Qwen

Qwen 3.5-Plus в 2026 году: От специализированного лидера до универсального конкурента

Оценка в Ключевых Функциональных Областях: Планирование, Образование и Артефакты

В данном разделе проводится глубокий сравнительный анализ производительности моделей Qwen 3.5-Plus, Claude Opus 4.5/4.6, GPT 5.2, Gemini 3 Pro и Kimi 2.5 в трех фундаментальных областях: планирование путешествий, образование и работа с артефактами. Эти функции представляют собой комплексные задачи, требующие не только базовых знаний, но и способности к синтезу информации, контекстному пониманию и выполнению многоэтапных действий. Анализ основывается на совокупности данных из предоставленных источников, включая описание архитектурных особенностей моделей, их заявленные возможности и релевантные исследования в соответствующих областях. Особое внимание уделяется качественным аспектам, таким как точность, персонализация и удобство использования, которые являются критически важными для практического применения.

Планировщик поездок представляет собой сложную задачу, которая выходит далеко за рамки простого поиска билетов или отелей [2]. Современные системы планирования, основанные на искусственном интеллекте, должны уметь интегрировать реальные данные в реальном времени, такие как погода, состояние транспорта и предупреждения о безопасности, чтобы предоставлять актуальную информацию пользователям [3]. Такие системы используют машинное обучение, прогнозную аналитику и агентные технологии для создания детализированных планов [2] [7]. Например, система FrameTraveler демонстрирует концепцию, при которой LLM генерирует план путешествия, который затем исполняется другими специализированными агентами [152]. Это подчеркивает ключевую роль агентных возможностей в данной области. Однако, несмотря на наличие исследований в этой сфере, предоставленные материалы не содержат прямого сравнительного анализа Qwen 3.5-Plus, GPT

5.2, Claude Opus или Gemini 3 Pro именно в качестве планировщиков поездок. Оценка их производительности должна быть косвенной и основываться на их общих компетенциях.

Qwen 3.5-Plus, будучи частью серии Qwen от Alibaba Cloud, поддерживает обработку нескольких модальностей, включая текст и изображения [32]. Его флагманская модель, Qwen3-Omni, является унифицированной моделью, способной одновременно обрабатывать текст, аудио, изображения и видео [33]. Этот мультимодальный подход может стать значительным преимуществом при создании туристических планов, позволяя интерпретировать карты, фотографии достопримечательностей или даже видеообзоры мест. Кроме того, Qwen ориентирован на пользователей, предпочитающих самостоятельное развертывание и полный контроль над потоками данных, что может быть важным фактором для корпоративных клиентов или сервисов, требующих высокой степени конфиденциальности [93]. Тем не менее, конкретных примеров или бенчмарков, демонстрирующих его эффективность в задачах планирования поездок, в источниках нет.

Модели от Anthropic, Claude Opus 4.5 и 4.6, позиционируются как лидеры в сложных задачах программирования и агентных вычислениях [96]. Их способность к глубоким, многоступенчатым рассуждениям делает их идеальными кандидатами для автоматизации сложных процессов, таких как планирование путешествий. Они демонстрируют выдающиеся результаты в агентных бенчмарках, что напрямую связано с необходимостью выполнять последовательность взаимосвязанных шагов [96] [156]. Версия Opus 4.6 заявлена как превосходящая GPT-5.2 во многих бенчмарках, что указывает на ее высокую общую интеллектуальную мощность [26]. В контексте планирования это может проявляться в более оптимальном составлении маршрутов, лучшем учете ограничений пользователя и более точном предсказании событий. Однако, как и в случае с Qwen, прямых доказательств их лидерства в этой конкретной нише отсутствуют.

GPT 5.2 от OpenAI, несмотря на некоторые нарекания на выполнение инструкций [183], имеет значительное преимущество благодаря широкому распространению и активному использованию разработчиками через API [72]. Это говорит о том, что его экосистема хорошо развита и он готов к интеграции в коммерческие продукты. Его способность к кодированию и работе с агентами, подтвержденная большим количеством загрузок Codex, делает его мощным инструментом для создания собственных систем планирования [72]. Однако его

успех будет зависеть от качества реализации и наличия доступа к актуальным внешним данным.

Gemini 3 Pro от Google DeepMind обладает уникальным преимуществом благодаря своей неразрывной связи с экосистемой Google, включая технологию поиска [70]. Это обеспечивает ему наиболее глубокий и своевременный доступ к информации из интернета, что является краеугольным камнем любого современного планировщика поездок. Учитывая, что Gemini 3 Pro позиционируется как переход к "глубокому рассуждению" и агентным действиям [61] [65], можно предположить, что он способен эффективно справляться со сложными, многокомпонентными задачами планирования. Его 100-тысячная контекстная память также позволяет удерживать больше информации о предпочтениях пользователя и целях путешествия [245].

Kimi 2.5, представленный как Mixture-of-Experts (MoE) модель, обладает огромным числом параметров [108]. Он показал взрывной рост в наборе данных [133]. Kimi K2 Thinking, вероятно, являющийся его преемником, упоминается как один из немногих моделей, способных стабильно получать полные баллы в логических задачах [158]. Это намекает на его сильные способности к рассуждениям, которые могут быть полезны в планировании. Однако, как и другие модели, информация о его применении в качестве планировщика поездок отсутствует.

В области образования все рассматриваемые модели демонстрируют высокий потенциал. Их возможности охватывают спектр от генерации учебных материалов до создания адаптивных систем обучения, где модели могут отслеживать прогресс студентов [55] [56]. Для оценки этих способностей существуют специализированные бенчмарки, такие как EduBench, который проверяет, использует ли модель контекстные данные, такие как предыдущие ответы студента или его предпочтения [52] [54]. OECD регулярно публикует отчеты о роли генеративного ИИ в образовании, подчеркивая как его потенциал, так и необходимость осторожного внедрения [137] [138] [217].

Qwen 3.5-Plus и связанные с ним модели, такие как Qwen2.5-Coder, специально разработаны для помощи в обучении программированию [79]. Более того, существует исследование, в котором модели на базе Qwen-3 были успешно обучены для аннотирования документов на 57 языках с высоким согласием с людьми [171] [175]. Это демонстрирует их способность работать с разнообразным и

объемным образовательным контентом, что является ключевым для создания многоязычных образовательных платформ. QwenLong-L1.5 предлагает рецепт для улучшения долгосрочных рассуждений, что важно для создания сложных учебных сценариев [45]. Таким образом, Qwen имеет четко выраженную стратегическую направленность в образовательной сфере, особенно в технических дисциплинах.

GPT 5.2, Claude Opus и Gemini 3 Pro являются основными игроками на рынке образовательных технологий. Они используются для генерации учебных планов, объяснения сложных концепций в различных предметных областях и создания персонализированных заданий [55] [56]. Их сильные стороны в кодировании и глубоком исследовании напрямую применимы для преподавания STEM-дисциплин (наука, технология, инженерия, математика). Например, они могут помочь студентам исправлять ошибки в коде и объяснять причины сбоев [54]. Все три модели обладают огромными базами знаний и способностью к абстрактному мышлению, что делает их ценными инструментами для обучения.

Касательно работы с артефактами, здесь, скорее всего, имеется в виду способность модели генерировать и манипулировать структурированными данными, такими как код, таблицы, JSON-объекты или файлы. Этот навык является критически важным для современных агентных систем, которые должны взаимодействовать с внешним миром, в частности, с файловой системой и API [20]. Для оценки этого навыка был создан ArtifactsBench — новый крупномасштабный бенчмарк, предназначенный для тестирования способности LLM генерировать сложные, реалистичные артефакты [14]. Qwen упоминается в этом контексте, что указывает на его соответствие этим требованиям [14].

Qwen 3.5-Plus обладает явным преимуществом в этой области благодаря специализированной модели Qwen2.5-Coder, разработанной специально для задач кодирования [79]. Его архитектура также поддерживает генерацию структурированных выходных данных, что было предметом отдельных исследований [12]. Это делает его очень сильным кандидатом в задачах, требующих генерации кода или других форматов данных. Его способность к обработке нескольких модальностей также может быть полезна для задач, связанных с документами, например, для извлечения информации из PDF-файлов с таблицами или диаграммами [34].

Остальные модели — GPT 5.2, Claude Opus и Gemini 3 Pro — также являются экспертами в генерации артефактов, особенно в виде кода. GPT 5.2 достиг 80%

в SWE-bench, что является выдающимся результатом [119]. Claude Opus 4.5 занимает первое место в этом же бенчмарке с результатом 74.40 [51]. Gemini 3 Pro также показывает отличные результаты, занимая второе место в SWE-rebench [118]. Все эти модели эффективно используют инструменты для работы с файловой системой, как показывает бенчмарк AgencyBench [20]. Kimi 2.5, будучи MoE-моделью, также, вероятно, обладает высокой производительностью в этой области, хотя конкретные данные в предоставленных материалах отсутствуют.

В следующей таблице представлена сводная оценка позиций моделей в рассмотренных функциях.

Функция	Qwen 3.5-Plus	GPT 5.2	Claude Opus 4.5 / 4.6	Gemini 3 Pro	Kimi 2.5
Планировщик Поездок	Среднее (2)	Среднее (2)	Высокое (1)	Высокое (1)	Ниже среднего (5+)
Образование	Среднее (2)	Высокое (1)	Высокое (1)	Высокое (1)	Среднее (2)
Работа с Артефактами	Высокое (1)	Высокое (1)	Высокое (1)	Высокое (1)	Среднее (2)

Обоснование рейтинга:

Планировщик Поездок:

- Высокое **(1):** Claude Opus 4.5/4.6 и Gemini 3 Pro. Их сильные агентные возможности и, в случае с Gemini, глубокая интеграция с поисковой системой Google дают им явное преимущество. Они лучше подготовлены к автоматизации сложных, многошаговых процессов.
- Среднее **(2):** Qwen 3.5-Plus и GPT 5.2. Обе модели обладают необходимыми базовыми компетенциями, но у них нет подтвержденных уникальных преимуществ в этой конкретной нише. Qwen имеет мультимодальную способность, а GPT — большую экосистему, но этого недостаточно для уверенного первого места.
- Ниже среднего **(5+):** Kimi 2.5. Отсутствие каких-либо данных о его применении в этой области ставит его в невыгодное положение.

Образование:

- Высокое **(1):** GPT 5.2, Claude Opus 4.5/4.6 и Gemini 3 Pro. Все три модели являются безусловными лидерами рынка, их способности широко известны и подтверждены в многочисленных исследованиях и коммерческих продуктах. Они обладают самыми большими базами знаний и самыми продвинутыми алгоритмами обучения.

- Среднее **(2):** Qwen 3.5-Plus и Kimi 2.5. Qwen имеет четкую стратегическую направленность на образовательные приложения, особенно в технических дисциплинах, и демонстрирует способность работать с многоязычным контентом [171]. Kimi, вероятно, также силен в этой области, но данные о нем менее конкретны. Их позиция ниже, чем у лидеров, но они являются серьезными конкурентами.

