ModelComparator · Финальная сводка

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

Краткий вывод

Итоговый отчёт
Проект: Демо проект
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)

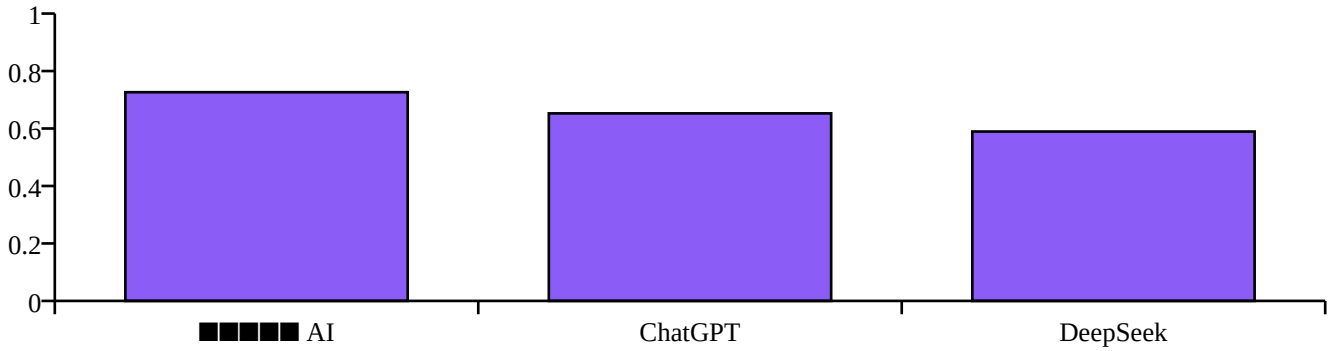
Условия, при которых выбор может измениться:
- Если повысить веса критериев, где сильнее ChatGPT (Логичность и структура, Глубина и полнота ответа, Понимание технической специфики), лидер может измениться.

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 запро
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 специфики	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

Критерий	ChatGPT	DeepSeek	Алиса AI
Точность ответа	1	2	5
Логичность и структура	5	2	3
Глубина и полнота ответа	5	2	3
Гибкость в интерпретации	3	4	4
Адекватность формата	1	3	3
Чувствительность к контексту	4	5	1
Креативность	5	5	5
Скорость ответа	4	1	3
Адаптивность к стилю общения	1	2	5
Обработка сложных запросов	5	3	5
Восприятие неоднозначности	1	2	4
Оценка релевантности источников	5	4	2
Эффективность исправления ошибок	1	1	1
Понимание технической специфики	5	4	2
Устойчивость к когнитивной нагрузке	3	3	5
Интерактивность	5	1	5
Анализ и синтез информации	5	5	5
Поддержка мультимодальных запросов	2	4	2
Работа с ограниченными данными	2	5	4
Контекстная согласованность	1	4	1

Сильные и слабые стороны

Модель	Сильные стороны	Слабые стороны	Пропуски
Алиса AI	Точность ответа (оценка 5, вклад 1.50); Гибкость в интерпретации (оценка 4, вклад 0.80); Креативность (оценка 5, вклад 0.25)	Логичность и структура (оценка 3, вклад 0.60); Глубина и полнота ответа (оценка 3, вклад 0.60); Адекватность формата (оценка 3, вклад 0.60); Чувствительность к контексту (оценка 1, вклад 0.05); Скорость ответа (оценка 3, вклад 0.60); Адаптивность к стилю общения (оценка 5, вклад 1.00); Обработка сложных запросов (оценка 5, вклад 1.00); Восприятие неоднозначности (оценка 4, вклад 0.80); Оценка релевантности источников (оценка 2, вклад 0.40); Эффективность исправления ошибок (оценка 1, вклад 0.05); Понимание технической специфики (оценка 2, вклад 0.40); Устойчивость к когнитивной нагрузке (оценка 5, вклад 1.00); Интерактивность (оценка 5, вклад 1.00); Анализ и синтез информации (оценка 5, вклад 1.00); Поддержка мультимодальных запросов (оценка 2, вклад 0.40); Работа с ограниченными данными (оценка 4, вклад 0.80); Контекстная согласованность (оценка 1, вклад 0.05)	
ChatGPT	Логичность и структура (оценка 5, вклад 1.00); Глубина и полнота ответа (оценка 5, вклад 1.00); Адекватность формата (оценка 1, вклад 0.05); Чувствительность к контексту (оценка 4, вклад 0.80); Креативность (оценка 5, вклад 0.25); Скорость ответа (оценка 4, вклад 0.80); Адаптивность к стилю общения (оценка 1, вклад 0.05); Обработка сложных запросов (оценка 5, вклад 1.00); Восприятие неоднозначности (оценка 1, вклад 0.05); Оценка релевантности источников (оценка 5, вклад 1.00); Эффективность исправления ошибок (оценка 1, вклад 0.05); Понимание технической специфики (оценка 5, вклад 1.00); Устойчивость к когнитивной нагрузке (оценка 3, вклад 0.60); Интерактивность (оценка 5, вклад 1.00); Анализ и синтез информации (оценка 5, вклад 1.00); Поддержка мультимодальных запросов (оценка 2, вклад 0.40); Работа с ограниченными данными (оценка 2, вклад 0.40); Контекстная согласованность (оценка 1, вклад 0.05)		
DeepSeek	Точность ответа (оценка 2, вклад 0.60); Логичность и структура (оценка 2, вклад 0.40); Глубина и полнота ответа (оценка 2, вклад 0.40); Гибкость в интерпретации (оценка 4, вклад 0.80); Адекватность формата (оценка 3, вклад 0.60); Чувствительность к контексту (оценка 5, вклад 1.00); Креативность (оценка 5, вклад 0.25); Скорость ответа (оценка 1, вклад 0.05); Адаптивность к стилю общения (оценка 2, вклад 0.40); Обработка сложных запросов (оценка 3, вклад 0.60); Восприятие неоднозначности (оценка 2, вклад 0.40); Оценка релевантности источников (оценка 4, вклад 0.80); Эффективность исправления ошибок (оценка 1, вклад 0.05); Понимание технической специфики (оценка 4, вклад 0.80); Устойчивость к когнитивной нагрузке (оценка 3, вклад 0.60); Интерактивность (оценка 1, вклад 0.05); Анализ и синтез информации (оценка 5, вклад 1.00); Поддержка мультимодальных запросов (оценка 4, вклад 0.80); Работа с ограниченными данными (оценка 5, вклад 0.50); Контекстная согласованность (оценка 4, вклад 0.80)		