

OpenCohost - Gráficas de validación runtime

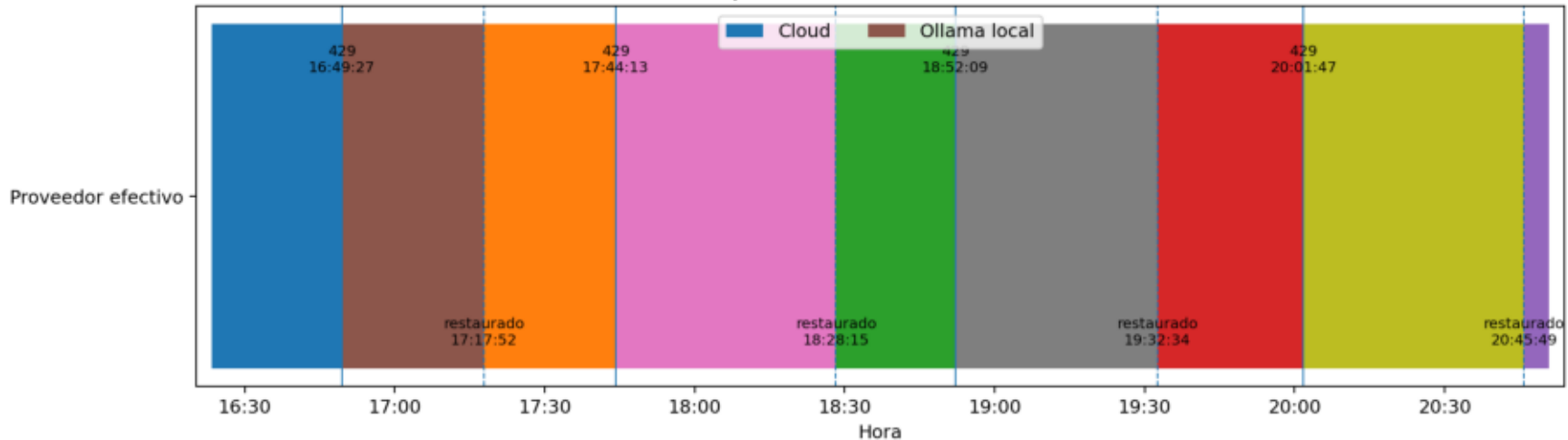
Sesión del 1 de agosto de 2026 · agenda 16:23:21-20:50:56

