

---

# Промт-инжиниринг для GenAI

*Паттерны надежных запросов  
для качественных результатов*

*Джеймс Феникс, Майк Тейлор*

**SPRINT**  
book

2025

Выпущено  
при поддержке

**КРОК**

---

# Краткое содержание

Предисловие .....	14
<b>Глава 1.</b> Пять принципов промтинга .....	21
<b>Глава 2.</b> Введение в большие языковые модели .....	61
<b>Глава 3.</b> Стандартные приемы генерации текста с помощью ChatGPT.....	77
<b>Глава 4.</b> Продвинутое приемы генерации текста с помощью LangChain .....	150
<b>Глава 5.</b> Векторные базы данных с использованием FAISS и Pinecone.....	213
<b>Глава 6.</b> Автономные агенты с памятью и инструментами.....	253
<b>Глава 7.</b> Введение в диффузионные модели генерирования изображений.....	301
<b>Глава 8.</b> Стандартные практики создания изображений с помощью Midjourney .....	316
<b>Глава 9.</b> Продвинутое способы создания изображений с помощью Stable Diffusion.....	352
<b>Глава 10.</b> Создание приложений на базе ИИ.....	403
Об авторах .....	428
Иллюстрация на обложке.....	429

---

# Оглавление

Отзывы о книге.....	12
Предисловие.....	14
Требования к программному обеспечению для работы с книгой.....	15
Условные обозначения.....	18
Использование примеров кода.....	19
Благодарности.....	19
От издательства.....	20
<b>Глава 1. Пять принципов промтинга.....</b>	<b>21</b>
Обзор пяти принципов промтинга.....	25
1. Задайте направление.....	28
2. Укажите формат.....	34
3. Приведите примеры.....	37
4. Оцените качество.....	41
5. Разделите задачу на подзадачи.....	53
Резюме.....	60
<b>Глава 2. Введение в большие языковые модели.....</b>	<b>61</b>
Что такое модели генерации текста.....	61
Векторные представления: числовая сущность языка.....	62
Архитектура трансформеров: организация контекстных отношений.....	63
Вероятностная генерация текста: механизм принятия решений.....	65
Исторические предпосылки: появление архитектуры трансформеров.....	66
Предварительно обученные генеративные трансформеры OpenAI.....	68
GPT-3.5-turbo и ChatGPT.....	69
GPT-4.....	71
Google Gemini.....	71
Meta Llama и открытый исходный код.....	72
Использование квантования и LoRA.....	73
Mistral.....	74
Anthropic: Claude.....	74
GPT-4V(ision).....	75
Сравнение моделей.....	75
Резюме.....	76

<b>Глава 3. Стандартные приемы генерации текста с помощью ChatGPT.....</b>	<b>77</b>
Генерирование списков.....	77
Генерирование иерархического списка.....	79
Когда следует избегать регулярных выражений.....	85
Генерирование данных в формате JSON.....	85
YAML.....	87
Фильтрация текста в формате YAML.....	89
Обработка недопустимых полезных данных в YAML.....	90
Генерирование разных форматов с помощью ChatGPT.....	93
Пример данных CSV.....	94
Объясни это пятилетнему ребенку.....	95
Универсальный перевод с помощью LLM.....	96
Возможность запросить контекст.....	98
Разделение текстовых стилей.....	101
Выявление текстовых характеристик.....	102
Создание нового контента с использованием извлеченных характеристик.....	103
Извлечение текстовых характеристик с помощью LLM.....	104
Обобщение.....	104
Обобщение с учетом ограниченного окна контекста.....	105
Разбиение текста на части.....	107
Преимущества разбиения текста на части.....	108
Сценарии разбиения текста на части.....	108
Пример плохого разбиения.....	109
Стратегии разбиения на части.....	110
Обнаружение предложений с использованием spaCy.....	112
Создание простого алгоритма разбиения на Python.....	113
Разбиение методом скользящего окна.....	115
Пакеты для разбиения текста на части.....	117
Разбиение текста на части с помощью tiktoken.....	117
Кодировки.....	118
Что такое токенизация строк.....	118
Оценка числа токенов, используемых в вызовах Chat API.....	119
Анализ настроек.....	122
Методы улучшения анализа настроек.....	123
Ограничения и сложности анализа настроек.....	124
От меньшего к большему.....	124
Планирование архитектуры.....	125
Реализация отдельных функций.....	125
Добавление тестов.....	126
Преимущества метода «от меньшего к большему».....	127
Сложности метода «от меньшего к большему».....	128

Рольевые запросы .....	128
Преимущества рольевых запросов .....	130
Сложности рольевых запросов .....	130
Когда использовать рольевые запросы.....	130
Тактика запросов к GPT .....	131
Как избежать галлюцинаций с помощью референсов.....	131
Предоставление модели GPT «времени на размышления» .....	133
Тактика внутреннего монолога .....	134
Оценка моделью своих собственных ответов.....	136
Классификация с помощью LLM.....	138
Построение модели классификации .....	139
Классификация голосованием.....	140
Критерии оценки.....	141
Метапромтинг.....	145
Резюме .....	149
<b>Глава 4. Продвинутое приемы генерации текста с помощью LangChain .....</b>	<b>150</b>
Введение в LangChain .....	150
Настройка среды.....	152
Модели чатов .....	154
Потоковые модели чата .....	155
Создание нескольких ответов от LLM.....	156
Шаблоны запросов в LangChain.....	157
Язык выражений LangChain .....	157
Использование PromptTemplate с моделями чата .....	159
Парсеры вывода .....	160
Оценивание с помощью LangChain.....	164
Вызов функций OpenAI.....	172
Параллельный вызов функций.....	175
Вызов функций в LangChain.....	177
Извлечение данных с помощью LangChain .....	179
Планирование запросов.....	179
Создание шаблонов запросов с несколькими примерами .....	181
Включение нескольких примеров фиксированной длины.....	181
Форматирование примеров .....	182
Выбор нескольких примеров по длине.....	183
Ограничения обучения с несколькими примерами.....	185
Сохранение и загрузка подсказок LLM .....	186
Включение данных .....	187
Загрузчики документов.....	189

