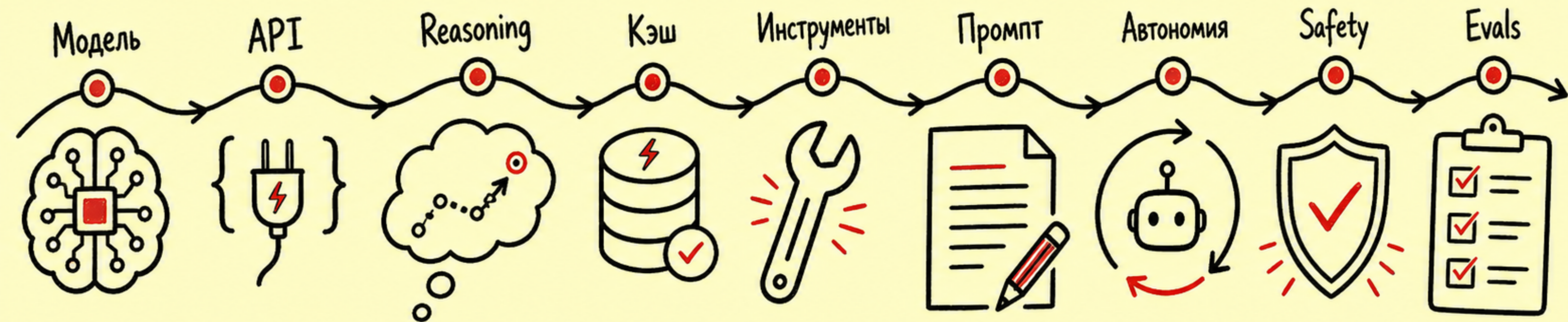


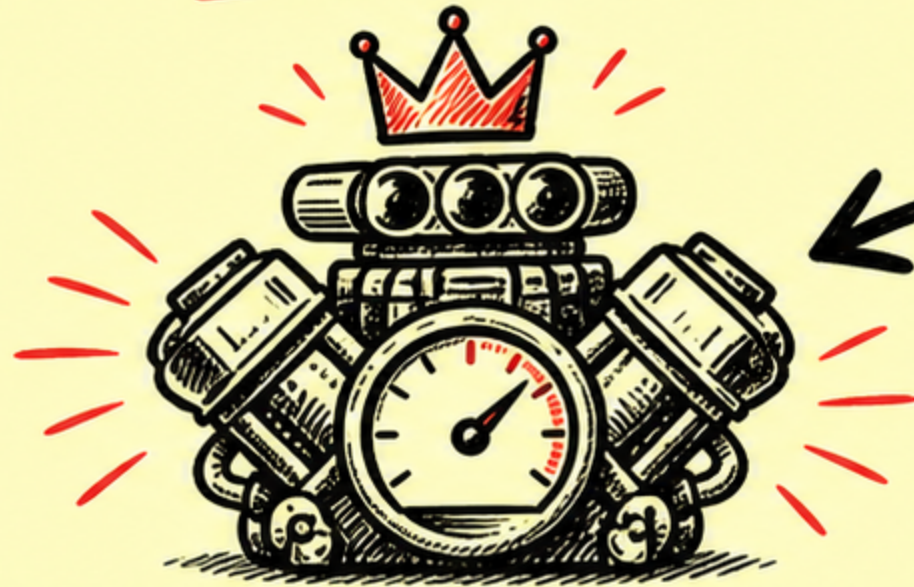
GPT-5.6: карта решений

17 практических рекомендаций OpenAI



Выбирайте модель под нагрузку

gpt-5.6-sol



Максимум возможностей

gpt-5.6-luna



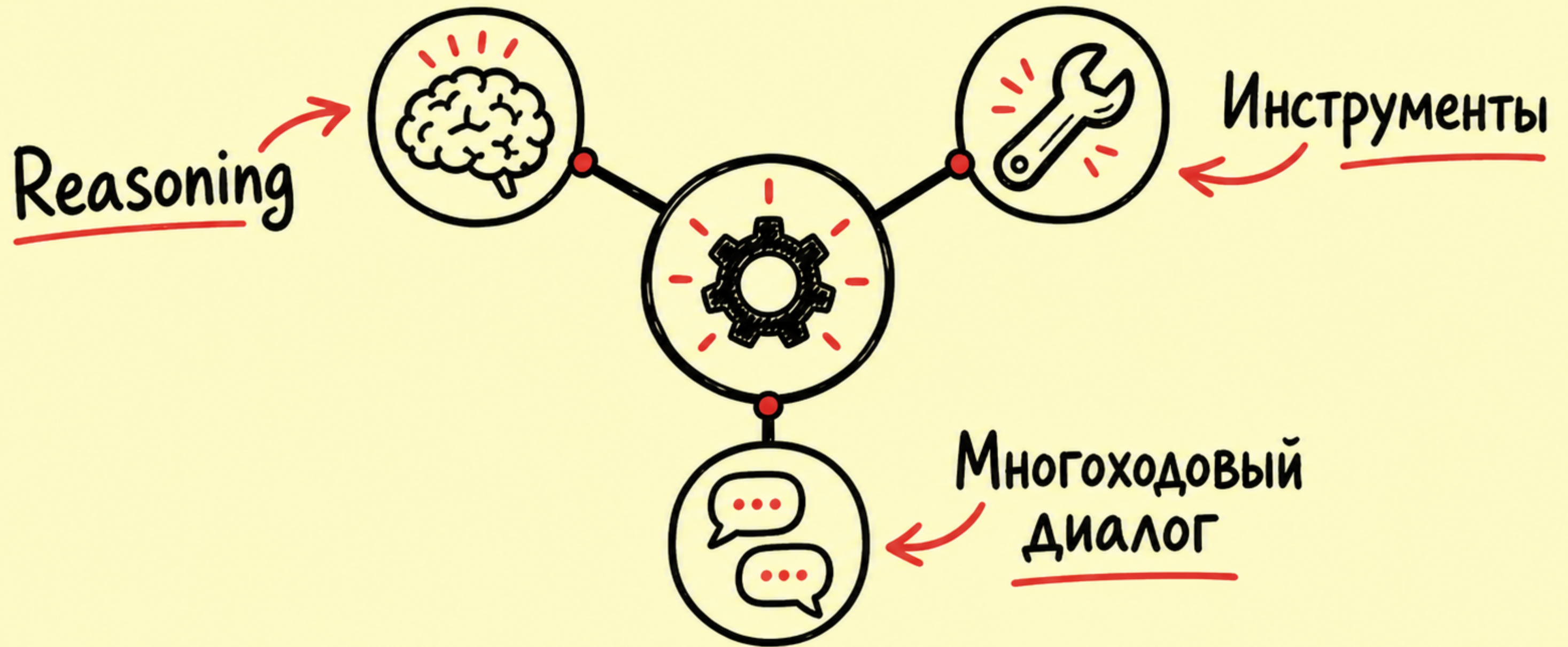
Высокообъёмные задачи

gpt-5.6-terra



