

ИСПОЛЬЗУЕМ ИИ В СВОИХ ИНТЕРЕСАХ

ЧАСТЬ 2

Практическое руководство по работе с

Qwen, **GigaChat**, **YandexGPT** и **Kandinsky**



Автор:

НИКИТА КУЛИКОВ

Никита Куликов

Используем ИИ в своих интересах. Часть 2

<https://litres.ru/74584248>

SelfPub; 2026

Аннотация

Вы открываете ChatGPT. Пишете запрос. Получаете ответ. Он неплохой, но не совсем то, что нужно. Вы переписываете запрос. Ответ становится лучше, но вы уже потратили двадцать минут на то, что должно было занять пять.

Или другая ситуация. Вы слышали, что DeepSeek дешевле и быстрее. Пробуете там тот же запрос – и получаете совершенно другой результат. Не хуже, но иначе. Вы не понимаете, почему.

Или третья. Вы загружаете договор в Gemini, но модель путается в деталях. Вы пробуете Claude – и получаете точный анализ за минуту. Почему?

Ответ на все три вопроса один: разные модели устроены по-разному, и работать с ними нужно по-разному.

Используем ИИ в своих интересах - это серия книг которая обучит вас понимать ключевые принципы работы с ИИ. В этой части вы получите практическое руководство по работе с Qwen, GigaChat, YandexGPT, Kandinsky.

После прочтения вы будете знать:

- какую систему выбрать для конкретной задачи;
- как формулировать запросы;
- какие модели в чем сильны;

Содержание

ВВЕДЕНИЕ	6
Часть I. Ландшафт: что доступно в России	9
Глава 1. Четыре категории инструментов	9
Глава 2. Что изменилось в 2026 году	15
Глава 3. Как выбирать систему в российских условиях	21
Глава 4. Что такое Qwen и почему он стал основным инструментом в России	36
Глава 5. Как получить доступ к Qwen	40
Глава 7. Как формулировать запросы в Qwen	55
Конец ознакомительного фрагмента.	60

Никита Куликов

Используем ИИ в своих

интересах. Часть 2

ТИТУЛЬНЫЙ ЛИСТ

Используем ИИ в своих интересах.
Часть 2

**Практическое руководство по работе с Qwen,
GigaChat, YandexGPT, Kandinsky**

Никита Куликов

ВВЕДЕНИЕ

В первой книге мы разбирали DeepSeek, ChatGPT, Claude и Gemini. Это глобальные лидеры. Но есть проблема, о которой редко пишут в руководствах: **большинство из них в России официально недоступны.**

ChatGPT и Claude не работают без VPN. Для доступа нужна зарубежная карта. Это неудобно. Это рискованно. Это не для всех.

Gemini тоже требует обхода. Даже если вы найдёте способ подключиться – вы нарушаете условия использования. Ваш аккаунт могут заблокировать в любой момент.

Но это не значит, что ИИ в России недоступен. Это значит, что ландшафт другой. Другие инструменты. Другие правила. Другие возможности. Эта книга – о том, что реально работает. Не в теории. На практике.

Оглавление

ВВЕДЕНИЕ

Часть I. Ландшафт: что доступно в России

Глава 1. Четыре категории инструментов

Глава 2. Что изменилось в 2026 году

Глава 3. Как выбирать систему в российских условиях

Часть II. Qwen: китайская модель, которая говорит по-русски

Глава 4. Что такое Qwen и почему он стал основным инструментом в России

Глава 5. Как получить доступ к Qwen

Глава 6. Выбор модели: Max, Plus, Flash, Long

Глава 7. Как формулировать запросы в Qwen

Глава 8. Сценарии использования Qwen

Глава 9. Ограничения Qwen и как их обходить

Часть III. GigaChat – модель, которая работает по российским правилам

Глава 10. Что такое GigaChat и почему он важен

Глава 11. Как получить доступ к GigaChat

Глава 12. Линейка моделей: Lite, Pro, Max, Ultra

Глава 13. Как формулировать запросы в GigaChat

Глава 14. Сценарии использования GigaChat

Глава 15. Ограничения GigaChat и как их обходить

Часть IV. YandexGPT. Почему он есть везде

Глава 16. Что такое YandexGPT и почему он везде

Глава 17. Как получить доступ к YandexGPT

Глава 18. Линейка моделей: Lite, Pro, Alice AI

Глава 19. Как формулировать запросы в YandexGPT

Глава 20. Сценарии использования YandexGPT

Глава 21. Ограничения YandexGPT и как их обходить

Часть V. Kandinsky: генерация изображений и видео

Глава 22. Что такое Kandinsky и почему он важен

Глава 23. Как получить доступ к Kandinsky

Глава 24. Возможности Kandinsky

Глава 25. Как формулировать запросы для Kandinsky

Глава 26. Сценарии использования Kandinsky

Глава 27. Ограничения Kandinsky и как их обходить

Часть VI. Комбинирование российских и китайских моделей

Глава 28. Зачем комбинировать Qwen, GigaChat, YandexGPT и Kandinsky

Глава 29. Четыре паттерна комбинирования

Глава 30. Практические примеры комбинирования

Глава 31. Как построить собственный пайплайн

Глава 32. Будущее комбинирования

Заключение.

Приложение А. Сравнительная таблица: Qwen, GigaChat, YandexGPT, Kandinsky

Приложение Б. Чек-лист перед началом работы

Приложение В. Словарь терминов

Приложение Г. Шаблоны промптов для русскоязычных моделей

Приложение Д. FAQ: доступ, оплата, 152-ФЗ

Часть I. Ландшафт: что доступно в России

Глава 1. Четыре категории инструментов

1.1. Почему картина изменилась

Первое: китайские модели стали доступны.

DeepSeek и Qwen работают в России без VPN. Они понимают русский язык. Они бесплатны или почти бесплатны. Они не требуют зарубежных карт.

По данным опроса, проведённого в 2026 году, китайские модели стали основным инструментом для российских разработчиков, программистов и аналитиков. DeepSeek называют «бесплатным инструментом номер один» для тех, кто работает с кодом и данными.

Второе: российские модели догнали.

GigaChat и YandexGPT перестали быть «игрушками». Они стали полноценными инструментами для бизнеса. Они работают на серверах в России. Они соответствуют 152-ФЗ. Они принимают оплату в рублях.

GigaChat 3.5 Ultra, представленный в 2026 году, почти

вдвое компактнее предыдущей версии и до четырёх раз быстрее работает с объёмными документами. YandexGPT 5.1 Pro лучше справляется с задачами, работает с системным промптом и избегает ошибок.

Третье: агрегаторы упростили доступ.

Российские операторы связи и облачные платформы начали предлагать доступ к западным моделям через свои сервисы. «МегаФон», «Билайн», МТС, Cloud.ru. Без VPN. С русскоязычной поддержкой. С оплатой в рублях.

