

Эффективный разговорный ИИ

СОЗДАЕМ ЧАТ-БОТОВ, КОТОРЫЕ
ДЕЙСТВИТЕЛЬНО РАБОТАЮТ

ЭНДРЮ ФРИД, КАРИ ДЖЕЙКОБС, ЭНИКО РОЖА
ПРЕДИСЛОВИЕ ХЕСУСА МАНТАСА



Санкт-Петербург • Москва • Минск

2026

Краткое содержание

Часть 1

Как усовершенствовать разговорный ИИ

Глава 1. Как работает разговорный ИИ?	30
Глава 2. Построение разговорного ИИ	54
Глава 3. Планирование улучшений	77

Часть 2

Паттерн: ИИ не понимает пользователя

Глава 4. Понимание потребностей пользователей	116
Глава 5. Улучшаем понимание традиционного ИИ	147
Глава 6. Повышение качества ответов с помощью RAG	179
Глава 7. Дополнение намерений с помощью генеративного ИИ	219

Часть 3

Паттерн: ИИ слишком сложный

Глава 8. Оптимизация сложных сценариев	244
Глава 9. Влияние контекста на адаптивность виртуального помощника	260
Глава 10. Снижение сложности системы с помощью генеративного ИИ	289

Часть 4

Паттерн: разрешение конфликтов

Глава 11. Сокращение числа отказов	310
Глава 12. Резюмирование при передаче диалога оператору	341

Оглавление

От издательства	15
О научном редакторе русского издания	15
Предисловие	16
Введение.....	18
Благодарности.....	20
О книге.....	23
Для кого эта книга	23
Структура книги	23
О коде в книге	25
Форум liveBook	25
Об авторах	26
Иллюстрация на обложке	28

Часть 1

Как усовершенствовать разговорный ИИ

Глава 1. Как работает разговорный ИИ?	30
1.1. Введение в разговорный ИИ.....	31
1.1.1. Для чего нужен разговорный ИИ?	32
1.1.2. Как работает разговорный ИИ?.....	34
1.1.3. Как построить разговорный ИИ.....	35
1.2. Генеративный ИИ в разговорных системах: основные понятия	39
1.2.1. Что такое генеративный ИИ.....	39
1.2.2. Ограничители генеративного ИИ.....	40

1.2.3. Эффективное использование генеративного ИИ в разговорных системах	42
1.3. Непрерывное улучшение разговорных ИИ	45
1.3.1. Зачем нужно непрерывное улучшение	45
1.3.2. Цикл непрерывного улучшения	46
1.3.3. Объясняем концепцию непрерывного улучшения стейкхолдерам	49
1.4. Начало работы	52
Итоги	53
Глава 2. Построение разговорного ИИ	54
2.1. Создание FAQ-бота	55
2.1.1. Основа FAQ-бота	55
2.1.2. Статический режим вопросов и ответов	57
2.1.3. Динамические ответы на вопросы	63
2.2. Маршрутизирующие агенты и процессно-ориентированные боты	66
2.2.1. Маршрутизирующие агенты	66
2.2.2. От маршрутизирующего агента к процессно- ориентированному боту	68
2.3. Ответы пользователю с помощью генеративного ИИ	71
2.3.1. Интеграция с LLM	71
2.3.2. Маршрутизация запросов к LLM	74
Итоги	76
Глава 3. Планирование улучшений	77
3.1. Когда необходимо улучшение	78
3.2. Ваша межфункциональная команда	80
3.3. Движение к общей цели	83
3.3.1. Пересмотр бизнес-целей	84
3.3.2. Эффективность	87
3.3.3. Охват	97
3.4. Выявление и устранение проблем	100
3.4.1. Поиск проблем	100
3.4.2. Групповой обзор	104
3.4.3. Определение критериев приемлемости	109

3.5. Разработка и внедрение исправлений	112
3.5.1. Планирование спринта.....	112
3.5.2. Повторная оценка	112
Итоги	113