Lectura principal

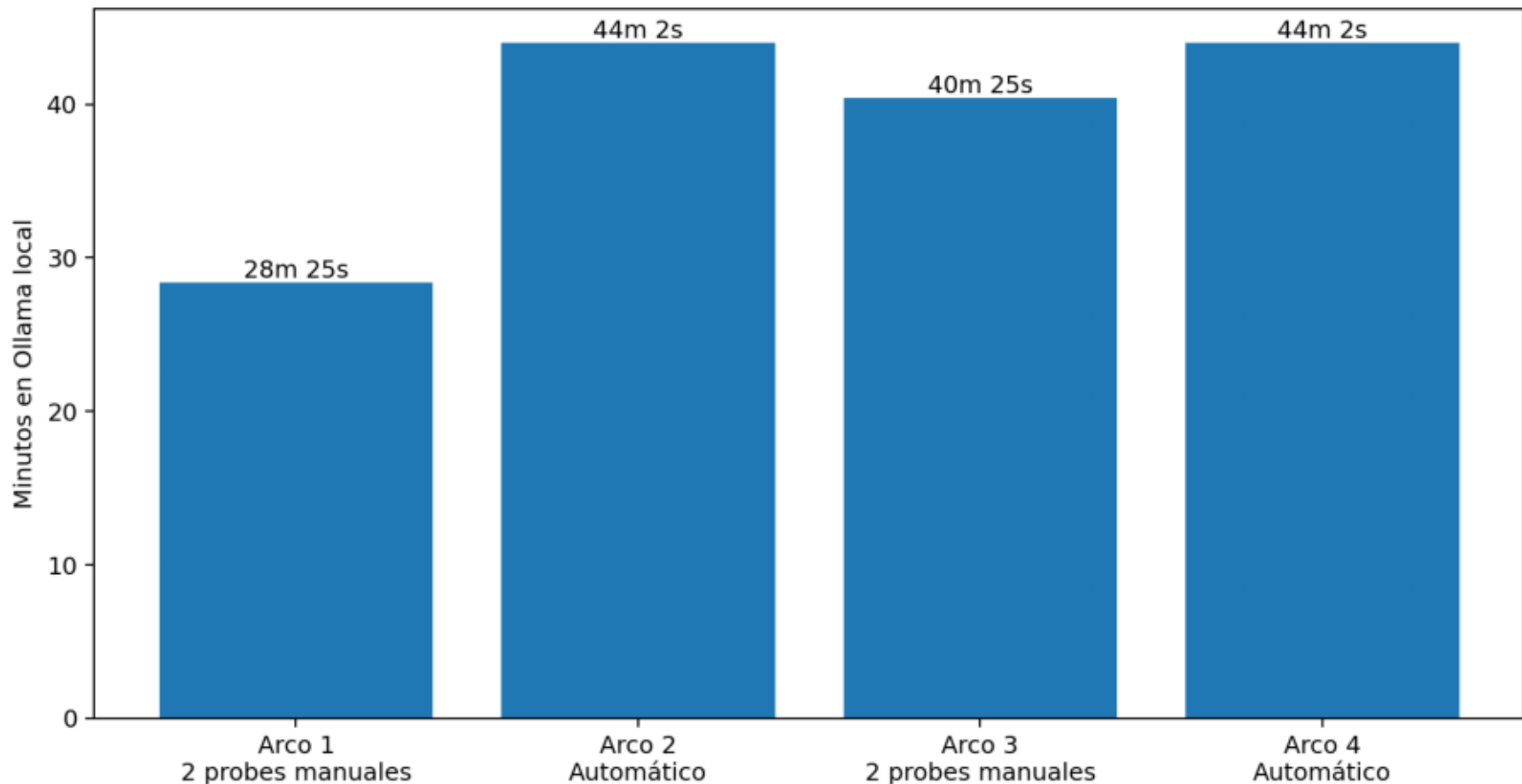
El cohorte principal de 429 reales, operó 58.6% de la ventana activa en fallback local, restauró cloud cuatro veces y mantuvo continuidad mediante prefetch. La medición directa ya separa espera en cola de generación, por lo que el costo conversacional es visible.

Ventana de agenda	4 h 27 min 35 s
Generaciones trazadas	285
Generaciones cloud	120
Generaciones locales	165
Tiempo en fallback local	2.62 h (58.6%)
Arcos de fallback	4
Restauraciones automáticas puras	2
Preguntas directas	10
Mediana direct total	30.2 s
Máximo direct total	75.9 s
Prefetch usados	194
Mediana de gap prefetch	46 ms
Pipelines TTS	245
Mediana primer fragmento	0.38 s
Mediana duración TTS	63.38 s
Evicciones de contexto	165
Pares expulsados	1279
Guardrails/rechazos	32

