

Шаблон брифа для инженера

Заполните вместе с командой, передайте в Инком интер для предварительной оценки конфигурации сервера.

Как пользоваться шаблоном. Пройдите пять факторов, отметьте варианты и заполните поля своими данными. Не обязательно отвечать на всё: то, что пока неизвестно, оставьте пустым, мы обсудим эти пункты вместе. Готовый бриф даёт базовое ТЗ, из которого получается конфигурация под конкретный сценарий, а не под «средний» проект.

1 Задача

ТИП СЦЕНАРИЯ

- ☐ Внутренний ассистент по документации ☐ Суммаризация длинных текстов
- ☐ Поиск по внутренней базе ☐ Code-помощник в IDE ☐ Классификация обращений
- ☐ Автоматизация процессов ☐ Другое (укажите ниже)

РЕЖИМ РАБОТЫ

- ☐ Интерактивный (пользователь ждёт ответ) ☐ Пакетный (задачи в фоне)
- ☐ Смешанный (оба сценария)

ХАРАКТЕР РАБОТЫ МОДЕЛИ

- ☐ Генерация нового текста ☐ Анализ существующего ☐ Оба

СВОБОДНОЕ ОПИСАНИЕ ЗАДАЧИ И ЕЁ КОНТЕКСТА В БИЗНЕСЕ

2 Concurrency (одновременность)

СРЕДНЕЕ ЧИСЛО ОДНОВРЕМЕННЫХ ЗАПРОСОВ (ТИПОВАЯ РАБОЧАЯ СИТУАЦИЯ)

ПИКОВОЕ ЧИСЛО ОДНОВРЕМЕННЫХ ЗАПРОСОВ (МАКСИМУМ ПОД НАГРУЗКОЙ)

ПРОФИЛЬ НАГРУЗКИ В СУТКАХ

- ☐ Равномерная в течение рабочего дня ☐ Пиковые часы (утро, после обеда)
☐ Круглосуточная (24/7)

ЦЕЛЕВОЙ КВАНТИЛЬ ЗАДЕРЖКИ ПОД ПИКОМ

- ☐ p95 (95 из 100 запросов укладываются) ☐ p99 (99 из 100) ☐ Не определено

3 Время отклика

ЦЕЛЕВАЯ ЗАДЕРЖКА ПЕРВОГО ТОКЕНА (TIME-TO-FIRST-TOKEN)

- ☐ До 0,5 сек ☐ До 1 сек ☐ До 2 сек ☐ До 5 сек ☐ Не критично

ДОПУСТИМОЕ ПОЛНОЕ ВРЕМЯ ГЕНЕРАЦИИ ОТВЕТА

- ☐ До 5 сек ☐ До 15 сек ☐ До 60 сек ☐ Больше приемлемо

КОММЕНТАРИИ ПО ВОСПРИЯТИЮ ОТКЛИКА ПОЛЬЗОВАТЕЛЯМИ

4 Класс модели

КЛАСС ПЛАНИРУЕМОЙ МОДЕЛИ

- ☐ Малая (единицы миллиардов параметров) ☐ Средняя (десятки миллиардов)
☐ Большая (сотни миллиардов) ☐ Не определено

ОЖИДАЕМАЯ ДЛИНА КОНТЕКСТА (В СРЕДНЕМ, В ПРЕДЕЛАХ)

ПЛАНЫ НА КВАНТОВАНИЕ

- ☐ FP16 (исходное качество) ☐ INT8 (VRAM вдвое меньше) ☐ INT4 (VRAM вчетверо меньше)
☐ Не определено, обсудить

Если для сценария критично качество ответа (юридический ассистент, точная терминология), квантование лучше тестировать на своих данных перед выбором.

5 Горизонт роста

ОЖИДАЕМОЕ ЧИСЛО ПОЛЬЗОВАТЕЛЕЙ ЧЕРЕЗ 12 МЕСЯЦЕВ

ОЖИДАЕМОЕ ЧИСЛО ПОЛЬЗОВАТЕЛЕЙ ЧЕРЕЗ 24 МЕСЯЦА

ПЛАНЫ ПО КЛАССУ МОДЕЛИ ЧЕРЕЗ 12–24 МЕСЯЦА

- ☐ Останемся в текущем классе ☐ Перейдём в следующий (средняя, большая)
☐ Расширим набор моделей ☐ Не определено

СЦЕНАРИЙ РАЗВИТИЯ (РАСШИРЯЕМ АССИСТЕНТА, НОВЫЕ ЗАДАЧИ, РОСТ КОМПАНИИ)

+ Дополнительно

ТРЕБОВАНИЯ К ОТКАЗОУСТОЙЧИВОСТИ

- ☐ Один сервер приемлемо ☐ Резервирование на уровне сервера ☐ Кластер с горячим резервом

ОГРАНИЧЕНИЯ ПО СТОЙКЕ (ПИТАНИЕ, ОХЛАЖДЕНИЕ, МЕСТО В ЦОДЕ)

ОЖИДАЕМЫЙ БЮДЖЕТНЫЙ ДИАПАЗОН (ПО ЖЕЛАНИЮ)

СРОКИ РЕАЛИЗАЦИИ

КОНТАКТ ДЛЯ ОБРАТНОЙ СВЯЗИ

Готовый бриф отправьте нам

Мы вернёмся с предварительной экспертной оценкой требований и возможной конфигурации

Серверы для бизнеса — Инком интер