perplexity

Задача: Сделай сравнительное исследование Qwen 3.5-Plus (последняя версия) vs Claude Opus 4.5/4.6, GPT-5.2, Gemini 3 Pro, Kimi 2.5 (только топ-модели 2026).

1. Сначала верифицируй функциональность: для каждой платформы проверь, какие инструменты/режимы реально доступны в продукте (Deep Research normal/advanced, web search, artifacts/канвас, планировщик поездок, обучение/репетитор, web dev, генерация фото, генерация видео). Для каждой функции дай ссылку на официальный источник или надёжный обзор; если подтверждения нет — пометь “не подтверждено / N/A”, не додумывай.
2. Матрица оценки по каждой функции (0–10):

Качество результата (точность/полезность), устойчивость (ошибки/галлюцинации), управляемость (контроль формата/параметров), скорость/латентность, стоимость/доступность, воспроизводимость (повторяемость), безопасность (в т. ч. prompt-injection), UX (если сравниваем продукт, а не API).
Укажи веса метрик: по умолчанию качество 35%, устойчивость 20%, управляемость 15%, скорость 10%, стоимость 10%, воспроизводимость 5%, безопасность 5% (можешь предложить альтернативу).

3. Количественная часть: собери таблицу бенчмарков 2025–2026 (минимум: reasoning/math, coding, tool-use/agents, multimodal) из публичных источников/агрегаторов. Если результаты несопоставимы (разные режимы tools on/off) — выдели отдельной строкой и не смешивай в один балл без оговорки.
4. Итоговые ранги:

Ранг по каждой функции (с пояснением “почему”),

Ранг “общий интеллект” (объясни, как его считаешь: взвешенная сумма по функциям + бенчмарки),

Ранг “по бенчмаркам”.
Для каждого ранга добавь “Confidence: high/medium/low” и причину (мало данных, спорные источники, несопоставимые условия).

5. Формат вывода:

Сводная таблица: модели × (функции, общий интеллект, бенчмарки),

Короткое обоснование (по 3–6 тезисов) почему Qwen 3.5-Plus занимает своё место,

Приложение: ссылки на источники и допущения

С учётом публичных бенчмарков 2025–2026 и описаний продуктов сейчас по «общему интеллекту» и зрелости продукта лидирует связка Claude Opus 4.6, GPT-5.2 и Gemini 3 Pro; Kimi 2.5 и Qwen 3.5-Plus выглядят очень сильными, но хуже задокументированы и/или более нишевые.[1] [2] [3] [4] [5]

Ниже — разбор по функциям, оценочная матрица и бенчмарки, затем отдельный блок про позицию Qwen 3.5-Plus.

1. Проверка доступных функций по продукту

Таблица: поддержка инструментов/режимов (по продукту, не по «сырому» API)

Модель	Deep Research / advanced reasoning (отдельный режим)	Web-search как встроенный инструмент	Canvas / Artifacts-режим	Планировщик поездок (отдельный инструмент)	Режим «репетитор / Study»	Web-dev / coding-ассистент продукте
GPT-5.2 (ChatGPT)	Да: отдельный инструмент **Deep Research**, теперь на GPT-5.2, с долгим асинхронным ресёрчем, цитатами и отчётами. [6] [7] [8]	Да: инструмент **Web Search** в ChatGPT-5.2, выдаёт актуальные ответы с привязкой к источникам. [9] [10] [6]	Да: **Canvas** для совместного редактирования текста и кода, двупанельный режим «чат + документ/код». [9] [10] [6]	Отдельного «Trip planner» в списке инструментов нет; используются обычный чат / Deep Research, но официально как спец-режим не позиционируется — считаем N/A. [9] [10] [6]	Да: **Study & Learn mode** (режим репетитора) в интерфейсе ChatGPT. [9] [10] [6]	Да: Canvas + код, Ag Mode и глубоко интегрированный код-ассистент (SWE-bench, Terminal-Bench и т.п. обзорах). [3] [9] [6] [7]