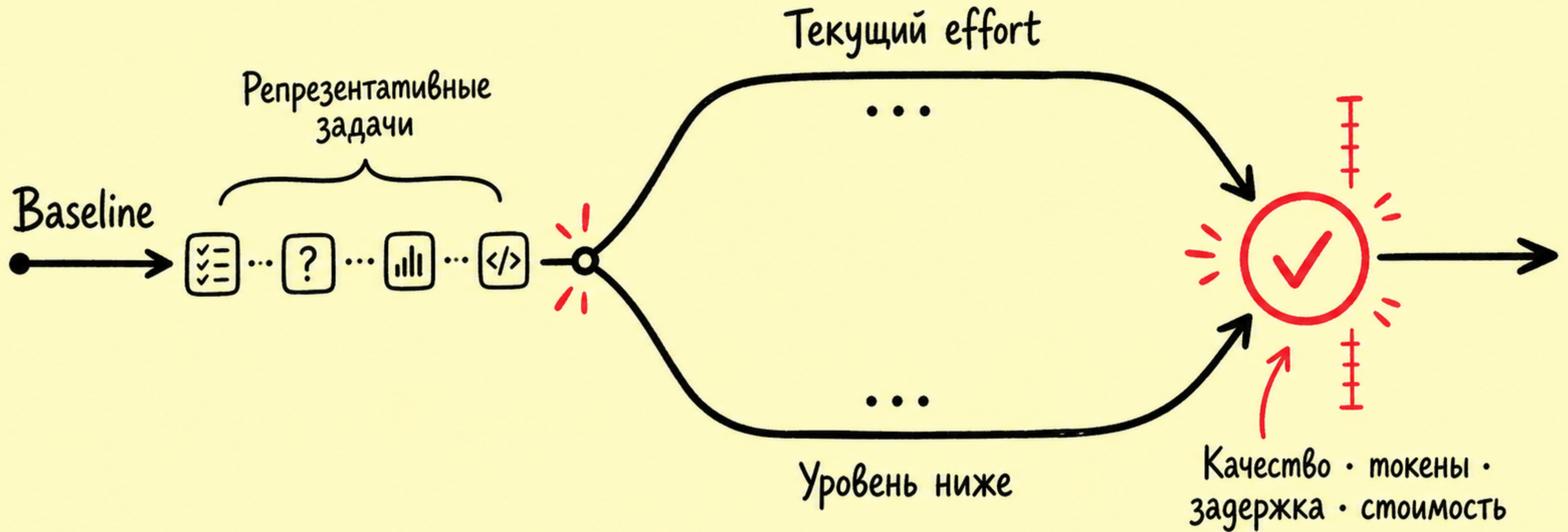
Баланс интеллекта и стоимости

Для сложных сценариев – Responses API



➤ Один операционный контур ➤

Мигрируйте через измеримый эксперимент



Задавайте reasoning.effort намеренно

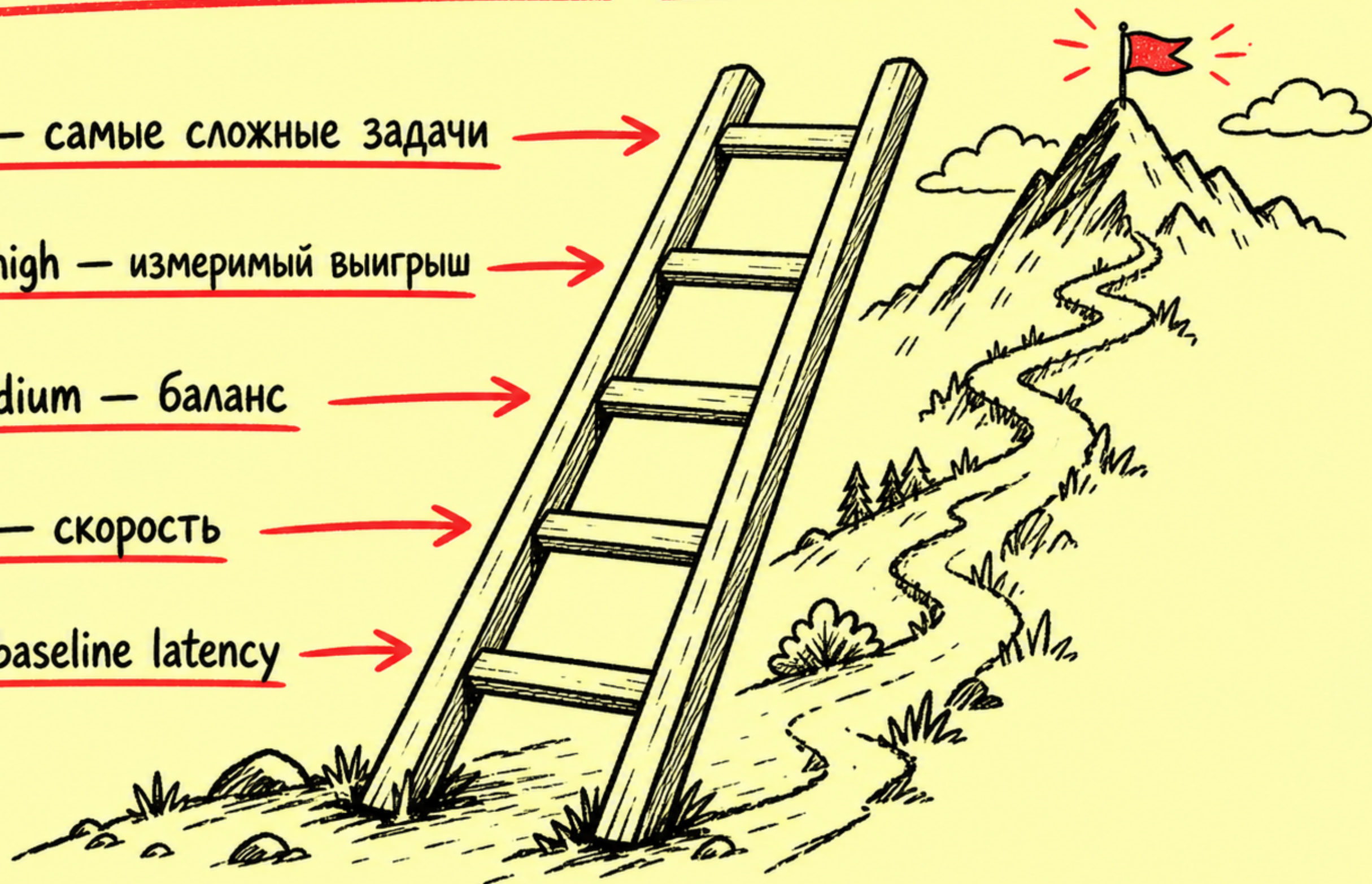
max — самые сложные задачи

high / xhigh — измеримый выигрыш

