# 01. Для каких случаев какую модель применять?

## Сравнение популярных моделей ИИ и их режимов

> **Для кого этот материал:** для людей без технического образования, которые делают первые шаги в изучении инструментов искусственного интеллекта. Здесь нет сложных терминов — только практические объяснения, примеры из жизни и конкретные рекомендации.

---

## Часть 1. Какие модели вообще существуют?

Из основных моделей у нас есть **ChatGPT** (OpenAI), **Google Gemini**, **Anthropic Claude**, **Grok** (xAI, это компания Илона Маска), **Perplexity**, **DeepSeek**, **Qwen** — это китайские модели. Если вам вдруг мало представленных вариантов, можно зайти на **HuggingFace** — там почти миллион моделей (740+ тысяч и постоянно прибавляется). Причём многие из них специализированные: одни заточены под медицину, другие — под юриспруденцию, третьи — под генерацию изображений. Их можно прямо там запускать и экспериментировать сколько угодно, совершенно бесплатно.

Но если взять подавляющее большинство публики — условно 99% всех людей, которые пользуются искусственным интеллектом, — то основная масса запросов звучит как «дай мне рецепт горохового супа», «помоги написать письмо», «объясни, что такое инфляция». И для этих целей какой-нибудь **ChatGPT** в платной версии — это, как говорится, **best choice**. Голову себе ломать особо не нужно. Важная оговорка: я говорю только о платных тарифах, потому что они дают минимально нормальный уровень возможностей. Бесплатные версии — это лотерея: поиграться можно, но серьёзно рассчитывать на них не стоит.

---

## Часть 2. Быстрый обзор основных платформ

### ChatGPT (OpenAI)

Самый популярный сервис в мире. На момент написания этого материала — более **5 миллиардов посещений в месяц**, что делает его одним из топ-5 сайтов в мире по трафику. Для понимания масштаба: ещё недавно он уступал Twitter, а сейчас уже обогнал его.

Что важно знать о тарифах:

- **Бесплатно** — ограниченный доступ к GPT-4o, без продвинутых режимов
- **ChatGPT Plus (~$20/мес.)** — полноценный доступ к GPT-4o и другим моделям, режимы глубокого исследования, генерация изображений
- **ChatGPT Pro ($200/мес.)** — для тех, кто работает с ИИ профессионально, с доступом к самым мощным и «думающим» режимам

**Google Gemini**

У Google всё хитро: они скрывают реальный трафик, потому что Gemini работает под доменом Google и встроен почти везде — в Gmail, в Google Docs, в Android. Поэтому оценить охват сложно, но он огромен. Тариф **Gemini Advanced (Gemini 2.5 Pro)** — примерный аналог ChatGPT Plus. Там тоже написано «быстрая помощь с любыми задачами», есть режим **DeepResearch**. Стоит около $20/мес., иногда идёт бесплатно в комплекте с Google One.

Главная фишка Gemini — **гигантское контекстное окно** (об этом ниже), хорошая интеграция с сервисами Google и возможность работать с видео.

**Grok (xAI / Илон Маск)**

Есть два варианта доступа: через Twitter/X (особенно если у вас платный аккаунт — Grok идёт в комплекте) и через отдельный сайт grok.com. Внутри есть несколько режимов:

- **Обычный чат** — просто разговариваем
- **DeepSearch** — глубокое исследование с выходом в интернет
- **Think** — режим рассуждений, когда модель «думает вслух» перед ответом
- **Big Brain** — ещё более глубокий анализ, требует больше времени

По охвату — пока меньше 200 миллионов посещений, но активно растёт. Честно говоря, если бы у Маска не было Twitter — этот Grok никто бы особо не заметил. Но сейчас это уже серьёзный игрок.

**Claude (Anthropic)**

Тут нужно быть честным: Claude пользуются примерно **96 миллионов посещений в месяц** — и это в разы меньше, чем у ChatGPT. Среди программистов и разработчиков есть определённый культ «Клауда», причём особенно активный в области программирования. Насколько это оправдано — вопрос отдельный (посмотрим на бенчмарки в следующей части).

По интерфейсу — если честно, там всё не идеально: иногда всё подтягивается медленно, кое-что не работает с первого раза. Что до режима «Проект» и возможности загружать документы — это есть и работает, но надо привыкнуть.

**Perplexity**

Специализированный инструмент для поиска с подтверждением — то есть он даёт ответ **со ссылками на источники**. Около 113 миллионов посещений в месяц. Если вам важно не просто «что-то умное», а конкретные факты с указанием, откуда они взяты — Perplexity отличный выбор. Они фактически **законодатели моды** в своей нише, их доля — около 85% рынка «ИИ-поиска со ссылками».

