OpenCohost

Informe de trazabilidad del test de estres de agenda

Sesion del 30 de julio de 2026 | Interfaz PDF + log de runtime

DICTAMEN
La agenda sobrevivio casi tres horas de ejecucion, continuo despues de un 429 cloud y sostuvo audio de forma muy densa. La arquitectura de continuidad esta validada. La interaccion directa y la identidad visible del motor no lo estan: todas las preguntas se respondieron, pero ninguna de inmediato y la espera fue de aproximadamente 14 a 29 minutos.

2 h 50 min 25 s Ventana activa	**139** Salidas TTS	**2 h 27 min 32 s** Audio acumulado
86.6% Ocupacion TTS	**169** Completions	**4/4 respondidas** Preguntas

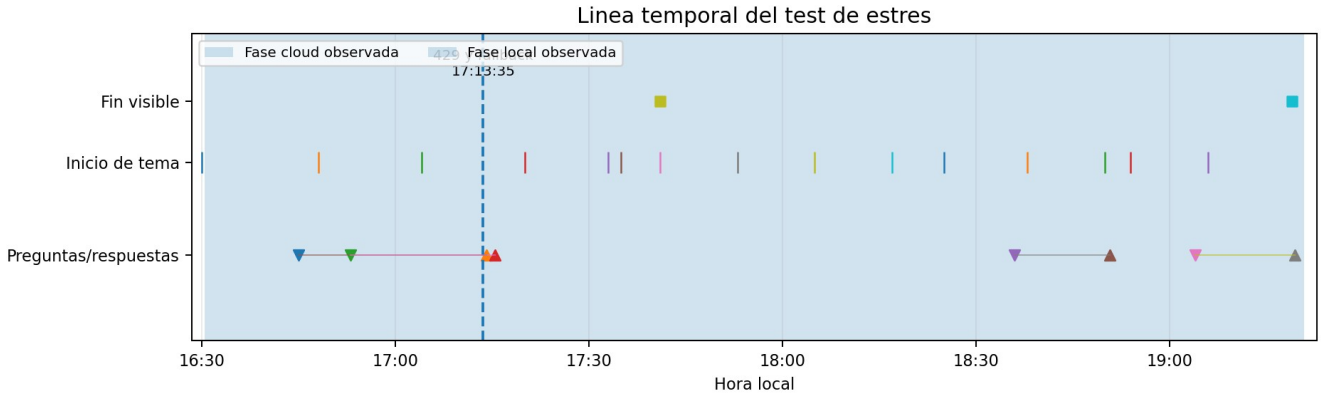

Figura 1. Fases, temas, preguntas, respuestas y marcadores de cierre.

1. Alcance y contrato de evidencia

Este informe funciona como auditoria de trazabilidad. No investiga la causa raiz en el repositorio, no corrige el bug de numeracion y no califica la personalidad o la factualidad general de Kira. Clasifica cada conclusion segun lo que realmente soportan los dos artefactos.

Clase	Definicion
Demostrado	Evento visible en PDF o log con hora y secuencia reproducible.
Demostrado con limitaciones	La evidencia existe, pero no permite concluir proveedor real, audio exacto o fidelidad interna.
Observacion del operador	Dato escuchado o medido durante runtime, sin telemetria o grabacion en los artefactos.
Hipotesis	Patron compatible con los datos que requiere verificacion en codigo o una nueva prueba.

2. Resultado de arquitectura

El log cubre 2 h 53 min 59 s. La agenda entra a la cola a las 16:30:26 y el ultimo pipeline de audio termina a las 19:20:51: 2 h 50 min 25 s de actividad entre ambos puntos.

La fase cloud termina con HTTP 429 a las 17:13:35. El sistema ejecuta el fallback automatico, prepara `gemma4:e4b` local en 19.36 s, reproduce un aviso corto de dos fragmentos y despues procesa dos preguntas directas pendientes antes de reanudar la agenda.