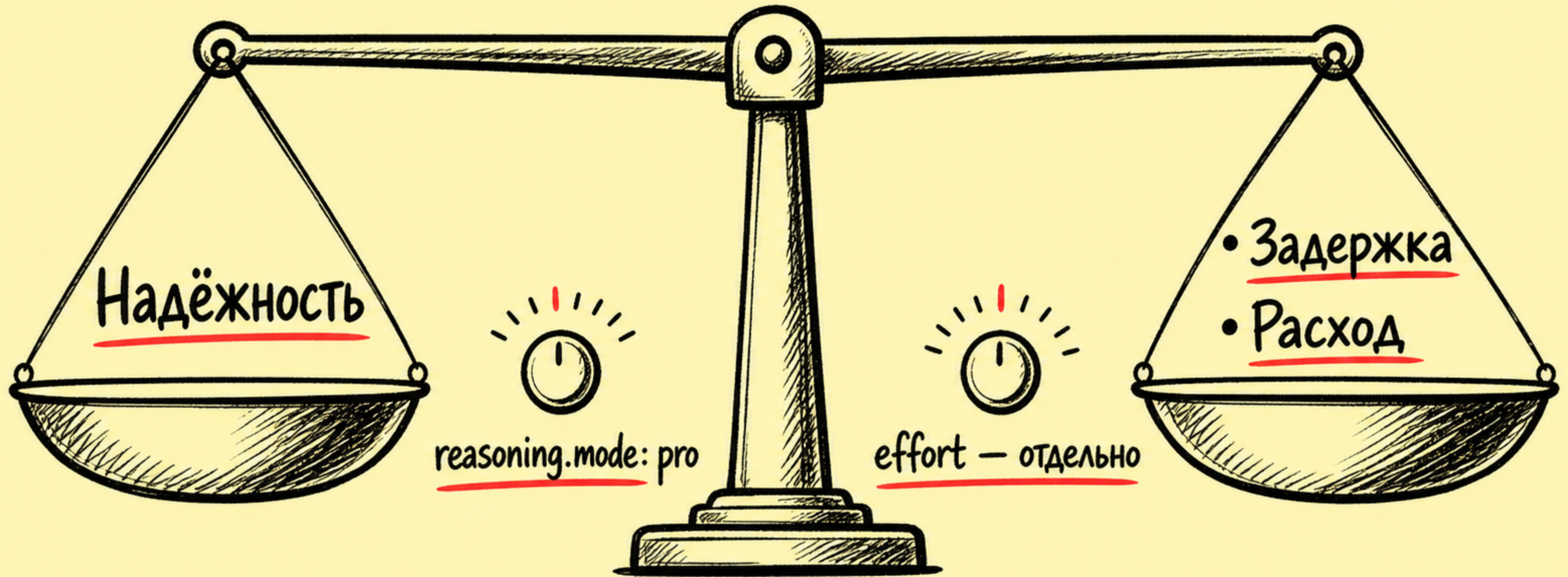
medium — баланс

low — скорость

none — baseline latency

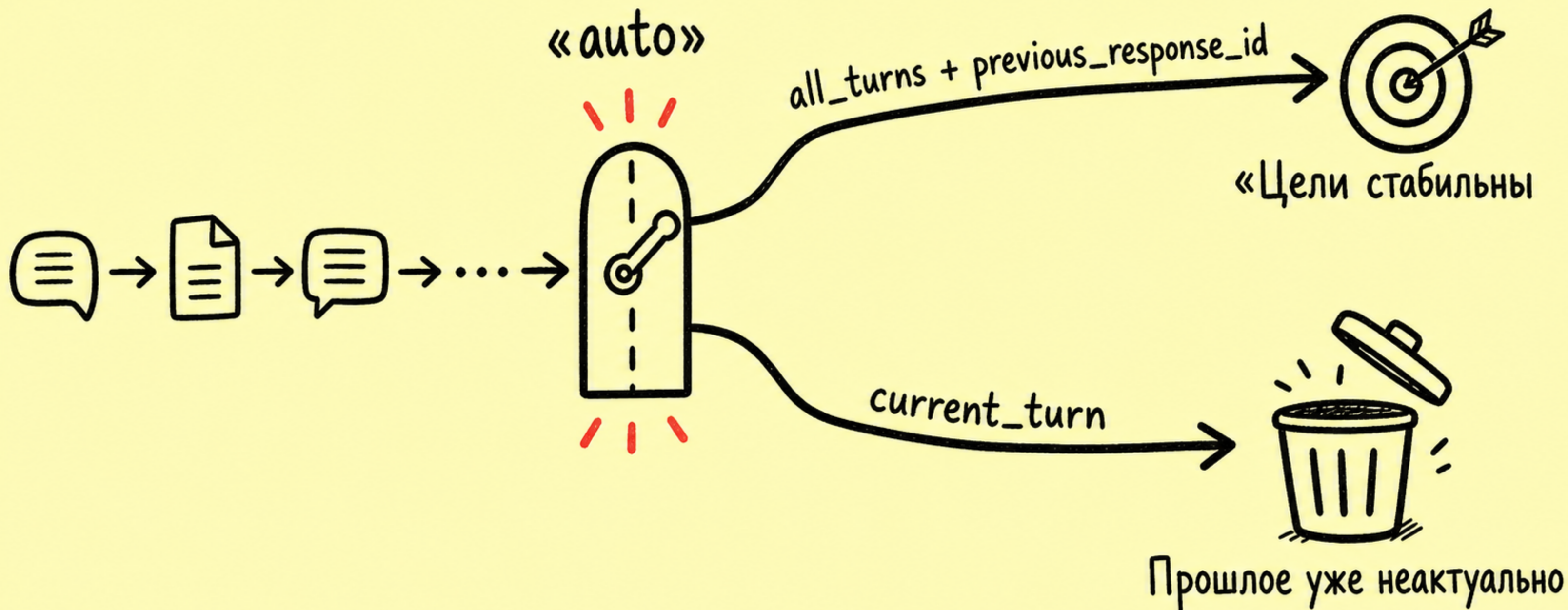


Pro mode — когда цена качества высока

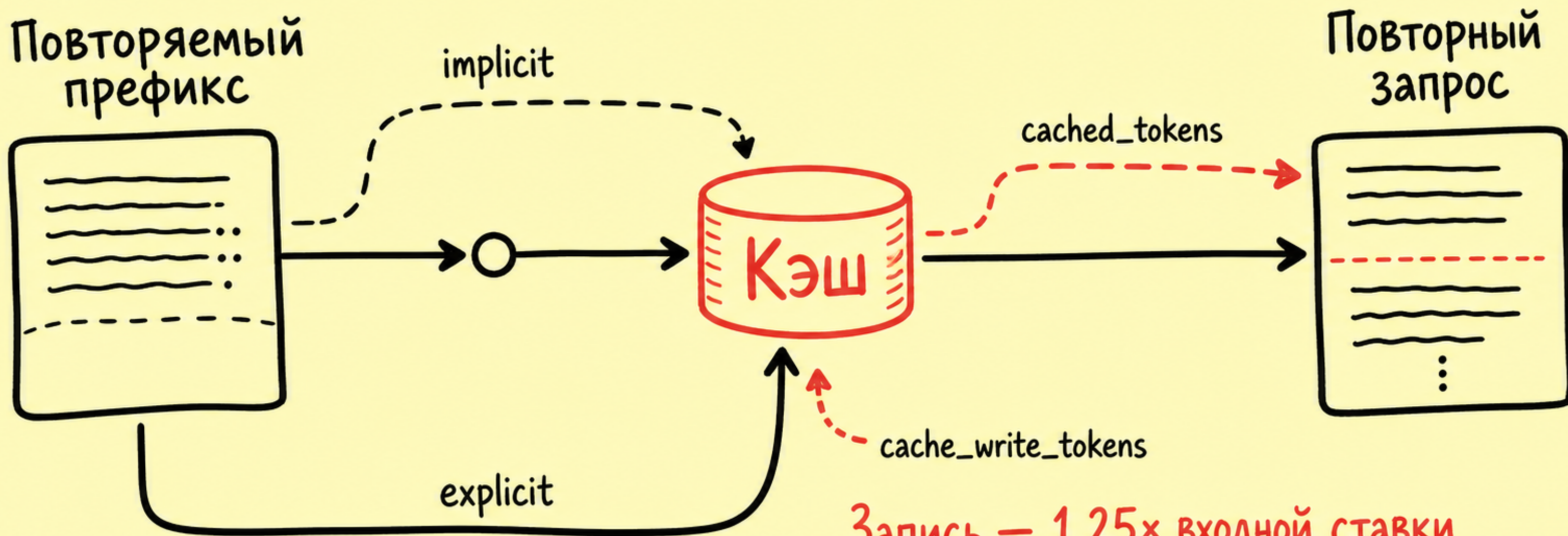


Сравнивайте со standard

«Сохраняйте reasoning, когда контекст стабилен»