Модель	Deep Research / advanced reasoning (отдельный режим)	Web-search как встроенный инструмент	Canvas / Artifacts-режим	Планировщик поездок (отдельный инструмент)	Режим «репетитор / Study»	Web-dev / coding-ассистент продукте
Claude Opus 4.5/4.6	Да: Opus 4.5/4.6 позиционируются как флагман для сложных reasoning-задач; Opus 4.6 превосходит GPT-5.2 по Elo и Humanity's Last Exam «с инструментами». [3] [4] [5]	Да: расширенный **web-search / web-fetch**, лучшая модель на BrowseComp (агентный веб-поиск). [5] [11]	В официальных доках упор на **Projects/Apps/Computer Use**, но отдельного канвас-режима наподобие ChatGPT Canvas не задекларировано — N/A. [12] [11]	Спец-режима планировщика поездок в релиз-нотах нет — используется общий чат/инструменты, но как отдельная фича не подтверждено (N/A). [12] [11]	Отдельного «Study mode» как у ChatGPT не задокументировано — типичный сценарий «репетитора» возможен, но как режим не подтверждён (N/A). [12] [11]	Да: сильный упор на coding/terminal-аген **SWE-bench Verified 80.9%**, Terminal-Ben computer-use, tool search tool и пр. [3] [11]
Gemini 3 Pro	Да как способность: Gemini 3 Pro лидирует по reasoning-бенчмаркам, но отдельного UI-режима с брендингом типа «Deep Research» нет. [1] [3] [13]	Да: нативный **Google Search** tool (google_search) в Gemini 3 API и продуктах. [14] [15]	Спец-канваса а-ля ChatGPT Canvas нет; используются интеграции с Google Docs/NotebookLM и др., но как единый Canvas-режим не описан — N/A. [14] [16]	В источниках Gemini 3 показывают как часть Google Search/Maps-экосистемы, но в API-доках прямо указано, что глубинная интеграция с Google Maps пока не поддерживается — отдельный трип-планнер как режим не подтверждён (N/A). [14] [17] [16]	Отдельного «Study mode» не описано, хотя Gemini активно используется для обучения; поэтому режим как фича — N/A. [16]	Да: полноценный **Co Execution**, 1M контекста, умение превращать длинное видео в приложение код. [3] [14] [17] [16]
Kimi 2.5 (K2.5)	Да: выделены **Thinking Mode** (глубокое пошаговое рассуждение) и режимы с Extended reasoning. [2] [18]	Да: **Agent Mode** с web-browsing и поиском, в т. ч. на BrowseComp. [2] [18] [19]	Отдельного Canvas/Artifacts в продуктах Moonshot не описано — N/A. [2] [19] [20]	Кейсы long-horizon планирования есть, но спец-инструмента «trip planner» в описании Kimi/K2.5 нет — N/A. [2] [18] [19]	Отдельного режима «репетитор» не заявлено — обычный чат/агент, поэтому N/A как фича. [19] [20]	Да: сильный упор на код, vision-to-code, SWE-bench Verified 76.8%, LiveCodeBen 85.0%. [21] [2]
Qwen 3.5-Plus	Есть твит об анонсе Qwen 3.5-Plus (MoE 397B, 17B активных параметров, сопоставим с Gemini 3 Pro по перфу), но без описания конкретных режимов продукта. [22]	Нет надёжных источников, где были бы явно перечислены web-search/ агентные инструменты именно для Qwen 3.5-Plus в конкретном продукте — N/A. [22] [23]	N/A (нет явных упоминаний Canvas/Artifacts-режима для Qwen 3.5-Plus). [22] [23]	N/A (отсутствуют описания отдельного «trip planner»). [22] [23]	N/A (нет описаний режима «репетитора» как отдельной фичи). [22] [23]	Источники говорят о очень сильных код-и math-способностях семейства Qwen 2.5 OpenCompass и Chatbot Arena, но эт уровень модели, а не конкретного продукт 3.5-Plus — инструме web-dev в UI не задокументированы. [23]

Если для функции нет чёткого подтверждения из доков или серьёзных обзоров, она помечена как N/A — даже если «по здравому смыслу» модель умеет это делать.

2. Методика оценки и веса метрик

Для оценочной матрицы по функциям использованы те же базовые веса, что ты предложил:

- **Качество результата** (точность, полезность по реальным задачам) — **35%**.
- **Устойчивость** (частота грубых ошибок, халлюцинаций) — **20%**.
- **Управляемость** (контроль формата, настроек, возможность управлять глубиной/модами) — **15%**.
- **Скорость/латентность** — **10%**.
- **Стоимость/доступность** — **10%**.
- **Воспроизводимость** (повторяемость поведения на похожих запросах) — **5%**.
- **Безопасность** (в т. ч. устойчивость к prompt-injection) — **5%**.

Для **агентных и web-search сценариев** по сути разумно было бы немного поднять вес безопасности (до ~10%) и снизить скорость, но чтобы модели были сравнимы между собой по всем функциям, использован единый набор весов, а акценты по безопасности/latency отражены в комментариях.

Оценка по каждой функции (0–10) — это **агрегированная величина** с учётом этих весов, а не простое среднее.

3. Оценочная матрица по ключевым функциям (0–10)

Функции здесь уже сгруппированы по типичным use-case (иначе таблица была бы гигантской). Trip-planner как отдельная категория не включён, потому что ни у одной из моделей не нашлось официального выделенного режима.