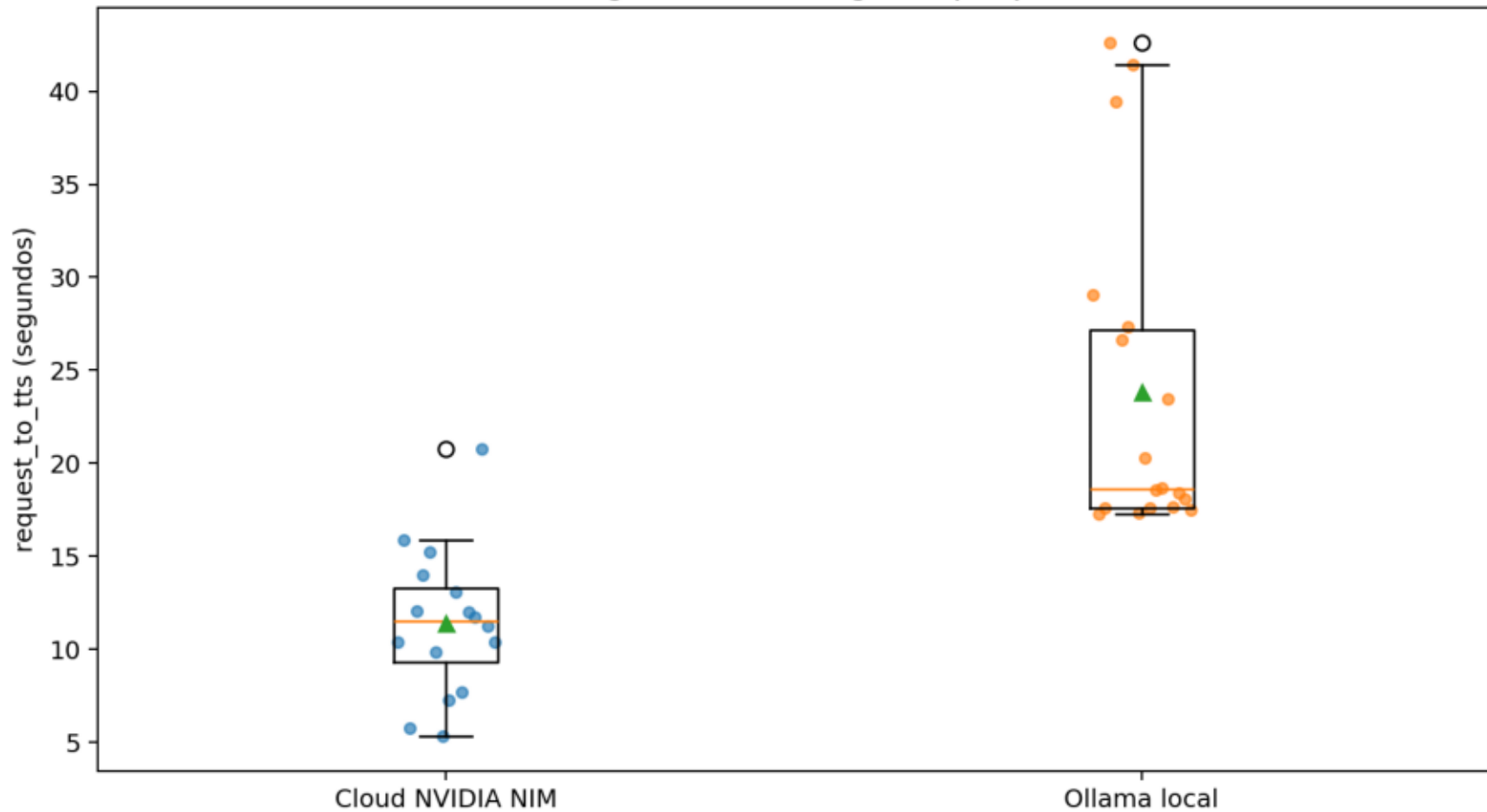
Timeline del proveedor efectivo - sesión 2026-08-01