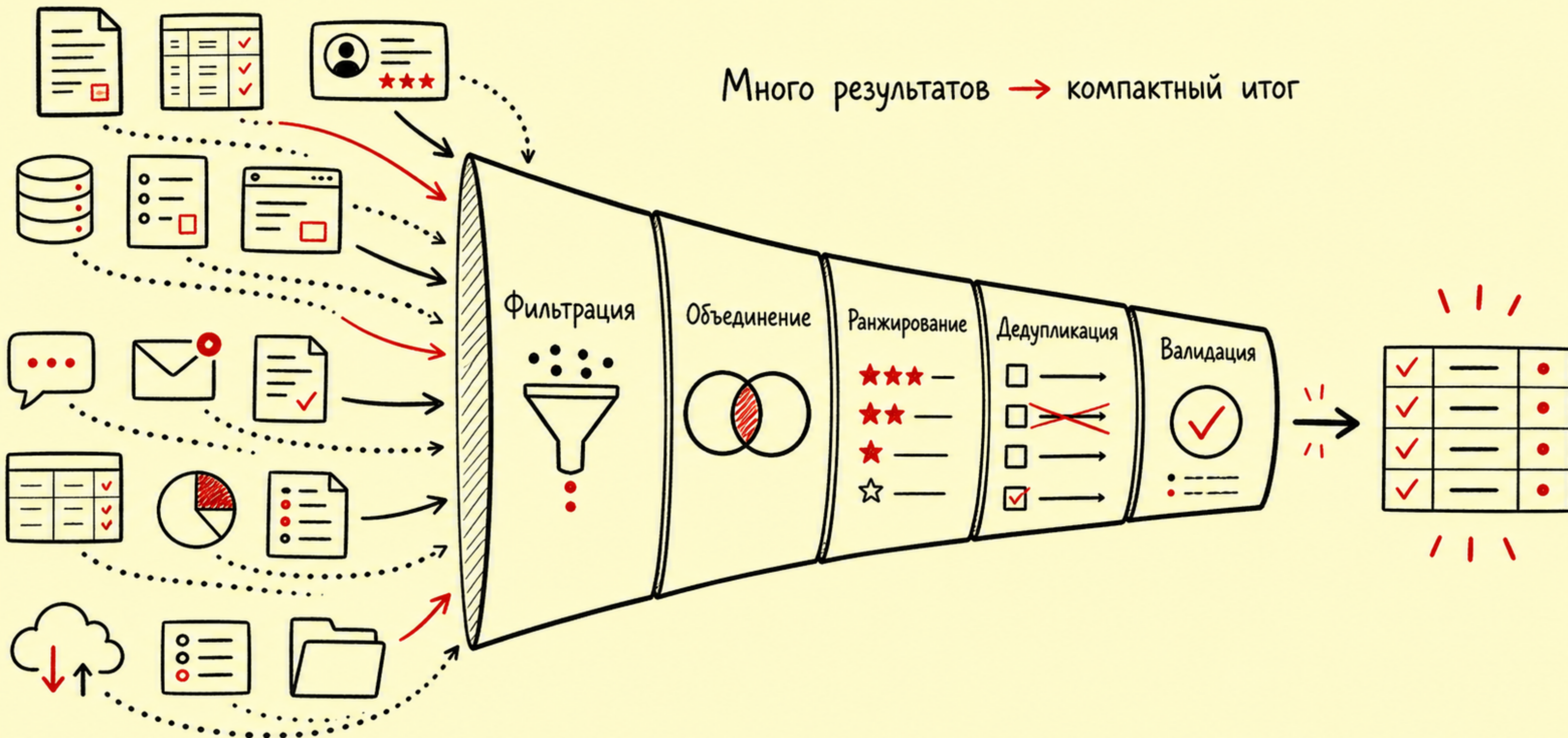


Кэшируйте осознанно

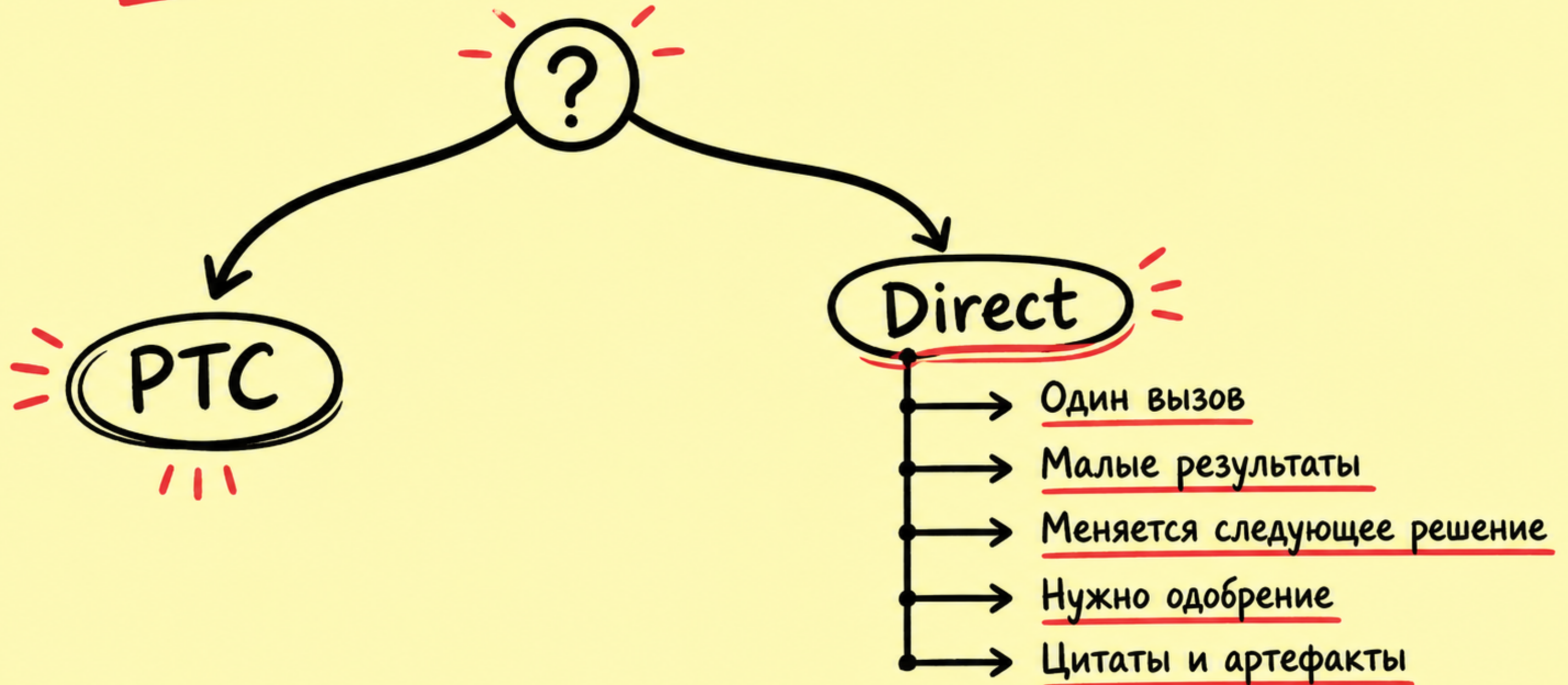


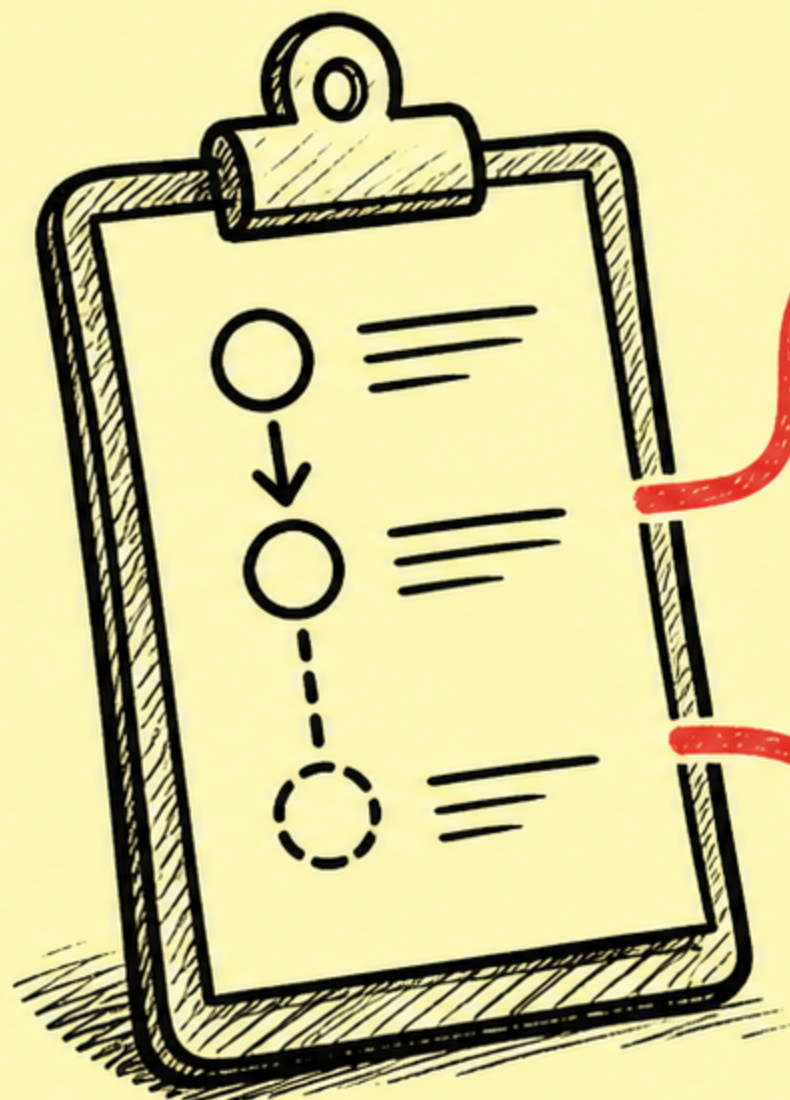
Запись — 1,25x входной ставки

РТС — для ограниченной обработки данных



Не включайте PTC просто из-за нескольких вызовов