Часть 2

Паттерн: ИИ не понимает пользователя

Глава 4. Понимание потребностей пользователей.....	116
4.1. Основы понимания.....	117
4.1.1. Последствия слабого понимания	117
4.1.2. Причины слабого понимания	118
4.1.3. Как обеспечить понимание в традиционном разговорном ИИ?	120
4.1.4. Как обеспечить понимание в генеративном ИИ?.....	122
4.2. Как измерить понимание?	124
4.2.1. Измерение понимания для традиционного (основанного на классификации) ИИ.....	125
4.2.2. Измерение понимания для генеративного ИИ	127
4.2.3. Измерение понимания с помощью обратной связи от пользователей	129
4.3. Оценка текущего состояния системы	129
4.3.1. Оценка традиционного (основанного на классификации) ИИ-решения	129
4.3.2. Оценка генеративного ИИ-решения.....	131
4.4. Извлечение тестовых данных из логов и их подготовка	131
4.4.1. Извлечение продакшен-логов	132
4.4.2. Рекомендации по отбору кандидатных тестовых высказываний.....	133
4.4.3. Подготовка и очистка данных для использования в итеративных улучшениях	138
4.4.4. Процесс разметки.....	139
4.5. Что говорят данные?	142
4.5.1. Анализ размеченных логов для традиционного ИИ (классификатора).....	142
4.5.2. Анализ размеченных логов для генеративного ИИ.....	144

4.5.3. Необходимость итеративного улучшения	144
Итоги.....	145
Глава 5. Улучшаем понимание традиционного ИИ.....	147
5.1. План улучшений.....	148
5.1.1. Определение проблемных паттернов в неправильно интерпретированных высказываниях.....	148
5.1.2. Поэтапные улучшения	153
5.1.3. С чего начать: выявление самых серьезных проблем.....	153
5.2. Решение проблемы «намерение неверное».....	159
5.2.1. Улучшение показателя полноты для одного намерения	159
5.2.2. Улучшение показателя точности для одного намерения	161
5.2.3. Улучшение оценки F1 для одного намерения	163
5.2.4. Улучшение показателей точности и полноты для нескольких намерений.....	164
5.3. Решение проблемы «намерение не найдено»	168
5.3.1. Определение высказываний для новых намерений.....	169
5.3.2. Когда лучше перестать добавлять намерения.....	173
5.4. Дополнение традиционного ИИ генеративным контентом.....	175
5.4.1. Сочетание традиционного и генеративного ИИ при обработке одного намерения	176
5.4.2. Промпты для демонстрации понимания	177
Итоги.....	178
Глава 6. Повышение качества ответов с помощью RAG.....	179
6.1. За пределами намерений: роль поиска в разговорном ИИ	180
6.1.1. Использование поиска в системах разговорного ИИ	181
6.1.2. Преимущества традиционного поиска	183
6.1.3. Недостатки традиционного поиска	184
6.2. За пределами поиска: генерация ответов с помощью RAG.....	185
6.2.1. Использование RAG в системах разговорного ИИ.....	185
6.2.2. Преимущества RAG.....	187
6.2.3. Сочетание RAG с другими моделями генеративного ИИ.....	190
6.2.4. Сравнение подходов: намерения, поиск и RAG.....	191

6.3. Реализация RAG.....	192
6.3.1. Общая схема реализации.....	192
6.3.2. Подготовка базы документов для RAG	194
6.4. Дополнительные аспекты реализации RAG	197
6.4.1. Почему нельзя просто использовать LLM напрямую?	197
6.4.2. Поддержание актуальности и релевантности ответов с помощью RAG	198
6.4.3. Как настроить пайплайн загрузки данных?	199
6.4.4. Управление задержками	204
6.4.5. Когда использовать резервный механизм и когда выполнять поиск	205
6.5. Оценка и анализ эффективности RAG	206
6.5.1. Метрики индексации	207
6.5.2. Метрики поиска.....	209
6.5.3. Метрики генерации	211
6.5.4. Сравнение эффективности решений для индексации и эмбедингов в RAG	213
Итоги.....	218
Глава 7. Дополнение намерений с помощью генеративного ИИ.....	219
7.1. Начало работы.....	220
7.1.1. Зачем это нужно: преимущества и недостатки.....	221
7.1.2. Что для этого нужно.....	222
7.1.3. Как использовать дополненные данные	223
7.2. Улучшение существующих намерений.....	225
7.2.1. Творческий подход к подбору синонимов.....	226
7.2.2. Генерация новых грамматических конструкций.....	230
7.2.3. Создание намерений на основе вывода LLM	233
7.2.4. Получение новых примеров с помощью шаблонов	237
7.3. Повышение креативности.....	239
7.3.1. Генерация дополнительных намерений	240
7.3.2. Проверка на наличие неоднозначностей	240
Итоги	242