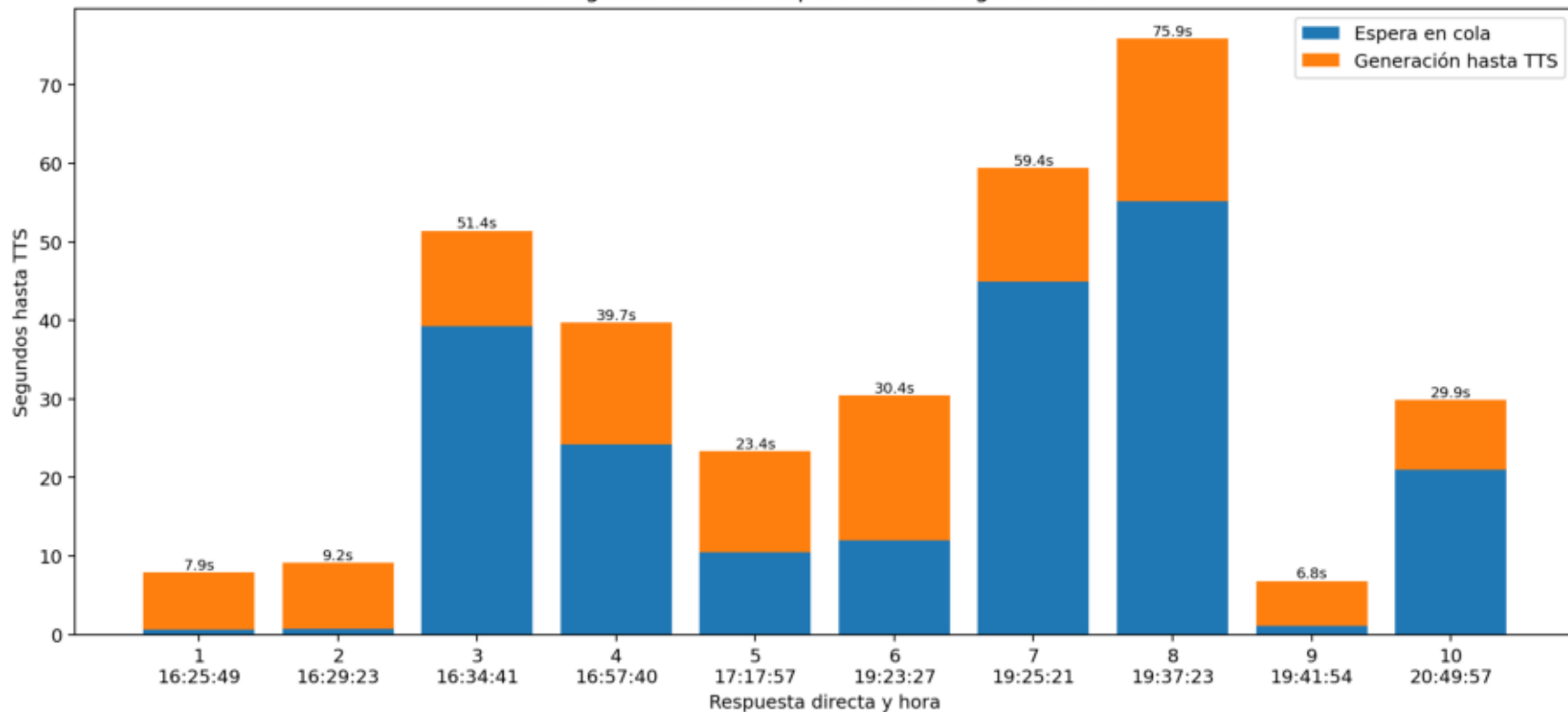
Duración de los cuatro arcos de fallback



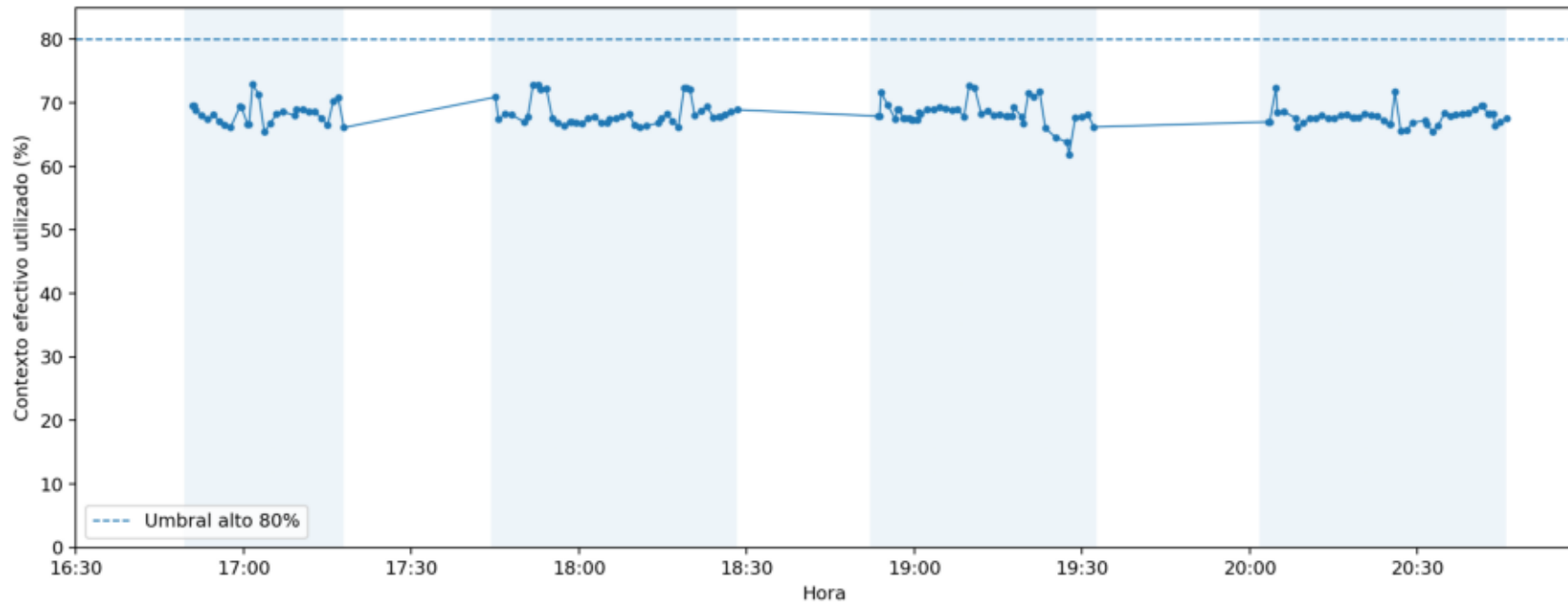
Latencia de generación de agenda por proveedor



