

Чек-лист пилота DeepSeek-V4-Flash-0731

Чек-лист нужен для команды, которая хочет отделить быстрый тест API от проверки self-hosting и производных квантованных сборок. Подготовить для скачивания отдельный PDF-файл <downloads/deepseek-v4-flash-0731-pilot-checklist.pdf>.

Перед запуском API-пилота

- Зафиксировать целевые задачи: длинные документы, код, RAG, tool calls или агентный workflow.
- Собрать контрольный набор запросов и ожидаемых форматов ответов.
- Отдельно измерять качество, задержку, длину ответа и стоимость задачи.
- Проверить влияние cache hit и cache miss на повторяющиеся префиксы.

Перед self-hosting

- Определить доступные GPU, сеть, дисковую подсистему и ограничения тестового контура.
- Проверить поддержку нужного runtime, KV-cache, MoE-параллелизма и DSpark-механизмов.
- Измерить latency, throughput, расход памяти и поведение при конкурентной нагрузке.
- Настроить логирование, мониторинг ошибок и повторяемый сценарий развертывания.

Перед использованием порта или квантованной сборки

- Зафиксировать происхождение сборки и базовую ревизию модели.
- Проверить tokenizer, chat template и параметры квантования.
- Сравнить качество с базовой версией на одном контрольном наборе.
- Провести smoke tests, regression tests и нагрузочную проверку до интеграции в сервис.