ШАМИМ БХУЯН | ТИМУР ИСАЧЕНКО

ГЕНЕРАТИВНЫЙ ИИ

С ОБУЧЕНИЕМ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ (LLM)

ДЛЯ ДЖУНОВ

БОМБОРА
ИЗДАТЕЛЬСТВО
Москва

УДК 004.8
ББК 32.813
Б94

Shamim Bhuiyan and Timur Isachenko
Generative AI with local LLM
A comprehensive roadmap for building AI-Driven applications with local LLMs

Бхуян, Шамим.

Б94 Генеративный ИИ с обучением больших языковых моделей (LLM) для джунов / Шамим Бхуян, Тимур Исаченко ; [перевод О. И. Перфильева]. — Москва : Эксмо, 2025. — 320 с. — (Путеводитель по GPT и AI).

ISBN 978-5-04-220583-5

Это практическое руководство по созданию приложений на основе генеративного искусственного интеллекта и больших языковых моделей (LLM). Особое внимание уделяется прикладным аспектам: промпт-инжинирингу, работе с локальными LLM, тонкой настройке моделей на частных данных и созданию автономных AI-агентов. Приводятся примеры реальных решений, таких как интеллектуальная обработка SQL-запросов и автоматизация работы с изображениями.

Подходит для разработчиков и аналитиков данных с базовыми знаниями Python, желающих освоить генеративный ИИ.

УДК 004.8
ББК 32.813

ISBN 978-5-04-220583-5

ОГЛАВЛЕНИЕ

Моему любящему сыну Мишелю,
чьи любопытные вопросы о том,
почему я так много времени
провожу за компьютером, ежедневно
вдохновляют меня.

ПРЕДИСЛОВИЕ

В последнее время тема искусственного интеллекта (ИИ) обрела невероятную популярность и освещается повсюду — от школьных газет до дебатов в Сенате США. Эта быстро развивающаяся область привлекает внимание как экспертов в сфере машинного обучения, так и обычных людей. Кое-кто с беспокойством рассуждает о том, в каком направлении движутся современные технологии, а кто-то предлагает такие экстремальные меры, как уничтожение центров обработки данных. Свое мнение о будущем ИИ высказывают даже представители Администрации президента.

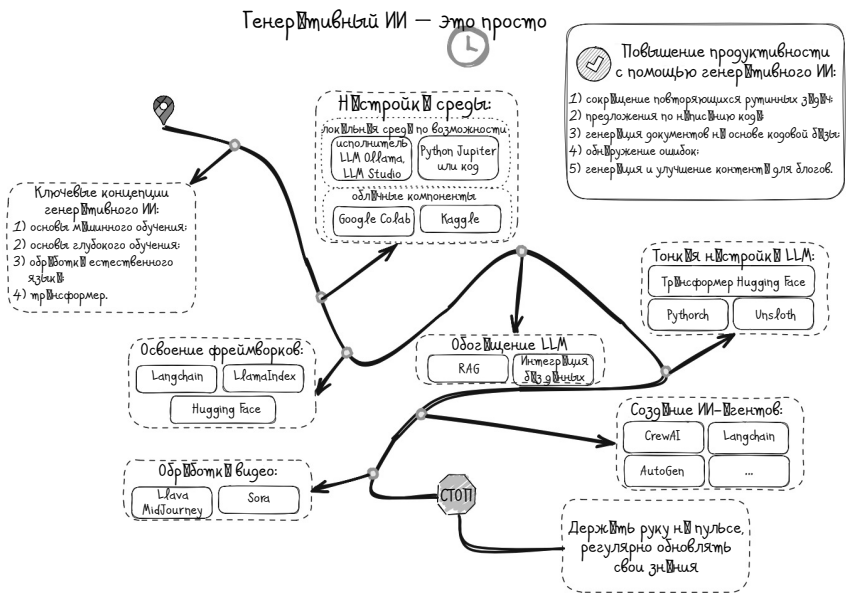
Но мы считаем, что гораздо продуктивнее не жить в страхе перед возможными злоупотреблениями в данной сфере, а сосредоточиться на полезных сторонах ИИ. То есть на том, что он может улучшить нашу повседневную жизнь — с помощью Siri или других ИИ-ассистентов.

Искусственному интеллекту, особенно генеративному ИИ и большим языковым моделям, посвящены бесчисленные статьи и публикации, рассчитанные на читателей с самым разным уровнем подготовки. К сожалению, для полного понимания большинства из них требуется определенный уровень знаний.

Вместе с тем не существует какого-то общепринятого пособия, с помощью которого можно было бы легко овладеть этой темой, поэтому мы и написали эту книгу. По нашему замыслу, она должна стать четкой и понятной дорожной картой, которая провела бы но-

вичков через все терминологические и концептуальные сложности. Чтобы с ее помощью разобраться в некоторых принципах машинного обучения, вовсе не обязательно быть экспертом в области компьютерных технологий, поэтому не бойтесь приступать к обучению!