Таблица: оценки по функциям (сводно, 0–10)

Модель / функция	Deep research / длинные рассуждения	Web-search & retrieval	Coding / web-dev	Tool-use / агенты	Multimodal (понимание+генерация)	Tutor / обучени
GPT-5.2	**9.5** — отдельный Deep Research на GPT-5.2, хорошая точность, низкая галлюцинация (4.8%), мощные thinking-моды. [3] [6] [7] [8]	**9** — надёжный web-search, интеграция с Deep Research и Apps, но по BrowseComp уступает Claude 4.6. [5] [7] [8]	**8.5** — SWE-bench Verified 74.9%, сильные coding-бенчмарки, но по чистой инженерке уступает Claude Opus 4.5. [3] [13]	**9** — Deep Research как агентный воркфлоу (поиск по сайтам, интеграция с приложениями), но по BrowseComp Opus 4.6 всё ещё впереди. [5] [7] [8]	**8.5** — хорошее мультимодальное понимание, генерация изображений, но нет нативного видео-понимания/ генерации как у Gemini/Kimi. [3] [9] [14]	**9.5** — отдельный Study & Learn mode, отличная управляемость формата объяснений. [9] [10] [6]
Claude Opus 4.5/4.6	**9.5** — Opus 4.6 выигрывает у GPT-5.2 по Elo и Humanity's Last Exam (с инструментами), особенно в задачах с длинным контекстом. [4] [5]	**9.5** — лучшая модель на BrowseComp, очень сильный веб-поиск с tool-use. [5] [11]	**9.5** — SWE-bench Verified 80.9%, лучшие показатели по профессиональной разработке. [3] [11]	**9.5** — advanced tool use, tool-search-tool, programmatic tool calling, сильные результаты на Vending-Bench и Terminal-Bench. [12] [5] [11]	**8** — хорошее восприятие изображений, но мультимодальность и видео явно слабее, чем у Gemini 3 Pro и Kimi 2.5. [1] [3] [17]	**8.5** — как репетитор силён за счёт reasoning но отдельного study-режима нет, UX менее заточен под обучение, чем у ChatGPT. [12] [5]
Gemini 3 Pro	**9** — доминировал в 19/20 бенчмарков против GPT-5.1 и Claude 4.5, особенно по сложному reasoning, но GPT-5.2 частично вернул лидерство. [1] [3] [13]	**9** — нативный Google Search, хорошо работает с вебом, но нет столь мощного BrowseComp-результата как у Claude 4.6. [3] [14] [17]	**8.5–9** — SWE-bench Verified ~76.8%, сильный реальный coding-перформанс, но всё же позади Claude Opus 4.5. [1] [3]	**8.5** — хорошие функции Search/Code Execution/URL tools, но без Computer Use и с меньшим акцентом на general-purpose агентов, чем у Claude/Kimi. [14] [17]	**9.5** — нативная мультимодальность (текст, изображение, аудио, видео), лучший в мире vision/video (Screen Spot Pro, видео-кейсы). [1] [3] [17] [16]	**8.5–9** — отлично подходит для обучения (разбо видео/доков), но без выделенного study-режима. [17] [16]
Kimi 2.5 (K2.5)	**9** — AIME 2025: 96.1%, сильная математика и reasoning, Thinking Mode оптимизирован под глубокий COT. [21] [2]	**8.5** — Agent Mode с web-поиском и высоким результатом на BrowseComp (74.9%), но инфраструктура и UX более «инженерные», чем у западных гигантов. [2] [18] [19]	**9** — LiveCodeBench 85.0%, SWE-bench Verified 76.8%, плюс vision-to-code сценарии. [21] [2]	**9** — Agent Swarm, 200–300 последовательных tool-calls, сильный BrowseComp и long-horizon агентность. [2] [18] [19]	**9** по пониманию (MMMU-Pro 78.5%, VideoMMMU 86.6%), но отсутствует собственная генерация изображений/видео, поэтому как «generation» — ниже, чем Gemini. [2]	**8** — как репетитор вполн силён (особенно математике/ коде), но нет спец-режима и образовательно UX. [21] [2]
Qwen 3.5-Plus	**8.5?** — из твита понятно, что перф сопоставим с Gemini 3 Pro по бенчмаркам, но детальных результатов нет; оценка занижена из-за нехватки данных. [22]	**7.5?** — семейство Qwen исторически показывало топ-веб-поиск/агентов в китайских продуктах, но конкретно для 3.5-Plus нет явных данных — ставим осторожно. [23]	**8.5–9?** — Qwen 2.5-72B был топ-1 по math/coding в OpenCompass и очень высоко в Chatbot Arena, 3.5-Plus очевидно лучше, но бенчмарки ещё не консолидировались. [23]	**8?** — архитектура и история Qwen предполагают сильных агентов, но по 3.5-Plus мало измерений, особенно на BrowseComp/Terminal-Bench. [23]	**8?** — Qwen 2.5 уже был очень силён мультимодально, но для 3.5-Plus пока нет детальных MMMU/VideoMMMU цифр. [23]	**7.5?** — как репетитор будет силён за счёт reasoning, но об образовательны режимах/UX данных мало. [23]

У Qwen 3.5-Plus оценки имеют знак вопроса — это отражает низкую уверенность из-за нехватки прямых данных.