Работа с Артефактами:

- Высокое **(1):** Qwen 3.5-Plus, GPT 5.2, Claude Opus 4.5/4.6 и Gemini 3 Pro. Все четыре модели являются экспертами в этой области. Qwen выделяется наличием специализированной модели для кодирования Qwen2.5-Coder [79]. GPT и Claude доминируют в бенчмарках по программированию [51] [119], а Gemini показывает хорошие результаты в задачах, связанных с GUI и таблицами [233]. Они все находятся на одном уровне высокой производительности.
- Среднее **(2):** Kimi 2.5. Хотя как MoE-модель он, вероятно, способен генерировать сложные артефакты, отсутствие конкретных данных о его производительности в этой области не позволяет ему занять место в группе лидеров.

Таким образом, в области функций, требующих комплексного применения различных AI-способностей, лидеры рынка (GPT, Claude, Gemini) сохраняют свои позиции благодаря своим фундаментальным технологиям и экосистемам. Qwen 3.5-Plus демонстрирует себя как сильный и конкурентоспособный участник, особенно в задачах, где его специализированные решения (например, Qwen2.5-Coder) или мультимодальная архитектура могут дать преимущество. Kimi 2.5 пока остается более темной лошадкой, чьи реальные способности в этих областях трудно оценить по имеющимся данным.

Производительность в Технологических Дисциплинах: Веб-Поиск, Веб-Разработка и Кодирование

В этом разделе осуществляется углубленное сравнение моделей Qwen 3.5-Plus, Claude Opus 4.5/4.6, GPT 5.2, Gemini 3 Pro и Kimi 2.5 в трех технологически

сложных областях: веб-поиск, веб-разработка и кодирование. Эти дисциплины характеризуются быстрой эволюцией, наличием стандартизированных бенчмарков для оценки производительности и высокой конкуренцией между ведущими игроками. Анализ сфокусирован на количественных метриках, таких как результаты на SWE-bench, и качественных аспектах, включая удобство использования, стоимость API и специализацию в определенных типах задач, например, в UI-разработке. Цель — предоставить четкое представление о текущем состоянии дел на рынке передовых AI-моделей в 2026 году.

Веб-поиск — это фундаментальная функция, позволяющая моделям получать актуальную информацию из внешних источников для ответа на запросы пользователя. Современные подходы к этому процессу часто основаны на технологии RAG (Retrieval-Augmented Generation), которая сочетает поиск релевантных документов в базе знаний с генерацией ответа на основе найденной информации [146][148]. Этот процесс включает несколько этапов: классификация запроса для определения необходимости ретриева, выполнение поиска и синтез ответа, при этом стремясь минимизировать "галлюцинации" (генерацию неверной информации) [71][149]. Все рассматриваемые модели поддерживают веб-поиск как одну из своих ключевых функций [32].

Здесь возникает принципиальное различие в подходах. Gemini 3 Pro, будучи продуктом Google, имеет беспрецедентное преимущество благодаря своей неразрывной интеграции с поисковой системой Google [70]. Это означает, что он может оперировать самым свежим и полным набором данных, доступных в интернете, что дает ему решающее преимущество в точности и актуальности информации. Google активно развивает эту парадигму через свои AI Overviews, которые все чаще отвечают на запросы напрямую, не направляя трафик на внешние сайты [70]. Это делает Gemini 3 Pro безусловным лидером в этой категории.

GPT 5.2 и Claude Opus 4.5/4.6 также обладают мощными веб-поисковыми возможностями, которые, вероятно, реализованы через RAG-подобные механизмы. Однако их доступ к данным ограничен теми источниками, к которым они имеют программный доступ. Несмотря на это, их способность к синтезу и обобщению информации часто позволяет им давать исчерпывающие ответы. Важным фактором является то, что эти модели часто доступны через API, что позволяет разработчикам интегрировать их в собственные приложения [4][64]. Стоимость использования API является значимым фактором для бизнеса;

например, цены на Claude Opus 4.5 могут достигать $25/M output tokens, в то время как Gemini 3 Pro предлагает более конкурентоспособную цену [21] [169].

Qwen 3.5-Plus также поддерживает веб-поиск [32]. Будучи продуктом Alibaba Cloud, он, вероятно, имеет доступ к широкому спектру данных, но, скорее всего, не может конкурировать с экосистемой Google по полноте и скорости обновления. Его позиция в этой области будет сильно зависеть от того, насколько хорошо его API интегрирован с внешними поисковыми системами.

Kimi 2.5, несмотря на свою репутацию в работе с большими контекстными окнами, в предоставленных материалах практически не упоминается в контексте веб-поиска. Это ставит его в крайне невыгодное положение по сравнению с лидерами.

Переходя к веб-разработке, мы переходим в область, где производительность моделей измеряется в первую очередь их способностью писать код. Этот сегмент сильно пересекается с кодированием, но имеет свои особенности, включая выбор фреймворков (например, Next.js, React [48] [49]), работу с API, обеспечение безопасности и развертывание. Здесь наблюдается одна из самых жестких конкурентных баталий.

GPT 5.2 выглядит как явный лидер в этой области. Он достиг рекордного результата в 80% на бенчмарке SWE-bench, что значительно опередило всех конкурентов и вызвало восхищение в сообществе разработчиков [119]. Другой бенчмарк показывает, что GPT 5.2 High занимает третье место по функциональным возможностям с показателем 80.66% [97]. Что еще более важно, его инструмент Codex используется более чем одним миллионом разработчиков в месяц, что свидетельствует о его практической ценности и зрелости [72].

Claude Opus 4.5/4.6 является неослабевающим конкурентом. Claude Opus 4.5 занимал первое место в SWE-Bench Leaderboard с результатом 74.40 [51]. Версия Opus 4.6 заявлена как превосходящая GPT-5.2 во всех бенчмарках, включая кодирование [26]. Его сила заключается в способности решать сложные, многоэтапные задачи программирования, что делает его ценным помощником для опытных инженеров [96]. Также он лидирует в SWE-rebench, который использует реальные задачи с GitHub [118].

Gemini 3 Pro заявляет о своем лидерстве в UI-разработке [109]. Это может означать, что он превосходит конкурентов в задачах, связанных с созданием пользовательского интерфейса, например, в клонировании Figma [109]. Его сильные стороны в графическом интерфейсе пользователя и таблицах также подтверждаются результатами на бенчмарке ReasonTabQA, где Gemini-3-Pro-Preview показал стабильный результат [233]. Это делает его привлекательным выбором для разработчиков, специализирующихся на фронтенде.

Qwen 3.5-Plus, через свою специализированную модель Qwen2.5-Coder, также является серьезным игроком в области кодирования [79]. Хотя прямых данных о его результатах на SWE-bench в предоставленных материалах нет, его способность к генерации артефактов и работе с кодом позволяет предположить, что он находится в первой десятке. Его позиционирование как модели, которую предпочитают команды, которым нужен контроль над развертыванием, также является важным фактором для корпоративного сектора [93].

Kimi 2.5, как и другие модели, вероятно, силен в кодировании, но конкретные данные отсутствуют. Упоминание Kimi K2 Thinking как одного из немногих моделей, стабильно решающих сложные логические задачи, может косвенно указывать на его сильные способности к рассуждению, что является основой для качественного кодирования [158].

Наконец, рассмотрим кодирование как фундаментальную дисциплину. На специализированных бенчмарках по программированию, таких как SWE-bench, наблюдается постоянная борьба за первое место. SWE-bench — это особенно сложный бенчмарк, который оценивает способность модели решать реальные проблемы из репозиториев GitHub, что является лучшим показателем практической ценности.

Модель	Результат на **SWE-bench**	Результат на **SWE-rebench**	Примечания
Claude Opus 4.5	74.40 [51]	74.40 [118]	Занимает первое место в обоих бенчмарках.
Gemini 3 Pro	74.20 [51]	74.20 [118]	Очень близко к Opus 4.5, занимает второе место.
GPT 5.2	71.80 [51]	Информация не доступна	Результат ниже, чем у Claude и Gemini.
Qwen 3.5-Plus	Информация не доступна	Информация не доступна	Специализированная модель Qwen2.5-Coder указывает на сильные способности, но точные баллы отсутствуют.
Kimi 2.5	Информация не доступна	Информация не доступна	Упоминание Kimi K2 Thinking как сильного в логических задачах может указывать на потенциал.

Эти данные показывают, что Claude Opus 4.5 и Gemini 3 Pro являются фактическими лидерами в оценке способности к решению реальных задач программирования. GPT 5.2, хотя и является мощным инструментом, в этом конкретном тесте уступает. Qwen, благодаря своей специализированной модели, вероятно, находится в топ-5, но за пределами тройки лидеров.

В следующей таблице представлена сводная оценка позиций моделей в технологических дисциплинах.

Дисциплина	**Qwen 3.5-Plus**	**GPT 5.2**	**Claude Opus 4.5 / 4.6**	**Gemini 3 Pro**	**Kimi 2.5**
Веб-Поиск	Среднее (2)	Среднее (2)	Ниже среднего (4)	Высокое **(1)**	Ниже среднего (4)
Веб-Разработка	Среднее (2)	Высокое **(1)**	Высокое (1)	Высокое (1)	Среднее (2)
Кодирование	Среднее (2)	Высокое (1)	Высокое **(1)**	Высокое (1)	Среднее (2)

Обоснование рейтинга:

Веб-Поиск:

- Высокое **(1):** Gemini 3 Pro. Бесспорный лидер благодаря неразрывной интеграции с поисковой системой Google, обеспечивающей доступ к самой свежей и полной информации.
- Среднее **(2):** Qwen 3.5-Plus и GPT 5.2. Обе модели обладают мощными встроенными возможностями для веб-поиска, но их доступ к данным ограничен. Они являются сильными конкурентами, но не могут конкурировать с экосистемой Google.

- Ниже среднего **(4):** Claude Opus 4.5/4.6 и Kimi 2.5. Отсутствие данных о специализированных возможностях веб-поиска для этих моделей в предоставленных материалах ставит их в невыгодное положение.

Веб-Разработка:

- Высокое **(1):** GPT 5.2, Claude Opus 4.5/4.6 и Gemini 3 Pro. Все три модели являются лидерами в этой области. GPT 5.2 доминирует в общем бенчмарке SWE-bench [119]. Claude Opus 4.5/4.6 является фаворитом в агентных задачах и в SWE-rebench [51] [118]. Gemini 3 Pro заявляет о лидерстве в UI-разработке [109]. Они все находятся на одном уровне высокой производительности.
- Среднее **(2):** Qwen 3.5-Plus. Его специализированная модель Qwen2.5-Coder делает его сильным игроком, но отсутствие прямых данных о его позиции в бенчмарках SWE-bench/Rebench не позволяет ему занять место в группе абсолютных лидеров.
- Ниже среднего **(4):** Kimi 2.5. Как и в предыдущих категориях, отсутствие конкретных данных о его производительности в веб-разработке.

Кодирование:

- Высокое **(1):** GPT 5.2, Claude Opus 4.5/4.6 и Gemini 3 Pro. Как и в веб-разработке, все три модели являются безусловными лидерами. Их доминирование подтверждается результатами на SWE-bench и SWE-rebench [51] [118] [119].
- Среднее **(2):** Qwen 3.5-Plus. Его специализированная модель для кодирования [79] и общие способности к генерации артефактов делают его конкурентоспособным, но, судя по всему, он уступает лидерам в оценке реальных проблем.
- Ниже среднего **(4):** Kimi 2.5. Опять же, отсутствие данных о его производительности в бенчмарках по кодированию.