«Давайте РТС точный контракт маршрутизации»

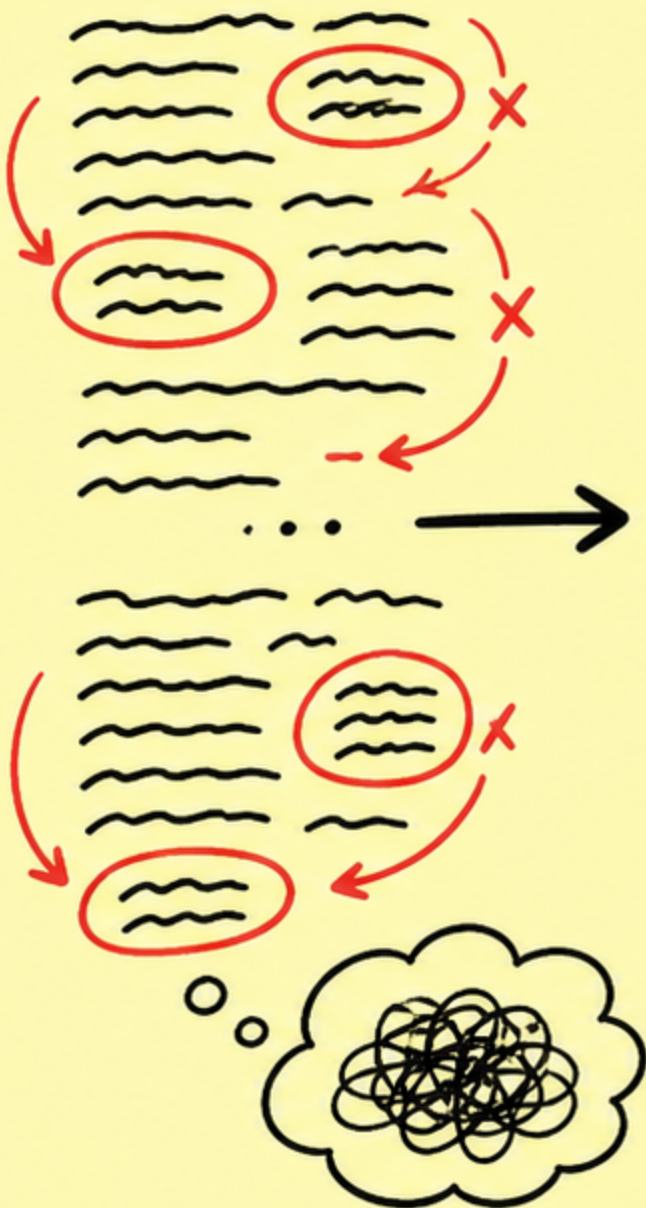
✓	Этап	≡
🔧	Допустимые инструменты	≡
📄	Схема результата и evidence	≡
🔄	Concurrency · retry · stop	≡
🛡️	Direct: judgment · approval · validation	≡