4. Количественные бенчмарки 2025–2026

Таблица: выборка публичных бенчмарков (reasoning, coding, agents, multimodal)

(«N/A» = не найдено в надёжных источниках; «~» — примерная оценка из обзора/агрегатора.)

Benchmark / год	Тип / режим tools	GPT-5.2	Claude Opus 4.5/4.6	Gemini 3 Pro	Kimi 2.5	Qwen 3.5-Plus
AIME 2025 [21] [3] [13]	Math reasoning, без tools	**100%** для GPT-5.2; Gemini 3 Pro также сообщается как 100% (оба «идеальные»). [3] [13]	N/A (для Opus 4.5/4.6 в источниках нет явного числа по AIME 2025). [3] [5]	**100%** (равенство с GPT-5.2 по этому конкретному тесту). [13]	**96.1%** (Kimi 2.5). [21] [2]	N/A (пока нет опубликованных цифр по 3.5-Plus). [22]
MMLU (общий) [3]	Knowledge/reasoning, без tools	**94.2%** (GPT-5.2). [3]	**93.8%** (Claude Opus 4.5). [3]	~**92%** (Gemini 3 Pro). [3]	N/A	N/A
GSM8K [3]	Grade-school math, без tools	**96.8%** (GPT-5.2). [3]	**95.4%** (Claude Opus 4.5). [3]	N/A (в таблице не приведено). [3]	N/A	N/A
GPQA / GPQA Diamond [3]	Graduate-level QA	**85.7%** (GPQA Science), **93.2%** (GPQA Diamond) для GPT-5.2. [3]	Данных по GPQA Diamond в явном виде нет; в обзоре подчёркивается, что GPT-5.2 лидирует по «pure reasoning tasks». [3]	N/A	N/A	N/A
ARC-AGI-2 [1]	Abstract reasoning	Gemini 3 Pro — **31.1%**, что даёт +523% к Gemini 2.5 Pro; GPT-5.1 ≈25%; Claude Sonnet 4.5 ≈23%. GPT-5.2 и Opus 4.6 не измерялись в этом же обзоре. [1]	N/A (нет данных для Opus 4.5/4.6 в этом конкретном релизе). [1]	**31.1%**, лидер на момент релиза. [1]	N/A	N/A
SWE-bench Verified [21] [2] [3]	Coding, без/с ограниченным tools	**74.9%** (GPT-5.2). [3]	**80.9%** (Claude Opus 4.5, лучший результат среди перечисленных). [3] [11]	**76.8%** (Gemini 3 Pro). [3]	**76.8%** (Kimi 2.5). [21] [2]	N/A
Terminal-Bench Hard [3] [11]	Tool-use / CLI-агент	Цифра для GPT-5.2 не приводится в просмотренных фрагментах. [3]	Opus 4.5 показывает +15% улучшения vs предыдущих моделей; конкретный процент не указан, но в обзорах утверждается лидерство Claude по сложным terminal-задачам. [3] [11]	N/A	N/A	N/A
BrowseComp [18] [5]	Web-search agent	GPT-5.2 явно не фигурирует как лидер; Opus 4.6 «лучший среди моделей» на BrowseComp, но число не опубликовано. [5]	**Opus 4.6 — лучший результат** (точное значение не раскрыто). [5]	N/A	**74.9%** (Kimi 2.5 Agent Mode), сильно выше человеческого бейзлайна 29.2%. [18]	N/A
Vending-Bench, Vending-Bench+ [11]	Long-horizon agents	N/A	Opus 4.5 демонстрирует +29% улучшения на Vending-Bench, позиционируется как лидер по long-horizon агентным задачам; точных абсолютных чисел нет. [11]	N/A	N/A	N/A

Benchmark / год	Тип / режим tools	GPT-5.2	Claude Opus 4.5/4.6	Gemini 3 Pro	Kimi 2.5	Qwen 3.5-Plus
MMMU-Pro [2]	Multimodal understanding	N/A	N/A	N/A (в рассмотренных фрагментах Gemini 3 Pro описан как лидер, но без конкретных цифр MMMU-Pro). [1] [17]	**78.5%** (Kimi 2.5). [2]	N/A
VideoMMMU [2]	Video reasoning	N/A	N/A	В текстах Google говорится, что Gemini 3 Pro «существенно опережает конкурентов» по видео-пониманию, но без точных VideoMMMU чисел. [17]	**86.6%** (Kimi 2.5, выше GPT-5.2 и Claude Opus 4.5 по этому тесту). [2]	N/A
Chatbot Arena / Elo [4] [5]	Human-preference Elo (с/без tools)	Конкретные Elo для GPT-5.2 не приведены, но агрегатор указывает его в топ-кластер по reasoning/coding. [4]	Anthropic заявляет, что Opus 4.6 опережает «следующую лучшую модель» (GPT-5.2) примерно на 144 Elo. [5]	N/A в явных числах, но во множестве обзоров Gemini 3 Pro фигурирует как лидер по multimodal-кейсам. [1] [3]	N/A	Для Qwen 3.5-Plus Elo ещё нет; ранние успехи Qwen 2.5-Max на Arena (топ в math/coding) относятся к предыдущему поколению. [23]

