Шпаргалка по топологии развёртывания AI

Latency-бюджет × Топология × Real-time vs batch — для каждой видео-AI-фичи.

1. Latency-бюджеты по UX-классам

Live-субтитры (partial)	300 мс
Live-субтитры (final)	1.0 с
Live-перевод	800 мс
AI-агент в звонке	200–500 мс
Размытие фона, на кадр	33 мс
In-call action items	3–10 с
Резюме встречи	30 с – 5 мин
Highlight-ролик	1–10 мин
Индекс поиска по архиву	часы / nightly

2. Слои топологии — round-trip + envelope стоимости минуты

СЛОЙ	ROUND-TRIP	СТОИМОСТЬ / МИН	ТИПОВАЯ ФИЧА
On-Device	0 мс	$0 / мин	Размытие, шумодав
Edge	5–30 мс	$0.0003–0.003	Субтитры, перевод, агент
In-Region Cloud	30–100 мс	$0.001–0.05	Резюме, главы, gen-video
Cross-Region Cloud	100–300 мс	$0.001–0.05	Только регуляторика

3. Пятишаговое дерево решений

1. Какой UX-latency-бюджет у этой фичи?
2. Max допустимый round-trip = бюджет − inference window.
3. Какой слой топологии подходит размеру модели?
4. Envelope стоимости минуты при ожидаемом concurrency?
5. Streaming partial или batched single answer?

4. Три повторяющихся ошибки

On-device-как-cloud: размытие в облачном GPU, когда on-device — $0.
Live-резюме: резюме mid-call, когда batched post-call даёт лучше.
Cross-region по умолчанию: us-east-1 для EU — фатально для real-time.