Fase	Ventana	Completions	TTS	Rasgo
Cloud	16:30:26-17:13:35	29	29	Inferencia y TTS mayormente seriales; loaded=unknown.
Local	17:13:35-19:20:51	140	110	Prefetch concurrente, contexto efectivo 4096 y modelo loaded=gemma4:e4b.

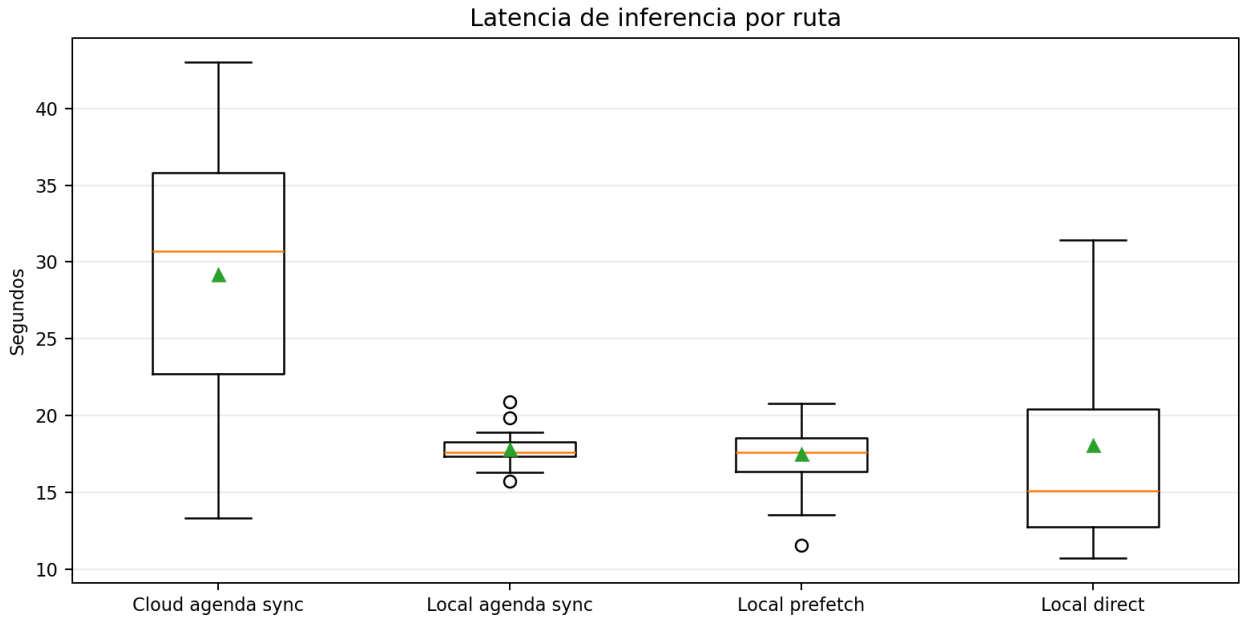

Figura 2. Distribucion de latencias por ruta de inferencia.

3. Identidad del motor y trazabilidad

Afirmacion	Clasificacion	Evidencia / limite
El modelo local fue gemma4:e4b	Demostrado	Inicio del log, preparacion local y MODEL_TRACE cargado.
El proveedor cloud fue GLM 5.2	Observacion del operador	El log usa el alias gemma4:e4b y registra cloud_llm_usage, pero no revela el backend remoto efectivo.
La interfaz mostro un estado obsoleto	Observacion del operador	El PDF impreso no preserva de forma fiable el indicador superior durante cada fase.
Existio una brecha de trazabilidad cloud	Demostrado	29 MODEL_MISMATCH_WARNING con desired=active=generation y loaded=unknown.

Conclusion: el runtime distingue la ruta suficientemente para recuperarse, pero no ofrece una identidad de proveedor/modelo que el usuario pueda auditar sin leer logs.

4. Preguntas directas y cola

Las cuatro preguntas visibles se emparejan semanticamente con cuatro respuestas directas posteriores y conservan el orden. Ninguna modifica la respuesta de agenda inmediatamente posterior. La conducta observada es "aceptar, retener y responder tarde", no conversacion concurrente.