**DeepSeek и Qwen**

Китайские модели, популярные среди русскоязычной аудитории прежде всего потому, что для них **не нужен VPN**. У DeepSeek есть режим DeepThink (рассуждения) и Search (выход в интернет). Qwen — примерно то же самое. Если честно, интерфейсы у них практически идентичны: всё друг у друга воруют, как это и принято в отрасли.

---

## Часть 3. Что происходит с рынком прямо сейчас (актуально на 2025–2026 год)

Вот это важная часть, которую нельзя пропустить, если вы хотите понимать, куда всё движется.

**ИИ вытесняет Google**

Долгое время казалось, что Google — это вечный монополист. С 2016 года его доля в мировом поиске держалась на уровне **90%** и не менялась. Но затем случилось то, чего многие не ожидали: впервые с 2015 года доля Google в поиске устойчиво опустилась ниже отметки 90% — и это происходило на протяжении большей части 2025 года.

Казалось бы, меньше одного процента — и что такого? Но именно устойчивость этого тренда говорит о многом. Трафик ИИ-чат-платформ за тот же период вырос на 80% год к году, тогда как общий трафик поисковых систем фактически стоял на месте.

Аналитическая компания Gartner, которая занимается прогнозированием рынков, ещё в 2024 году сделала резкое заявление: к 2026 году объём традиционного поиска сократится на 25%, поскольку ИИ-чатботы и виртуальные агенты перехватывают пользователей. Проще говоря — люди всё чаще идут с вопросом не в Google, а напрямую к ИИ-ассистенту.

Почему? Потому что ИИ не просто находит ссылки — он **объясняет**. Для старшего поколения это особенно заметно: они используют ИИ как замену поисковику именно потому, что тот сразу даёт ответ, а не список из десяти сайтов, где ещё надо разбираться.

**Конкуренция между платформами обострилась до предела**

В начале 2026 года произошла волна крупных обновлений: OpenAI выпустил GPT-5.5, Anthropic запустил Claude Opus 4.7, xAI раскатал Grok 4.3 Beta, а Google продолжал улучшать Gemini 2.5 Pro. Причём всё это происходит практически одновременно — компании буквально следят друг за другом и выпускают обновления чуть ли не день в день.

---

## Часть 4. Три типа задач — три разных подхода

Все задачи, с которыми вы обращаетесь к ИИ, можно разделить на три большие категории. Это важно понять, потому что именно от типа задачи зависит, какой инструмент и какой режим вам нужен.

**Тип 1: Обычный разговор и простые вопросы**

Рецепт горохового супа, объяснение термина, помощь с письмом, идеи для подарка, перевод текста — всё это первый тип. Здесь никаких особых режимов не нужно. **GPT-4o**, **Gemini 2.5 Flash**, **Claude Sonnet** — любой из них справится на отлично. Не надо ничего включать, не надо думать о моделях. Просто открываете и пишете.

Есть даже интересная статистика: молодёжь до 25 лет использует ИИ по полной — как инструмент для работы, учёбы, творчества. Люди среднего возраста (35–45) чаще применяют его как советника. Те, кто постарше, — как замену Google. И для всех этих сценариев первого типа вполне хватает базовых моделей.

**Тип 2: Глубокое исследование со ссылками**

Вот вам нужно разобраться в каком-то юридическом вопросе. Или понять, какие лекарства взаимодействуют между собой. Или написать аналитическую записку с данными и источниками. Это уже второй тип — и тут нужен специальный режим.

Практически у всех крупных платформ он есть, называется по-разному:

- В **ChatGPT** — «Проведите глубокое исследование» (Deep Research)
- В **Google Gemini** — тоже Deep Research, кнопка внизу экрана
- В **Grok** — режим DeepSearch
- В **Perplexity** — это вообще основная специализация сервиса

Что делает этот режим? По сути, вы нанимаете аналитика. Вы даёте задание — он идёт в интернет, собирает источники, синтезирует данные, строит таблицы, добавляет ссылки. Это буквально обнулило работу многих маркетинговых и исследовательских агентств — зачем платить аналитику за неделю работы, если ИИ делает похожее за 5 минут?

Важно: этот режим доступен, как правило, только в **платных** тарифах. Но $20 в месяц — это несерьёзная сумма по сравнению с тем, что вы получаете.

**Тип 3: Сложные рассуждения, код, анализ алгоритмов**

Вот здесь начинается самое интересное. Когда задача нетривиальная — например, вы пишете код и получаете сообщение об ошибке, которое нужно разобрать, или вам нужно построить сложную логическую цепочку, проанализировать большой договор, найти противоречия в данных — обычная модель может не справиться.