1.2. Четыре категории

Все инструменты, доступные в России, делятся на четыре категории.

Категория первая: китайские модели. DeepSeek. Qwen. GLM. Kimi. Они работают без VPN. Они бесплатны или очень дешевы. Они понимают русский язык. Они стали основным инструментом для российских разработчиков программного обеспечения.

Категория вторая: российские модели. GigaChat (Сбер). YandexGPT (Яндекс). Они работают на серверах в России. Они принимают оплату в рублях. Они соответствуют 152-ФЗ. Они созданы для российского рынка.

Категория третья: западные модели через агрегаторы. ChatGPT. Claude. Gemini. Midjourney. Они доступны через российских операторов связи и облачные платформы. «МегаФон», «Билайн», МТС, Cloud.ru предлагают доступ к

ним через свои сервисы.

Категория четвёртая: специализированные инструменты. Kandinsky. Шедеврум. GigaCode. SourceCraft. Они решают конкретные задачи: генерация изображений, написание кода, создание презентаций.

1.3. Китайские модели: главный выбор разработчиков

Почему российские разработчики выбрали китайские модели?

Причина первая: доступность. DeepSeek и Qwen не требуют VPN. Они не требуют зарубежных карт. Они не требуют сложной регистрации. Вы открываете сайт – и работаете.

Причина вторая: цена. DeepSeek V4-Flash стоит около \$0,14 за миллион входных токенов. Qwen3.5 Flash – \$0,10. Для сравнения: GPT-5.4 стоит в разы дороже.

Причина третья: качество. DeepSeek V3.2 показывает результаты, близкие к GPT-5, в задачах кодирования и математики. Qwen3.5 397B конкурирует с GPT-5.4 на бенчмарках. Разрыв сокращается быстрее, чем разрыв в цене.

Причина четвёртая: русский язык. Qwen работает с русским и китайским текстом лучше, чем DeepSeek. На структурированных задачах – извлечение данных, классификация, суммаризация – результат стабильный.

Причина пятая: открытость. DeepSeek и Qwen – от-

крытые модели. Их можно развернуть локально. Их можно дообучить. Их можно адаптировать под конкретные задачи.

1.4. Российские модели: для бизнеса и государства

Почему российские компании выбирают GigaChat и YandexGPT?

Причина первая: 152-ФЗ. Закон требует, чтобы персональные данные россиян хранились на серверах в России. GigaChat и YandexGPT соответствуют этому требованию. Китайские модели – нет.

Причина вторая: интеграция с российскими сервисами. GigaChat интегрирован с экосистемой Сбера. YandexGPT – с экосистемой Яндекса. Это упрощает работу.

Причина третья: поддержка. Русскоязычная документация. Русскоязычная поддержка. Оплата в рублях.

Причина четвёртая: цены. GigaChat Lite стоит 65 рублей за миллион токенов. YandexGPT Lite – от 0,1 рубля за тысячу токенов. Это конкурентные цены.

1.5. Западные модели через агрегаторы

Если вам нужны именно ChatGPT или Claude – есть путь.

«МегаФон Нейросети». «МегаФон» запустил сервис, который даёт доступ к 25 зарубежным моделям через веб-интерфейс. Без VPN. Без зарубежной карты. С русскоязычной поддержкой.

Cloud.ru. Облачная платформа предлагает доступ к мо-

делям Alibaba, DeepSeek, Z.ai через единый API. Плюс – к западным моделям через партнёров.

МТС, «Билайн». Операторы связи также предлагают доступ к ИИ-сервисам через свои платформы.

1.6. Специализированные инструменты

Не всё можно сделать текстовой моделью. Иногда нужны специализированные инструменты.

Kandinsky. Генерация изображений и видео. Сбер + AIRI. Бесплатно. Русскоязычные запросы. Интегрирован в GigaChat.

Шедеврум. Генерация изображений и коротких видео. Яндекс. На базе YandexART. До 70 картинок в день на сайте.

GigaCode. ИИ-помощник для программистов. Сбер. Русскоязычный аналог GitHub Copilot. Поддерживает 30 языков.

SourceCraft. ИИ-помощник для программистов. Яндекс. Дописывает код, ищет ошибки, генерирует тесты.

1.7. Что это значит для вас

Если вы разработчик или программист: DeepSeek и Qwen – ваши основные инструменты. Бесплатно. Без VPN. Понимают русский. DeepSeek V3.2 показывает результаты, близкие к GPT-5, в задачах кодирования.

Если вы бизнес: GigaChat и YandexGPT – для задач, где важны 152-ФЗ и хранение данных в России. Для остального

– китайские модели через облачные платформы.

Если вы обычный пользователь: GigaChat через Telegram-бот. YandexGPT через Алису. DeepSeek через сайт. Все бесплатно. Все работает без VPN.

Если вам нужны западные модели: «МегаФон», «Билайн», МТС, Cloud.ru. Они дают доступ к ChatGPT, Claude, Gemini через свои сервисы.

1.8. Что будет дальше

Российский рынок ИИ быстро меняется.

Тренд первый: китайские модели становятся основными. MTS Web Services сообщила, что клиенты сгенерировали 400 миллиардов токенов через китайские модели за первую половину 2026 года. Это в 11 раз больше, чем за весь 2025 год.

Тренд второй: российские модели догоняют. GigaChat 3.5 Ultra и YandexGPT 5.1 Pro улучшили качество. Они становятся конкурентными не только для бизнеса, но и для обычных пользователей.

Тренд третий: агрегаторы упрощают доступ. «МегаФон», Cloud.ru, МТС. Они дают доступ к моделям, которые раньше были недоступны.

Тренд четвёртый: специализация. Появляются модели для конкретных задач: код, изображения, видео, аудио.

Глава 2. Что изменилось в 2026 году

2.1. Год, когда всё сдвинулось

К середине 2026 года картина изменилась радикально. Китайские модели стали основным инструментом для российских компаний. Российские модели догнали по качеству. Агрегаторы упростили доступ к западным решениям. И самое главное – ИИ перестал быть экзотикой. Он стал рабочим инструментом.

2.2. Китайские модели: взрывной рост

Самое заметное изменение 2026 года – массовый переход российского бизнеса на китайские языковые модели.

Цифры, которые говорят сами за себя: В MTS Web Services зафиксировали рост использования китайских нейросетей **в 11 раз** – до 400 миллиардов токенов за первую половину 2026 года. Для сравнения: за весь 2025 год этот показатель составлял 39,1 миллиарда токенов.

На платформе Yandex AI Studio объём потребления вырос **в 18 раз** и достиг 597 миллиардов токенов. Второе и третье места среди зарубежных моделей заняли Qwen (21,2%) и DeepSeek (13,5%).

