14

Модель отвечает не так — что крутить?

Prompt engineering или fine-tuning: в чём разница и что выбрать

Что вы сможете: выбрать рычаг по симптому

Таир Султанов · Newgate Team

Три рычага



Промпт

Как отвечать: инструкция, формат, примеры



RAG

Что модель знает: фрагменты ваших документов в запросе



Fine-tuning

Как она ведёт себя: стиль и формат на примерах

Fine-tuning — не способ «залить в модель базу знаний». Для актуальных фактов лучше RAG.

Диагноз по симптому

СИМПТОМ	РЫЧАГ
Не знает наши данные	RAG
Оформляет не так или не тот тон	Сначала промпт с примерами, потом fine-tuning
Дорого и медленно на большом объёме	Малая модель + fine-tuning

Порядок всегда один



1. Промпт

Дёшево, быстро, обратимо



2. RAG

Когда не хватает знаний



3. Fine-tuning

Когда первые два упёрлись в потолок на ваших тестах



Движение слева направо: дешевле → дороже.

Демо: бот про условия доставки

БЕЗ ДАННЫХ

```
Сколько идёт доставка в Астану?
```

«Обычно 1–2 рабочих дня.»

Выдумка: модели неоткуда брать факты. Это не проблема промпта и не проблема обучения.

С ДОКУМЕНТОМ

```
Правила доставки (вымышленная
компания)
— Алматы: 1 день
— Другие города: 3-5 дней
— Заказ после 18:00 — на следующий
день
Отвечай только по документу.
```

«В Астану — 3-5 дней (раздел „Другие города“).»

Нет данных в документе — модель честно отказывается. **Никакого обучения.**

Когда всё-таки fine-tuning

1 **Сотни–тысячи чистых примеров**
«Вход → эталонный ответ»

2 **Данные однородные и чистые**
Грязные данные — модель выучит путаницу

3 **Есть тест-набор из урока 9**
Иначе не узнаете, стало лучше или нет

Опишите симптом → выберите рычаг → проверьте на тестах

ДЕЙСТВИЕ

Одним предложением опишите, что именно не устраивает в ответах модели

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