Практический пример: если вы работаете с кодом, вставляете его в редактор — и он выдаёт ошибку. Даёте ChatGPT 4o сообщение об ошибке, просите найти и исправить. В 90% случаев он не справляется. А вот если ту же задачу дать **O4-mini-high**, **O4-mini** или **O3** — они в 100% случаев находят ошибку и сразу исправляют. Прямо очевидная разница.

Для этих задач нужны **модели с расширенным мышлением** (Thinking / Reasoning):

- OpenAI: **O3, O4-mini, O4-mini-high**
- Google: **Gemini 2.5 Pro** с режимом Thinking
- Anthropic: **Claude** с режимом Extended Thinking
- Grok: режим **Think** и **Big Brain**

Почему они лучше? Потому что такие модели не сразу выдают ответ — они сначала «рассуждают вслух», проверяют себя, перебирают варианты, и только потом дают результат. Видно буквально, как модель пыжится и думает — иногда это занимает 30–60 секунд, зато результат принципиально лучше.

---

## Часть 5. Что такое контекстное окно и почему это важно

Один из самых важных технических параметров модели — это **контекстное окно**, то есть сколько текста модель может «удержать» в голове одновременно. Измеряется в **токенах** (примерно 3/4 слова на токен для русского языка).

Для понимания масштаба:

- **128 000 токенов** — это примерно 400–500 страниц текста
- **1 000 000 токенов** — это примерно **2–3 миллиона символов**, или три-четыре полноценных романа. Две «Войны и мира», грубо говоря

GPT-4.1 и Gemini 2.5 Pro имеют контекстное окно в 1 миллион токенов — это примерно 750 000 слов. GPT-5.5 также поддерживает контекстное окно в 1 миллион токенов, что вдвое больше, чем у его предшественника GPT-5.

Зачем это нужно на практике? Представьте, что вы хотите загрузить в ИИ целый договор на 100 страниц и попросить найти в нём все потенциальные риски. Или скормить ему всю переписку с клиентом за полгода и попросить сделать выжимку. Без большого контекстного окна модель просто не сможет это обработать — она «забудет» начало, пока читает конец.

Для 99,9% повседневных задач даже 128K токенов — это огромный запас. Но если вы работаете с большими документами, аналитикой или кодовыми базами — разница становится принципиальной.

---

## Часть 6. Какие модели сейчас на вершине? (Актуально на 2025–2026)

Ситуация в мире ИИ меняется примерно раз в квартал, иногда быстрее. Вот что происходит сейчас:

**«Гонка вооружений» в полном разгаре**

Все четыре крупнейших игрока выпустили свои флагманские модели примерно в одно время: Google DeepMind запустил Gemini 2.5 Pro в конце марта 2025 года, xAI представил Grok 4 в июле, Anthropic анонсировал Claude Opus 4.1 буквально за несколько дней до того, как OpenAI наконец выкатил долгожданный GPT-5. Это не совпадение — компании буквально следят друг за другом.

**Кто в чём силён?**

На бенчмарке по математическим рассуждениям Grok 4 достиг идеального результата на тесте AIME 2025, GPT-5 занял второе место с показателем 94,6%, тогда как Gemini 2.5 Pro показал 88%.

В области реального программирования (бенчмарк SWE-bench) результаты удивительно близки у всех лидеров — и именно здесь Claude демонстрирует сопоставимые с лидерами показатели, что говорит о том, что Anthropic специально оптимизировал модель под точную и безопасную генерацию кода.

Gemini 2.5 Pro доминирует в работе с длинными текстами — его контекстное окно в 1 миллион токенов делает его идеальным для анализа больших документов, юридических обзоров и масштабного исследовательского синтеза.

Если говорить совсем просто:

| Задача | Лучший выбор |
|---|---|
| Повседневные вопросы, написание текстов | GPT-5 / GPT-4o, Gemini 2.5 Flash |

| | |
|---|---|
| Математика, логика, сложные рассуждения | Grok 4, GPT-5 (режим Thinking) |
| Программирование, код | Claude Opus 4, GPT-5, Grok 4 |
| Анализ огромных документов | Gemini 2.5 Pro (контекст 1M токенов) |
| Поиск с источниками и ссылками | Perplexity, любая модель в режиме Deep Research |
| Актуальные события, Twitter-контекст | Grok (встроен в X/Twitter) |
| Без VPN для России | DeepSeek, Qwen |

**Важная оговорка:** ни одна модель не является лучшей во всём. Вместо одного «победителя» мы видим специализированное превосходство: Claude — для кода, Grok — для рассуждений, Gemini — для мультимодальных задач, DeepSeek — для экономичного использования. Выбор зависит от вашей конкретной задачи.