За несколько недель, проведенных в исследованиях и составлении планов, мы разработали надежное практическое пособие по генеративному искусственному интеллекту. И рады поделиться им с вами.



В целом полный путь освоения темы генеративного ИИ состоит из семи взаимосвязанных этапов.

1. **Знакомство с ключевыми понятиями в сфере генеративного ИИ**

   - Искусственный интеллект. Основные понятия и принципы его работы, включая то, что он собой представляет и как применяется.
   - Машинное обучение. Основные концепции в данной области, включая обучение с учителем и без учителя, алгоритмы и обучение моделей.
   - Глубокое обучение. Основы глубокого обучения, нейронные сети и принципы их работы, включая такие популярные архи-

тектуры, как CNN и RNN (сверточные и рекуррентные нейронные сети).
- Обработка естественного языка (NLP, Natural Language Processing). Знакомство с обработкой естественного языка, подразумевающей взаимодействие между компьютерами и людьми, дающими команды и инструкции на естественном языке.

Эти ключевые понятия помогут вам перейти к более сложным темам, связанным с ИИ.

2. Настройка среды для локальных LLM

- Запуск локальных моделей. Эксперименты с их запуском с помощью таких инструментов, как *Ollama* или других программ с открытым исходным кодом. Знания о том, как настраивать и использовать модели на локальной машине.
- Настройка локальной среды Python с инструментами Miniconda и JupyterLab, которые помогут вам начать путешествие в мир генеративного ИИ.
- Платформы для обмена моделями. Знакомство с такими платформами, как *Hugging Face*, предлагающими доступ к предварительно обученным моделям и наборам данных. Они позволят вам экспериментировать с широким спектром моделей генеративного ИИ.

3. Освоение фреймворков

LangChain. Создание приложений, использующих языковые модели. Так вы узнаете, как можно внедрить генеративный ИИ в реальные приложения.

- LlmIndex. Индексирование и поиск по большим языковым моделям.
- Python. Повышение навыков программирования на Python, самом распространенном языке для разработки ИИ. Основное внимание уделяется таким библиотекам, как TensorFlow, PyTorch и Transformers от *Hugging Face*.

4. Тонкая настройка/обогащениеLLM

- Методы тонкой настройки. Как настраивать предварительно обученные модели для лучшего решения конкретных задач или для конкретных наборов данных.
- Методы «обогащения». Метод RAG (Retrieval-Augmented Generation, «расширенная генерация данных»), позволяющий улучшать качество выходных данных моделей.

5. Создание проектов и агентов

- Небольшие проекты. Создание небольших проектов по выполнению таких задач, как генерация текста, изображений или других медиафайлов, с целью закрепления полученных знаний.
- Агенты. Разработка агентов, которые смогут выполнять некоторые повседневные задачи, — например, искать информацию в Интернете и генерировать качественный текст для блога на определенную тему.

6. Продолжение обучения

- Интерес к области ИИ. Регулярное знакомство с последними достижениями в области ИИ, просмотр соответствующих блогов, каналов YouTube и онлайн-курсов.
- Сотрудничество и обмен информацией. Взаимодействие с онлайн-сообществами, участие в решении задач в области ИИ и обмен проектами на таких платформах, как GitHub.

7. Продвинутые темы

- Овладев основами, вы сможете погрузиться в такие продвинутые темы, как обучение с подкреплением, продвинутые архитектуры моделей и передовые исследования в области ИИ.

Описанный выше курс заложит прочный фундамент знаний в области генеративного ИИ и позволит вам постепенно наращивать свой опыт с применением его в реальных сценариях.

О чем эта книга

Данная книга представляет собой развернутое руководство по использованию генеративного искусственного интеллекта и больших языковых моделей (LLM) в локальной среде разработки. Она знакомит читателей с основами генеративного ИИ и такими продвинутыми методами, как тонкая настройка моделей, обогащение их частными наборами данных и использование в практических сценариях. Например, для обработки SQL-запросов и изображений, а также разработки агентов. Если ваша продукция в какой-то степени связана с ИИ, если вы разработчик приложений, специалист по анализу данных или просто энтузиаст своего дела, увлекающийся данной темой, то эта книга вооружит вас знаниями и инструментами, необходимыми для эффективного использования ИИ в своих проектах.

Глава 1. Начало работы с локальными LLM
Прежде чем приступить к реализации ИИ-моделей, нужно создать надежную локальную среду их разработки. В этой главе мы расскажем вам о том, как настроить компьютер для разработки ИИ-приложений, в том числе установить необходимое программное обеспечение, настроить среду Python и выбрать подходящее оборудование (например, графический процессор, или GPU) для оптимальной производительности. Глава также познакомит вас с такими инструментами, как блокноты Jupyter, которые помогут упростить рабочий процесс разработки ИИ.

Глава 2. Погружение в теорию генеративного ИИ
Наше путешествие в мир генеративного ИИ начинается с основ, то есть с объяснения того, как и почему работает генеративный ИИ, что это такое, и какую революцию он совершил в различных областях. В главе рассматриваются фундаментальные понятия ИИ, машинного обучения и глубокого обучения. В ней же закладывается основа для дальнейшего понимания того, как LLM генерируют контент — текст, код и изображения. Вы познакомитесь с различными типами генеративных моделей, в том числе с механизмом самовнимания,

с архитектурой «кодировщик-декодировщик» и трансформерами, а также с их применением в различных областях.