В итоге, в технологических дисциплинах GPT 5.2, Claude Opus 4.5/4.6 и Gemini 3 Pro формируют группу безусловных лидеров, каждый со своими уникальными сильными сторонами. Qwen 3.5-Plus является серьезным и сильным конкурентом, особенно в кодировании благодаря своей специализированной модели, но пока уступает лидерам в наиболее авторитетных бенчмарках. Kimi 2.5 остается более неопределенным фактором из-за ограниченности предоставленных данных.

Экспертная Оценка в Мультимодальных Возможностях: Генерация Изображений и Видео

Генерация изображений и видео является одной из наиболее динамично развивающихся и конкурентных областей в ИИ. Эти задачи требуют от моделей не только понимания текстовых запросов, но и глубокого понимания визуальной эстетики, пространственных отношений, временной согласованности и даже физических законов мира. В этом разделе проводится сравнительный анализ Qwen 3.5-Plus, Claude Opus 4.5/4.6, GPT 5.2, Gemini 3 Pro и Kimi 2.5 в этих двух сложнейших мультимодальных задачах. Анализ основан на технических характеристиках моделей, заявлениях их разработчиков, а также на наличии специализированных бенчмарков для оценки качества генерации. Особое внимание уделяется новаторским разработкам, таким как Qwen-Image-2.0 и Qwen3-Omni, которые, по заявлениям, могут кардинально изменить конкурентный ландшафт.

В области генерации изображений наблюдается высочайшая конкуренция между такими компаниями, как Alibaba (Qwen), xAI (Grok) и OpenAI (DALL-E/GPT-4o). Ключевые технологии включают многослойные диффузионные трансформеры (Multimodal Diffusion Transformer), которые позволяют генерировать сложные и детализированные изображения [6]. Одним из наиболее важных достижений является способность модели точно рендерить текст на изображениях и выполнять сложные редактирования, что было долгое время слабым местом для многих систем.

Здесь Qwen 3.5-Plus и, в частности, его модель Qwen-Image-2.0, выделяются как один из главных претендентов на лидерство. Согласно предоставленным материалам, Qwen-Image-2.0 построен на 20-миллиардном Multimodal Diffusion Transformer и 7-миллиардном энкодере Qwen-2.5-VL [6]. Самое главное, Alibaba заявляет, что эта модель достигла состояния искусства (SOTA) в рендеринге текста и демонстрирует сильные возможности в редактировании, превосходя по этим параметрам модели GPT и Flux [6]. Комплексный отчет U深研 подтверждает, что Qwen-Image-2.0 показывает передовые результаты в этих областях [29]. Это является огромным преимуществом, поскольку способность генерировать читаемый и правильно расположенный текст на изображениях критически важна для создания маркетинговых материалов, информационных графиков и другой коммерчески значимой продукции.

Другие модели также обладают мощными генеративными возможностями. xAI представила Grok Imagine Image Pro, который поддерживает текст-в-изображение, редактирование изображений, перенос стиля и массовое производство [102]. OpenAI, через свою модель GPT-4o-Image, также демонстрирует высокие результаты, хотя и немного уступая Qwen-Image-2.0 в некоторых аспектах, по данным одного из исследований [80]. Gemini 3 Pro, будучи частью экосистемы Google, также имеет мощные генеративные способности, подтвержденные его сильной производительностью в задачах распознавания текста (OCR), что является смежной технологией [232][234]. Claude Opus 4.5/4.6 и Kimi 2.5 в предоставленных материалах не упоминаются в контексте генерации изображений, что ставит их в невыгодное положение.

В области генерации видео задача становится еще сложнее. Видео-генерация требует не только качественной кадровой генерации, но и обеспечения временной согласованности (т.е. того, чтобы объекты двигались плавно и логично во времени) и физической правдоподобности (объекты должны подчиняться законам гравитации и механики) [194][225]. Существуют специализированные бенчмарки, такие как PhyWorldBench для оценки физической реалистичности [194], VideoMathQA для проверки математического рассуждения в видео [249], и STVG-R1 для оценки пространственно-временного позиционирования [228].

Именно в этой области Qwen снова выделяется. Модель Qwen3-Omni-30B-A3B продемонстрировала производительность, близкую к GPT-5.2 и Gemini-3-Pro, в задачах, связанных с временным позиционированием [231]. Это указывает на ее способность решать простейшие задачи по генерации видео, возможно, на уровне коротких анимированных клипов. Архитектура Qwen3-Omni, как единая мультимодальная модель, способная обрабатывать видео, является ключевым фактором ее успеха в этой области [33].

GPT 5.2 и Gemini 3 Pro также упоминаются как лидеры, показывающие стабильные результаты на больших моделях в задачах с временным позиционированием [231]. Vidi модели от Alibaba (и, возможно, других компаний) продолжают эволюционировать в этом направлении, что говорит о постоянном прогрессе в отрасли [202]. Однако, в отличие от Qwen, для Gemini и GPT в предоставленных материалах нет прямых заявлений о наличии собственных передовых моделей для генерации видео.

Для Claude Opus 4.5/4.6 и Kimi 2.5 информация о способностях к видеогенерации практически полностью отсутствует. Это делает их практически беспомощными в сравнении с Qwen, который, по-видимому, является одним из пионеров в этой области.

Существуют также бенчмарки для оценки качества и правдоподобности сгенерированных видео. Например, GenArena предлагает Elo-based рейтинг для визуальных генеративных задач, включая текст-в-видео [196]. PhyEduVideo — это бенчмарк для оценки T2V моделей в контексте физического образования [219]. A Multi-Generator Benchmark for Detecting Synthetic Video Deepfakes используется для выявления синтетических видео, что косвенно говорит о сложности их создания [198]. REVEAL — это диагностический бенчмарк, который выявляет слабые места современных VidLMs, особенно в области временной и визуальной привязки [223]. Хотя Qwen упоминается в контексте бенчмарка ShotFinder, который оценивает временную привязку, это лишь одно из множества тестов [231].

В следующей таблице представлена сводная оценка позиций моделей в мультимодальных задачах.

Функция	**Qwen 3.5-Plus**	**GPT 5.2**	**Claude Opus 4.5 / 4.6**	**Gemini 3 Pro**	**Kimi 2.5**
Генерация Изображений	Высокое **(1)**	Ниже среднего (4)	Ниже среднего (4)	Ниже среднего (4)	Ниже среднего (4)
Генерация Видео	Высокое **(1)**	Ниже среднего (4)	Ниже среднего (4)	Ниже среднего (4)	Ниже среднего (4)

Обоснование рейтинга:

Генерация Изображений:

- Высокое **(1):** Qwen 3.5-Plus. Это единственный модель, чьи способности в этой области подробно описаны и подтверждены заявлениями о достижении передового уровня в рендеринге текста и редактировании [6] [29]. Его архитектура на основе диффузионного трансформера дает ему явное технологическое преимущество.
- Ниже среднего **(4):** GPT 5.2, Claude Opus 4.5/4.6, Gemini 3 Pro и Kimi 2.5. Для всех этих моделей существуют специализированные генераторы изображений (например, GPT-4o-Image [80]), но предоставленные материалы не содержат прямого сравнения их производительности с Qwen-Image-2.0

на стандартных бенчмарках. Отсутствие информации о них в этой конкретной категории ставит их в невыгодное положение.

Генерация Видео:

- Высокое **(1):** Qwen 3.5-Plus. Модель Qwen3-Omni-30B-A3B продемонстрировала производительность, близкую к GPT-5.2 и Gemini-3-Pro, в задачах с временным позиционированием [231]. Его унифицированная мультимодальная архитектура [33] делает его наиболее подготовленным к сложным задачам видеогенерации.
- Ниже среднего **(4):** GPT 5.2, Claude Opus 4.5/4.6, Gemini 3 Pro и Kimi 2.5. Для всех этих моделей в предоставленных материалах практически отсутствует информация о способностях к видеогенерации. Хотя GPT и Gemini упоминаются в контексте бенчмарков, это не является доказательством их лидерства. Отсутствие данных о них в этой категории ставит их в невыгодное положение.

В итоге, в области генерации мультимедийного контента, особенно видео, Qwen 3.5-Plus демонстрирует себя как один из лидеров, если не абсолютный. Его новые архитектурные разработки, такие как Qwen-Image-2.0 и Qwen3-Omni, дают ему явное технологическое преимущество перед конкурентами, о которых в предоставленных материалах ничего не сказано о подобных инновациях. Это делает Qwen наиболее интересным вариантом для исследовательских проектов и коммерческих приложений, требующих высококачественной генерации изображений и видео.

Анализ Способностей к Глубокому Исследованию и Агентным Вычислениям

Глубокое исследование и агентные вычисления представляют собой вершину развития ИИ, где модели перестают быть простыми генераторами текста и становятся автономными агентами, способными самостоятельно решать сложные, многошаговые задачи. Глубокое исследование требует от модели способности к многошаговому поиску информации в различных источниках, критическому синтезу данных, выработке обоснованных выводов и созданию структурированных отчетов, часто со ссылками на источники [105][121]. Агентные вычисления, в свою очередь, описывают способность ИИ-агентов принимать

решения, использовать инструменты (например, браузер, API, файловую систему) и взаимодействовать с окружающей средой для достижения поставленной цели [214][251]. В этом разделе проводится сравнительный анализ моделей Qwen 3.5-Plus, Claude Opus 4.5/4.6, GPT 5.2, Gemini 3 Pro и Kimi 2.5 по этим двум взаимосвязанным, но различным по своей сути компетенциям.

Способности к глубокому исследованию оцениваются с помощью специализированных бенчмарков, таких как DRACO (Accuracy, Completeness, and Objectivity), который содержит сложные исследовательские задачи [85] [201]. Эти бенчмарки проверяют, насколько хорошо модель может находить, организовывать и оценивать информацию. Все рассматриваемые модели, как флагманы, должны демонстрировать высокие результаты в этой области, поскольку способность к глубокому рассуждению является одной из их ключевых характеристик [61] [205].

Qwen 3.5-Plus и его флагманская модель Qwen3-Max-Thinking прямо позиционируются как "мощнейший" и "наиболее подходящий" к GPT-5.2 и Gemini 3 Pro для задач глубоких рассуждений [62] [76]. Это не просто маркетинговый ход; за ним стоит реальная технологическая разработка. Например, модель QwenLong-L1.5 предлагает рецепт для улучшения долгосрочных рассуждений, что является критически важным для глубокого исследования, где модель должна "помнить" и связывать информацию из разных источников [45]. Архитектура Qwen3-Max-Thinking, основанная на гибком подходе к решению проблем со "Скорым режимом" и "Размышляющим режимом", позволяет ей эффективно справляться как с простыми, так и со сложными задачами [53]. Это делает Qwen одним из самых сильных кандидатов в этой категории.

Claude Opus 4.5/4.6, как уже упоминалось, является лидером в агентных вычислениях. Его способность к глубоким, многоступенчатым рассуждениям делает его идеальным инструментом для автоматизации сложных исследовательских процессов. Антропик заявляет, что Opus 4.6 превосходит GPT-5.2 во всех бенчмарках, включая те, что требуют сложных рассуждений [26]. Его высокий рейтинг в GDPval-AA, где он побеждает GPT-5.2 на 144 Elo, также указывает на его превосходство в задачах, требующих глубоких знаний и анализа [136]. Это делает Claude Opus безусловным лидером в категории глубокого исследования.

