

ModelComparator · Финальная сводка

Проект: Демо проект (резерв копия)
Сценарий: Творческий сценарий
Описание сценария: Придумай песню про разработчика
Моделей: 3 · Критериев: 20 · Оценок: 60

Краткий вывод

Итоговый отчёт
Проект: Демо проект (резерв копия)
Сценарий: Творческий сценарий

Рекомендуемая модель: Алиса AI
Интегральная оценка (K): 0.726
Интерпретация: Приемлемая модель, требуется доработка.

Почему именно эта модель:

- Ключевые критерии вклада: Точность ответа, Логичность и структура, Глубина и полнота ответа
- Отрыв от второго места: 0.074
- Выбор выглядит стабильным.
- Ближайший конкурент: ChatGPT
- Наибольшее преимущество по критериям: Точность ответа, Адаптивность к стилю общения, Устойчивость к когнитивной нагрузке
- Где лидер уступает: Логичность и структура, Глубина и полнота ответа, Понимание технической специфики
- Сильные стороны лидера: Точность ответа (оценка 5, вклад 1.50), Логичность и структура (оценка 3, вклад 0.60), Глубина и полнота ответа (оценка 3, вклад 0.60)
- Слабые стороны лидера: Чувствительность к контексту (оценка 1, вклад 0.05), Эффективность исправления ошибок (оценка 1, вклад 0.05), Контекстная согласованность (оценка 1, вклад 0.05)

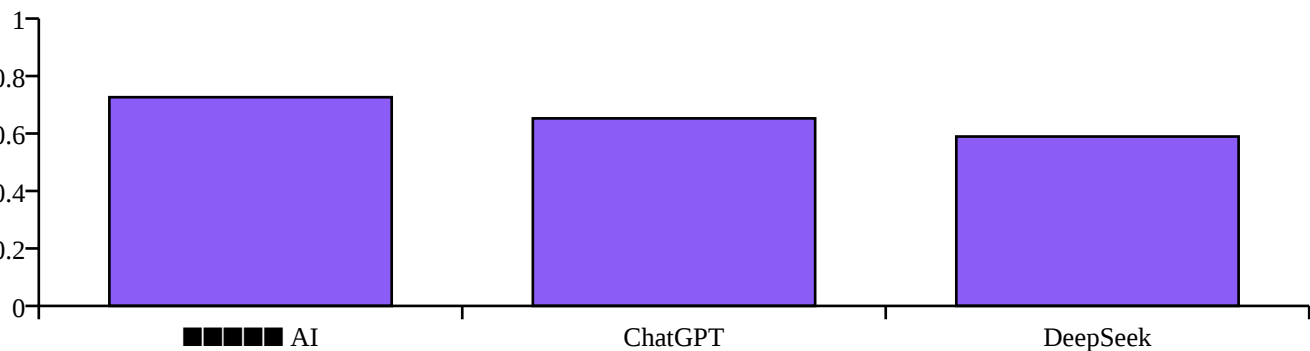
Условия, при которых выбор может измениться:

- Если повысить веса критериев, где сильнее ChatGPT (Логичность и структура, Глубина и полнота ответа, Понимание технической специфики), лидер может измениться.

Примечание: расчёт основан на включённых критериях и текущих весах.
Рейтинг моделей

Место	Модель	K	Взвешенная сумма	Пропуски
1	Алиса AI	0.726	6.900	0
2	ChatGPT	0.653	6.200	0
3	DeepSeek	0.589	5.600	0

График K по моделям



Критерии и веса

Группа	Критерий	Вес	Описание
Точность	Точность ответа	0.300	Насколько точно и уместно модель отвечает на запрос.
Точность	Логичность и структура	0.200	Логичность, структурированность и последовательность ответа.
Точность	Глубина и полнота ответа	0.200	Насколько полно и глубоко раскрыта тема запроса.
Точность	Гибкость в интерпретации	0.100	Работа с неоднозначными или неточно сформулированными запросами.
Точность	Адекватность формата	0.050	Соответствие формата ответа задаче (таблица, код, шаги и т.д.).
Контекст	Чувствительность к контексту	0.050	Учет предыдущих сообщений и контекста диалога.
Контекст	Креативность	0.050	Способность предложить нестандартные решения или идеи.
Контекст	Скорость ответа	0.050	Скорость реакции модели на запрос.
Контекст	Адаптивность к стилю общения	0.050	Подстройка под стиль и тон пользователя.
Устойчивость	Обработка сложных запросов	0.100	Работа с многоэтапными и сложными задачами.
Устойчивость	Восприятие неоднозначности	0.050	Работа с двусмысленными и открытыми запросами.
Контекст	Оценка релевантности источников	0.050	Наличие ссылок на релевантные и достоверные источники.
Устойчивость	Эффективность исправления ошибок	0.050	Исправление опечаток без потери смысла запроса.
Точность	Понимание технической спецификации	0.100	Корректная работа со специализированными запросами.
Устойчивость	Устойчивость к когнитивной нагрузке	0.100	Стабильность при больших объёмах данных и сложных вводах.
Контекст	Интерактивность	0.050	Умение поддерживать диалог и задавать уточняющие вопросы.
Точность	Анализ и синтез информации	0.100	Объединение данных из разных источников и обобщение.
Контекст	Поддержка мультимодальных запросов	0.100	Работа с текстом, изображениями и другими форматами.
Устойчивость	Работа с ограниченными данными	0.100	Осмысленные ответы при минимуме исходной информации.
Контекст	Контекстная согласованность	0.050	Сохранение связности и согласованности контекста.

Метрики LLM

Модель	Accuracy	Precision	Recall	F1	BLEU	ROUGE-L	BERTScore	PPL
ChatGPT	0.640	0.714	0.600	0.652	0.616	0.628	0.637	33.8
DeepSeek	0.620	0.629	0.615	0.622	0.598	0.609	0.618	35.4
Алиса AI	0.680	0.714	0.662	0.687	0.652	0.666	0.676	30.6

Матрица оценок

