-е годы)

С появлением методов Word2Vec и GloVe в 2013–2015 годах обработка естественного языка (Natural Language Processing, NLP) сделала значительный шаг вперед. Эти методы предложили новый способ представления слов в виде векторов, что позволило моделям понимать семантические и синтаксические связи между словами. Такой подход открыл новые возможности для обработки текста и стал важной основой для развития современных языковых моделей.

Word2Vec — это метод, предложенный Томашем Миколовым (Tomas Mikolov) и его командой в 2013 году. Это нейросетевая модель, которая преобразует слова в векторы фиксированной длины, сохраняя их семантические и синтаксические свойства. Векторы, созданные Word2Vec, называются эмбеддингами слов (word embeddings).

Word2Vec основывается на гипотезе распределения: "Слова, которые встречаются в похожих контекстах, имеют схожее значение". Например, слова "кофе" и "чай" часто встречаются в схожих предложениях ("пить кофе", "налить чай"), поэтому их векторные представления должны быть близки друг к другу.

Преимущества и недостатки Word2Vec представлены в таблице 6.

|  |  |
| --- | --- |
| Преимущества | Недостатки |
| Семантическая близость | Не учитывает глобальную структуру текста |
| Эффективность | Не масштабируется для больших текстовых коллекций |

Таблица 6

GloVe (Global Vectors for Word Representation) был предложен в 2014 году исследователями Стэнфордского университета. В отличие от Word2Vec, GloVe использует статистический подход для создания эмбеддингов, комбинируя локальную и глобальную информацию о словах.

GloVe строит эмбеддинги, анализируя совместное появление слов в тексте. Это означает, что модель учитывает, как часто два слова встречаются вместе в одном контексте по сравнению с тем, как часто они встречаются по отдельности.

Преимущества и недостатки GloVe представлены в таблице 7.

|  |  |
| --- | --- |
| Преимущества | Недостатки |
| Глобальная информация | Требует больше памяти |
| Эффективность | Не подходит для потоковых данных |

Таблица 7

### Революция трансформеров (2017-по настоящее время)

Настоящая революция в области языковых моделей произошла с появлением архитектуры трансформеров.

#### Появление трансформеров (2017)

В 2017 году исследователи Google представили архитектуру трансформеров в статье "Attention is All You Need". Трансформеры заменили рекуррентные сети механизмом внимания (attention), который позволяет модели эффективно фокусироваться на ключевых частях текста, независимо от его длины.

Преимущества трансформеров:

* Параллельная обработка данных (ускорение обучения).
* Учет долгосрочного контекста.
* Гибкость для различных задач NLP.

#### BERT (2018)

В 2018 году Google представил модель BERT (Bidirectional Encoder Representations from Transformers), которая впервые использовала двунаправленный контекст. Это означало, что модель могла учитывать не только слова до текущего, но и слова после него. BERT стал стандартом для обработки текста и улучшил результаты во многих NLP-задачах.

BERT основан на энкодере трансформера (Transformer Encoder).

Архитектура BERT, ключевые особенности:

* Маскированное языковое моделирование (Masked Language Modeling, MLM)
* Предсказание следующего предложения (Next Sentence Prediction, NSP)

Преимущества и недостатки BERT представлены в таблице 8.

|  |  |
| --- | --- |
| Преимущества | Недостатки |
| Двунаправленный контекст | Высокая вычислительная сложность |
| Универсальность | Ограниченность в генерации текста |
| Качество |  |

Таблица 8

#### GPT (2018-по настоящее время)

GPT был разработан OpenAI, и его первая версия была выпущена в 2018 году. В отличие от BERT, GPT ориентирован на генерацию текста, что делает его идеальным для задач, связанных с созданием осмысленного и связного текста.

GPT использует однонаправленный контекст, анализируя текст слева направо. Это позволяет модели предсказывать следующее слово в последовательности, основываясь на предыдущих словах.

GPT основан на декодере трансформера (Transformer Decoder).

Архитектура GPT, Ключевые особенности:

* Авторегрессионное обучение
* Обучение на больших корпусах

Преимущества и недостатки GPT-моделей представлены в таблице 9.

|  |  |
| --- | --- |
| Преимущества | Недостатки |
| Генерация текста | Однонаправленный контекст |
| Гибкость | Сложность обучения |
| Масштабируемость | Риск генерации некорректной информации |

Таблица 9