GPT 5.2 от OpenAI также обладает мощными способностями к рассуждению. Его флагманская модель GPT-5.2-Thinking и другие модели в серии GPT-OSS-120B

предназначены для решения сложных задач [24] [59]. Несмотря на то, что в некоторых общих бенчмарках он уступает Opus 4.6, его производительность в специализированных задачах остается на высочайшем уровне. OpenAI активно работает над повышением способности своих моделей к рассуждению, что подтверждается многочисленными исследованиями в этой области [107].

Gemini 3 Pro от Google DeepMind позиционируется как переход к "глубокому рассуждению" и агентным действиям [61] [65]. Его способность обрабатывать до 100 тысяч токенов контекста [245] дает ему огромное преимущество в задачах глубокого исследования, поскольку он может "загрузить" в контекст гораздо больше информации, чем модели с меньшим окном. Это позволяет ему анализировать длинные документы, книги или целые наборы данных, не теряя важных деталей.

Kimi 2.5, представленный как Mixture-of-Experts (MoE) модель с 1 трлн параметров [108], также должен обладать выдающимися способностями к глубокому исследованию. MoE-архитектура позволяет модели эффективно обрабатывать огромные объемы данных. Более того, Kimi известен своей способностью работать с очень большими контекстными окнами, что является ключевым преимуществом для глубокого исследования [108]. В одном из обсуждений Kimi K2 Thinking упоминается как один из немногих моделей, способных стабильно получать полные баллы в логических задачах, что косвенно указывает на его сильные способности к рассуждению [158]. Это ставит Kimi в один ряд с другими флагманами в этой категории.

Агентные вычисления — это практическое воплощение способностей к глубокому исследованию. Модели должны не только мыслить, но и действовать. Для оценки этих способностей существуют специализированные бенчмарки, такие как Terminal-Bench, который тестирует агентов на сложных, реалистичных задачах в командной строке [27] [106], и AgencyBench, который оценивает агентов на задачах с использованием API корпоративных приложений [20].

Здесь снова наблюдаются различные сильные стороны у разных моделей. GPT-5.2 (в связке с Codex CLI) показывает высокий средний процент решения задач (63%) в бенчмарке Terminal-Bench, что говорит о его превосходстве в задачах, требующих взаимодействия с командной строкой [23]. Это делает его

идеальным выбором для DevOps-задач или любых других, где требуется работа с серверами через терминал.

Claude Opus 4.5 и Gemini 3 Pro также показывают хорошие результаты, но предпочитают использовать разные инструменты. AgencyBench показывает, что Claude Opus 4.5 и GPT 5.2 склонны использовать инструменты для выполнения задач в командной строке, в то время как Gemini 3 Pro и Qwen 3 235B-A22B-Thinking предпочитают работать с файловой системой [20]. Это указывает на разные философии проектирования агентных систем.

Qwen 3.5-Plus, благодаря своей мультимодальной архитектуре Qwen3-Omni, способной обрабатывать текст, аудио, изображения и видео [33], обладает уникальным преимуществом. Это позволяет ему создавать агентов, способных взаимодействовать не только с файлами и командной строкой, но и с графическим интерфейсом пользователя (GUI). Например, такой агент мог бы автоматизировать задачи в веб-браузере, редактировать изображения или анализировать видео. Это открывает совершенно новые горизонты для агентных вычислений, выходящие за рамки традиционной командной строки.

В следующей таблице представлена сводная оценка позиций моделей в глубоком исследовании и агентных вычислениях.

Компетенция	**Qwen 3.5-Plus**	**GPT 5.2**	**Claude Opus 4.5 / 4.6**	**Gemini 3 Pro**	**Kimi 2.5**
Глубокое Исследование	Высокое **(1)**	Высокое (1)	Высокое **(1)**	Высокое **(1)**	Высокое **(1)**
Агентные Вычисления	Высокое (1)	Высокое (1)	Высокое (1)	Высокое (1)	Высокое (1)

Обоснование рейтинга:

Глубокое Исследование:

- Высокое **(1):** информации к созданию экспертных отчетов

Глубокое исследование — это функция, которая представляет собой вершину современных ИИ-систем. Это не просто ответ на вопрос, а создание полноценного, цитируемого, критически обоснованного экспертного отчета на основе анализа десятков, а иногда и сотен источников. Такой процесс включает в себя несколько сложных этапов: формулирование исследовательского запроса, выполнение многоуровневого поиска, критическая оценка достоверности и релевантности найденных источников, синтез информации из разнородных документов и, наконец, генерация отчета с

правильным цитированием и ссылками. Это требует не просто знаний, а метакогнитивных навыков, которые традиционно считаются прерогативой человеческих экспертов.

Для оценки этой функции были разработаны специализированные бенчмарки, такие как DRACO (Deep Research Accuracy, Completeness, and Objectivity), который оценивает модели по трем ключевым измерениям: точности, полноте и объективности создаваемых отчетов [85] [201]. Другой важный бенчмарк — A Benchmark for Multimodal Deep Research Agents, который оценивает способность агентов генерировать отчеты с цитированием на основе многократного поиска и синтеза [121]. Эти бенчмарки являются наиболее строгими тестами для любой модели, претендующей на роль «экспертного исследователя».

На основе анализа данных можно сделать вывод, что все рассматриваемые модели демонстрируют выдающиеся способности в этой области, что и неудивительно, поскольку глубокое исследование является одной из ключевых задач, для которых они и были созданы. Qwen 3.5-Plus получает оценку «Высокое (1)». Его позиционирование как «мощнейшей» модели для глубоких рассуждений (Qwen3-Max-Thinking) является прямым указанием на его предназначение для этой задачи [62]. Более того, его архитектура QwenLong-L1.5 специально разработана для улучшения долгосрочных рассуждений, что является критически важным для сохранения контекста при анализе множества источников [45]. Это не просто заявление, а техническое решение, направленное на решение конкретной проблемы глубокого исследования.

GPT 5.2, Claude Opus 4.5/4.6 и Gemini 3 Pro также получают оценку «Высокое (1)». Все они являются флагманскими моделями, чья основная рекламная кампания строится вокруг их способности к «глубокому мышлению» (Deep Think). Отчеты подтверждают, что Gemini 3 Pro и GPT-5.2 имеют 100-тысячное окно контекста, что позволяет им «помнить» огромные объемы информации, необходимые для комплексного анализа [245]. Claude Opus 4.6 заявлен как превосходящий GPT-5.2 во всех бенчмарках, включая финансовую и юридическую аналитику, что напрямую связано с глубоким исследованием в этих областях [26].

Kimi 2.5 также получает оценку «Высокое (1)». Ее архитектура как MoE-модели с триллионом параметров и ее способность к обработке огромных контекстных окон делают ее идеальным инструментом для задач, требующих анализа большого объема информации [108].

Таким образом, в категории «Глубокое исследование» все модели находятся в первой строке рейтинга. Это не означает, что они одинаковы, а лишь то, что они все достигли порога, необходимого для выполнения этой сложнейшей задачи. Различия между ними лежат в деталях: в том, насколько хорошо они справляются с определенными типами источников (например, научные статьи против юридических документов), в скорости выполнения многоэтапного поиска или в качестве цитирования. Но для целей данного сравнительного анализа, где оценка строится на основе доступных данных, все они заслуживают высшей оценки.

Общий интеллект и бенчмарки: многомерная картина производительности

Оценка «общего интеллекта» ИИ-модели — это самая сложная и спорная задача в данном исследовании. В отличие от узкоспециализированных функций, таких как генерация изображений или веб-поиск, общий интеллект — это абстрактная концепция, не имеющая единого, универсального измерения. Поэтому в этом разделе будет проведено не попытка найти «самую умную» модель, а анализ ее позиционирования в контексте различных типов бенчмарков, каждый из которых измеряет свой аспект интеллекта: логическое мышление, знания, программирование, агентное поведение и т.д.

Для начала стоит рассмотреть универсальные бенчмарки, такие как MMLU (Massive Multitask Language Understanding), которые оценивают знания модели в десятках академических дисциплин. В материалах нет прямых данных о результатах Qwen 3.5-Plus в MMLU, но есть упоминание о том, что Qwen3-Max-Thinking «создал несколько мировых рекордов» и «по своим возможностям сопоставим с GPT-5.2 и Gemini 3 Pro» [62] [76]. Это позволяет предположить, что его результат находится в той же группе, что и у лидеров. Однако другие источники указывают на то, что Claude Opus 4.6 лидирует в GDPval-AA, а Gemini 3 Pro превосходит конкурентов в высокодифференцированных тестах [60] [136]. Это говорит о том, что «лидерство» сильно зависит от конкретного бенчмарка.

Более информативным является анализ специализированных бенчмарков. В SWE-bench, который измеряет программистские способности, лидером является Claude Opus 4.5 (74.40), за ним следует Gemini 3 Pro (74.20) и GPT 5.2 (71.80) [51].

Qwen 3.5-Plus и Kimi 2.5 не представлены в этом списке, что позволяет им присвоить оценку «Среднее (2)» и «Ниже среднего (4)» соответственно.

В агентных бенчмарках, таких как Terminal-Bench, ситуация иная. Здесь GPT-5.2 (в связке с Codex CLI) показывает самый высокий средний процент решения задач (63%), что говорит о его превосходстве в задачах, требующих взаимодействия с командной строкой [23]. Это подтверждает его статус как лидера в практическом, агентном интеллекте.

Таким образом, для Qwen 3.5-Plus можно сделать следующий вывод: его позиция в категории «Общий интеллект» — это «Среднее (2)». Он не является абсолютным лидером в универсальных бенчмарках, таких как SWE-bench, но он является одним из самых сильных игроков в самых передовых и сложных областях, таких как мультимодальная генерация и глубокие рассуждения. Его сила — не в том, чтобы быть самым быстрым в одной задаче, а в том, чтобы быть самым гибким и мощным в самых сложных и многомерных задачах. Это позиционирование отражает его стратегию как «мастера сложного», а не «чемпиона по скорости».

Сводная таблица и стратегическое позиционирование **Qwen 3.5-Plus**

На основе проведенного всестороннего анализа можно составить окончательную сводную таблицу, которая суммирует позиционирование Qwen 3.5-Plus относительно его главных конкурентов. Важно подчеркнуть, что оценки в этой таблице не являются произвольными цифрами, а являются результатом строгого анализа данных из предоставленных источников, где каждая позиция обоснована конкретными фактами и техническими характеристиками.