**Что нового появилось в 2025–2026 году**

**GPT-5 и GPT-5.5** — OpenAI сделал принципиальный шаг вперёд. GPT-5.5 позиционируется как «самая умная и интуитивная модель» компании, причём это не просто чат — это агентная система, которая может планировать многошаговые задачи, использовать инструменты, проверять себя и доводить дело до конца без постоянного контроля пользователя. Это уже не просто «ответить на вопрос» — это «сделать за вас задачу от начала до конца».

**Grok 4 Heavy** — необычная архитектура: в версии Heavy запускаются параллельные цепочки рассуждений, из которых в конце выбирается наиболее уверенный ответ. По сути, это как если бы несколько аналитиков одновременно решали одну задачу, а потом сравнивали результаты. Минус — работает медленнее (4–7 раз дольше обычного). Плюс — точность выше на сложных задачах.

**Gemini 2.5 Pro** теперь умеет работать с видео и аудио напрямую — не просто смотреть картинки, а анализировать видеозаписи, расшифровывать встречи, сравнивать визуальный контент.

---

## Часть 7. Про бенчмарки — почему не стоит слепо им верить

Когда вы читаете в интернете «модель X набрала 94.6% на бенчмарке AIME» — это звучит очень солидно. Но давайте честно: что это значит для вас на практике?

**Бенчмарк** — это стандартизированный тест, что-то вроде ЕГЭ для нейросетей. Он проверяет способность решать математические задачи олимпийского уровня, писать код, отвечать на научные вопросы. Полезная информация для понимания возможностей модели — но не прямая инструкция «используйте X для написания письма маме».

Важнее понимать другое: все крупные модели сейчас **примерно одинаково хороши** для 95% повседневных задач. Разница проявляется в нюансах:

- насколько модель «галлюцинирует» (придумывает несуществующие факты)
- как она держит контекст длинного разговора
- насколько понимает русский язык
- насколько удобен интерфейс

Если OpenAI опубликует таблицу, где OpenAI выглядит лучше всех — не удивляйтесь. Если Google — то Gemini будет на первом месте. У всех есть свои «любимые» тесты. Истина, как всегда, где-то посередине.

---

## Часть 8. API — для тех, кто хочет пойти глубже

Если вы захотите использовать ИИ не через браузер, а в своих автоматизациях, скриптах или приложениях — для этого есть **API** (Application Programming Interface). Это специальный «вход» в модель через программный код.

Почти у всех платформ он есть:

- **OpenAI** — самая проработанная инфраструктура. Есть Playground для экспериментов, сравнение моделей, подробная документация. Есть раздел, где можно поставить лимит трат — например, $30 в месяц, и система автоматически останавливается.
- **Google AI Studio** — аналогичный инструмент для работы с Gemini через API
- **Anthropic** — API к Claude, с хорошей документацией
- **Grok/xAI** — тоже есть документация и стандартный API
- **Perplexity, DeepSeek, Qwen** — всё то же самое

Важно понимать: при использовании API **каждый запрос стоит денег**. Стоимость измеряется в токенах — сколько текста вы отправили и сколько получили в ответ. Для экспериментов это копейки (буквально $4–5 в месяц при умеренном использовании), но если запустить большую автоматизацию без лимита — можно неожиданно накрутить счёт. Ставьте лимиты заранее.

---

## Часть 9. Low-code инструменты: автоматизация без программирования

Если вы хотите использовать ИИ не просто в чате, а встроить его в какой-то рабочий процесс — например, «пришло письмо → ИИ его анализирует → создаёт задачу в менеджере задач» — для этого не нужно быть программистом.

Есть так называемые **no-code / low-code** платформы. Самые популярные сейчас:

- **Make** (раньше Integromat) — визуальный конструктор, с которого удобнее всего начинать. Есть готовые интеграции с OpenAI, Gemini, Anthropic, Perplexity, DeepSeek.
- **Zapier** — примерно то же самое, чуть больше интеграций, чуть проще интерфейс.
- **n8n** — более гибкий и мощный, любимец разработчиков. Если нужны сложные сценарии — это сюда.

Суть: вы строите цепочку из блоков — «при событии X сделай Y через ИИ и передай результат в Z». Никакого кода. Если задачи несложные — Make более чем достаточно. Если нужна гибкость — n8n.

---

## Часть 10. Сколько это стоит и стоит ли платить?

Вопрос честный, отвечаю честно.