#	Entrada	Respuesta	Espera	Inferencia	Pregunta
1	04:45 p. m.	05:14 p. m.	28:08-29:07 min	31.42 s	y como haces para decidir quien eres, que es lo que haces y por que lo haces o...
2	04:53 p. m.	05:15 p. m.	21:26-22:25 min	10.75 s	espera acabas de darme la clasificacion, por meses o años organizados, imagina...
3	06:36 p. m.	06:50 p. m.	13:47-14:46 min	16.78 s	Cual es el resumen de toda la sesion o tu punto de vista mas acertado?
4	07:04 p. m.	07:19 p. m.	14:27-15:26 min	13.47 s	y donde esta la gracia de todo esto?

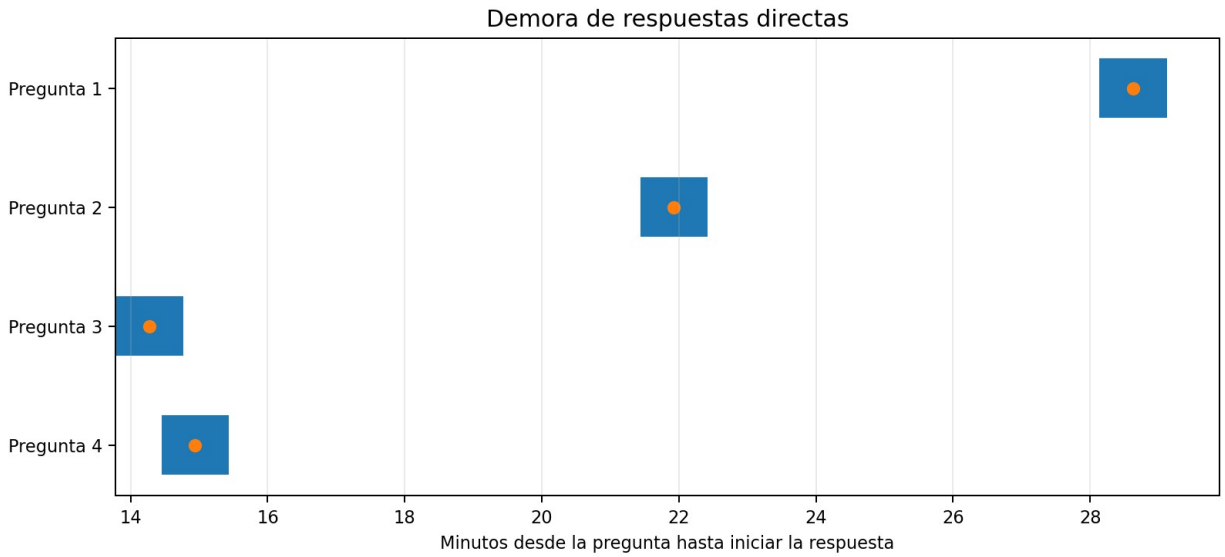

Figura 3. Rango de demora: el PDF solo muestra el minuto de entrada.

Clasificacion: ALTA. Para un cohost de streaming, responder finalmente no equivale a reaccionar. El sistema necesita acuse de recibo, politica de interrupcion o estado de cola visible.

5. TTS y continuidad audible

Se completaron 139 pipelines, 1625 fragmentos y 2 h 27 min 32 s acumulados. Esto equivale al 86.6% de la ventana activa, aunque parte de la sintesis y la inferencia se solapan.