Signature



Упрощайте промпты, не требования

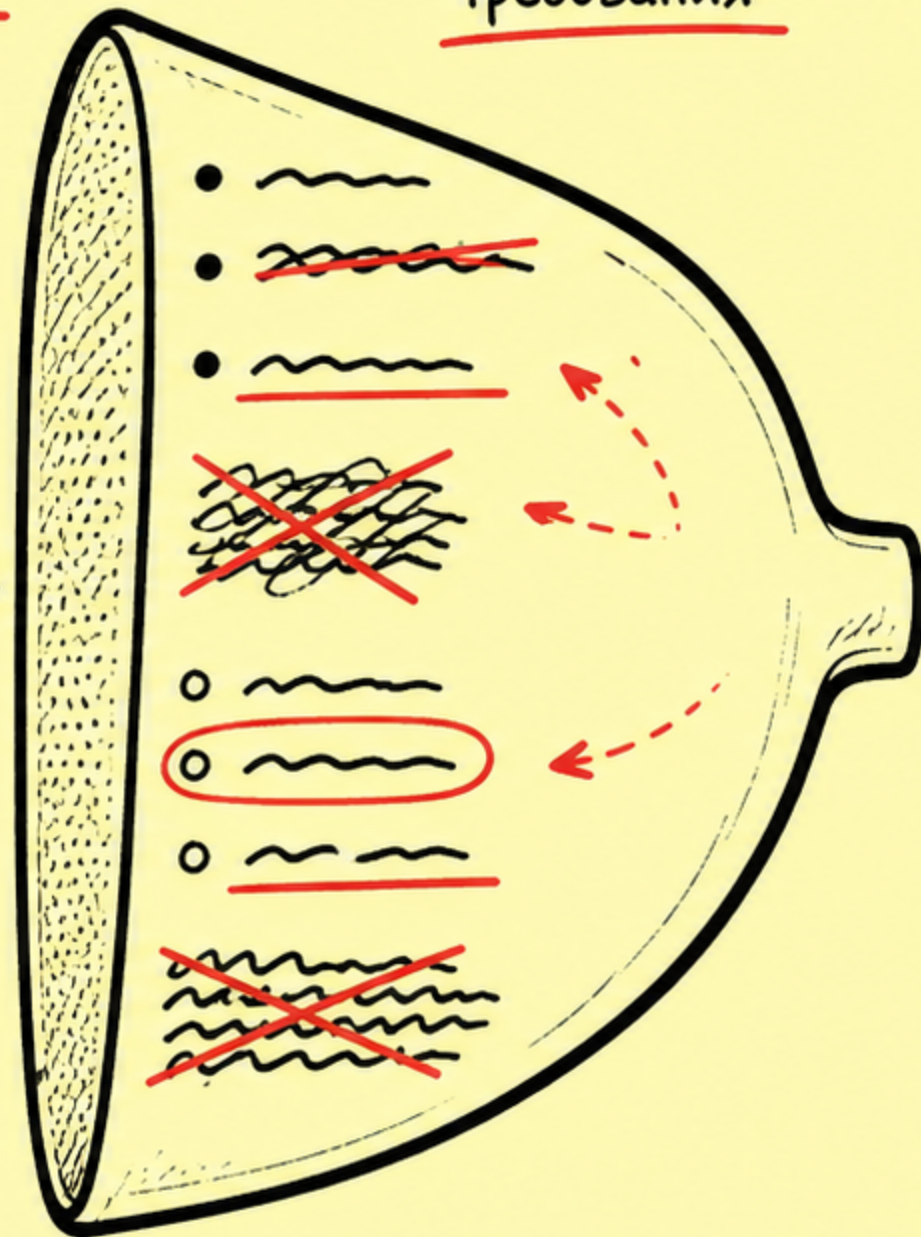
① Убрать повторы



② Убрать лишние примеры



③ Оставить продуктивные требования



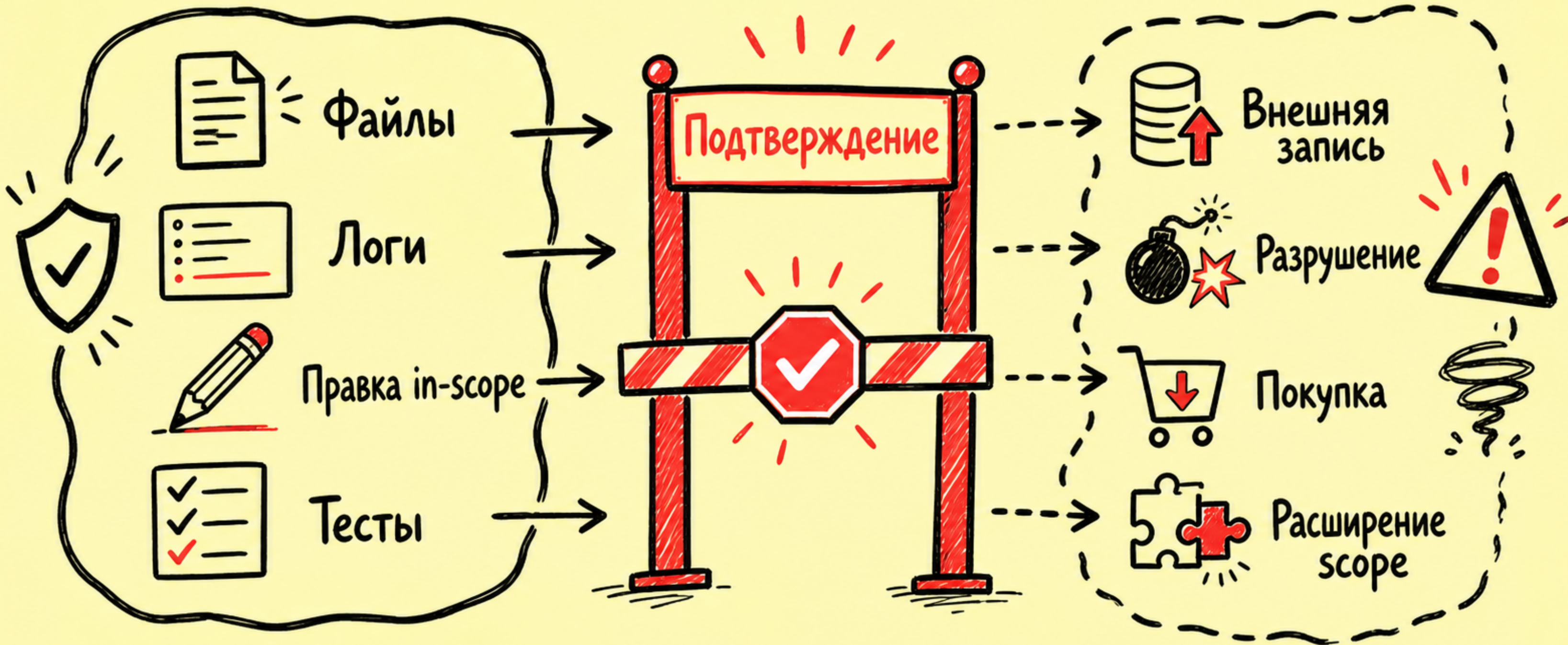
④ Оставить релевантные инструменты



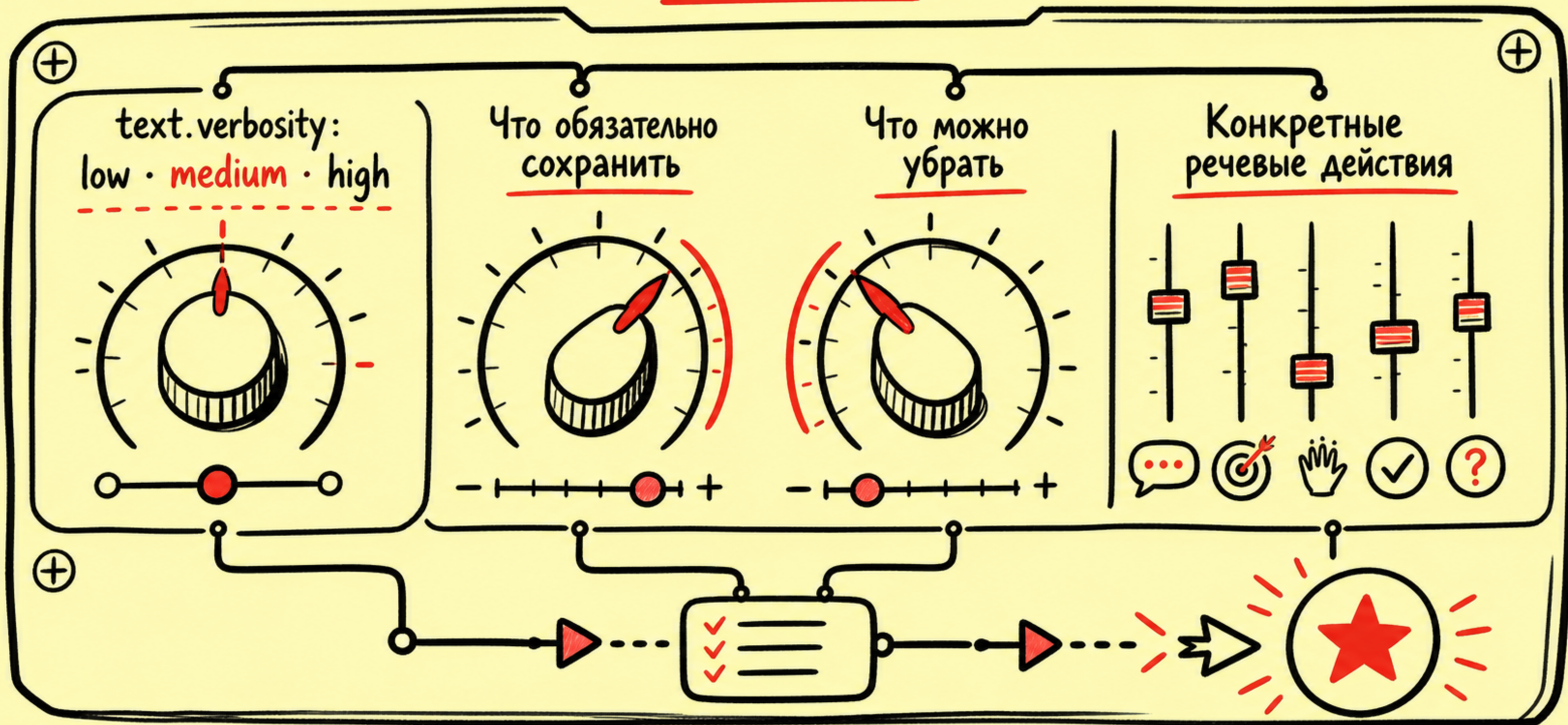
⑤ Прогнать те же evals



Ясно проведите границу автономии