**Базовые подписки (~$20/мес.):**

- ChatGPT Plus — $20
- Google Gemini Advanced — $20 (иногда включён в Google One)
- Claude Pro (Anthropic) — $20
- Grok SuperGrok — $30

Это, по сути, стоимость пары чашек кофе в кафе. Если вы заходите в ИИ хотя бы раз в день — разница между бесплатной и платной версией огромная. Платная версия даёт доступ к более умным моделям, глубокому исследованию, генерации изображений и прочим инструментам.

**Продвинутые подписки:**

- ChatGPT Pro — $200/мес. (для профессионального использования с доступом к самым мощным режимам)
- Claude Max — от $100/мес.
- Grok SuperGrok Heavy — $300/мес.

Это уже для тех, кто активно использует ИИ в бизнесе или профессиональной деятельности.

**Лайфхак:** если вам нужно провести одно большое исследование или поработать с конкретным инструментом — можно подписаться на месяц, сделать что нужно, и отписаться. Никто вас не держит.

---

## Резюме: как принять решение без боли

Запомните простую схему:

**1. Для простых задач (99% случаев)** → Используйте **GPT-4o**, **Gemini 2.5 Flash** или любой аналог в обычном режиме. Никаких дополнительных опций включать не нужно. Это покроет 99% ваших запросов.

**2. Нужна точность, ссылки, фактура?** → Включайте режим **Deep Research** (есть в GPT, Gemini, Grok) или используйте **Perplexity**. Это критично, если работаете с юридическими, медицинскими или научными данными.

**3. Нужен выход в интернет?** → Включайте режим **Web Search** или **DeepSearch**. У большинства платформ это отдельная кнопка. Без этой опции модель отвечает только на основе своих знаний, без актуальных данных.

**4. Сложные рассуждения, код, анализ?** → Переключайтесь на **O3, O4-mini, O4-mini-high** (OpenAI), **Gemini 2.5 Pro**, **Claude Opus** с режимом Extended Thinking или **Grok Think / Big Brain**. Они медленнее, но принципиально точнее на сложных задачах.

**5. Хотите понять, что лучше для вашего случая?** → Зайдите в документацию OpenAI → раздел «Сравнение моделей». Выбираете две модели слева и справа и видите: скорость, контекст, стоимость, уровень «умности». Всё сразу.

**6. Без VPN (для России)?** → **DeepSeek** или **Qwen** — работают напрямую.

---

## Бонус: три вещи, которые изменились с 2024 года

1. **Контекстные окна выросли кратно.** Если раньше модели «забывали» начало разговора через несколько страниц текста, то сейчас миллион токенов — это стандарт для топовых моделей.
2. **Режимы «глубокого мышления» появились везде.** То, что год назад было экспериментальной фишкой, сейчас доступно в каждой крупной платформе в платной подписке.
3. **ИИ начал вытеснять традиционный поиск.** ChatGPT теряет долю рынка среди чат-ботов (–22 пп за год), тогда как Google Gemini почти учетверил

свою долю — с 5,7% до 21,5% — благодаря встройке в поиск, Android и Workspace. Рынок перетрясается прямо на наших глазах.

---

## Итог для начинающих

не нужно разбираться во всех моделях сразу. Начните с ChatGPT Plus или Google Gemini Advanced. Освойте базовый режим, потом попробуйте Deep Research, потом — режим рассуждений. Каждый следующий шаг будет логичным продолжением предыдущего. ИИ — это не ракетостроение. Это инструмент. Как кофемашина: сначала кажется сложно, потом — не представляешь, как жил без неё.

- В общем, давайте еще раз все зафиксируем. С простыми повседневными штуками отлично справляются базовые версии: GPT-4.0, Gemini, Claude и прочие. Их возможностей хватит за глаза.
- Нужно копнуть тему посерьезнее? Включаем соответствующие тумблеры. В GPT, например, есть отдельная опция для глубокого ресерча. Если модель должна шерстить интернет, активируйте режим поиска. Где-то он называется Web Search, где-то DeepSearch, суть одна. Без него никуда, если вы ищете конкретные факты, собираете ссылки или пруфы. Для юристов, ученых и аналитиков данных это вообще обязательное условие.
- А вот когда задача требует выстраивания сложной логики, написания кода или разбора алгоритмов, тут уже достаем тяжелую артиллерию. Переходим на продвинутые модели вроде GPT-4.1, O4-mini, O4-mini-high или O3.
- Если хочется разобраться в нюансах еще лучше, просто откройте режимы сравнения. Покрутите настройки, посмотрите на объемы контекста, цены и скорость работы. После этого обычно все встает на свои места.