В Cloud.ru потребление увеличилось **в 18 раз** за первые восемь месяцев 2026 года.

Кто лидирует: Qwen – безусловный лидер. 261,1 миллиарда токенов от российских пользователей. Далее следуют GLM (91,2 миллиарда) и Kimi (81,4 миллиарда) .

Почему это произошло: Причина первая – цена. «Для абсолютного большинства прикладных задач разница в качестве между флагманом за условный доллар и китайской моделью за два цента либо не видна, либо не важна», – отмечают участники рынка.

Причина вторая – доступность. Китайские модели работают без VPN. Они не требуют зарубежных карт. Их можно развернуть на российской инфраструктуре.

Причина третья – безопасность. «Основной формат работы – развёртка модели на российской инфраструктуре. При таком формате данные остаются в контуре компании», – говорит директор ИИ-вертикали Selectel Александр Тугов.

2.3. Российские модели: догоняют

Пока китайские модели захватывали рынок, российские разработчики не стояли на месте.

GigaChat 3.5 Ultra. Сбер представил флагманскую модель в 2026 году. Она почти вдвое компактнее предыдущей версии и до четырёх раз быстрее работает с объёмными документами. Контекстное окно – 128 000 токенов.

Важное преимущество: GigaChat – единственная модель, у которой есть всё сразу: место в независимом рейтинге MERA, опубликованный прайс, наличие в реестре россий-

ского ПО и работа как через облако, так и локально.

YandexGPT 5.1 Pro / Alice AI. Яндекс провел ребрендинг YandexGPT в Alice AI в октябре 2025 года. В июне 2026 модель обновилась – теперь она умеет запоминать факты из предыдущих диалогов и подстраиваться под стиль общения. Контекстное окно – 32 000 токенов.

Открытые веса YandexGPT-5-Lite доступны на Hugging Face. По данным карточки модели, она превосходит сопоставимые по размеру Llama и Qwen в знании русской культуры и фактов.

Остальные игроки. Cotype от МТС – в реестре российского ПО. A-Vibe от Авито и T-Pro от Т-Технологий – бесплатные открытые модели, но требуют собственных серверов.

2.4. Агрегаторы: доступ к западным моделям

Если вам нужны именно ChatGPT или Claude – в 2026 году это стало проще.

Российский рынок платформ, предоставляющих легальный корпоративный доступ к зарубежным языковым моделям, достиг **3 миллиардов рублей в год**.

«МегаФон Нейросети». Доступ к 25 зарубежным моделям через веб-интерфейс. Без VPN. Без зарубежной карты. С русскоязычной поддержкой.

Cloud.ru. Облачная платформа предлагает доступ к моделям Alibaba, DeepSeek, Z.ai через единый API. Плюс – к

западным моделям через партнёров.

Yandex AI Studio. Галерея моделей включает не только YandexGPT, но и DeepSeek, Qwen, GPT-OSS. Модели можно подключать к сайтам и приложениям по API.

MWS GPT Model Hub. MTS Web Services расширила каталог до 17 моделей. Среди них – GLM 5.2, Kimi K2.6, Qwen3.6, Gemma 4, GPT OSS. Все доступны через единый OpenAI-совместимый API.

2.5. 78% россиян уже используют ИИ

Самое важное изменение – не технологическое, а человеческое.

78% интернет-пользователей в России хотя бы раз работали с нейросетями. Почти **60% обращаются к ним минимум раз в неделю.**

Для сравнения: в начале 2026 года опыт взаимодействия с ИИ имели 73% опрошенных. Регулярность использования выросла с 51% до 58% за полгода.

Как люди используют ИИ:

- 31% – для поиска информации и быстрых ответов;
- 24% – для работы или учёбы;
- 18% – для создания текстов, изображений и презентаций;
- 15% – для планирования дел;
- 12% – для развлечений.

27% россиян используют нейросети для рабочих за-

дач. Среди них 40% обращаются к ИИ ежедневно, 43% – несколько раз в неделю.

Что это значит: ИИ перестал быть технологией для избранных. Он стал массовым инструментом. Если вы ещё не используете его – вы в меньшинстве.

2.6. Бизнес: от пилотов к внедрению

2026 год стал годом, когда российский бизнес перешёл от экспериментов к реальному внедрению.

86% крупных компаний полноценно используют или тестируют большие языковые модели. **79%** – интеллектуальных чат-ботов.

Половина опрошенных российских компаний планирует нарастить инвестиции в ИИ в ближайший год.

Где сосредоточен спрос: Финансовый сектор, телеком, ритейл, промышленность, госсектор. Основные сценарии: клиентский сервис, обработка документов, поиск по корпоративным знаниям, аналитика, разработка ПО.

Проблема, о которой не говорят: 80% ИИ-проектов в мире терпят неудачу. В России доля неудачных проектов может быть выше. Причины: неправильная постановка задач, дефицит качественных данных, отсутствие метрик эффективности.

2.7. Рынок: 19,6 миллиарда рублей

Российский рынок больших языковых моделей по итогам

2026 года вырастет на **35%** и составит около **19,6 миллиарда рублей**.

Облачный сегмент – предоставление доступа к LLM через SaaS и API – достигнет почти **1,5 миллиарда рублей**.

Вице-премьер Дмитрий Григоренко прогнозирует кратный рост рынка искусственного интеллекта до **770 миллиардов рублей к 2030 году**.

2.8. Что это значит для вас

Если вы ещё не используете ИИ: Вы в меньшинстве. 78% россиян уже работают с нейросетями. Начните с GigaChat через Telegram-бот. Или с DeepSeek через сайт. Бесплатно. Без VPN.

Если вы используете только одну модель: Попробуйте другие. Qwen для кода. GigaChat для бизнес-задач. YandexGPT для поиска. Kandinsky для изображений. Каждая модель имеет сильные стороны.

Если вы бизнес: Китайские модели – для задач, где важна цена и скорость. Российские – для задач, где важны 152-ФЗ и хранение данных в России. Комбинируйте.

Глава 3. Как выбирать систему в российских условиях

3.1. Почему стандартные критерии не работают

В первой книге мы разбирали три вопроса: цена ошибки, формат входа, бюджет. Эти критерии универсальны. Но в России к ним добавляются новые.

Первый – **доступность**. Можно ли вообще использовать эту систему? Без VPN? Без зарубежной карты?

Второй – **соответствие 152-ФЗ**. Можно ли обрабатывать персональные данные? Не нарушаете ли вы закон?

Третий – **оплата**. Можно ли заплатить рублями? Есть ли счёт для бухгалтерии?

Четвёртый – **приватность**. Где хранятся данные? Кто имеет к ним доступ?

