
Промт-инжиниринг для LLM

*Искусство построения приложений
на основе больших языковых моделей*

Джон Берриман, Альберт Циглер

Краткое содержание

Предисловие	11
-------------------	----

ЧАСТЬ I ОСНОВЫ

Глава 1. Введение в промт-инжиниринг	16
Глава 2. Понимание LLM.....	30
Глава 3. Переходим к чатам.....	60
Глава 4. Проектирование приложений на базе LLM.....	83

ЧАСТЬ II ОСНОВНЫЕ ТЕХНИКИ

Глава 5. Содержимое промта.....	106
Глава 6. Составление промта.....	143
Глава 7. Приручение модели	169

ЧАСТЬ III НАВЫКИ ЭКСПЕРТА

Глава 8. Диалоговый агент	192
Глава 9. Рабочие потоки LLM.....	225
Глава 10. Оценка приложений на базе LLM	251
Глава 11. Обозримое будущее	274
Об авторах	283
Иллюстрация на обложке.....	284

Оглавление

Предисловие	11
Для кого эта книга	11
Чему вы научитесь	12
Условные обозначения	13
Благодарности	14
От Джона	14
От Альберта	14
От издательства	14
О научном редакторе русскоязычного издания	14

ЧАСТЬ I ОСНОВЫ

Глава 1. Введение в промт-инжиниринг	16
Большие языковые модели — это волшебство	17
Языковые модели: как мы к этому пришли	19
Первые языковые модели	21
GPT	24
Промт-инжиниринг	26
Резюме	29
Глава 2. Понимание LLM	30
Что такое LLM	31
Завершение документа	33
Человеческое мышление в сравнении с LLM-обработкой	34
Галлюцинации	36
Как большие языковые модели видят мир	37
Отличие 1: LLM используют детерминированные токенизаторы	39
Отличие 2: LLM не могут остановиться и присмотреться к буквам	39
Отличие 3: LLM видят текст иначе	42
Подсчет токенов	43

По токену за раз	44
Авторегрессионные модели	44
Паттерны и повторяемость	46
Температура и вероятности	47
Архитектура трансформера	52
Резюме	59
Глава 3. Переходим к чатам	60
Обучение с подкреплением на основе обратной связи от человека	61
Процесс создания модели с помощью RLHF	62
Обеспечение честности LLM	66
Исключение индивидуальных особенностей поведения	67
RLHF позволяет получить хорошую отдачу от вложенных усилий	67
Учитывайте издержки выравнивания	67
Переходим от инструкций к чату	68
Instruct-модели	68
Чат-боты	70
Изменяющийся API	73
API чата на основе завершений	73
Сравнение модели чата с моделью завершений	76
Переход от чатов к инструментам	78
Промт-инжиниринг как написание пьесы	79
Резюме	81
Глава 4. Проектирование приложений на базе LLM	83
Структура цикла	83
Задача пользователя	85
Преобразование задачи пользователя в понятную для модели форму	86
Выбор LLM для завершения промта	92
Преобразование обратно в предметную область задачи пользователя	93
Разбор механизма прямого распространения	94
Создание простого механизма прямого распространения	95
Изучение сложности цикла	97
Оценка качества приложений на базе LLM	102
Офлайн-оценка	102
Онлайн-оценка	103
Резюме	104

ЧАСТЬ II

ОСНОВНЫЕ ТЕХНИКИ

Глава 5. Содержимое промта	106
Источники содержимого	108
Статическое содержимое	109
Уточнение вопроса	109
Few-shot-промтинг	111
Динамическое содержимое	120
Поиск динамического контекста для промта	123
Дополнение промта поисковыми данными (RAG)	126
Использование резюме в промте	138
Резюме	142
Глава 6. Составление промта	143
Структура идеального промта	143
Форматы документов для запросов в LLM	147
Помощь в формате диалога	147
Аналитический отчет	150
Структурированный документ	152
Форматирование фрагментов	157
Подробнее об инертности	159
Форматирование few-shot-примеров	160
Гибкие фрагменты	160
Связи между элементами промта	162
Позиция	162
Важность	162
Зависимость	163
Собираем все воедино	165
Резюме	168
Глава 7. Приручение модели	169
Структура идеального ответа LLM	169
Преамбула	170
Узнаваемые начало и конец	173
Послесловие	174
За гранью текста: логарифмическая вероятность	175
Оценка качества ответа	176
LLM для классификации	177
Критические места промта	180
Выбор модели	182
Резюме	189

ЧАСТЬ III НАВЫКИ ЭКСПЕРТА

Глава 8. Диалоговый агент	192
Использование инструментов	192
LLM, обученные использовать инструменты	194
Руководство по определению инструментов	203
Рассуждение	206
Цепочка рассуждений	206
ReAct: итеративное рассуждение и действие	208
За рамками ReAct	211
Контекст для взаимодействий в рамках задачи	212
Источник контекста	213
Выбор и организация контекста	215
Создание диалогового агента	217
Управление беседами	217
Пользовательский опыт	221
Резюме	224
Глава 9. Рабочие потоки LLM	225
Достаточно ли диалогового агента?	227
Базовые рабочие потоки LLM	230
Задачи	231
Сборка рабочего потока	239
Пример рабочего потока: продвижение плагинов для Shopify	242
Продвинутые рабочие потоки на базе LLM	245
Разрешение LLM-агенту управлять рабочим потоком	246
Агенты, хранящие состояния задач	247
Роли и делегирование работы	248
Резюме	250
Глава 10. Оценка приложений на базе LLM	251
Что мы вообще будем тестировать	252
Офлайн-оценка	253
Набор примеров	253
Поиск примеров	257
Оценка решений	259
SOMA-анализ	264
Онлайн-оценка	268
А/В-тестирование	269
Метрики	270
Резюме	273

Глава 11. Обозримое будущее	274
Мультиmodalность.....	274
Пользовательский опыт и интерфейс	276
Интеллект	278
Резюме	280
Об авторах	283
Иллюстрация на обложке.....	284