

Чек-лист: проверка Dynamic-сборки перед локальным пилотом

PDF-версия предназначена для сохранения и использования при подготовке стенда, сравнении квантизованных вариантов и переносе в рабочий контур.

Происхождение и совместимость

- Зафиксировать базовую модель DeepSeek-V3, имя сборки и ревизию файла.
- Проверить поддержку архитектуры и формата весов выбранным runtime.
- Зафиксировать версии backend, драйверов, зависимостей и параметры генерации.
- Проверить шаблон чата, системные инструкции и стоп-последовательности.

Качество и производительность

- Использовать единый набор промптов и ожидаемых результатов для сравнения конфигураций.
- Отдельно проверить структурированный вывод, инструменты и код, если они используются.
- Измерить обработку входа, генерацию, задержку и потребление памяти.
- Проверить поведение с целевой длиной контекста и параллельной нагрузкой.

Решение о переносе

- Сопоставить экономию ресурсов с качеством на прикладных задачах.
- Отделить результаты пилота от требований к production-сервису.
- Запускать smoke-тест после обновления зависимостей.
- Не считать успешный инференс подтверждением готовности к full fine-tuning.