Эти четыре фактора меняют картину. То, что теоретически лучше, может быть недоступно. То, что доступно, может не соответствовать требованиям. То, что соответствует, может не подходить по цене.

3.2. Фактор первый: доступность

Самый простой вопрос: **можете ли вы открыть этот сервис?**

Работают без VPN:

- GigaChat (Сбер);
- YandexGPT (Яндекс);
- DeepSeek;
- Qwen;
- Kandinsky;
- Шедеврум;
- GigaCode;
- SourceCraft.

Требуют VPN:

- ChatGPT;
- Claude;
- Gemini (напрямую).

Доступны через агрегаторы:

– ChatGPT, Claude, Gemini – через «МегаФон Нейросети», Cloud.ru, ProxyAPI, AITunnel, Polza.AI.

Практический вывод: Если вы не хотите возиться с VPN – используйте GigaChat, YandexGPT, DeepSeek, Qwen.

Если вам нужны именно ChatGPT или Claude – используйте агрегаторы. Это легально. Это удобно. Это дороже, но не настолько, чтобы отказываться.

3.3. Фактор второй: 152-ФЗ

Это самый важный фактор для бизнеса. И самый сложный.

Что требует закон: Персональные данные россиян должны храниться на серверах в России. Это требование ло-

кализации.

Любой, кто обрабатывает персональные данные – включая самозанятых – должен зарегистрироваться в реестре операторов Роскомнадзора .

Согласие на обработку данных оформляется как отдельный документ.

Штрафы за нарушения – серьёзные. От 100 000 до 300 000 рублей для юридических лиц за отсутствие уведомления. От 1 миллиона рублей – за невыполнение требований локализации.

Что это значит на практике: если вы используете GigaChat или YandexGPT – данные хранятся в России. Вы соответствуете 152-ФЗ. Если вы используете DeepSeek или Qwen – данные уходят в Китай. Вы не соответствуете 152-ФЗ. Для персональных данных это нарушение.

Если вы используете ChatGPT через агрегатор – данные уходят за границу. Даже если агрегатор российский, инференс происходит на серверах OpenAI.

Как обходить: для задач, связанных с персональными данными – используйте только GigaChat и YandexGPT. Для задач без персональных данных – используйте что угодно.

Если нужны китайские модели для персональных данных – разверните их локально. Тогда данные не покидают вашу инфраструктуру.

Пример: Вы – интернет-магазин. Обрабатываете заказы клиентов. Это персональные данные.

Плохо: использовать DeepSeek для анализа заказов. Данные уходят в Китай.

Хорошо: GigaChat. Данные остаются в России.

Или: развернуть DeepSeek локально. Данные не покидают ваш сервер.

3.4. Фактор третий: оплата

Можно ли заплатить? Как? Есть ли счёт?

Принимают рубли:

- GigaChat;
- YandexGPT;
- Kandinsky;
- DeepSeek (через российские платформы);
- Qwen (через российские платформы);
- агрегаторы (ProхуAPI, AITunnel, Polza.AI).

Требуют зарубежную карту:

- ChatGPT напрямую;
- Claude напрямую;
- Gemini напрямую.

Дают счёт для бухгалтерии:

- GigaChat (через cloud.ru);
- YandexGPT (через Yandex Cloud);
- Cloud.ru;
- MTS Web Services;
- ProхуAPI (для корпоративных клиентов);
- AITunnel (для корпоративных клиентов).

Практический вывод: если вы физическое лицо – ис-

пользуйте GigaChat, YandexGPT, DeepSeek. Бесплатно или дёшево. Оплата рублями.

Если вы бизнес – используйте GigaChat, YandexGPT через облачные платформы. Или агрегаторы для доступа к зарубежным моделям.

3.5. Фактор четвёртый: приватность

Где хранятся данные? Кто имеет к ним доступ?

Данные в России:

- GigaChat – серверы Сбера;
- YandexGPT – серверы Яндекса.

Данные в Китае:

- DeepSeek – серверы DeepSeek;
- Qwen – серверы Alibaba.

Данные уходят за границу:

- ChatGPT – серверы OpenAI (США);
- Claude – серверы Anthropic (США);
- Gemini – серверы Google (США).

Через агрегаторы: Данные проходят через серверы агрегатора, затем уходят к провайдеру модели. Где именно происходит инференс – зависит от агрегатора и модели.

Важно понимать: «Не логируем» на стороне агрегатора не означает, что никто в цепочке не обрабатывает запрос. Агрегатор – это транзит. Провайдер модели – это тот, кто видит данные.

Как обходить: Для конфиденциальных данных – ис-

пользуйте GigaChat и YandexGPT.

Для особо чувствительных данных – разверните модель локально. Для остального – используйте что угодно.

3.6. Четыре вопроса перед началом

Теперь у нас шесть вопросов. Три из первой книги. Три новых.

Старые вопросы из первой книги:

Какова цена ошибки?

Какой формат у входа?

Каков бюджет?

Новые вопросы которые мы добавляем:

Доступна ли система без VPN?

Соответствует ли она 152-ФЗ?

Можно ли оплатить рублями?

Что на это можно ответить:

Вопрос 1: цена ошибки.

Высокая → GigaChat Pro/Max, YandexGPT Pro, Claude через агрегатор.

Низкая → GigaChat Lite, YandexGPT Lite, DeepSeek, Qwen.

Вопрос 2: формат входа.

Текст → любая система.

Изображения → GigaChat (Kandinsky), Qwen-VL, Gemini

через агрегатор.

Видео → Kandinsky, Gemini через агрегатор.

Вопрос 3: бюджет.

Ограничен → DeepSeek, Qwen, GigaChat Lite.

Не ограничен → GigaChat Pro, YandexGPT Pro, Claude через агрегатор.

Вопрос 4: доступность.

Нужен VPN → ChatGPT, Claude, Gemini.

Без VPN → GigaChat, YandexGPT, DeepSeek, Qwen, Kandinsky.

Вопрос 5: 152-ФЗ.

Персональные данные → GigaChat, YandexGPT, локальное развёртывание.

Без персональных данных → любая система.

Вопрос 6: оплата.

Рубли → GigaChat, YandexGPT, DeepSeek, Qwen, агрегаторы.

Зарубежная карта → ChatGPT, Claude, Gemini.

