ModelComparator · Финальная сводка

Проект: Демо проект (резерв копия)
Сценарий: Логический сценарий
Описание сценария: Как приготовить свиные крылышки?
Моделей: 3 · Критериев: 10 · Оценок: 30

Краткий вывод

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

Рекомендуемая модель: DeepSeek
Интегральная оценка (K): 0.792
Интерпретация: Хорошая модель, подходит для большинства задач.

Почему именно эта модель:
- Ключевые критерии вклада: Точность ответа, Логичность и структура, Глубина и полнота ответа
- Отрыв от второго места: 0.217
- Выбор выглядит стабильным.
- Ближайший конкурент: ChatGPT
- Наибольшее преимущество по критериям: Глубина и полнота ответа, Логичность и структура, Обработка сложных запросов
- Где лидер уступает: Креативность, Гибкость в интерпретации, Понимание технической специфики
- Сильные стороны лидера: Точность ответа (оценка 4, вклад 1.20), Логичность и структура (оценка 5, вклад 1.00), Глубина и полнота ответа (оценка 5, вклад 1.00)
- Слабые стороны лидера: Креативность (оценка 1, вклад 0.05), Гибкость в интерпретации (оценка 1, вклад 0.10), Чувствительность к контексту (оценка 4, вклад 0.20)

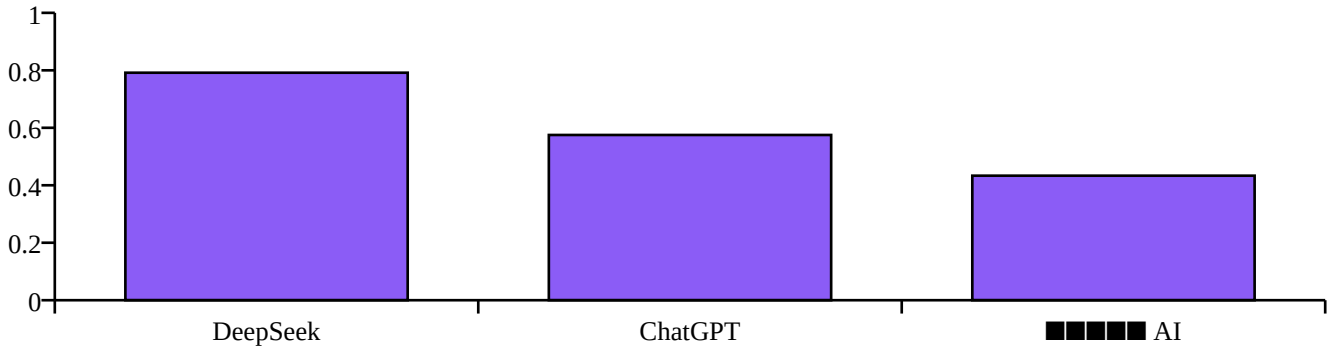
Условия, при которых выбор может измениться:
- Если повысить веса критериев, где сильнее ChatGPT (Креативность, Гибкость в интерпретации, Понимание технической специфики), лидер может измениться.

Примечание: расчёт основан на включённых критериях и текущих весах.

Рейтинг моделей

Место	Модель	K	Взвешенная сумма	Пропуски
1	DeepSeek	0.792	4.750	0
2	ChatGPT	0.575	3.450	0
3	Алиса AI	0.433	2.600	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.
Устойчивость	Обработка сложных запросов	0.100	Работа с многоэтапными и сложными задачами.
Точность	Понимание технической специфики	0.000	Корректная работа со специализированными запросами.
Контекст	Контекстная согласованность	0.050	Сохранение связности и согласованности контекста.

Метрики LLM

Модель	Accuracy	Precision	Recall	F1	BLEU	ROUGE-L	BERTScore	PPL
ChatGPT	0.600	0.533	0.700	0.605	0.580	0.590	0.598	37.0
DeepSeek	0.740	0.767	0.700	0.732	0.706	0.723	0.735	25.8
Алиса AI	0.400	0.467	0.300	0.365	0.400	0.400	0.402	53.0

Матрица оценок

Критерий	ChatGPT	DeepSeek	Алиса AI
Точность ответа	4	4	1
Логичность и структура	3	5	5
Глубина и полнота ответа	1	5	3
Гибкость в интерпретации	2	1	1
Адекватность формата	2	5	3
Чувствительность к контексту	5	4	2
Креативность	5	1	2
Обработка сложных запросов	1	4	1
Понимание технической специфики	4	3	1
Контекстная согласованность	3	5	1

Сильные и слабые стороны

Модель	Сильные стороны	Слабые стороны	Пропуски
DeepSeek	Точность ответа (оценка 4, вклад 1.20); Логичность и структура (оценка 5, вклад 1.00); Гибкость в интерпретации (оценка 1, вклад 0.10); Адекватность формата (оценка 5, вклад 0.25); Чувствительность к контексту (оценка 4, вклад 0.20); Креативность (оценка 1, вклад 0.05); Обработка сложных запросов (оценка 4, вклад 0.40); Понимание технической специфики (оценка 3, вклад 0.00); Контекстная согласованность (оценка 5, вклад 0.25)	—	—
ChatGPT	Точность ответа (оценка 4, вклад 1.20); Адекватность формата (оценка 2, вклад 0.10); Чувствительность к контексту (оценка 5, вклад 0.25); Креативность (оценка 5, вклад 0.25); Понимание технической специфики (оценка 4, вклад 0.00); Контекстная согласованность (оценка 3, вклад 0.15)	Логичность и структура (оценка 3, вклад 0.60); Глубина и полнота ответа (оценка 1, вклад 0.20); Гибкость в интерпретации (оценка 2, вклад 0.20); Обработка сложных запросов (оценка 1, вклад 0.10)	—
Алиса AI	Логичность и структура (оценка 5, вклад 1.00); Глубина и полнота ответа (оценка 3, вклад 0.60); Гибкость в интерпретации (оценка 1, вклад 0.10); Адекватность формата (оценка 3, вклад 0.15); Чувствительность к контексту (оценка 2, вклад 0.10); Креативность (оценка 2, вклад 0.10); Обработка сложных запросов (оценка 1, вклад 0.10); Понимание технической специфики (оценка 1, вклад 0.00); Контекстная согласованность (оценка 1, вклад 0.05)	—	—