Инструменты для разбиения текста .....	192
Разбиение текста по количеству символов и токенов .....	193
Рекурсивное разбиение по символам .....	194
Декомпозиция задач .....	196
Объединение запросов в цепочки .....	198
Последовательная цепочка .....	199
itemgetter и извлечение ключей словаря .....	201
Структурирование цепочек LCEL .....	206
Цепочки документов .....	207
Наполнение .....	210
Уточнение .....	210
Отображение и свертка (Map Reduce) .....	210
Отображение с повторным ранжированием (Map Re-rank) .....	211
Резюме .....	212
<b>Глава 5. Векторные базы данных с использованием FAISS и Pinecone</b> .....	<b>213</b>
Генерация с дополненной выборкой .....	216
Введение в эмбединги .....	218
Загрузка документа .....	227
Извлечение данных из памяти с помощью FAISS .....	230
Реализация RAG с использованием LangChain .....	235
Векторные базы данных на хостинге от Pinecone .....	236
Собственные запросы .....	245
Альтернативные механизмы извлечения .....	250
Резюме .....	252
<b>Глава 6. Автономные агенты с памятью и инструментами</b> .....	<b>253</b>
Метод рассуждения с цепочкой мысли .....	253
Агенты .....	255
Рассуждения и действие (Reason and Act, ReAct) .....	258
Реализация паттерна ReAct .....	260
Использование инструментов .....	266
Использование LLM в качестве API (функции OpenAI) .....	269
Сравнение типов агентов ReAct и функции OpenAI .....	273
Варианты использования функций OpenAI .....	274
ReAct .....	274
Варианты использования ReAct .....	275
Наборы инструментов для агентов .....	275
Настройка стандартных агентов .....	277
Пользовательские агенты в LCEL .....	279

Память и ее использование .....	281
Долгосрочная память .....	281
Краткосрочная память .....	282
Краткосрочная память в диалоговых агентах типа вопрос-ответ.....	283
Память в LangChain .....	283
Сохранение состояния .....	284
Запрос состояния.....	285
ConversationBufferMemory.....	285
Другие популярные типы памяти в LangChain.....	288
ConversationBufferWindowMemory .....	288
ConversationSummaryMemory .....	288
ConversationSummaryBufferMemory .....	289
ConversationTokenBufferMemory.....	289
Агент функций OpenAI с памятью.....	290
Продвинутое фреймворки для агентов .....	292
Агенты планирования и выполнения.....	292
Дерево рассуждений .....	293
Обратные вызовы .....	295
Глобальные обратные вызовы (в вызове конструктора).....	297
Обратные вызовы, ограниченные запросами.....	297
Аргумент verbose .....	297
Когда и какой подход использовать .....	298
Подсчет токенов с помощью LangChain.....	298
Резюме .....	300
<b>Глава 7.</b> Введение в диффузионные модели генерирования изображений.....	301
OpenAI DALL-E.....	304
Midjourney .....	307
Stable Diffusion.....	310
Google Gemini .....	312
Преобразование текста в видео .....	313
Сравнение моделей .....	313
Резюме .....	314
<b>Глава 8.</b> Стандартные практики создания изображений с помощью Midjourney .....	316
Управление форматом.....	316
Управление художественным стилем .....	320
Реконструирование запросов.....	322
Усилители качества .....	323
Отрицательные запросы .....	324
Взвешенные понятия.....	327

Запросы с изображениями.....	330
Изменение фрагментов изображений.....	333
Дорисовывание обрамления изображений.....	336
Получение разных изображений одного персонажа.....	338
Переделка запросов.....	340
Выделение мемов.....	342
Картирование мемов.....	347
Анализ запросов.....	350
Резюме.....	351
<b>Глава 9. Продвинутые способы создания изображений с помощью Stable Diffusion.....</b>	<b>352</b>
Запуск Stable Diffusion.....	352
Веб-интерфейс AUTOMATIC1111.....	360
Img2Img.....	367
Увеличение разрешения изображений.....	370
Реконструирование запроса по изображению.....	373
Изменение фрагментов и добавление обрамления.....	373
ControlNet.....	377
Segment Anything Model (SAM).....	387
Тонкая настройка DreamBooth.....	390
Уточняющая модель в Stable Diffusion XL.....	397
Резюме.....	401
<b>Глава 10. Создание приложений на базе ИИ.....</b>	<b>403</b>
Разработка блога на основе ИИ.....	403
Исследуемая тема.....	405
Интервью с экспертом.....	407
Генерирование плана публикации.....	409
Генерирование текста.....	411
Стиль письма.....	414
Оптимизация заголовков.....	417
Генерирование иллюстраций для блога.....	418
Пользовательский интерфейс.....	424
Резюме.....	426
Об авторах.....	428
Иллюстрация на обложке.....	429