3.7. Матрица выбора для России

Задача

Первый выбор

Альтернатива

Почему

Код

Qwen

DeepSeek

Бесплатно, без VPN, качество

Код-ревью

Claude через агрегатор

GigaChat Pro

Точность

Анализ договора

GigaChat Pro

Claude через агрегатор

152-ФЗ, качество

Массовая обработка

DeepSeek

Qwen Flash

Цена

Анализ изображений

Qwen-VL

GigaChat (Kandinsky)

Мультимо-дальность

Анализ видео

Kandinsky

Gemini через агрегатор

Доступность

Генерация изображений

Kandinsky

Шедевр

Бесплатно, русский язык

Написание текстов

GigaChat

YandexGPT

Русский язык, 152-ФЗ

Автоматизация

GigaChat (Function calling)

YandexGPT (RAG)

Интеграции

Научные исследования

Claude через агрегатор

GigaChat Pro

Воспроизводимость

Медицинские вопросы

Claude через агрегатор

GigaChat Pro

Надёжность

Юридические вопросы

GigaChat Pro

Claude через агрегатор

152-ФЗ

Бизнес-задачи

GigaChat

YandexGPT

152-ФЗ, интеграции

Повседневные вопросы

GigaChat (Telegram)

YandexGPT (Алиса)

Бесплатно, удобство

Критические решения

Claude через агрегатор

GigaChat Ultra

Максимальная надёжность

3.8. Примеры выбора

Пример первый: вы пишете код.

Задача: написать функцию.

Цена ошибки: низкая.

Формат: текст.

Бюджет: ограничен.

Доступность: без VPN.

152-ФЗ: не важно.

Оплата: рубли.

Выбор: Qwen. Бесплатно, без VPN, качество.

Пример второй: вы анализируете договор с персональными данными.

Задача: анализ договора.

Цена ошибки: высокая.

Формат: текст.

Бюджет: не критичен.

Доступность: без VPN.

152-ФЗ: важно.

Оплата: рубли.

Выбор: GigaChat Pro. 152-ФЗ, качество, доступность.

Пример третий: вы обрабатываете 10 000 отзывов.

Задача: классификация.

Цена ошибки: низкая.

Формат: текст.

Бюджет: ограничен.

Доступность: без VPN.

152-ФЗ: не важно.

Оплата: рубли.

Выбор: DeerpSeek. Почему? Цена, скорость, качество.

Пример четвёртый: вы создаёте изображения для блога.

Задача: генерация изображений.

Цена ошибки: низкая.

Формат: текст → изображение.

Бюджет: ограничен.

Доступность: без VPN.

152-ФЗ: не важно.

Оплата: бесплатно.

Выбор: Kandinsky. Бесплатно, русский язык, качество.

Пример пятый: вы анализируете финансовый отчёт.

Задача: финансовый анализ.

Цена ошибки: высокая.

Формат: текст.

Бюджет: не критичен.

Доступность: без VPN.

152-ФЗ: важно.

Оплата: рубли.

Выбор: GigaChat Pro. 152-ФЗ, качество, доступность.

Пример шестой: вы нуждаетесь в максимальной точности.

Задача: критическое решение.

Цена ошибки: очень высокая.

Формат: текст.

Бюджет: не критичен.

Доступность: через агрегатор.

152-ФЗ: не важно (данные обезличены).

Оплата: рубли.

Выбор: Claude через агрегатор. Максимальная надёжность.

3.9. Как выбрать, если не уверены

Если вы не уверены – начните с GigaChat. Почему? Потому что GigaChat доступен без VPN. Принимает рубли. Соответствует 152-ФЗ. Бесплатен для базового использования. Имеет Telegram-бот.

Если GigaChat не справляется – попробуйте YandexGPT. Если нужна экономия – DeepSeek или Qwen.

Если нужна максимальная точность – Claude через агрегатор.

Если нужны изображения – Kandinsky.

Простой принцип: начните с GigaChat. Если не работает – меняйте.

3.10. Как выбрать, если бюджет ограничен

Если бюджет ограничен – начните с бесплатных инструментов.

Бесплатно:

- GigaChat (базовый доступ);
- YandexGPT Lite;
- DeepSeek;
- Qwen;
- Kandinsky;
- Шедеврум.

Дёшево:

- GigaChat Lite – 65 рублей за миллион токенов;
- YandexGPT Lite – от 0,1 рубля за тысячу токенов.

Платно, но доступно:

- GigaChat Pro/Max;
- YandexGPT Pro;
- DeepSeek API;
- Qwen API.

Простой принцип: используйте бесплатные инструменты для всего. Платные – только для критического.

3.11. Как выбрать для бизнеса

Если вы бизнес – три требования.

Первое: 152-ФЗ. Используйте GigaChat или YandexGPT. Или разверните модель локально.

Второе: счёт для бухгалтерии. GigaChat через cloud.ru.

YandexGPT через Yandex Cloud. MTS Web Services. Cloud.ru.

Третье: поддержка. Русскоязычная документация. Русскоязычная поддержка. Оплата в рублях.

Простой принцип: для бизнеса – GigaChat и YandexGPT. Для остального – комбинируйте.

Часть

II

. Qwen: китайская модель, которая говорит по-русски

Глава 4. Что такое Qwen и почему он стал основным инструментом в России

4.1. Модель, о которой говорят все

Если вы спросите российского разработчика какой ИИ он использует чаще всего, ответ будет предсказуемым: Qwen. Не ChatGPT. Не Claude. Не DeepSeek. Именно Qwen.

Почему? Потому что Qwen даёт то, что нужно российскому пользователю: доступность без VPN, бесплатный или почти бесплатный доступ, понимание русского языка и качество, сопоставимое с западными флагманами.

4.2. История: Alibaba и её ставка на открытость

Qwen – продукт Alibaba Cloud, облачного подразделения китайского гиганта Alibaba Group. Первая версия модели вышла в 2023 году. С тех пор семейство разрослось до десятков моделей: от компактных до флагманских, от текстовых до мультимодальных.

Ключевое отличие Qwen от многих конкурентов – **открытость**. Alibaba публикует веса моделей. Любой может скачать их, развернуть локально, дообучить под свои задачи. Это стратегическое решение. Alibaba не пытается зарабо-

тать только на API. Она строит экосистему. Открытые модели привлекают разработчиков. Разработчики создают приложения. Приложения приводят пользователей. Пользователи платят за облачные сервисы Alibaba.

Для российского пользователя это означает простое. **Qwen можно использовать бесплатно.** Через веб-версию. Через API с щедрыми лимитами. Или локально, если есть оборудование.

4.3. Почему разработчики выбрали Qwen

В 2026 году Qwen стал лидером среди китайских моделей в России. 261,1 миллиарда токенов от российских пользователей за первую половину 2026 года. Это больше, чем у GLM (91,2 миллиарда) и Kimi (81,4 миллиарда) вместе взятых. Почему?

Причина первая: доступность. Qwen работает без VPN. Регистрация из России. API через Alibaba Cloud. Бесплатный доступ к базовым моделям.

Причина вторая: качество. Qwen3.5-397B-A17B – флагманская открытая модель. 397 миллиардов общих параметров, 17 миллиардов активных. Это уровень, который раньше был доступен только закрытым моделям.

