Чек-лист: артефакты моделей и закупка ИИ

К статье «Форматы артефактов моделей и решение open vs closed»

Шаг 1 — Фильтр privacy и compliance

- ☐ Есть ли в обработке регулируемые данные (152-ФЗ, HIPAA, PCI, GDPR special category)?
- ☐ Подтверждён ли регион API, где вы планируете звать closed-модель, по договору о данных?
- ☐ Просмотрел ли юрист обязательства Article 50 EU AI Act, вступающие 2 августа 2026?
- ☐ При self-host: есть ли документированная политика происхождения моделей и сканирования?

Шаг 2 — Фильтр capability gap

- ☐ Задача требует frontier reasoning (уровень GPQA) или это production (суммаризация, ASR, CV)?
- ☐ Прогнан ли хотя бы один open-weight на образце ВАШИХ данных в бенчмарке?
- ☐ При closed: задокументирован ли план fallback, если вендор удалит модель из API?
- ☐ Зафиксирована ли лицензия (Apache 2.0 vs custom) для каждого open-weight в стеке?

Шаг 3 — Crossover по стоимости

- ☐ Прогноз tokens/day на старте ________; через 12 месяцев ________.
- ☐ Closed API: стоимость за миллион токенов (input + output, blended) $________.
- ☐ GPU-час в вашем регионе (L4 / A100 / H100) $________.
- ☐ Ожидаемый batch size и throughput на GPU ________.
- ☐ Точка пересечения (closed = open total) рассчитана и зафиксирована?

Шаг 4 — Налог на инженерную ёмкость

- ☐ Доступные роли: ML-инженер ___, DevOps ___, SRE ___, security ___.
- ☐ Месячная полная стоимость open-weights стека $________.
- ☐ Runbook на дефицит GPU / падение региона / краш vLLM написан?
- ☐ Процесс обновления моделей и A/B-тестов для новых open-weights релизов описан?

Шаг 5 — Гибридный routing

- ☐ Описано правило routing для 'лёгких 80%' (open) и 'сложных 20%' (closed)?
- ☐ Per-request телеметрия по стоимости (closed tokens, open GPU-seconds) включена?
- ☐ Квартальный пересчёт стоимости в календаре?
- ☐ План отката, если open-weights качество просядет после bump-а модели?

Чек-лист по формату артефакта

- ☐ Мобильное приложение: модель сконвертирована в Core ML (iOS) или LiteRT / ONNX (Android)?
- ☐ Браузер: ONNX (WebGPU) или llama.cpp WASM (GGUF)? Размер бандла замерен?
- ☐ Edge-сервер: ONNX Runtime + TensorRT EP или vLLM Safetensors? Throughput замерен?
- ☐ Cloud LLM serving: Safetensors через vLLM или скомпилированный TensorRT-LLM engine?
- ☐ Если только pickle: отказались или просканировали picklescan и положили в карантин?
- ☐ Каждый выбор квантизации (Q4_K_M, Q5_K_M, Q8_0) подкреплён бенчмарком качества?