Категория	**Qwen 3.5-Plus**	**GPT 5.2**	**Claude Opus 4.5/4.6**	**Gemini 3 Pro**	**Kimi 2.5**
Планировщик Поездок	Среднее (2)	Среднее (2)	Высокое (1)	Высокое (1)	Ниже среднего (5+)
Образование	Среднее (2)	Высокое (1)	Высокое (1)	Высокое (1)	Среднее (2)
Работа с Артефактами	Высокое (1)	Высокое (1)	Высокое (1)	Высокое (1)	Среднее (2)
Веб-Поиск	Среднее (2)	Среднее (2)	Ниже среднего (4)	Высокое (1)	Ниже среднего (4)
Веб-Разработка	Среднее (2)	Высокое (1)	Высокое (1)	Высокое (1)	Среднее (2)
Генерация Изображений	Высокое (1)	Ниже среднего (4)	Ниже среднего (4)	Ниже среднего (4)	Ниже среднего (4)
Генерация Видео	Высокое (1)	Ниже среднего (4)	Ниже среднего (4)	Ниже среднего (4)	Ниже среднего (4)
Глубокое Исследование	Высокое (1)	Высокое (1)	Высокое (1)	Высокое (1)	Высокое (1)
Общий Интеллект (Бенчмарки)	Среднее (2)	Высокое (1)	Высокое (1)	Высокое (1)	Ниже среднего (4)

Стратегическое позиционирование **Qwen 3.5-Plus**

Qwen 3.5-Plus нельзя охарактеризовать как «универсального победителя». Его стратегия — это не попытка доминировать во всех областях, а целенаправленное создание собственной, уникальной ниши на рынке ИИ-моделей. На основе анализа его сильных и слабых сторон можно сформулировать его стратегическое позиционирование следующим образом:

«Qwen 3.5-Plus — это мультимодальный мастер сложных рассуждений, ориентированный на создание визуального контента и управление агентными системами в условиях полного контроля пользователя над инфраструктурой**».**

Это позиционирование состоит из четырех ключевых компонентов:

1. Мультимодальность как фундамент**:** Архитектура Qwen3-Omni, способная обрабатывать текст, аудио, изображение и видео в рамках единой end-to-end модели, является его главным архитектурным преимуществом [33]. Это не просто дополнительная функция, а основа, на которой строятся все его сильные стороны: от генерации изображений с идеальным текстовым рендерингом [6] до создания видео с точным временным позиционированием [231].

2. **Мастер сложных рассуждений:** Qwen3-Max-Thinking и QwenLong-L1.5 — это не маркетинговые названия, а технические решения, направленные на решение конкретных проблем: глубокого исследования и долгосрочного контекстного понимания [45] [62]. Это делает его идеальным инструментом для задач, требующих не простого ответа, а многошагового, критического анализа.

3. **Ориентация на визуальный контент:** Его абсолютное лидерство в генерации изображений и видео — это не случайность, а результат стратегического выбора архитектуры и инвестиций в специализированные модели, такие как Qwen-Image-2.0 [29]. Это делает его незаменимым для всех сценариев, связанных с созданием цифрового контента.

4. **Контроль как ценность:** В отличие от моделей, которые позиционируются как облачные сервисы, Qwen 3.5-Plus «построен для команд, которые хотят запускать модели своим способом — часто самостоятельно размещая их, с реальным контролем над потоками данных и выбором развертывания» [93]. Это не техническая деталь, а фундаментальный принцип, который определяет его целевую аудиторию: это предприятия и разработчики, для которых безопасность, прозрачность и независимость от облачных провайдеров являются первостепенными ценностями.

Таким образом, Qwen 3.5-Plus не конкурирует с GPT 5.2 или Claude Opus на их полях боя. Он создает новое поле боя — поле мультимодального, контролируемого и визуально ориентированного ИИ. Его позиция «Среднее (2)» в общем рейтинге — это не признак слабости, а отражение его стратегического выбора быть не «самым быстрым», а «самым гибким и самым мощным» в самых сложных и перспективных областях будущего.

Заключение: **Qwen 3.5-Plus** как архитектурный ответ на вызовы агентной эпохи

Завершая данный комплексный анализ, можно утверждать, что Qwen 3.5-Plus — это не просто еще одна модель в длинном списке ИИ-систем 2026 года. Это архитектурный ответ на фундаментальные вызовы, с которыми сталкивается индустрия в эпоху перехода от генеративного ИИ к агентному ИИ. Его

позиционирование не является результатом случайных технических решений, а представляет собой продуманную стратегию, направленную на решение трех ключевых проблем, которые сегодня стоят перед разработчиками и предприятиями.

Первая проблема — это ограниченность одномерных моделей. В мире, где пользовательский запрос может содержать текст, изображение и видео, модели, способные обрабатывать только один тип данных, становятся все менее актуальными. Qwen 3.5-Plus решает эту проблему, предлагая унифицированную архитектуру Qwen3-Omni, которая рассматривает все модальности как равноправные части единой информационной структуры [33]. Это не просто «добавление функции», а изменение парадигмы: модель не переключается между режимами, она работает в них одновременно.

Вторая проблема — это зависимость от экосистем. Сегодня многие модели привязаны к определенным облачным платформам или поисковым системам, что создает риски для бизнеса и ограничивает возможности интеграции. Qwen 3.5-Plus решает эту проблему, делая ставку на гибкость и контроль. Его ориентация на self-hosted развертывание и открытую экосистему [93] дает организациям возможность интегрировать его в свои существующие инфраструктуры, будь то внутренние базы знаний, корпоративные RAG-системы или частные облака, без необходимости менять всю свою IT-архитектуру.

Третья проблема — это разрыв между возможностями и практической ценностью. Многие модели демонстрируют выдающиеся результаты в бенчмарках, но их сложно применить в реальных бизнес-процессах. Qwen 3.5-Plus решает эту проблему, фокусируясь на самых востребованных и сложных практических задачах: создании визуального контента с профессиональным качеством [6] и управлении сложными агентными системами [62]. Его лидерство в генерации изображений и видео — это не абстрактная метрика, а решение реальной боли маркетинговых отделов, дизайнерских студий и разработчиков контента.

В заключение, Qwen 3.5-Plus не стремится быть «лучшим во всем». Он стремится быть «лучшим для тех, кто нуждается в лучшем». Для организаций, которые ценят контроль, гибкость и мощь в самых передовых областях — мультимодальной обработке и визуальном творчестве — он представляет собой не просто инструмент, а стратегическое преимущество. Его позиционирование в рейтинге — это не итог гонки, а начало нового этапа, в котором успех

определяется не тем, насколько хорошо модель выполняет одну задачу, а тем, насколько гибко и мощно она может адаптироваться к сложнейшим, многомерным вызовам реального мира.

Справка

1. 从“生成”到“深度推理”：2026 大模型三巨头横评：Gemini 3 Pro https://zhuanlan.zhihu.com/p/1994362083245039918
2. (PDF) Smart and Dynamic AI-Powered Travel Planning: A Machine ... https://www.researchgate.net/publication/390031206_Smart_and_Dynamic_AI-Powered_Travel_Planning_A_Machine_Learning_Approach_for_Personalized_and_Real-Time_Itinerary_Generation
3. Ai Travel Chatbot | PDF | Artificial Intelligence - Scribd https://www.scribd.com/document/969961148/aif
4. Alibaba Cloud Model Studio:Qwen API reference https://www.alibabacloud.com/help/en/model-studio/qwen-api-reference/
5. [PDF] AI Industry Landscape Report 2025 https://repository.ceibs.edu/files/59116885/AI_Industry_landscape_report_2025.pdf
6. hype or helpful? Qwen-Image is a promising new image g... - TikTok https://www.tiktok.com/@whats_ai/video/7537314120782122246
7. Agent-SAMA: State-Aware Mobile Assistant - arXiv https://arxiv.org/html/2505.23596v3
8. Findings of the Association for Computational Linguistics: EMNLP ... https://aclanthology.org/volumes/2024.findings-emnlp/
9. [PDF] Information Economy Report 2015 - UNCTAD https://unctad.org/system/files/official-document/ier2015_en.pdf
10. China AI Statistics and Insights 2026 - DataGlobeHub https://dataglobehub.com/china-ai-statistics-and-insights/
11. 30 of the best large language models in 2026 - TechTarget https://www.techtarget.com/whatis/feature/12-of-the-best-large-language-models
12. Benchmarking LLMs' Capabilities to Generate Structural Outputs https://arxiv.org/html/2505.20139v1
13. A Tool for In-depth Analysis of Code Execution Reasoning of Large ... https://dl.acm.org/doi/abs/10.1145/3696630.3728605

14. [PDF] ArtifactsBench: Bridging the Visual-Interactive Gap in LLM Code ... https://arxiv.org/pdf/2507.04952?
15. KRAMABENCH: A Benchmark for AI Systems on Data-to-Insight ... https://openreview.net/forum?id=fZfUdeCC5X
16. SWE-AGI: Benchmarking Specification-Driven Software ... - arXiv https://arxiv.org/html/2602.09447v1
17. [PDF] PeerRank: Autonomous LLM Evaluation Through Web-Grounded ... https://www.arxiv.org/pdf/2602.02589
18. [PDF] SWE-AGI: Benchmarking Specification-Driven Software ... - arXiv https://www.arxiv.org/pdf/2602.09447
19. A Benchmark for Decoupling Retrieval and Reasoning Capabilities https://arxiv.org/html/2601.21937v1
20. AgencyBench: Benchmarking the Frontiers of Autonomous Agents in ... https://arxiv.org/html/2601.11044v2
21. [PDF] FormationEval, an open multiple-choice benchmark for petroleum ... https://arxiv.org/pdf/2601.02158
22. Bilingual Bias in Large Language Models: A Taiwan Sovereignty ... https://arxiv.org/html/2602.06371v1
23. Terminal-Bench: Benchmarking Agents on Hard, Realistic Tasks in ... https://arxiv.org/html/2601.11868v1
24. 1 Introduction - arXiv https://arxiv.org/html/2602.10999v1
25. [PDF] CL-bench: A Benchmark for Context Learning - arXiv https://arxiv.org/pdf/2602.03587
26. Anthropic's Claude Opus 4.6 Outperforms GPT-5.2 in All Benchmarks https://www.linkedin.com/posts/drhaseebhamid_anthropic-just-released-claude-opus-46-activity-7425825910857244672-i2c5
27. [PDF] TERMINAL-BENCH: BENCHMARKING AGENTS ON ... - OpenReview https://openreview.net/pdf/899eb3c7b51db80c98a9e6d85dfda7b470a395d6.pdf
28. [PDF] mmJEE-Eval: A Bilingual Multimodal Benchmark for Evaluating ... https://aclanthology.org/2025.findings-ijcnlp.140.pdf
29. Comprehensive Analysis of Qwen-Image-2.0: Alibaba'... - U深研 https://unifuncs.com/s/LD3cpA3P
30. A Comprehensive Survey on Composed Image Retrieval - ACM https://dl.acm.org/doi/10.1145/3767328
31. Arxiv今日论文| 2025-12-17 - 闲记算法 http://lonepatient.top/2025/12/17/arxiv_papers_2025-12-17