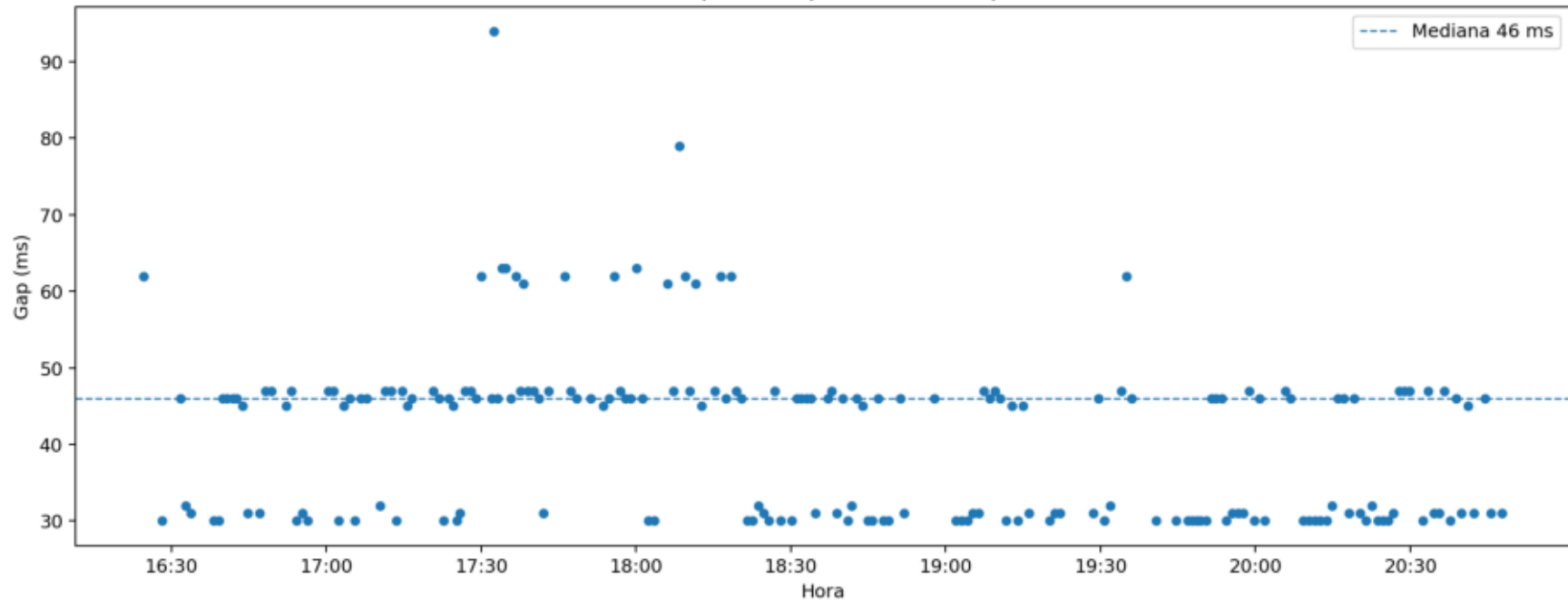
Preguntas directas: espera en cola vs generación