Часть 3

Паттерн: ИИ слишком сложный

Глава 8. Оптимизация сложных сценариев.....	244
8.1. Бремя сложности.....	245
8.1.1. Влияние сложности на конечного пользователя	245
8.1.2. Влияние сложности на бизнес-метрики	247
8.1.3. Дополнительные издержки и выгоды от снижения сложности для пользователя.....	249
8.2. Упрощение и оптимизация пути пользователя	251
8.2.1. Определение сложных диалоговых сценариев.....	251
8.2.2. Использование уже известной информации о клиенте.....	251
8.2.3. Соответствие ожиданиям пользователя	253
8.2.4. Гибкость в ожидаемых ответах пользователя	254
8.2.5. Поддержка сценариев самообслуживания с помощью API и бэкэнд-процессов.....	257
Итоги.....	259
Глава 9. Влияние контекста на адаптивность виртуального помощника.....	260
9.1. Роль контекста в работе виртуальных помощников	261
9.1.1. Влияние контекста на взаимодействие с пользователем	263
9.1.2. Что такое контекстная информация?	267
9.2. Понятие модальности.....	272
9.2.1. Сравнение модальностей.....	273
9.2.2. Влияние модальности на сценарии виртуальных помощников	274
9.2.3. Примеры того, как модальность влияет на пользовательский опыт	276
9.2.4. Особенности проектирования голосовых ботов	278
9.3. Повышение контекстной осведомленности и улучшение общего пользовательского опыта с помощью RAG.....	280
9.3.1. Проектирование адаптивных потоков с использованием RAG.....	281
9.3.2. Стратегии поиска и генерации контекстно-релевантных ответов.....	283

9.3.3. Сопровождение и обновление адаптивных потоков	285
Итоги.....	287
Глава 10. Снижение сложности системы с помощью генеративного ИИ	289
10.1. Разработка процессных потоков с поддержкой ИИ.....	290
10.1.1. Генерация диалоговых потоков с помощью генеративного ИИ.....	291
10.1.2. Улучшение диалогового потока с помощью генеративного ИИ.....	295
10.2. Выполнение процессных потоков с поддержкой ИИ	297
10.2.1. Выполнение диалоговых потоков с помощью генеративного ИИ.....	297
10.2.2. Использование LLM для поиска	300
10.3. Тестирование потоков с помощью ИИ.....	303
10.3.1. Настройка генеративного ИИ для роли пользователя.....	304
10.3.2. Настройка тестирования диалога	306
Итоги	308

Часть 4

Паттерн: разрешение конфликтов

Глава 11. Сокращение числа отказов.....	310
11.1. Что вызывает отказы?	311
11.1.1. Причины мгновенных отказов	311
11.1.2. Причины более поздних отказов	312
11.1.3. Сбор данных об отказе	313
11.2. Сокращение числа мгновенных отказов.....	316
11.2.1. Начните с хорошего: приветствия и представления	316
11.2.2. Обозначьте возможности и установите ожидания.....	318
11.2.3. Подкрепляйте мотивацию	319
11.2.4. Задействуйте пользователя.....	320
11.3. Сокращение числа отказов, возникающих позже	322
11.3.1. Старайтесь понять пользователя	322
11.3.2. Старайтесь быть понятыми	322
11.3.3. Будьте гибкими и готовыми подстроиться.....	323
11.3.4. Показывайте прогресс.....	325

11.3.5. Предугадывайте потребности пользователя	325
11.3.6. Будьте вежливы	326
11.4. Удержание пользователя при отказе	327
11.4.1. Начните со сбора данных об отказе	328
11.4.2. Реализация потока удержания при отказе	328
11.5. Улучшение диалога с помощью генеративного ИИ	331
11.5.1. Улучшение сообщений об ошибках с помощью генеративного ИИ	332
11.5.2. Улучшение приветствия с помощью генеративного ИИ	334
11.6. Иногда эскалация необходима	339
Итоги	340
Глава 12. Резюмирование при передаче диалога оператору	341
12.1. Резюмирование: основные сведения	342
12.1.1. Зачем нужно резюмирование	342
12.1.2. Структура эффективного резюме	343
12.2. Подготовка чат-бота к резюмированию	347
12.2.1. Использование встроенных элементов	347
12.2.2. Инструментирование чат-бота для создания транскриптов	349
12.2.3. Инструментирование чат-бота (для сбора данных)	352
12.3. Улучшение резюме с помощью генеративного ИИ	354
12.3.1. Генерация текстового резюме транскрипта с помощью резюмирующих промптов	355
12.3.2. Генерация структурированного резюме транскрипта с помощью экстрактивных промптов	359
Итоги	364