10

Не «лучшая модель», а лучшая под вашу задачу

Как понять, какие AI-модели использовать и для каких задач

Что вы сможете: выбрать модель за час — по данным

Таир Султанов · Newgate Team

Три оси выбора и рамка ограничений



ОГРАНИЧЕНИЯ

- Можно ли отправлять данные в облако
- Как модель работает с русским и казахским
- Нужен ответ за секунду или можно подождать

Три класса моделей

КЛАСС	ДЛЯ ЧЕГО	ЦЕНА И СКОРОСТЬ
Флагманские	Сложные рассуждения	Дороже и медленнее
Средние и быстрые	Большинство бизнес-задач: классификация, пересказ, ответы по базе	Золотая середина
Малые и open-source	Простые задачи, можно развернуть у себя	Дёшево, но слабее на сложном

Названия и цены меняются каждый квартал — метод остаётся.

Метод из четырёх шагов

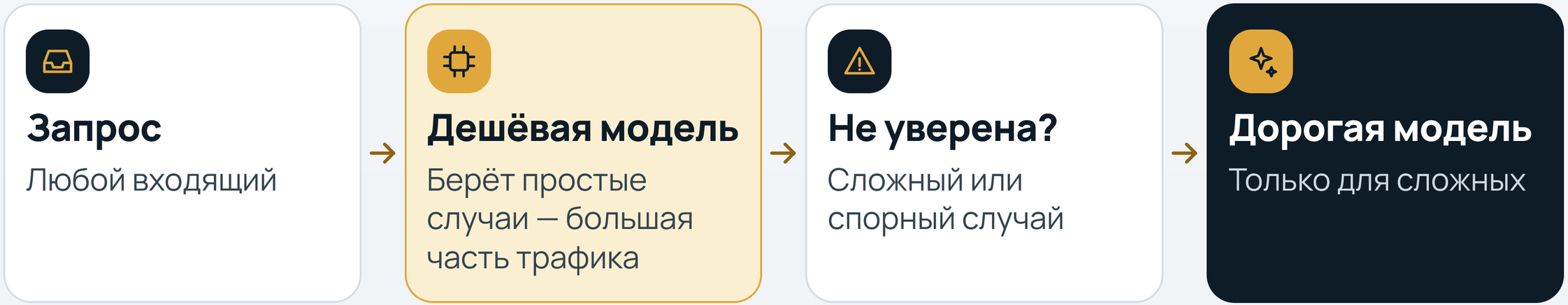
1. 10–20 реальных примеров вашей задачи
2. Прогон на 2–3 моделях с одним промптом
3. Оценка по понятному критерию: верно или нет
4. Цена за 1000 таких задач

Демо: три модели × 10 примеров

МОДЕЛЬ	ВЕРНО	СКОРОСТЬ	ЦЕНА ЗА 1000
Дорогая	10 из 10	Самая медленная	Самая высокая
Средняя	**9 из 10**	Быстрая	В разы ниже
Малая	**7 из 10**	Самая быстрая	Самая низкая

Условный пример: сортировка обращений по трём категориям. Свои цифры — из своего прогона.

Каскад: дорогая модель — только на сложное



Расходы падают, а качество на сложном сохраняется.

Соберите 10 своих примеров и прогоните на двух моделях

ЧЕРЕЗ НЕДЕЛЮ

Обновите тест на новых моделях — это 20 минут

Вопросы и примеры из ваших проектов — **@martelliad**