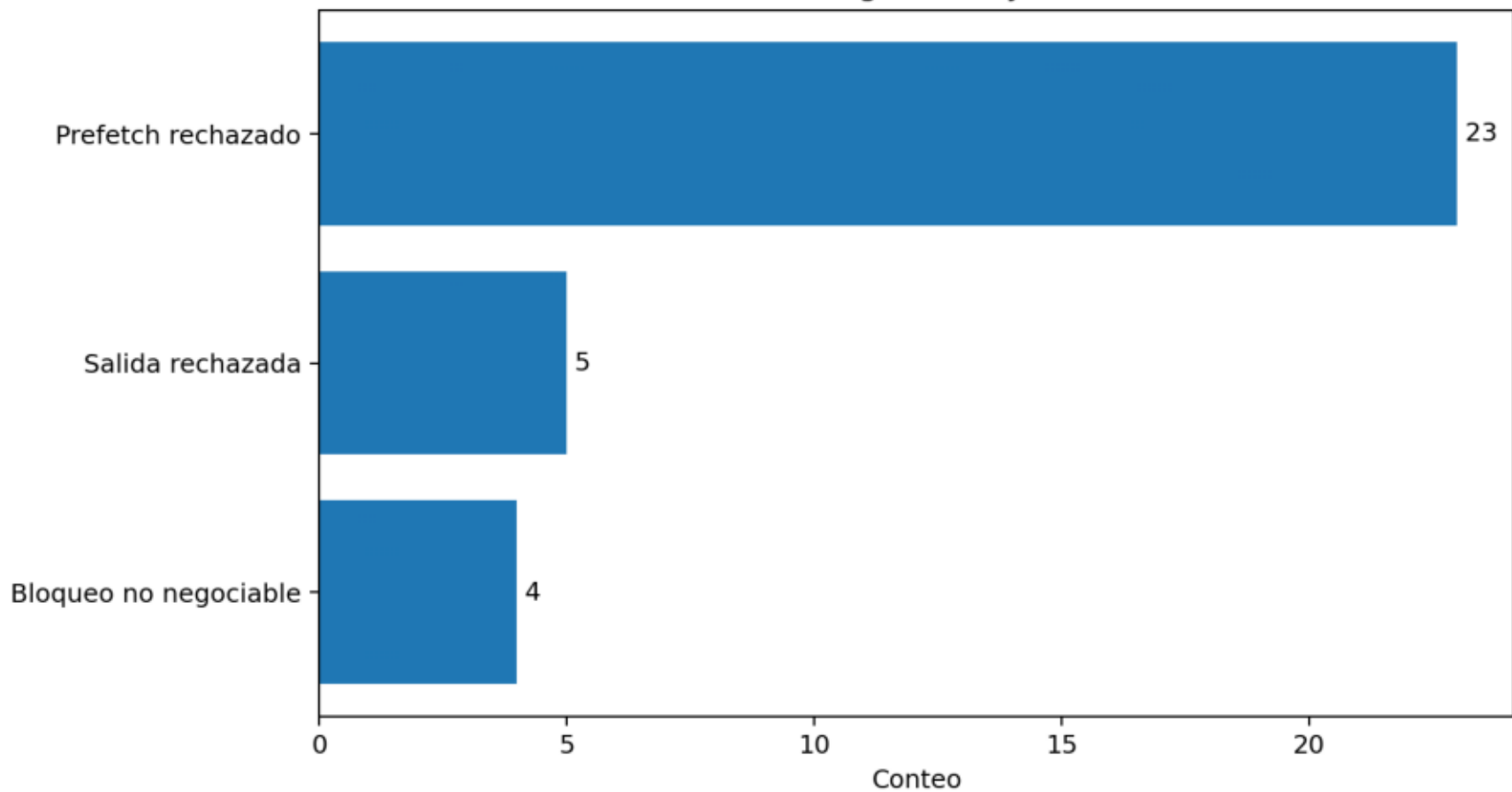
Utilización del contexto efectivo durante fallback local



Continuidad de prefetch: pausa entre bloques usados



Eventos de guardrail y rechazo



Tiempo hasta el primer fragmento TTS