Metrica	Valor
Primer fragmento medio	0.39 s
Primer fragmento mediano	0.37 s
Primer fragmento p95	0.54 s
Duracion TTS mediana	65.98 s
Duracion TTS maxima	112.45 s
Salidas de contenido	138

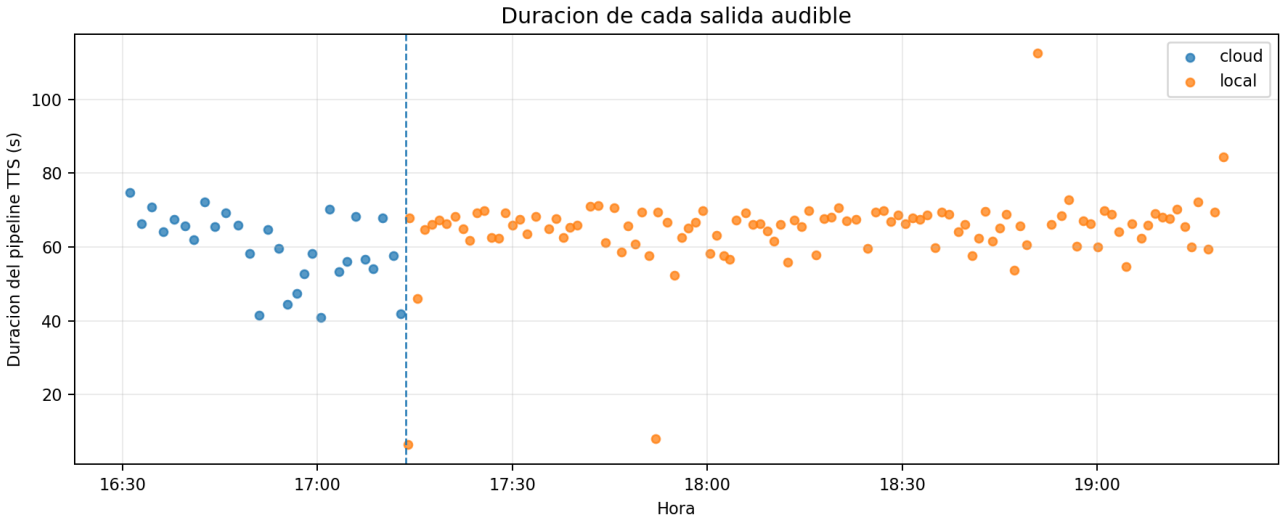

Figura 4. Duracion de cada pipeline de salida.

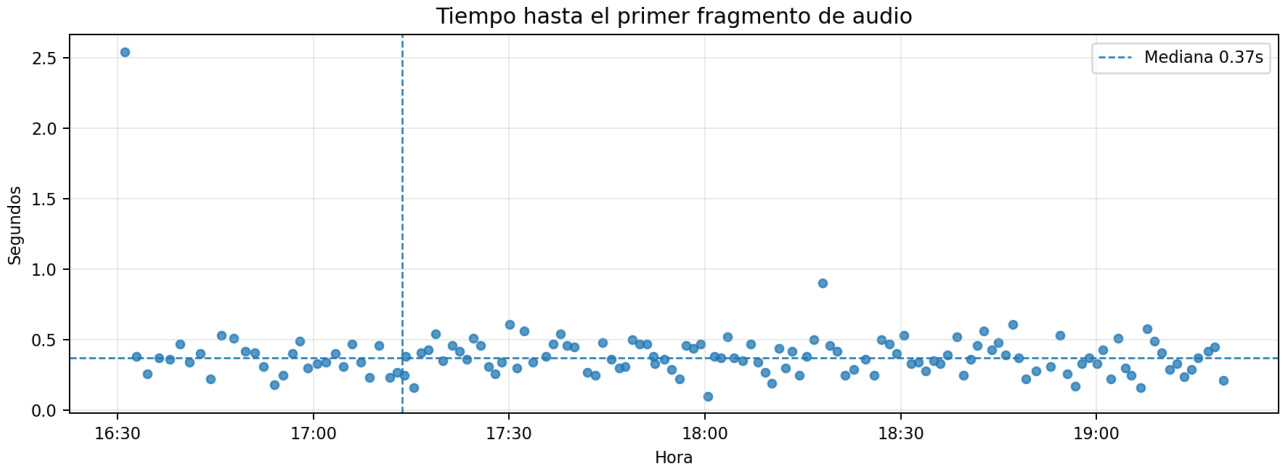

Figura 5. Inicio de audio estable, salvo el primer arranque cloud de 2.54 s.