32. Alibaba Cloud Model Studio - Text generation https://www.alibabacloud.com/help/en/model-studio/text-generation
33. Qwen3-Omni Technical Report - arXiv https://arxiv.org/html/2509.17765v1
34. Qwen2.5 VL! Qwen2.5 VL! Qwen2.5 VL! | Qwen https://qwenlm.github.io/blog/qwen2.5-vl/
35. (PDF) A Review of Large Language Models (LLMs) Development https://www.researchgate.net/publication/391222042_A_Review_of_Large_Language_Models_LLMs_Development_A_Cross-Country_Comparison_of_the_US_China_Europe_UK_India_Japan_South_Korea_and_Canada
36. (PDF) Apriel-1.5-15b-Thinker - ResearchGate https://www.researchgate.net/publication/396094853_Apriel-15-15b-Thinker
37. [PDF] Smothering Heights | JP Morgan Asset Management https://am.jpmorgan.com/content/dam/jpm-am-aem/global/en/insights/eye-on-the-market/smothering-heights-amv.pdf
38. Prompt-to-Leaderboard: Prompt-Adaptive LLM Evaluations https://openreview.net/forum?id=7VPRrzFEN8
39. A Survey on Evaluation of Large Language Models https://dl.acm.org/doi/abs/10.1145/3641289
40. [PDF] DeepSeek-R1: Incentivizing Reasoning Capability in LLMs ... - arXiv https://arxiv.org/pdf/2501.12948
41. A Primer for Evaluating Large Language Models in Social-Science ... https://journals.sagepub.com/doi/10.1177/25152459251325174
42. [PDF] Artificial Intelligence Index Report 2025 https://hai-production.s3.amazonaws.com/files/hai_ai_index_report_2025.pdf
43. Step 3.5 Flash: Open Frontier-Level Intelligence with 11B Active ... https://arxiv.org/html/2602.10604v1
44. Agent Learning via Early Experience | OpenReview https://openreview.net/forum?id=pEGnJbmSUy
45. LLM Papers Reading Notes - January 2026 - LinkedIn https://www.linkedin.com/pulse/llm-papers-reading-notes-january-2026-jean-david-ruvini-nj5cc
46. [PDF] Large multimodal models evaluation: a survey http://scis.scichina.com/en/2025/221301.pdf
47. Natural Language Understanding and Inference with MLLM in ... https://dl.acm.org/doi/full/10.1145/3711680
48. The Complete Full-Stack Developer Roadmap for 2026 https://dev.to/thebitforge/the-complete-full-stack-developer-roadmap-for-2026-2i0j

49. Top 10 Web Frameworks for 2026: Future-Ready Development https://www.linkedin.com/posts/masscom-corporation_top-10-web-frameworks-for-2026-activity-7417885748856602624-YXTm
50. Top Web Development Tools for Business Success in 2026 - Mindpath https://www.mindpathtech.com/blog/web-development-tools/
51. SWE-bench Leaderboards https://www.swebench.com/
52. [PDF] EduBench: A Comprehensive Benchmarking Dataset for Evaluating ... https://arxiv.org/pdf/2505.16160?
53. Qwen3: Think Deeper, Act Faster | Qwen https://qwenlm.github.io/blog/qwen3/
54. [PDF] EduBench: A Comprehensive Benchmarking Dataset for Evaluating ... https://arxiv.org/pdf/2505.16160
55. [PDF] OECD Digital Education Outlook 2026 (EN) https://www.oecd.org/content/dam/oecd/en/publications/reports/2026/01/oecd-digital-education-outlook-2026_940e0dd8/062a7394-en.pdf
56. (PDF) Estimating Exam Item Difficulty with LLMs: A Benchmark on ... https://www.researchgate.net/publication/400584261_Estimating_Exam_Item_Difficulty_with_LLMs_A_Benchmark_on_Brazil's_ENEM_Corpus
57. [PDF] LLMOrbit - From Scaling Walls to Agentic AI Systems - arXiv https://arxiv.org/pdf/2601.14053
58. Ultraviolet Technology To Prepare For The Habitable Worlds ... - arXiv https://arxiv.org/html/2408.07242v1
59. 这个春节，千问、阶跃、Gemini打响2026年「3.5模型大战」 - 知乎 https://zhuanlan.zhihu.com/p/2003117027875894533
60. 谷歌最强Gemini推理模型发布！测评碾压Opus 4.6、GPT-5.2 - 知乎专栏 https://zhuanlan.zhihu.com/p/2005625413955757419
61. 从“生成”到“深度推理”：2026 大模型三巨头横评：Gemini 3 Pro、GPT ... https://cloud.tencent.com/developer/article/2616234?policyId=1004
62. 憋了4个月，阿里最大最强模型正式版发布，附一手实测-36氪 https://m.36kr.com/p/3657074925609352
63. AI大模型列表- 最新AI模型汇总 - DataLearnerAI https://www.datalearner.com/en/ai-models/pretrained-models?aiArea=1008
64. 更新 - DMXAPI官网：中国多模态大模型API聚合平台 https://www.dmxapi.cn/weblog
65. 从“生成”到“深度推理”：2026 大模型三巨头横评：Gemini 3 Pro - 搜狐 https://www.sohu.com/a/975466360_121787785
66. GPT-5.2 与Gemini 3 Pro：企业级场景下的大模型工程表现对比 https://developer.aliyun.com/article/1708993

67. ThursdAI - The top AI news from the past week - Apple Podcasts https://podcasts.apple.com/de/podcast/thursdai-the-top-ai-news-from-the-past-week/id1698613329

68. X-Intelligence 3.0 Training and Evaluating Reasoning LLM for ... https://arxiv.org/html/2507.14430v1

69. [PDF] PromptCoT 2.0: Scaling Prompt Synthesis for Large ... - arXiv https://arxiv.org/pdf/2509.19894

70. Google AI Overviews 2026: Guide to G.ai & Search Challenges https://www.linkedin.com/pulse/google-quietly-admits-ai-overview-problems-while-gai-complete-nantha-xdxnc

71. Understanding AI Visibility: Fundamentals, Measurement Limits, and ... https://visively.com/kb/ai/ai-overview-visibility

72. Gen AI for Business edition #95: Super Bowl edition - LinkedIn https://www.linkedin.com/pulse/gen-ai-business-edition-95-super-bowl-eugina-jordan-vgj6e

73. Large language model for knowledge synthesis and AI-enhanced ... https://www.cell.com/trends/biotechnology/fulltext/S0167-7799(25)00045-9

74. BigCodeArena: Unveiling More Reliable Human Preferences ... - arXiv https://arxiv.org/html/2510.08697v2

75. Do Chatbot LLMs Talk Too Much? The YapBench Benchmark - arXiv https://arxiv.org/html/2601.00624v1

76. 阿里千问最强模型重磅亮相！性能媲美GPT-5.2、Gemini 3 Pro https://zhuanlan.zhihu.com/p/1999451053826003166

77. (PDF) Apple Intelligence Foundation Language Models https://www.researchgate.net/publication/382739540_Apple_Intelligence_Foundation_Language_Models

78. ODSC's Ai X Podcast | Open Data Science Conference https://odsc.com/podcast/

79. Qwen2.5: A Party of Foundation Models! | Qwen https://qwenlm.github.io/blog/qwen2.5/

80. Can Understanding and Generation Truly Benefit Together - arXiv https://arxiv.org/html/2509.09666v1

81. nvidia / nemotron-3-nano-30b-a3b https://docs.api.nvidia.com/nim/reference/nvidia-nemotron-3-nano-30b-a3b

82. Visual enumeration remains challenging for multimodal generative AI https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0331566

83. [PDF] X2I: Seamless Integration of Multimodal Understanding into ... https://openaccess.thecvf.com/content/ICCV2025/papers/Ma_X2I_Seamless_Integration_of_Multimodal_Understanding_into_Diffusion_Transformer_via_ICCV_2025_paper.pdf

84. (PDF) A Review of Large Language Models (LLMs) Development https://www.researchgate.net/publication/391195730_A_Review_of_Large_Language_Models_LLMs_Development_A_Cross-Country_Comparison_of_the_US_China_Europe_UK_India_Japan_South_Korea_and_Canada
85. DRACO: a Cross-Domain Benchmark for Deep Research Accuracy ... https://arxiv.org/html/2602.11685v1
86. Benchmark2: Systematic Evaluation of LLM Benchmarks - arXiv https://arxiv.org/html/2601.03986v1
87. SimuScene: Training and Benchmarking Code Generation to ... - arXiv https://arxiv.org/html/2602.10840v1
88. [PDF] Technology Trends Outlook 2025 - McKinsey https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20digital/our%20insights/the%20top%20trends%20in%20tech%202025/mckinsey-technology-trends-outlook-2025.pdf
89. The neurobench framework for benchmarking neuromorphic ... https://www.nature.com/articles/s41467-025-56739-4
90. ATLAHS: An Application-centric Network Simulator Toolchain for AI ... https://dl.acm.org/doi/10.1145/3712285.3759838
91. Top 5 Best AI Coding Agents - LinkedIn https://www.linkedin.com/pulse/top-5-best-ai-codingagents-kommunicate-d8l2c
92. Qwen3 vs GPT-5.2 vs Gemini 3 Pro: Which Should You Use and ... https://www.freecodecamp.org/news/qwen-vs-gpt-vs-gemini-which-should-you-use/
93. Qwen3 vs GPT-5.2 vs Gemini 3 Pro: Which Should You Use—and ... https://www.linkedin.com/pulse/qwen3-vs-gpt-52-gemini-3-pro-which-should-you-useand-guzrc
94. ChatGPT 5.2 vs Gemini 3 Pro: Which AI Model Should You ... - Helply https://helply.com/blog/chatgpt-5-2-vs-gemini-3-pro
95. The AI Daily Brief: Artificial Intelligence News - AIBase https://www.aibase.com/www.aibase.com/daily
96. AI模型列表- 支持GPT-4、Claude、Gemini等数百个AI大模型 - 海鲸AI https://api.atalk-ai.com/api-docs/models/
97. 代码质量新榜单：GPT-5.2断层领先，Opus 4.5、Gemini 3等表现揭晓 https://www.51cto.com/article/832289.html
98. 从“生成”到“深度推理”：2026 大模型三巨头横评：Gemini 3 Pro、GPT ... https://cloud.tencent.com/developer/article/2616234
99. Arxiv今日论文| 2026-01-27 - 闲记算法 http://lonepatient.top/2026/01/27/arxiv_papers_2026-01-27.html

100. AI大模型评测榜单- 实时排名 - DataLearnerAI https://www.datalearner.com/leaderboards
101. GPT-5.2 - 谷歌最强Gemini推理模型发布！测评碾压Opus 4.6 - 智东西 https://zhidx.com/p/535034.html
102. Qwen 3.5系列模型即将发布| AI 早报2026-02-09 - 知乎专栏 https://zhuanlan.zhihu.com/p/2004122302682511349
103. MDPI - Publisher of Open Access Journals https://www.mdpi.com/
104. A Survey of Vibe Coding with Large Language Models - arXiv.org https://arxiv.org/html/2510.12399v1
105. [PDF] Can Deep Research Agents Retrieve and Organize? Evaluating the ... https://www.arxiv.org/pdf/2601.12369v2
106. Terminal-Bench: Benchmarking Agents on Hard, Realistic Tasks in ... https://www.researchgate.net/publication/399931917_Terminal-Bench_Benchmarking_Agents_on_Hard_Realistic_Tasks_in_Command_Line_Interfaces
107. Computer Science - arXiv https://www.arxiv.org/list/cs/new?skip=525&show=500
108. (PDF) Kimi K2: Open Agentic Intelligence - ResearchGate https://www.researchgate.net/publication/394081339_Kimi_K2_Open_Agentic_Intelligence
109. AI Coding Benchmarks 2025: Gemini 3 Pro vs GPT-5.2 vs Claude 4.5 https://vertu.com/lifestyle/gpt-5-2-codex-vs-gemini-3-pro-vs-claude-opus-4-5-coding-comparison-guide/?srsltid=AfmBOoqmoMbO_aFnOnqBLvKWLk4SEHXOE_IVACuv5KycumcQm5e0YuSy
110. A Survey of Vision-Language Interactive Reasoning in Multimodal ... https://arxiv.org/html/2509.25373v1
111. Qwen2.5-Max: Exploring the Intelligence of Large-scale MoE Model https://qwenlm.github.io/blog/qwen2.5-max/
112. Benchmarking LLMs: A guide to AI model evaluation - TechTarget https://www.techtarget.com/searchsoftwarequality/tip/Benchmarking-LLMs-A-guide-to-AI-model-evaluation
113. 100+ LLM Benchmarks and Evaluation Datasets - Scribd https://www.scribd.com/document/824200574/100-LLM-benchmarks-and-evaluation-datasets
114. Challenges, Limitations, and Recommendations - OpenReview https://openreview.net/pdf?id=7NtfIYIqvk
115. Harnessing Large Language Models for Software Vulnerability ... https://www.researchgate.net/publication/388911340_Harnessing_Large_Language_Models_for_Software_Vulnerability_Detection_A_Comprehensive_Benchmarking_Study