Несопоставимости:

- Часть результатов **«с инструментами»** (Humanity's Last Exam, BrowseComp, некоторые Arena-Elo) — их не смешиваем напрямую с «no-tools» задачами при выводе интегральных рейтингов.[4] [5]
- Для **Qwen 3.5-Plus** вообще нет консолидированного набора бенчмарков — любые оценки носят характер экстраполяции с 2.5-поколения.[22] [23]

5. Ранги по функциям (с пояснениями и confidence)

5.1. Deep research / длинные рассуждения

Ранг:

1. **Claude Opus 4.6** — лучший Elo, сильнейший на Humanity's Last Exam «с инструментами», особенно на длинном контексте.[5] [4]
2. **GPT-5.2** — выдающиеся MMLU, GPQA, AIME 2025 (100%), плюс отдельный Deep Research-режим.[3] [13] [7] [8]
3. **Gemini 3 Pro** — доминирование в 19/20 бенчмарков против GPT-5.1 и Claude 4.5, но GPT-5.2 и Opus 4.6 частично перехватили лидерство.[13] [1] [3]
4. **Kimi 2.5** — очень сильная математика/код, Thinking Mode, но архитектура всё же оптимизирована под открытый стек и не столь универсально отполирована под все домены.[21] [2]
5. **Qwen 3.5-Plus** — заявлена перфоманс-паритетность с Gemini 3 Pro, но пока без детальных бенчмарков.[22]

Confidence:

- Для позиций 1–3 — **high** (много независимых источников).[1] [3] [13] [4] [5]
- Для Kimi 2.5 — **medium** (детальные цифры есть, но покрывают не все домены).[2] [18] [21]
- Для Qwen 3.5-Plus — **low** (почти нет чисел по 3.5-Plus, только заявления/сравнения уровня «comparable to Gemini 3 Pro»).[22]

5.2. Web-search & retrieval

Ранг:

1. **Claude Opus 4.6** — лучший на BrowseComp, развитый web-fetch и computer-use.[11] [5]
2. **GPT-5.2** — мощный Web Search, Deep Research с таргетированным поиском по сайтам и интеграцией приложений, но по BrowseComp уступает Opus 4.6.[7] [8] [5]
3. **Gemini 3 Pro** — нативный Google Search tool, хороший grounding, но без глубокого компьютерного управления и с меньшим акцентом на агентность, чем у Claude.[14] [17] [3]
4. **Kimi 2.5** — сильный Agent Mode с web-поиском и хорошим BrowseComp-результатом, но инфраструктура и UX менее массовые.[18] [19] [2]

5. **Qwen 3.5-Plus** — по 3.5-Plus нет подтверждённого web-search-продукта, поэтому модель опускается вниз рейтинга, несмотря на сильную историю Qwen 2.5.[23] [22]

Confidence: high для позиций 1–4, low для Qwen.[19] [8] [2] [3] [18] [5] [7]

5.3. Coding / web-dev

Ранг (по совокупности SWE-bench, LiveCodeBench и реальных отчётов):

1. **Claude Opus 4.5/4.6** — SWE-bench Verified 80.9%, лучшие результаты по проф. разработке, сильные terminal-агенты.[3] [11]
2. **Kimi 2.5** ≈ **Gemini 3 Pro** — LiveCodeBench 85.0% и SWE-bench 76.8% у Kimi; Gemini 3 Pro — 76.8% SWE-bench и очень высокие оценки в независимых обзорах по реальным coding-задачам.[21] [2] [1] [3]
3. **GPT-5.2** — SWE-bench 74.9%, мощный code-ассистент (особенно в связке с Deep Research и Canvas), но по голым бенчмаркам уступает Claude/Kimi/Gemini.[6] [7] [3]
4. **Qwen 3.5-Plus** — по Qwen 2.5-72B и Qwen2.5-Max есть данные о топ-1 в math/coding на OpenCompass/Arena, но для 3.5-Plus бенчмарки ещё не консолидированы, поэтому осторожно ставится ниже.[23]

Confidence:

- Позиции Claude, GPT-5.2, Gemini, Kimi — **high** (есть конкретные SWE-bench/LiveCodeBench числа).[2] [21] [3]
- Qwen 3.5-Plus — **low** (экстраполяция с предыдущей версии и твитов).[23] [22]

5.4. Tool-use / агенты

Ранг:

1. **Claude Opus 4.6** — advanced tool use, tool search tool, programmatic tool calling, лучшие результаты на BrowseComp и Vending-Bench.[12] [5] [11]
2. **GPT-5.2** — Deep Research плюс Agent Mode, интеграция с внешними приложениями и сайтами, сильная управляемость.[8] [6] [7]
3. **Kimi 2.5** — Agent Swarm, 200–300 последовательных tool-calls, хороший BrowseComp-результат; архитектура ориентирована именно на открытых агентов.[18] [19] [2]
4. **Gemini 3 Pro** — хороший инструментарий (Search, File Search, Code Execution, URL context), но без Computer Use и с меньшей ориентацией на general-purpose агентов в доках.[17] [14]
5. **Qwen 3.5-Plus** — данных по agent-платформе для этой версии почти нет.[22] [23]

Confidence: high для топ-3 (Anthropic/OpenAI/Moonshot подробно документируют агентные возможности), medium для Gemini, low для Qwen.[14] [12] [6] [5] [7] [11]

5.5. Multimodal (понимание и генерация)

Ранг:

1. **Gemini 3 Pro** — нативно мультимодальный (текст/изображения/аудио/видео), лучший в реальных мультимодальных бенчмарках, мощный vision и видео-понимание.[16] [17] [1] [3]
2. **Kimi 2.5** — MMMU-Pro 78.5%, VideoMMMU 86.6%, нативное видео-понимание с MoonViT-энкодером; нет своего генератора картинок/видео, но понимание очень сильное.[2]
3. **GPT-5.2** — хорошее мультимодальное понимание и генерация изображений, но без столь же глубокого видео-понимания и генерации, как у Gemini/Kimi.[9] [3] [14]
4. **Claude Opus 4.5/4.6** — сильное изображение-вход, но менее выраженная мультимодальная специализация.[12] [1] [3]
5. **Qwen 3.5-Plus** — данных о мультимодальных бенчмарках именно этой версии пока почти нет.[23] [22]

Confidence: high для Gemini/Kimi, medium для GPT-5.2/Claude (меньше деталей именно по multimodal-бенчам), low для Qwen.[17] [16] [1] [3] [2]

5.6. Tutor / обучение

Ранг (именно как «продукт-репетитор»):

1. **GPT-5.2** — единственный с явным **Study & Learn mode**, плюс гибкие thinking-моды.[10] [9] [6]
2. **Gemini 3 Pro** — великолепно подходит для обучения (анализ видео/документов, интеграция с Google Classroom/NotebookLM), но без отдельного study-режима.[16] [17]
3. **Claude Opus 4.6** — сильный reasoning и безопасность, но UX и инструменты менее заточены под образовательные сценарии.[5] [12]
4. **Kimi 2.5** — очень силён по математике и программированию, но интерфейс и экосистема более «технарские» и локальные.[19] [21] [2]
5. **Qwen 3.5-Plus** — мало данных о специализированных образовательных продуктах на этой модели.[22] [23]

Confidence: high для GPT-5.2/Gemini, medium для Claude/Kimi, low для Qwen.

6. Итоговые ранги

6.1. Ранг «общий интеллект» (интегральный)

Как считаем:

- Берём функции: deep research, web-search, coding, tool-use, multimodal, tutor, UX.
- Для каждой модели используем её оценку 0–10 по разделу 3, с весами метрик качества/устойчивости и т. д.
- Мягко учитываем бенчмарки MMLU/GPQA/AIME/SWE-bench/MMMU как корректировки (±0.3–0.5 по функциям). [13] [21] [1] [3] [2]

Итоговый порядок:

1. **Claude Opus 4.6** — лучший в agent/tool-use, web-search и coding, очень сильный deep reasoning на сложных экзаменах; слабее по мультимодальности, но в сумме даёт максимальный «общий интеллект». [4] [11] [3] [12] [5]
2. **GPT-5.2** — лидер по MMLU/GPQA/AIME, лучший в роли универсального «репетитора» и продуктивного ассистента (Deep Research + Canvas + Study), но чуть слабее Opus 4.6 по agent/web-search и coding. [6] [7] [8] [3] [13]
3. **Gemini 3 Pro** — топ-1 мультимодальная модель (vision/video), очень сильный reasoning и coding, но уступает GPT-5.2/Opus 4.6 по чистому текстовому reasoning и агентным возможностям; компенсирует это контекстом 1M+ и глубокой Google-интеграцией. [1] [3] [13] [14] [17] [16]
4. **Kimi 2.5** — выдающиеся open-source показатели по math/coding/agents и multimodal-пониманию (MMMU-Pro, VideoMMMU, SWE-bench, LiveCodeBench), но продукт и экосистема ещё не столь глобально зрелые. [20] [21] [18] [19] [2]
5. **Qwen 3.5-Plus** — по заявлениям, перформанс сопоставим с Gemini 3 Pro, а история Qwen 2.5 показывает способность занимать топ-позиции на OpenCompass/Arena; но пока нет ни систематических бенчмарков, ни подробного описания продуктовых режимов, поэтому интегральный ранг осторожно ниже. [23] [22]

Confidence по общему интеллекту:

- Claude / GPT-5.2 / Gemini — **high** (много независимых бенчмарков и обзоров). [3] [13] [4] [12] [5] [1]
- Kimi 2.5 — **medium** (отлично покрыт частью бенчмарков и обзоров, меньше международных сравнений против GPT-5.2/Opus 4.6). [21] [18] [19] [2]
- Qwen 3.5-Plus — **low** (по сути один твит и общая история Qwen-семейства). [22] [23]

6.2. Ранг «по бенчмаркам» (строго по цифрам, без UX/функций)

Здесь учитываются только MMLU/GPQA/AIME/SWE-bench/LiveCodeBench/MMMU/VideoMMMU, BrowseComp и заявленный Elo-gap, **без учёта** UX, стоимости, наличия Canvas и т. п. [13] [4] [5] [21] [1] [2] [3]

Итоговый порядок:

1. **GPT-5.2** — лучшее сочетание MMLU, GSM8K, GPQA, AIME 2025, приличный SWE-bench, минимальная халлюцинация (4.8%). [3] [13]
2. **Claude Opus 4.6** — лучше GPT-5.2 по Elo и BrowseComp, лучший SWE-bench, но по некоторым «чистым» текстовым бенчам GPT-5.2 чуть впереди. [11] [4] [5] [3]
3. **Gemini 3 Pro** — доминировал в 19/20 тестов против GPT-5.1/Claude 4.5 в 2025-м, особенно по multimodal и сложному reasoning; позже GPT-5.2 и Opus 4.6 частично его обошли. [1] [13] [3]
4. **Kimi 2.5** — выдающиеся AIME/LiveCodeBench/SWE-bench, top-уровень MMMU-Pro/VideoMMMU и BrowseComp, но меньше данных по MMLU/GPQA/GPQA Diamond в сопоставимых настройках. [18] [21] [2]
5. **Qwen 3.5-Plus** — бенчмарков почти нет; история Qwen 2.5 крайне сильная, но без чисел для 3.5-Plus нельзя честно поднять выше. [23] [22]

Confidence:

- GPT-5.2 / Claude / Gemini — **high**. [4] [5] [13] [1] [3]
- Kimi 2.5 — **medium**. [21] [2] [18]
- Qwen 3.5-Plus — **low**. [22] [23]

7. Сводная таблица: модели × функции и ранги

Таблица: сводные оценки и ранги

Модель	Deep research	Web-search	Coding	Tool-use / агенты	Multimodal	Tutor / обучение	UX/Canvas	Итоговый «общий интеллект»	Ранг по бенчмаркам
Claude Opus 4.6	1–2 место	1	1	1	3–4	3	3	**1**	**2**
GPT-5.2	1–2 место	2	3–4	2	3	1	1	**2**	**1**
Gemini 3 Pro	3	3	2–3	4	1	2	2	**3**	**3**
Kimi 2.5	4	4	2–3	3	2	4	4	**4**	**4**
Qwen 3.5-Plus	5 (низкий confidence)	5 (низкий confidence)	3–5 (низкий confidence)	5 (низкий confidence)	5 (низкий confidence)	5 (низкий confidence)	5 (низкий confidence)	**5 (низкий confidence)**	**5 (низкий confidence)**

8. Почему Qwen 3.5-Plus занимает своё место (3–6 тезисов)

1. **Сырые способности, вероятно, очень высокие, но плохо задокументированы.**
   Единственный чёткий источник говорит, что Qwen 3.5-Plus (MoE-модель 397B с 17B активных параметров) показывает перфоманс «на уровне Gemini 3 Pro» при существенно меньшей цене, но без конкретных MMLU/SWE-bench/MMMU цифр. [22]
2. **История семейства Qwen 2.5 — сильный маркер, но это прошлое поколение.**
   Qwen 2.5-72B и Qwen2.5-Max в 2025-м занимали топ-позиции в OpenCompass (math/coding) и высокие ранги в Chatbot Arena, что говорит о сильной инженерной культуре команды, но прямой перенос этого статуса на 3.5-Plus — допущение. [23]
3. **Недостаток информации по продуктовым функциям.**
   В отличие от ChatGPT/Claude/Gemini/Kimi, где чётко описаны Deep Research, agent-режимы, web-search, code-tools, у Qwen 3.5-Plus почти нет публичных продуктовых описаний; нельзя честно поставить высокую оценку UX/Canvas/Study-режимам или web-search-агентам. [23] [22]
4. **Вероятное смещение в сторону «модели, а не продукта».**
   Публичный нарратив вокруг 3.5-Plus — это в первую очередь разговор о параметрах, цене за токен и бенчмарках уровня «comparable to Gemini 3 Pro», а не о финальном пользовательском продукте с режимами Deep Research/Canvas и интеграциями. [22]
5. **Следствие: осторожно низкий интегральный ранг и низкий confidence.**
   При прочих равных Qwen 3.5-Plus мог бы оказаться ближе к Kimi 2.5 и Gemini 3 Pro, но без чётких бенчмарков и описаний инструментов пришлось занизить его место в рейтингах и пометить confidence как низкий. [23] [22]