На бенчмарках Qwen3.5 показывает результаты, близкие к GPT-5.2 и Claude 4.5 Opus. MMLU-Pro: 87.8 против 87.4 у GPT-5.2. GPQA: 88.4 против 87.0 у Claude. AIME 2026: 91.3%.

Причина третья: русский язык. Qwen работает с русским текстом лучше, чем DeepSeek. Это не значит, что он «русская модель». Это значит, что он достаточно хорошо понимает русский, чтобы быть полезным.

Причина четвёртая: мультимодальность. Qwen3.5 – нативная мультимодальная модель. Она понимает текст, изображения и видео. Это открывает сценарии, которые недоступны текстовым моделям.

Причина пятая: цена. Qwen3.8-Flash стоит 0,8 юаня за миллион входных токенов и 2,7 юаня за миллион выходных. Это дёшево. Qwen3.7-Flash ещё дешевле: 0,2 юаня за вход.

4.4. Экосистема моделей Qwen

Qwen – это не одна модель. Это семейство. Понимание различий – ключ к эффективному использованию.

Qwen3.5-397B-A17B – флагманская открытая модель. 397 миллиардов параметров, 17 миллиардов активных. Контекст 262 000 токенов, расширяемый до 1 миллиона. Мультимодальность. Открытые веса.

Qwen3-Max – самая мощная модель. 2,4 триллиона параметров. Контекст 1 миллион токенов. Для самых сложных задач.

Qwen3-Plus – баланс качества и цены. 350 миллиардов параметров. Контекст 1 миллион токенов. Для повседневной работы.

Qwen3-Flash – быстрая и дешёвая. 1250 миллиардов па-

параметров (60 миллиардов активных). Контекст 1 миллион токенов. Для массовых задач.

Qwen-Long – для документов. Контекст 10 миллионов токенов. Для анализа договоров, документации.

Qwen3-Coder – специализированная для кода. До 480 миллиардов параметров. Для генерации, ревью и рефакторинга кода.

Qwen3-VL – для изображений и видео. Мультимодальная. Для анализа визуального контента.

4.5. Что это значит для вас

Если вы разработчик: Qwen3-Coder – ваш инструмент. Бесплатно. Без VPN. Качество, близкое к Claude для кода.

Если вы аналитик: Qwen-Long – для больших документов. 10 миллионов токенов. Это примерно 5000 страниц текста за один раз.

Если вы обычный пользователь: Qwen3-Flash через веб-версию. Бесплатно. Быстро. Достаточно хорошо для большинства задач.

Если вы бизнес: Qwen3-Plus или Qwen3-Max через API. Дёшево. Качественно. Но – данные уходят в Китай. Для персональных данных используйте GigaChat или YandexGPT.

Глава 5. Как получить доступ к Qwen

5.1. Три пути

Qwen доступен тремя способами. Каждый – для своих задач.

Первый: веб-версия. Самый простой. Открываете сайт. Регистрируетесь. Работаете. Бесплатно.

Второй: API через Alibaba Cloud Model Studio. Для автоматизации. Оплата по токенам. Дёшево.

Третий: локальное развёртывание. Для приватности. Скачиваете веса. Разворачиваете на своём оборудовании. Бесплатно.

5.2. Веб-версия: для быстрого старта

Где: Tongyi Qianwen – веб-интерфейс Alibaba. Или Qwen Chat.

Как получить доступ: заходите на сайт. Регистрируетесь через email или телефон. Начинаете работать. Регистрация из России работает. VPN не нужен. Китайский номер не обязателен.

Что доступно: Базовые модели Qwen. Бесплатно. С ограничениями по количеству запросов.

Для чего подходит:

– быстрые вопросы;

- генерация текста;
- работа с кодом;
- анализ изображений (в мультимодальных моделях);
- знакомство с моделью.

Ограничения:

- лимиты на количество запросов;
- нет доступа к флагманским моделям (Qwen3.8);
- данные обрабатываются на серверах Alibaba.

5.3. API через Alibaba Cloud Model Studio

Что это: платформа Alibaba Cloud для доступа к моделям через API. Аналог OpenAI API.

Как получить доступ:

Шаг первый. Зарегистрируйтесь на Alibaba Cloud (alibabacloud.com).

Шаг второй. Активируйте Model Studio.

Шаг третий. Получите API-ключ.

Шаг четвёртый. Используйте API.

Оплата: Оплата по токенам. Принимаются международные карты. Есть бесплатные для новых пользователей.

Бесплатные: Для некоторых моделей – 1 миллион токенов на 90 дней.

Цены (за миллион токенов):

Qwen3.8-Max: 2,5 юаня вход, 10 юаней выход.

Qwen3.5-Plus: 0,8 юаня вход, 4,8 юаня выход.

Qwen3.8-Flash: \$0,15 вход, \$0,47 выход.

Qwen-Long: 0,5 юаня вход, 2 юаня выход.

Batch-вызовы: Скидка 50% на вход и выход.

Кэширование контекста: Скидка на входные токены.

Создание кэша – 125%, попадание – 10%.

5.4. Локальное развёртывание

Что это: скачиваете веса модели. Разворачиваете на своём оборудовании.

Где скачать: Hugging Face. ModelScope. Официальный сайт Qwen.

Что доступно:

Qwen3.8-Max: 2,4 триллиона параметров. Слишком большая для большинства.

Qwen3.8-27B: 27 миллиардов параметров. Требуется мощный GPU.

Qwen3.5-2B: 2 миллиарда параметров. Работает на### GPU.

Qwen3.5-35B-A3B: 35 миллиардов параметров, 3 миллиарда активных. МоЕ-архитектура.

Требования:

Для 27B: минимум 24 GB VRAM (квантизация).

Для 2B: 8 GB VRAM.

Для 35B-A3B: 16 GB VRAM.

Для чего подходит:

- приватность данных;
- кастомизация;

- дообучение;
- работа без интернета.

5.5. Что выбрать

Если вы обычный пользователь: Веб-версия. Бесплатно. Достаточно для большинства задач.

Если вы разработчик: API. Дёшево. Гибко. Много моделей.

Если вы бизнес с требованиями к приватности: Локальное развёртывание. Или API с гарантиями.

Если вы хотите максимальное качество: API. Доступ к Qwen3.8-Max. Флагманская модель.

5.6. Важное предупреждение

Qwen – китайская модель. Данные обрабатываются на серверах Alibaba. Для персональных данных это может быть нарушением 152-ФЗ.

Если вы работаете с персональными данными:

- используйте GigaChat или YandexGPT;
- или разверните Qwen локально;
- или анонимизируйте данные перед отправкой.

Если вы работаете с коммерческой тайной:

- локальное развёртывание;
- или GigaChat / YandexGPT.

Если данные не конфиденциальны:

- используйте Qwen без ограничений.