Глава 3. RAG, обогащение моделей LLM с помощью частных наборов данных

В этой главе рассматривается передовая концепция Retrieval-Augmented Generation (RAG), подразумевающая улучшение LLM моделей за счет использования частных наборов данных. Вы узнаете о том, как можно обогащать модели ИИ специфичными для конкретной области знаниями, благодаря чему они будут генерировать более точные и релевантные результаты. В главе описаны технические шаги по настройке RAG, включая создание векторных баз данных, индексацию частных данных и настройку модели для получения и генерации ответов на основе этой обогащенной информации.

Глава 4. Преобразование текста в SQL, улучшение ответов LLM за счет интеграции с базой данных

Большую ценность в последнее время приобретает автоматизация взаимодействия с базами данных на основе ИИ, и эта глава посвящена тому, как LLM помогают преобразовывать текст в SQL-запросы (так называемая задача «Text-to-SQL»). Вы узнаете, как пользоваться моделями, способными конвертировать запросы на естественном языке в команды SQL, специально адаптированные для взаимодействия с системой BigQuery. Глава содержит практические примеры и фрагменты кода, показывающие, как создавать системы для взаимодействия с базами данных без знания языка запросов SQL, что делает поиск данных более доступным и эффективным.

Глава 5. Файн-тюнинг (дообучение) LLM

В этой главе мы подробно рассмотрим процесс тонкой настройки LLM или файн-тюнинга (fine-tuning) — важнейший этап в создании LLM-моделей, предназначенных для решений конкретных задач. Вы узнаете, как адаптировать предварительно обученные модели к вашим данным и улучшать их производительность для задач, относящихся к нужной области. В главе рассматриваются технические аспекты тонкой настройки, включая подготовку данных, настройку гиперпараметров и оценку эффективности модели. К концу главы

вы сможете совершенствовать модели, повышая их точность и релевантность для своих приложений.

Глава 6. Обработка и генерация изображений с помощью LLM
Сфера генеративного ИИ не ограничивается текстом, он также играет важную роль в обработке и генерации изображений. В данной главе рассматривается использование LLM для создания и обработки изображений. В главе приведены практические примеры и описаны полезные инструменты, в том числе модели LLaVa и OpenJourney, которые помогут интегрировать генерацию изображений в ваши проекты.

Глава 7. Разработка и использование ИИ-агентов
Последняя глава посвящена одному из самых интересных применений LLM: разработке ИИ-агентов. Вы узнаете о назначении и преимуществах использования ИИ-агентов искусственного интеллекта для автоматизации повседневных задач. Мы рассмотрим архитектуру ИИ-агентов и их ключевые компоненты, а также опишем мультиагентный фреймворк, позволяющий пользователям без специальных технических знаний управлять ИИ-агентами и задачами с помощью интуитивно понятных высокоуровневых абстракций.

Примеры кода

Все примеры кода, скрипты и более подробные примеры можно найти в репозитории GitHub.

Целевая аудитория

Целевая аудитория данной книги — владельцы и разработчики цифровых продуктов, специалисты по анализу данных и энтузиасты в области искусственного интеллекта с минимальными навыками в области программирования. Никаких особых знаний для освоения материала книги не требуется, хотя неплохо иметь хотя бы поверхностное знакомство с Python, Java и такими инструментами, как Conda.

Форматирование и условные обозначения

1. Курсивом и жирным шрифтом выделяются новые термины, важные слова, ссылки на интернет-страницы (URL), имена файлов и расширения файлов.

2. Код блока задается следующим образом.

```
def process_image(image_file_path, prompt):
     print(f"\nProcessing {image_file_path} file \n")
     image = Image.open(image_file_path)
     display(image)

     with image as img:
          with BytesIO() as buffer:
          img.save(buffer, format='PNG')
          image_bytes = buffer.getvalue()

     # Генерация описания изображения
     for response in generate(model='llava:34b',
               prompt=prompt,
               images=[image_bytes],
               stream=True):
     # Вывести ответ на консоль и добавить к полному ответу
     print(response['response'], end='', flush=True)
```

3. Любая команда или вывод в командной строке печатаются следующим образом.

```
!pip install ollama
export OLLAMA_HOST="192.168.1.124"
Processing ./Text2SQL.png file
```

Совет

Значок указывает на совет или предложение.

Внимание

Значок указывает на предупреждение или предостережение.

Инфо

Значок указывает на общие сведения.

Отзывы читателей

Мы хотели бы услышать ваши комментарии — например, что понравилось или не понравилось в содержании книги. Эти отзывы помогут нам в следующий раз написать книгу получше, а другим — разобраться в освещаемых концепциях. Чтобы оставить свой отзыв, воспользуйтесь ссылкой для обратной связи.