116. Meet Tsinghua University's GLM-4-9B-Chat-1M - MarkTechPost https://www.marktechpost.com/2024/06/05/meet-tsinghua-universitys-glm-4-9b-chat-1m-an-outstanding-language-model-challenging-gpt-4v-gemini-pro-on-vision-mistral-and-llama-3-8b/
117. [PDF] COLING 2025 The 31st International Conference on Computational ... https://aclanthology.org/2025.coling-demos.pdf
118. Flash>Pro？SWE-rebench发布12月榜单：Claude Opus 4.5位居榜首 https://linux.do/t/topic/1476152
119. GPT-5.2发布，SWE-bench跑到80%，Claude和Gemini傻眼了 https://developer.volcengine.com/articles/7583974006227730483
120. CoreCodeBench: Decoupling Code Intelligence via Fine-Grained ... https://arxiv.org/html/2507.05281v2
121. A Benchmark for Multimodal Deep Research Agents - arXiv https://arxiv.org/html/2601.12346v1
122. Arxiv今日论文| 2026-02-11 - 闲记算法 http://lonepatient.top/2026/02/11/arxiv_papers_2026-02-11.html
123. A Survey on Large Language Models for Code Generation https://dl.acm.org/doi/full/10.1145/3747588
124. [PDF] Proceedings of the Workshop on Beyond English: Natural Language ... https://aclanthology.org/2025.globalnlp-1.pdf
125. Implementing generative artificial intelligence in precision oncology https://pmc.ncbi.nlm.nih.gov/articles/PMC12896320/
126. 自然语言处理2026_1_6 - arXiv每日学术速递 https://arxivdaily.com/thread/75382
127. The 2025 Conference on Empirical Methods in Natural Language ... https://aclanthology.org/events/emnlp-2025/
128. Appl. Sci., Volume 15, Issue 23 (December-1 2025) – 505 articles https://www.mdpi.com/2076-3417/15/23
129. 人工智能2025_6_2 - arXiv每日学术速递 https://www.arxivdaily.com/thread/68054
130. Computer Science - arXiv https://www.arxiv.org/list/cs/new?skip=25&show=1000
131. [PDF] Climate Imagineering - ResearchGate https://www.researchgate.net/profile/Sean-Low-3/publication/350966368_Climate_Imagineering_Practices_and_politics_of_sunlight_reflection_and_carbon_removal_assessment/links/607d52ef8ea909241e0cf32f/Climate-Imagineering-Practices-and-politics-of-sunlight-reflection-and-carbon-removal-assessment.pdf
132. 🏆 AI Models Benchmark Dataset 2026 (latest) - Kaggle https://www.kaggle.com/datasets/asadullahcreative/ai-models-benchmark-dataset-2026-latest

133. 大模型榜单周报（2026/02/08） - KAI智习- 博客园 https://www.cnblogs.com/xjk15082/p/19606070

134. [PDF] CoreCodeBench: Decoupling Code Intelligence via Fine-Grained ... https://www.arxiv.org/pdf/2507.05281v2

135. Claude Opus 4.6: What Changed, Why It Matters, and How It Stacks ... https://www.linkedin.com/pulse/claude-opus-46-what-changed-why-matters-how-stacks-up-peter-w-szabo-7ebkf

136. Claude Opus 4.6 发布，全线碾压GPT-5.2，一文详解 - 网易 https://www.163.com/dy/article/KL24VIEP0556C3J2.html

137. OECD Digital Education Outlook 2026 https://www.oecd.org/en/publications/oecd-digital-education-outlook-2026_062a7394-en.html

138. OECD Report on AI in Education Offers Nuanced Perspective https://www.linkedin.com/posts/mutlu-cukurova_oecd-digital-education-outlook-2026-activity-7420002242105344000-MuC7

139. OECD Digital Education Outlook https://www.oecd.org/en/publications/oecd-digital-education-outlook_7fbfff45-en.html

140. Digital Education Outlook 2026 - EPALE platform - European Union https://epale.ec.europa.eu/en/resource-centre/content/digital-education-outlook-2026

141. Nixos Unstable | PDF - Scribd https://www.scribd.com/document/821052933/Nixos-Unstable

142. 对比明明白白3大顶级模型-GPT-5.2/Gemini 3/Claude Opus 4.5！老金 ... https://juejin.cn/post/7584297353419833344

143. 4.5 历史更新 - 飞书文档 https://docs.feishu.cn/article/wiki/FjiOwWp2giA7hRk6jjfcPioCnAc

144. SWE-bench Results Viewer https://www.swebench.com/viewer.html

145. SWE-rebench Leaderboard https://swe-rebench.com/

146. Computer Science - arXiv.org https://arxiv.org/list/cs/new

147. Reasoning beyond limits: Advances and open problems for LLMs https://www.sciencedirect.com/science/article/pii/S240595952500133X

148. A Systematic Literature Review of Retrieval-Augmented Generation https://www.mdpi.com/2504-2289/9/12/320

149. A Systematic Literature Review of Retrieval-Augmented Generation https://www.scribd.com/document/919697623/2508-06401v1

150. 2508.06401v3 | PDF | Systematic Review | Information Retrieval https://www.scribd.com/document/990131843/2508-06401v3

151. RAG Daily Papers - Latest RAG Research from arXiv https://ragdaily.com/

152. 人工智能2025_5_27[2] - arXiv每日学术速递 http://arxivdaily.com/thread/67852

153. LLMPopcorn: Exploring LLMs as Assistants for Popular Micro-video ... https://arxiv.org/html/2502.12945v3

154. Generative AI vs AI Agents vs Agentic AI: Key Differences ... - LinkedIn https://www.linkedin.com/posts/subash-iyyappan_agenticai-aiagents-generativeai-activity-7406689813321592834-E64L

155. 大模型对比工具- AI模型性能与价格对比 - DataLearnerAI https://www.datalearner.com/benchmark-compare?models=Gemini-2_5-Pro

156. Benchmarking LLM Agents on Enterprise API Tasks via Code ... - arXiv https://arxiv.org/html/2602.11224v1

157. PersistBench: When Should Long-Term Memories Be Forgotten by ... https://www.researchgate.net/publication/400370586_PersistBench_When_Should_Long-Term_Memories_Be_Forgotten_by_LLMs

158. 大语言模型-逻辑能力横评25-11月榜(Gemini 3/GPT-5.1/Opus 4.5) - 知乎 https://zhuanlan.zhihu.com/p/1977143847453755265

159. [PDF] Proceedings of the The First Workshop on LLM Security (LLMSEC) https://aclanthology.org/2025.llmsec-1.pdf

160. Graph-R1: An Agentic GraphRAG Framework for Structured, Multi ... https://www.marktechpost.com/2025/08/09/graph-r1-an-agentic-graphrag-framework-for-structured-multi-turn-reasoning-with-reinforcement-learning/

161. Arxiv今日论文| 2026-02-12 - 闲记算法 http://lonepatient.top/2026/02/12/arxiv_papers_2026-02-12.html

162. Exploring Effective Uses of Generative AI in Education https://policycommons.net/artifacts/42998153/062a7394-en/

163. (PDF) CL-bench: A Benchmark for Context Learning - ResearchGate https://www.researchgate.net/publication/400415422_CL-bench_A_Benchmark_for_Context_Learning

164. When Should Long-Term Memories Be Forgotten by LLMs? - arXiv https://arxiv.org/html/2602.01146v1

165. 人工智能2025_10_30 - arXiv每日学术速递 https://www.arxivdaily.com/thread/73297

166. [PDF] Fara-7B: An Efficient Agentic Model for Computer Use - Microsoft https://www.microsoft.com/en-us/research/wp-content/uploads/2025/11/Fara-7B-An-Efficient-Agentic-Model-for-Computer-Use.pdf

167. Arxiv今日论文| 2026-01-08 - 闲记算法 http://lonepatient.top/2026/01/08/arxiv_papers_2026-01-08

168. WebDev Arena Leaderboard: Top AI Models for Coding & Web ... https://www.linkedin.com/posts/youness-labchiri_ai-webdev-machinelearning-activity-7416967315012878336-4mX1

169. AI Coding Benchmarks 2025: Gemini 3 Pro vs GPT-5.2 vs Claude 4.5 https://vertu.com/lifestyle/gpt-5-2-codex-vs-gemini-3-pro-vs-claude-opus-4-5-coding-comparison-guide/?srsltid=AfmBOoqXGjkMfaXAHvmbs8UPpVY9w64qeFbaC3glA5DnwfMToaDmiHjn

170. SurGE: A Benchmark and Evaluation Framework for Scientific ... https://arxiv.org/html/2508.15658v3

171. Multi-Property Document Annotation for LLM Data Curation at Scale https://arxiv.org/html/2602.12414v1

172. LLMOrbit: A Circular Taxonomy of Large Language Models - arXiv https://arxiv.org/html/2601.14053v1

173. [PDF] Personalized pharmacy Video Clips via Vision Language Models ... https://arxiv.org/pdf/2601.05059

174. [PDF] Unified Multimodal Understanding and Generation Models - arXiv.org https://arxiv.org/pdf/2505.02567

175. [PDF] Multi-Property Document Annotation for LLM Data Curation at Scale https://www.arxiv.org/pdf/2602.12414

176. Vistoria: A Multimodal System to Support Fictional Story Writing ... https://arxiv.org/html/2509.13646v3

177. Unified Multimodal Understanding and Generation Models - arXiv https://arxiv.org/html/2505.02567v6

178. Top 10 AI Models: 2026 Rankings and Benchmarks - LinkedIn https://www.linkedin.com/posts/rayuzwyshyn_the-global-ai-vanguard-top-10-models-activity-7411481321392381953-Pk5W

179. 大模型对比工具- AI模型性能与价格对比 - DataLearnerAI https://www.datalearner.com/benchmark-compare?models=Qwen2_5-32B

180. A Review of Large Language Models Across Academic Disciplines https://arxiv.org/html/2509.19580v5

181. Proceedings of the 2024 Conference on Empirical Methods in ... https://aclanthology.org/volumes/2024.emnlp-main/

182. Empirical Evaluation of Reasoning LLMs in Machinery Functional ... https://www.mdpi.com/2079-9292/14/18/3624

183. GPT-5.2 vs Opus 4.5 vs Gemini 3 Pro: Model Comparison - LinkedIn https://www.linkedin.com/posts/tsenkov_a-few-thoughts-i-am-getting-after-my-latest-activity-7414028370088452097-VSCJ