Управляйте длиной и тоном как спецификацией

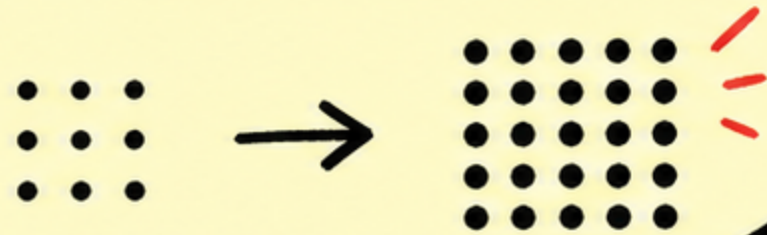


ДЕТАЛИЗАЦИЯ ИЗОБРАЖЕНИЙ — ПО ДЕЛУ

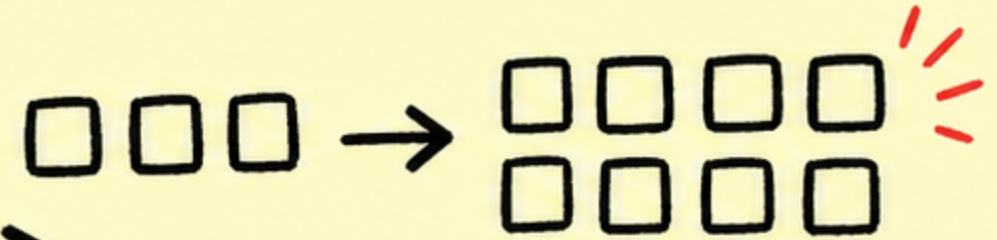
original / auto



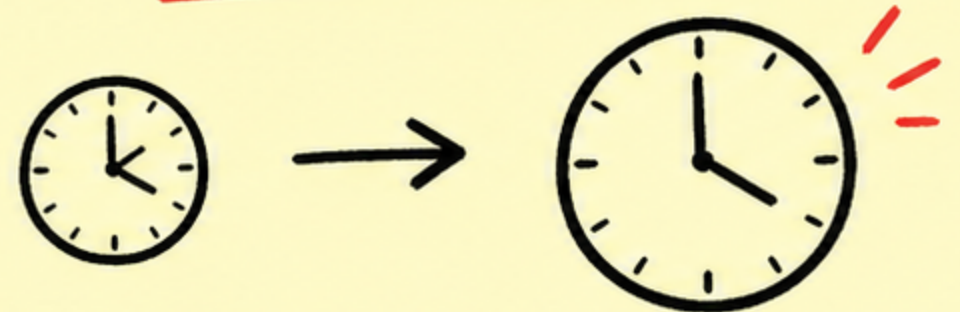
Больше деталей



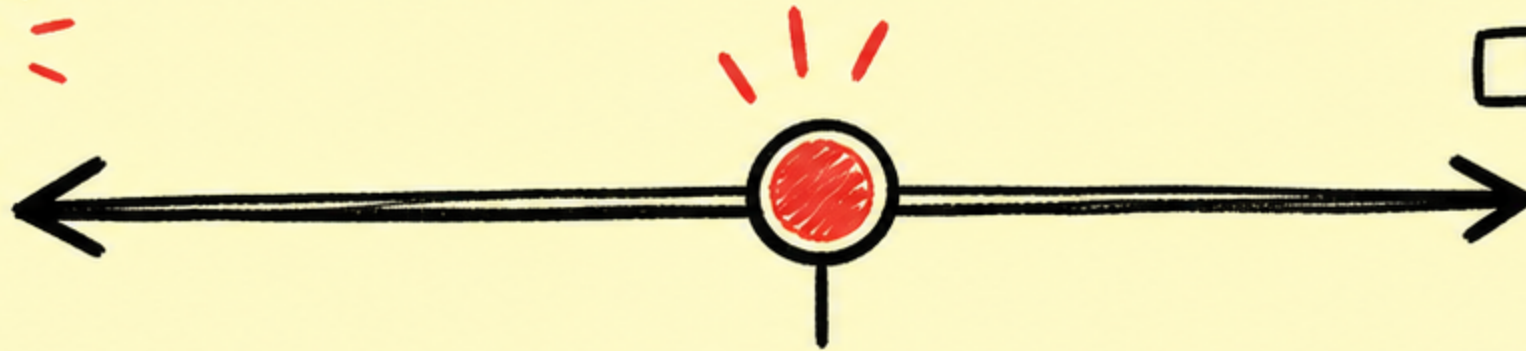
Больше токенов



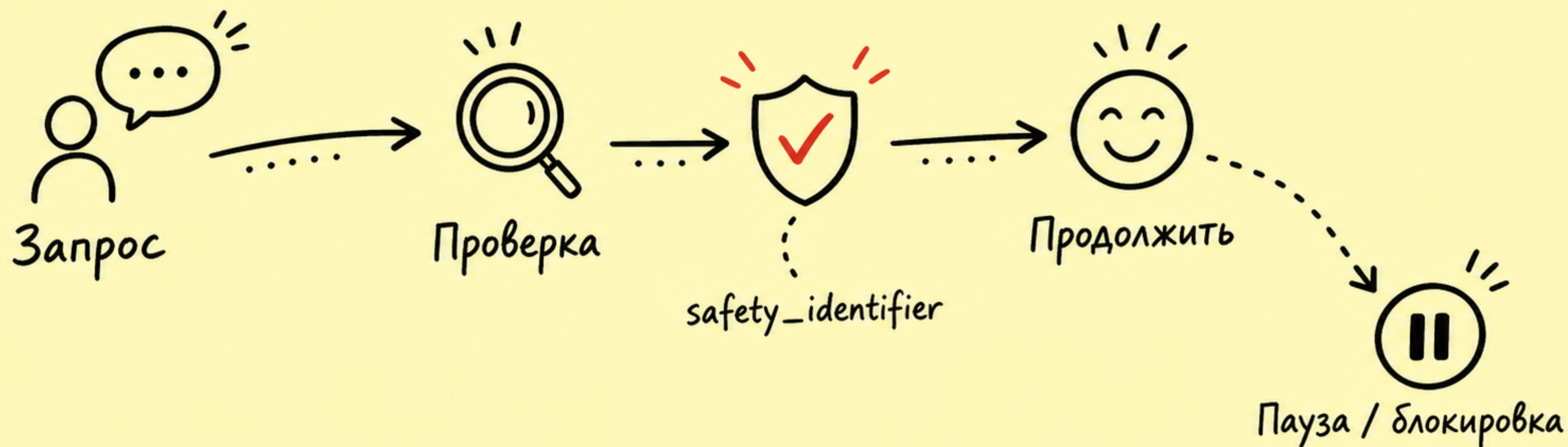
Выше задержка



