

Чек-лист rollout Gemini 3.8 Flash

Памятка для команды, которая сравнивает Gemini 3.8 Flash с текущей моделью и готовит ограниченный запуск в продакшене.

Перед переключением

- Зафиксировать baseline по текущей модели: TTFT, полное время задачи, токены, tool calls, ошибки и цену успешного результата.
- Проверить адаптер запросов: убрать неподдерживаемые параметры и заменить thinking_budget на thinking_level.
- Прогнать контрактные тесты многоходового function calling, включая call_id и name в FunctionResponse.
- Определить rollback-маршрут на предыдущую модель.

Во время эксперимента

- Сравнить low, medium и high на одном обезличенном наборе задач.
- Разделить измерения для чата, RAG, извлечения структурированных данных, кода и агентных сценариев.
- Не разрешать shadow-агенту исполнять внешние операции.
- Считать успехом только результат, прошедший бизнес-валидацию.

Для агентных действий

- Сохранять thought signatures вместе с историей взаимодействия.
- Проверять аргументы инструментов до исполнения.
- Требовать approval для платежей, удаления данных, изменения прав и публикаций.
- Не использовать preview computer use для необратимых операций без изолированного контура.