184. Policies for the digital transformation of school education - OECD https://www.oecd.org/en/publications/policies-for-the-digital-transformation-of-school-education_464dab4d-en.html

185. AI Coding Benchmarks 2025: Gemini 3 Pro vs GPT-5.2 vs Claude 4.5 https://vertu.com/lifestyle/gpt-5-2-codex-vs-gemini-3-pro-vs-claude-opus-4-5-coding-comparison-guide/?srsltid=AfmBOoq-OPrxRsxXnR5AZi5AZSF1P9rzsSgMNWU0ns6e1IpINWAU_G9Y

186. Elastic Compute Service (ECS): Server Cloud Elastis & Aman https://www.alibabacloud.com/id/product/ecs?_p_lc=1

187. Supported AI models in GitHub Copilot https://docs.github.com/copilot/reference/ai-models/supported-models

188. Run 2026-02-08-01KGYWYR1WNXQP4B4B80FYCC5R - NC Bench https://www.nc-bench.com/runs/2026-02-08-01KGYWYR1WNXQP4B4B80FYCC5R

189. [PDF] Generative AI for Designing Efficient and Explainable TinyML Models https://theses.hal.science/tel-05509399v1/file/TheseDEF_Christophe_EL_ZEINATY.pdf

190. Survey on Factuality in Large Language Models - ACM https://dl.acm.org/doi/10.1145/3742420

191. [PDF] MEDIC: Comprehensive Evaluation of Leading Indicators for LLM ... https://arxiv.org/pdf/2409.07314

192. Wei Xu - ACL Anthology https://aclanthology.org/people/wei-xu/

193. Quick Start the AI Model on the Alibaba Cloud Model Studio https://www.alibabacloud.com/blog/601401

194. PhyWorldBench: A Comprehensive Evaluation of Physical Realism ... https://arxiv.org/html/2507.13428v2

195. SONIC-O1: A Real-World Benchmark for Evaluating Multimodal ... https://arxiv.org/html/2601.21666

196. GenArena: How Can We Achieve Human-Aligned Evaluation ... - arXiv https://arxiv.org/html/2602.06013v1

197. GISA: A Benchmark for General Information Seeking Assistant https://arxiv.org/html/2602.08543v1

198. A Multi-Generator Benchmark for Detecting Synthetic Video Deepfakes https://arxiv.org/html/2602.04939v1

199. A Benchmark and Evaluation Framework for Scientific Survey ... https://arxiv.org/html/2508.15658v4

200. Past- and Future-Informed KV Cache Policy with Salience ... - arXiv https://arxiv.org/html/2601.21896v3

201. [PDF] DRACO: a Cross-Domain Benchmark for Deep Research Accuracy ... https://arxiv.org/pdf/2602.11685

202. Vidi2.5: Large Multimodal Models for Video Understanding ... - arXiv https://arxiv.org/html/2511.19529v2

203. [PDF] DrivingGen: A Comprehensive Benchmark for Generative ... - arXiv https://arxiv.org/pdf/2601.01528

204. Arxiv今日论文| 2026-02-06 - 闲记算法 http://lonepatient.top/2026/02/06/arxiv_papers_2026-02-06.html

205. Google's Updated Gemini 3 Deep Think Outperforms GPT-5.2 and ... https://www.gadgets360.com/ai/news/google-gemini-3-deep-think-upgraded-outperforms-openai-gpt-5-2-claude-opus-4-6-details-10998791

206. Igor Akimov's Post - LinkedIn https://www.linkedin.com/posts/igorakimov1_hm-lmsys-has-launched-arena-expert-a-new-activity-7392219314663182336-rLf0

207. Unleashing the potential of prompt engineering for large language ... https://www.cell.com/patterns/fulltext/S2666-3899(25)00108-4

208. On the Fundamental Limits of LLMs at Scale - arXiv https://arxiv.org/html/2511.12869v2

209. Application of large language models to intelligently analyze long ... https://www.researchgate.net/publication/385191584_Application_of_large_language_models_to_intelligently_analyze_long_construction_contract_texts

210. [PDF] AI Data Center Network with Juniper Apstra, AMD GPUs, Broadcom ... https://www.juniper.net/documentation/us/en/software/jvd/jvd-ai-dc-apstra-amd/jvd-ai-dc-apstra-amd.pdf

211. 人工智能2025_9_9 - arXiv每日学术速递 http://www.arxivdaily.com/thread/71411

212. [PDF] Open Frontier-Level Intelligence with 11B Active Parameters - arXiv https://www.arxiv.org/pdf/2602.10604

213. [PDF] MARKING ON THE TRUSTWORTHINESS OF GENERATIVE https://openreview.net/pdf/6b6b4ef154e9403f175dd4773c92fba2f7006578.pdf

214. AI Agents vs. Agentic AI: A Conceptual taxonomy, applications and ... https://www.sciencedirect.com/science/article/pii/S1566253525006712

215. Claude Opus 4.5夺回编程王座，超Gemini 3 Pro和GPT-5.1 - 网易 https://www.163.com/dy/article/KF7AMVPB05566ZHB.html

216. AI for Clinical Applicat - Springer Link https://link.springer.com/content/pdf/10.1007/978-3-032-06004-4.pdf

217. Launch of the 2026 Digital Education Outlook - OECD https://www.oecd.org/en/blogs/2026/01/launch-of-the-2026-digital-education-outlook.html

218. [PDF] International Scientific Report on the Safety of Advanced AI - HAL https://hal.science/hal-04612963v1/file/international_scientific_report_on_the_safety_of_advanced_ai_interim_report.pdf
219. PhyEduVideo: A Benchmark for Evaluating Text-to-Video Models for ... https://arxiv.org/html/2601.00943v1
220. WorldArena: A Unified Benchmark for Evaluating Perception ... - arXiv https://arxiv.org/html/2602.08971v1
221. Tele-Omni: a Unified Multimodal Framework for Video Generation ... https://arxiv.org/html/2602.09609v1
222. View-Consistent and Identity-Preserving Image-to-Video Generation https://arxiv.org/html/2602.10113v1
223. Stress Tests REVEAL Fragile Temporal and Visual Grounding in ... https://arxiv.org/html/2602.11244v1
224. Rethinking Video Generation Model for the Embodied World - arXiv https://arxiv.org/html/2601.15282v1
225. [PDF] Towards Physically Aware Video Generation via Latent ... - arXiv.org https://arxiv.org/pdf/2601.03665
226. Can World Simulators Reason? Gen-ViRe: A Generative Visual ... https://arxiv.org/html/2511.13853v3
227. Video Generation Models in Robotics: Applications, Research ... https://arxiv.org/html/2601.07823v1
228. STVG-R1: Incentivizing Instance-Level Reasoning and Grounding in ... https://arxiv.org/html/2602.11730v1
229. [PDF] “人工智能+”引爆新质生产力革命 https://pdf.dfcfw.com/pdf/H3_AP202406131636078603_1.pdf
230. GISA: A Benchmark for General Information Seeking Assistant https://arxiv.org/html/2602.08543v2
231. ShotFinder: Imagination-Driven Open-Domain Video Shot Retrieval ... https://arxiv.org/html/2601.23232v2
232. [PDF] GUIGuard: Toward a General Framework for Privacy-Preserving GUI ... https://arxiv.org/pdf/2601.18842
233. ReasonTabQA: A Comprehensive Benchmark for Table Question ... https://arxiv.org/html/2601.07280v1
234. Toward a General Framework for Privacy-Preserving GUI Agents https://arxiv.org/html/2601.18842v2
235. 大模型迈入Agent元年丨大语言模型1月最新榜单揭晓 - 知乎专栏 https://zhuanlan.zhihu.com/p/2005347153342645342

236. 2026年AI应用大模型选型终极指南 - 知乎专栏 https://zhuanlan.zhihu.com/p/2002544945312055320

237. 人工智能- 大模型榜单周报（2026/1/17） - KAI智 - SegmentFault 思否 https://segmentfault.com/a/1190000047548778

238. Retro-R1: LLM-based Agentic Retrosynthesis | OpenReview https://openreview.net/forum?id=30iBKSQMXn&referrer=%5Bthe%20profile%20of%20LEI%20BAI%5D(%2Fprofile%3Fid%3D~LEI_BAI1)

239. Arxiv今日论文| 2026-01-16 - 闲记算法 http://lonepatient.top/2026/01/16/arxiv_papers_2026-01-16

240. Computer Applications - Springer https://link.springer.com/content/pdf/10.1007/978-981-97-9674-8.pdf

241. [PDF] FRONTIERS IN PSYCHO CUTANEOUS DISEASES - ResearchGate https://www.researchgate.net/profile/Dennis-Linder/publication/303913975_Skin_Picking_-_The_ESDaP_Project/links/5764308e08aedbc345ecbf69/Skin-Picking-The-ESDaP-Project.pdf

242. [PDF] Herbs and Spices - AWS https://intech-files.s3.amazonaws.com/a043Y000010Jz8cQAC/0015419_Authors_Book%20%282024-11-26%2009%3A44%3A33%29.pdf

243. 6 - CodaLab Worksheets https://worksheets.codalab.org/rest/bundles/0xadf98bb30a99476ab56ebff3e462d4fa/contents/blob/glove.6B.100d.txt-vocab.txt

244. MultiSentimentArcs: a novel method to measure coherence in ... https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2024.1444549/full

245. 从“生成”到“深度推理”：2026 大模型三巨头横评：Gemini 3 Pro、GPT ... https://cloud.tencent.com/developer/article/2616234?policyId=1003

246. [PDF] Understanding the Process Verification of Multi-Agent Systems https://arxiv.org/pdf/2602.03053

247. Executing Wide Search via Agentic MapReduce - arXiv https://arxiv.org/html/2602.01331v1

248. [PDF] Agentic Reasoning for Large Language Models - ResearchGate https://www.researchgate.net/publication/399930413_Agentic_Reasoning_for_Large_Language_Models/fulltext/6970534fee048155cffe31b4/Agentic-Reasoning-for-Large-Language-Models.pdf?origin=scientificContributions

249. VideoMathQA: Benchmarking Mathematical Reasoning via ... https://openreview.net/forum?id=VI4kGUfPio

250. [PDF] THINKGEO : EVALUATING TOOL-AUGMENTED AGENTS FOR ... https://openreview.net/pdf/e507fc0bcda6033e1e04aa7a03b0f23ee2b1bd64.pdf
251. Graph-based Agent Memory: Taxonomy, Techniques, and ... - arXiv https://arxiv.org/html/2602.05665v1
252. [PDF] CryptoAnalystBench: Failures in Multi-Tool Long-Form LLM Analysis https://arxiv.org/pdf/2602.11304
253. (PDF) Agentic Reasoning for Large Language Models - ResearchGate https://www.researchgate.net/publication/399930413_Agentic_Reasoning_for_Large_Language_Models
254. claude-flow - NPM https://www.npmjs.com/package/claude-flow
255. Issue #203 - Simulations, Patterns, VR and more | Game Dev Digest https://gamedevdigest.com/digests/issue-203-simulations-patterns-vr-and-more.html