Глава 6. Выбор модели:

Max

,

Plus

,

Flash

,

Long

6.1. Зачем вообще выбирать

Когда вы открываете Alibaba Cloud Model Studio, перед вами не одна модель Qwen, а несколько. Это не маркетинговый ход. Это отражение реальности: разные задачи требуют разного уровня интеллекта.

Представьте, что у вас в гараже стоит грузовик, спортивный автомобиль и микроавтобус. Все едут. Но если нужно перевезти мебель – вы берёте грузовик. Если нужно быстро добраться до аэропорта – спортивный автомобиль. Если нужно перевезти группу людей – микроавтобус.

С Qwen то же самое. Четыре основные модели: **Max**, **Plus**, **Flash** и **Long**. Плюс специализированные: **Coder** для кода, **VL** для изображений, **Audio** для аудио. Разберём каждую.

6.2. Qwen3.8-Max: флагман

Основные характеристики:

- 2,4 триллиона общих параметров;
- 950 миллиардов активных параметров (MoE);

- контекст до 1 миллиона токенов;
- максимальный выход: 131 072 токена;
- нативная мультимодальность (текст, изображения, видео);
- поддержка режима;
- Function Calling, структурированный вывод.

Философия модели: решать самые сложные задачи.

Qwen3.8-Max – это тяжёлая артиллерия. Она используется для автономного программирования, сложного анализа, длительных агентных задач.

В внутренних тестах Alibaba Qwen3.8-Max завершил 16-дневный проект по разработке программного обеспечения автономно – от начала до конца, без вмешательства человека. Это уровень, который раньше был доступен только исследовательским лабораториям.

Когда выбирать Max:

- сложное программирование и автономная разработка;
- длительные агентные задачи (часы, дни);
- анализ данных с множеством переменных;
- юридическая проверка документов;
- финансовое моделирование;
- задачи, где качество важнее цены.

Когда НЕ выбирать Max:

- простые задачи (избыточно);
- массовая обработка (дорого);

- случаи, где Flash справляется не хуже.

Цена (за миллион токенов):

- вход: 12 юаней (около \$1,7);
- выход: 36 юаней (около \$5);
- кэш: 1,5 юаня.

6.3. Qwen3.7-Plus: рабочая лошадка

Основные характеристики:

- около 350 миллиардов общих параметров;
- около 170 миллиардов активных параметров;
- контекст до 1 миллиона токенов;
- максимальный выход: 131 072 токена;
- нативная мультимодальность;
- поддержка режима;
- Function Calling, структурированный вывод.

Философия модели: баланс качества и цены. Plus – рекомендуемая стартовая модель для большинства задач. Она не самая умная и не самая дешёвая, но даёт хороший результат за разумные деньги.

Когда выбирать Plus:

- ежедневная работа;
- генерация кода (не самый сложный);
- анализ документов;
- написание текстов;
- задачи средней сложности;
- случаи, где Max избыточен, а Flash недостаточен.

Когда НЕ выбирать Plus:

- максимально сложные задачи (Max лучше);
- массовая обработка (Flash дешевле).

Цена:

- вход: 1,6 юаня (около \$0,23);
- выход: 6,4 юаня (около \$0,9);
- кэш: 0,6 юаня.

Почему Plus – хороший выбор:

Согласно официальной документации Alibaba Cloud, для ботов, генерации контента, суммаризации и обработки документов рекомендуется именно Qwen3.7-Plus – «баланс способностей и стоимости, полная поддержка вызова инструментов, контекст 1М подходит для больших кодовых баз».

6.4. Qwen3.8-Flash: быстрая и дешёвая

Основные характеристики:

- 1250 миллиардов общих параметров;
- 60 миллиардов активных параметров;
- контекст до 1 миллиона токенов;
- максимальный выход: 131 072 токена;
- нативная мультимодальность;
- скорость: в 5+ раз быстрее Max;
- Function Calling, структурированный вывод.

Философия модели: скорость и экономия для высоконагруженных задач. Flash – для случаев, когда нужно быстро и дёшево. Она справляется с большинством задач, которые

не требуют максимальной глубины.

Когда выбирать Flash:

- массовая обработка;
- классификация;
- извлечение данных;
- чат-боты с высоким трафиком;
- задачи, где скорость критична;
- случаи, где Plus избыточен.

Когда НЕ выбирать Flash:

- сложные рассуждения;
- автономная разработка;
- задачи, требующие максимальной точности.

Цена:

- вход: 0,8 юаня (около \$0,11);
- выход: 2,7 юаня (около \$0,38).

Важно: для больших документов лучше использовать Long или Plus.

6.5. Qwen-Long: для документов

Основные характеристики:

- контекст до 10 миллионов токенов;
- максимальный выход: 8192 токена;
- только текст (без мультимодальности);
- без Function Calling;

Философия модели: читать много, писать мало. Long – специализированная модель для работы с огромными объ-

ёмами текста. 10 миллионов токенов – это примерно 5000 страниц.

Когда выбирать Long:

- анализ книг (целая книга за один раз);
- работа с сотнями договоров;
- анализ технической документации;
- извлечение данных из документов;
- задачи, где нужно «прочитать всё и найти нужное».

Когда НЕ выбирать Long:

- задачи, требующие мультимодальности;
- задачи, требующие Function Calling;
- задачи, где нужен длинный выход (максимум 8192 токена);
- сложные рассуждения.

Цена:

- вход: 0,5 юаня (около \$0,07);
- выход: 2 юаня (около \$0,28);
- batch: скидка 50%.

6.6. Специализированные модели

Помимо четырёх основных, Qwen предлагает специализированные модели.

Qwen3-Coder – для генерации, ревью и рефакторинга кода. До 480 миллиардов параметров. Доступна в вариантах Plus и Flash.

Qwen3-VL – для анализа изображений. Поддерживает классификацию, обнаружение объектов, описание, OCR, анализ графиков.

Qwen-Image-3.0 – для генерации изображений. Поддерживает до 4,5k токенов на вход, 12 языков, 20+ шрифтов. Текст до 10px остаётся читаемым.

Qwen-Audio-3.0 – для аудио. Синтез речи, понимание речи, эмоциональная выразительность.

6.7. Сравнительная таблица

Критерий

Max

Plus

Flash

Long

Общие параметры

2,4 трлн

~350 млрд

1250 млрд

Нет

Активные параметры

950 млрд

~170 млрд

60 млрд

Нет

Контекст

1M

1M

1M

10M

Макс. выход

131K

131K

131K

8K

Мультимо-дальность

Да

Да

Да

Нет

Thinking

Да

Да

Да

Нет

Function Calling

Да

Да

Да

Нет

Скорость

Медлен-ная

В 3 раза быстрее Мах

В 5+ раз быстрее Мах

Средняя

Цена (вход)

12 юаней

1,6 юаня

0,8 юаня

0,5 юаня

Цена (выход)

36 юаней

6,4 юаня

2,7 юаня

2 юаня

Лучшее применение

Сложные задачи
Ежедневная работа
Массовые задачи
документы

6.8. Как выбрать: практическое руководство

Шаг первый: оцените сложность задачи.