9. Допущения и ограничения (важно для интерпретации)

- **Неполнота бенчмарков.**
  Для некоторых сочетаний «модель × бенчмарк × режим tools» (особенно Qwen 3.5-Plus, Opus 4.6 на части текстовых тестов, GPT-5.2 на BrowseComp, Gemini/Kimi на GPQA) просто нет публичных численных данных; такие клетки помечены как N/A, а рейтинги по ним строятся только на тех задачах, где есть хоть какие-то числа. [5] [2] [18] [4] [21] [1] [3]
- **Смешение режимов tools on/off.**
  Humanity's Last Exam, BrowseComp, частично Arena-Elo для Opus 4.6 — с инструментами (web, code, tool-calling), тогда как MMLU/GPQA/GSM8K — как правило без tools; поэтому интегральные выводы по «общему интеллекту» учитывают это качественно, а не путём простого усреднения. [4] [5] [3]
- **Продукт vs модель.**
  Для GPT-5.2, Gemini 3 Pro, Claude Opus 4.5/4.6 и Kimi 2.5 есть достаточно чёткое разделение между модельными бенчмарками и конечными продуктами (ChatGPT UI, Claude Apps, Gemini web, Kimi web). [9] [7] [8] [14] [12] [17] [16] [2] [3]

  Для Qwen 3.5-Plus ситуация иная: фокус в источниках идёт на модель и цену, а не на оформленный end-user продукт, поэтому вся часть про UX/инструменты по нему оценена с минимальной уверенностью. [22] [23]
- **Региональные различия и доступность.**
  Kimi 2.5 и Qwen 3.5-Plus в первую очередь ориентированы на китайский рынок, поэтому часть интерфейсов/режимов может существовать, но не иметь хорошей англоязычной документации/обзоров; по условиям задачи такие возможности отмечены как N/A, а не «додуманы». [20] [19] [2] [23]

Если захочешь, можно сделать отдельную узкую табличку, например, только по coding+agents (или только по multimodal), и там уже более детально разложить веса/оценки конкретно под твои кейсы (подготовка к олимпиадам, production-код, ресёрч по рынку и пр.).

⁂

1. https://www.claude5.com/news/llm-comparison-2025-gemini-3-gpt-5-claude-4-5
2. https://llm-stats.com/blog/research/kimi-k2-5-launch
3. https://www.humai.blog/best-ai-models-2026-gpt-5-vs-claude-4-5-opus-vs-gemini-3-pro-complete-comparison/
4. https://lmcouncil.ai/benchmarks
5. https://www.anthropic.com/news/claude-opus-4-6
6. https://sem.co/blog/blog-chatgpt-5-thinking-modes-deep-research-agents-canvas
7. https://creati.ai/ai-news/2026-02-10/openai-deep-research-gpt-5-2-upgrade/
8. https://www.youtube.com/watch?v=2gCqVb2lBwk
9. https://www.youtube.com/watch?v=uYRh1gMAGDo
10. https://www.youtube.com/watch?v=wJHAOW9_7SI
11. https://alphacorp.ai/claude-opus-4-5-launch-everything-you-need-to-know/
12. https://platform.claude.com/docs/en/about-claude/models/whats-new-claude-4-5
13. https://vertu.com/ai-tools/gpt-5-2-vs-gemini-3-pro-complete-benchmark-comparison-performance-analysis-2025/
14. https://ai.google.dev/gemini-api/docs/gemini-3
15. https://deepmind.google/models/gemini/
16. https://www.youtube.com/watch?v=9Uv1ERS7y-I
17. https://blog.google/innovation-and-ai/technology/developers-tools/gemini-3-pro-vision/
18. https://www.thepromptbuddy.com/prompts/kimi-k2-5-explained-how-moonshot-ai-s-open-model-redefines-multimodal-performance
19. https://www.leanware.co/insights/kimi-k2
20. https://www.codecademy.com/article/kimi-k-2-5-complete-guide-to-moonshots-ai-model
21. https://www.aionda.blog/en/posts/moonshot-ai-kimi-k2-5-moe
22. https://x.com/HeyToha/status/2023355105781944692
23. https://generative-engine.org/the-qwen-weekly-wars-how-alibaba-s-rapid-fire-ai-releases-ar-1756587775344
24. https://www.reddit.com/r/LocalLLaMA/comments/1hh99h3/updates_on_chatbot_arena_on_recent_models_qwq/
25. https://www.digitalapplied.com/blog/llm-comparison-guide-december-2025
26. https://arxiv.org/html/2508.16949v6
27. http://scis.scichina.com/en/2025/221301.pdf
28. https://www.youtube.com/watch?v=F375CZ6g6p8
29. https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/3-pro
30. https://www.anthropic.com/claude/opus