Баланс

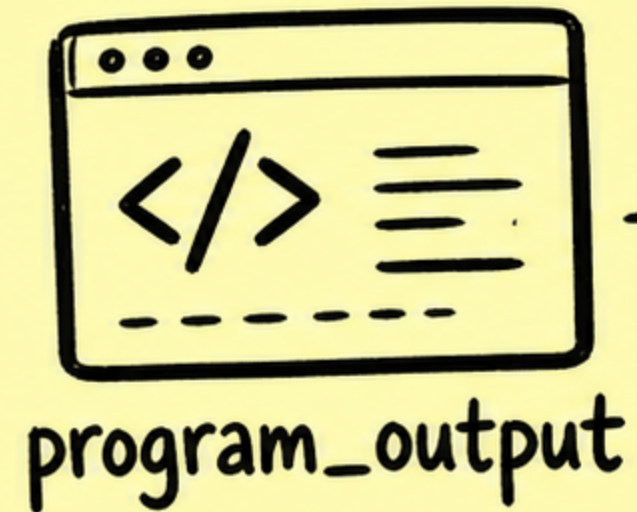


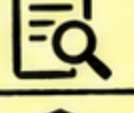
Стройте продукт с safeguards

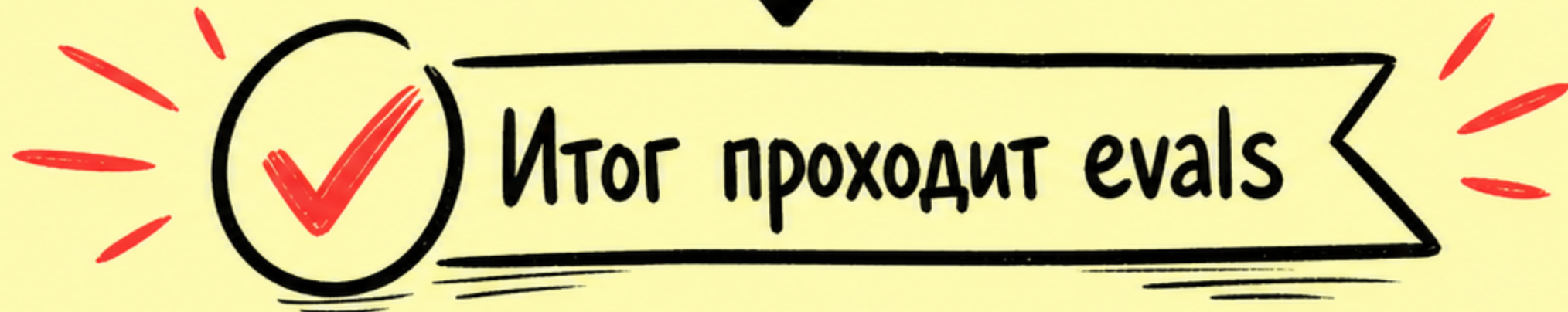


Легитимная защитная работа

Проверяйте финальный ответ, а не только выполнение



	Task success	✓
	Полнота	✓
	Evidence	✓
	Токены	✓
	Latency	✓
	Cost	✓



РАБОЧИЙ чек-лист внедрения