Простая задача – Flash.

Средняя – Plus.

Сложная – Max.

Документ – Long.

Шаг второй: оцените объём.

Массовая обработка – Flash.

Единичные задачи – Plus или Max.

Шаг третий: оцените бюджет.

Ограничен – Flash или Long.

Не ограничен – Max или Plus.

Шаг четвёртый: оцените формат.

Только текст – Long.

Текст + изображения – Max, Plus или Flash.

6.9. Примеры комбинаций

Комбинация 1: Flash + классификация

Задача: классифицировать 10 000 отзывов.

Почему: простая задача, большой объём, ошибка допустима.

Комбинация 2: Plus + анализ документов

Задача: проанализировать отчёт на 100 страниц.

Почему: средняя сложность, нужен хороший результат за разумные деньги.

Комбинация 3: Max + автономная разработка

Задача: написать сложный проект.

Почему: сложная задача, нужна автономность, ошибка стоит дорого.

Комбинация 4: Long + анализ книги

Задача: проанализировать книгу на 500 000 слов.

Почему: документ, нужен большой контекст.

Глава 7. Как формулировать запросы в Qwen

7.1. Почему это важно

Запрос – это единственный интерфейс между вами и моделью. Всё, что Qwen знает о вашей задаче, он знает из запроса. Если запрос плохой – результат будет плохим. Если хороший – результат будет хорошим.

Это звучит очевидно. Но большинство людей пишут запросы так, как будто модель умеет читать мысли.

«Напиши текст» – это не запрос. Это надежда.

7.2. Базовые принципы

Семь принципов из первой книги работают с Qwen так же, как с любой моделью. Но есть нюансы.

Принцип первый: конкретность. Чем конкретнее запрос, тем точнее результат.

Плохо: «Напиши статью о маркетинге».

Хорошо: «Напиши статью о маркетинге для малого бизнеса. 800 слов. Три раздела: социальные сети, email-рассылки, локальный SEO. Для каждого – два конкретных совета. Без воды. С примерами».

Принцип второй: контекст. Qwen не знает вашу ситуацию. Скажите ему.

Плохо: «Напиши письмо клиенту».

Хорошо: «Напиши письмо клиенту, который задержал оплату на две недели. Тон – вежливый, но настойчивый. Наша компания – поставщик оборудования. Клиент – строительная фирма. Цель письма – получить оплату до конца месяца. Объём – до 150 слов».

Принцип третий: формат. Если нужен конкретный формат – скажите.

«Оформи как таблицу». «Сделай список из пяти пунктов». «Напиши ровно 200 слов». «Выдели жирным ключевые термины».

Принцип четвёртый: роль. Qwen можно поставить в позицию.

«Ты – опытный юрист». «Ты – редактор с двадцатилетним стажем». «Ты – скептически настроенный инвестор».

Принцип пятый: ограничения. Скажите, чего делать не нужно.

«Не используй маркированные списки». «Не пиши вступление». «Не предлагай альтернативы». «Не извиняйся».

Принцип шестой: итерация. Первый ответ не идеален. Уточняйте.

Принцип седьмой: верификация. Проверяйте критически важные данные.

7.3. Особенности Qwen: русский язык

Qwen работает с русским языком. Но есть нюансы.

Нюанс первый: качество русского. Qwen3.5-397B показывает хорошие результаты на русском. Но он не «русская модель». Некоторые идиомы, культурные отсылки, специфические термины могут быть не поняты.

Что делать: если результат на русском не идеален – уточните. Или используйте GigaChat для чисто русскоязычных задач.

Нюанс второй: смешанные языки. Qwen хорошо работает с русско-китайскими задачами. Если вам нужно перевести с китайского или на китайский – Qwen справится лучше многих.

Что делать: используйте Qwen для китайского языка. Для европейских языков – ChatGPT или Claude через агрегатор.

Нюанс третий: токенизация. Русский текст токенизируется менее эффективно, чем английский. Это значит, что один и тот же текст на русском займёт больше токенов, чем на английском. Стоимость обработки русского текста выше.

Что делать: учитывайте это при оценке бюджета.

7.4. Особенности Qwen: Thinking режим

Qwen поддерживает Thinking режим – модель «думает» перед ответом. Показывает цепочку рассуждений.

Как включить:

В API: `enable_thinking=True`.

В веб-версии: кнопка «Deep Think».

Когда использовать:

- сложные математические задачи;
- логические рассуждения;
- многошаговые задачи;
- код со сложной логикой.

Когда НЕ использовать:

- простые задачи (избыточно);
- массовая обработка (медленно и дорого);
- задачи, где важна скорость.

Важно: не заставляйте Qwen думать пошагово. Модель уже умеет рассуждать. Навязывание цепочки может ухудшить результат.

7.5. Особенности Qwen: Function Calling

Qwen поддерживает Function Calling – вызов внешних функций.

Что это значит: Вы можете дать модели список функций. Модель решает, какую функцию вызвать, и с какими параметрами.

Пример: Вы создаёте чат-бота для интернет-магазина. Даете модели функции: «проверить наличие товара», «оформить заказ», «узнать статус доставки». Модель сама решает, какую функцию вызвать.

Как использовать: В API: передайте список функций в параметре tools.

Когда использовать:

- агентные задачи;
- автоматизация;
- интеграция с внешними системами.

7.6. Особенности Qwen: структурированный вывод

Qwen поддерживает структурированный вывод – JSON, таблицы, списки.

Как использовать: В API: параметр `response_format`.

Пример запроса: «Верни результат в формате JSON: {"name": "...", "age": ..., "city": "..."}».

Что вы получите: Структурированный JSON, который можно сразу использовать в коде.

7.7. Особенности Qwen:

Qwen поддерживает поиск в интернете.

Как включить:

в API: параметр `enable_search=True`.

Когда использовать:

- актуальные данные;
- новости;
- проверка фактов;
- задачи, требующие свежей информации.

Конец ознакомительного фрагмента.

Текст предоставлен ООО «Литрес».

Прочитайте эту книгу целиком, [купив полную легальную версию](#) на Литрес.

Безопасно оплатить книгу можно банковской картой Visa, MasterCard, Maestro, со счета мобильного телефона, с платежного терминала, в салоне МТС или Связной, через PayPal, WebMoney, Яндекс.Деньги, QIWI Кошелек, бонусными картами или другим удобным Вам способом.