
Оглавление

Предисловие	13
О чем эта книга.....	14
Чем книга не является	16
Для кого эта книга.....	17
Структура издания.....	18
Условные обозначения.....	20
Использование примеров кода	21
Благодарности.....	21
От издательства	24
О научном редакторе русскоязычного издания	24
Глава 1. Основы создания AI-приложений с использованием базовых моделей.....	25
Восход AI-инженерии.....	26
От языковых моделей к большим языковым моделям	26
От больших языковых моделей к базовым.....	33
От базовых моделей к AI-инженерии.....	37
Сценарии использования базовой модели	41
Написание кода	45
Генерация изображений и видео	47
Написание текстов	48
Образование.....	50
Боты-собеседники	52
Сбор информации	53
Организация данных.....	53
Автоматизация рабочих процессов.....	54
Планирование AI-приложений	55
Оценка сценариев использования.....	55
Определение ожиданий	60
Планирование основных этапов	60
Техническое обслуживание	61
Стек AI-инженерии	63
Три уровня стека AI	65
Сравнение AI-инженерии с ML-инженерией	67
Сравнение AI-инженерии и проектирования полного стека.....	75
Резюме	76

Глава 2. Знакомство с базовыми моделями	78
Обучающие данные	79
Многоязычные модели	81
Модели, связанные с конкретной предметной областью	85
Создание модели	88
Архитектура модели.....	88
Размер модели	98
Последующее обучение.....	110
Дообучение с учителем	112
Дообучение с учетом предпочтений	116
Сэмплирование.....	120
Основы сэмплирования	121
Стратегии сэмплирования.....	122
Вычисление во время инференса	128
Структурированный вывод.....	132
Вероятностная природа AI	137
Резюме	144
Глава 3. Методология оценки	146
Трудности оценки базовых моделей.....	147
Знакомство с метриками языкового моделирования	151
Энтропия.....	153
Кросс-энтропия	153
Количество битов в символе и количество битов на байт	154
Перплексия.....	155
Интерпретация и сценарии использования перплексии	156
Точная оценка.....	159
Функциональная корректность	159
Измерение сходства с эталонными данными.....	161
Знакомство с эмбедингом	168
AI в роли судьи	170
Зачем использовать AI в роли судьи	171
Как применять подход «AI в роли судьи»	173
Ограничения подхода «AI в роли судьи»	176
Какие модели могут выступать в роли судей	180
Ранжирование моделей путем сравнительной оценки	183
Проблемы сравнительной оценки	187
Будущее сравнительной оценки.....	191
Резюме	191

Глава 4. Оценка систем AI	193
Критерии оценки.....	194
Способности, специфичные для конкретной области.....	196
Способность к генерации нового контента.....	198
Способность следовать инструкциям.....	207
Стоимость и время отклика	213
Выбор модели.....	215
Рабочий процесс выбора модели	216
Разработать модель или купить.....	218
Разбираемся в общедоступных эталонных тестах	230
Спроектируйте собственный конвейер оценки	240
Шаг 1. Оцените все компоненты системы	240
Шаг 2. Создайте руководство по оценке.....	242
Шаг 3. Определите методы оценки и данные	244
Резюме	249
Глава 5. Промт-инжиниринг	251
Введение в промтинг.....	252
Контекстное обучение: без примеров и на нескольких примерах	253
Системные запросы и запрос пользователя.....	255
Длина и эффективность контекста	258
Лучшие практики промт-инжиниринга	261
Пишите четкие и ясные инструкции	261
Обеспечьте достаточный контекст	264
Разбейте сложные задачи на более простые подзадачи	265
Дайте модели время подумать	268
Итерируйте свои запросы.....	270
Оцените инструменты промт-инжиниринга	271
Организируйте и версионизируйте запросы	274
Оборонительный промт-инжиниринг.....	276
Проприетарные запросы и обратный промт-инжиниринг.....	278
Обход ограничений и внедрение запроса.....	280
Извлечение информации.....	285
Методы защиты от атак запросами	290
Резюме	294
Глава 6. RAG и агенты	295
RAG.....	295
Архитектура RAG.....	298
Алгоритмы поиска.....	299
Оптимизация поиска.....	311
Применение RAG для нетекстовых данных	317

Агенты	320
Обзор агентов.....	320
Инструменты.....	323
Планирование	327
Типы ошибок агента и оценка.....	344
Память.....	347
Резюме	351
Глава 7. Дообучение, или Тонкая настройка	353
Обзор дообучения.....	354
Когда проводить дообучение.....	359
Причины провести дообучение.....	359
Причины не проводить дообучение.....	360
Дообучение и RAG	364
Ограничения памяти	368
Обратное распространение и обучаемые параметры.....	369
Сколько нужно памяти	371
Числовые представления	374
Методы дообучения	382
Параметрически эффективное дообучение	382
LoRA.....	388
Слияние моделей и многозадачное дообучение	397
Тактики дообучения	408
Резюме	412
Глава 8. Проектирование наборов данных.....	414
Подготовка данных	416
Качество данных	419
Охват данных	422
Количество данных.....	425
Сбор и разметка данных.....	430
Аугментация и синтез данных	433
Зачем нужен синтез данных.....	434
Традиционные методы синтеза данных	436
Синтез данных с помощью AI	440
Дистилляция модели.....	449
Обработка данных.....	451
Проверьте данные.....	451
Выполните дедупликацию данных	453
Очистите и отфильтруйте данные	455
Отформатируйте данные.....	456
Резюме	458

Глава 9. Оптимизация вывода	460
Основы оптимизации вывода	461
Обзор выводов	461
Метрики эффективности вывода	467
Ускорители AI.....	475
Оптимизация вывода	483
Оптимизация модели	484
Оптимизация сервисов вывода	498
Резюме	506
Глава 10. Архитектура AI-инженерии и обратная связь от пользователей	508
Архитектура AI-инженерии.....	508
Шаг 1. Расширьте контекст.....	509
Шаг 2. Установите ограничивающие механизмы	511
Шаг 3. Добавьте маршрутизатор модели и шлюз.....	516
Шаг 4. Сократите время отклика с помощью кэш-памяти	521
Шаг 5. Добавьте шаблоны агентов	524
Мониторинг и наблюдаемость	526
Оркестрация конвейеров AI.....	534
Обратная связь от пользователей	536
Получение обратной связи из бесед	536
Проектирование обратной связи	542
Ограничения обратной связи.....	552
Резюме	554
Эпилог.....	556
Об авторе	557
Иллюстрация на обложке	558