6. Unicode y verbalizacion

El PDF contiene la secuencia CJK `承认` dos veces en una respuesta del primer tema. Son cuatro codepoints no latinos en total. El log confirma que esa respuesta entro al pipeline TTS, pero no registra la cadena normalizada ni el audio final.

Hecho	Clasificacion	Implicacion
Texto `承认` visible	Demostrado	La contaminacion se origina antes o durante la generacion textual.
TTS dijo "caracter chino"	Observacion del operador	No verificable sin audio o log de texto normalizado.
Solucion exacta	Fuera de alcance	La siguiente prueba debe registrar original, saneado y cadena enviada a TTS.

Clasificacion: MEDIA-ALTA. No rompe la continuidad, pero degrada de forma muy visible la naturalidad del cohost.

7. Prefetch y guardrails

El modelo produjo 169 completions registradas. 118 fueron prefetch; 31 se rechazaron y 87 se usaron. La aritmetica cierra: 51 completions sincronicas + 87 prefetch aceptados = 138 salidas de contenido. El pipeline TTS numero 139 es el aviso de carga local de dos fragmentos.

Etapa	Motivo	Eventos
output	exact_repetition	1
output	repeats_recent_line	1
output	skeleton_repetition	4
output_fallback	artificial_closure	1
prefetch	exact_repetition	1
prefetch	repeats_recent_line	4
prefetch	skeleton_repetition	26

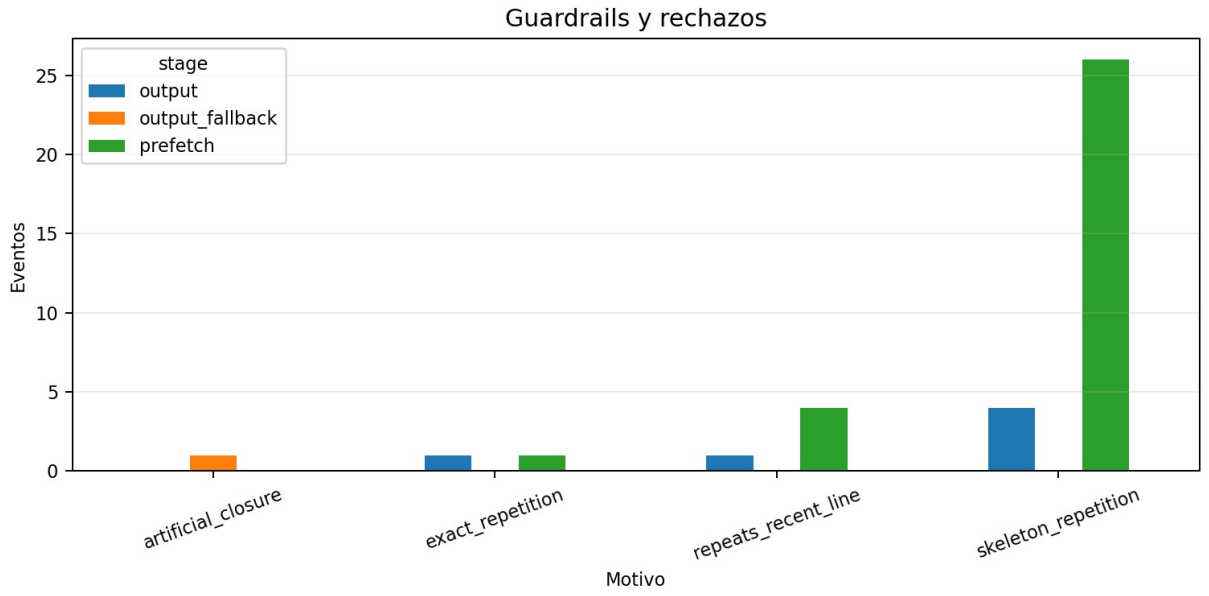

Figura 6. Rechazos por etapa y motivo.

De 87 borradores usados, 86 tuvieron un gap de 100 ms o menos. La mediana fue 46 ms. Existe un outlier de 9.06 s en una frontera `kira-agenda-stop`.

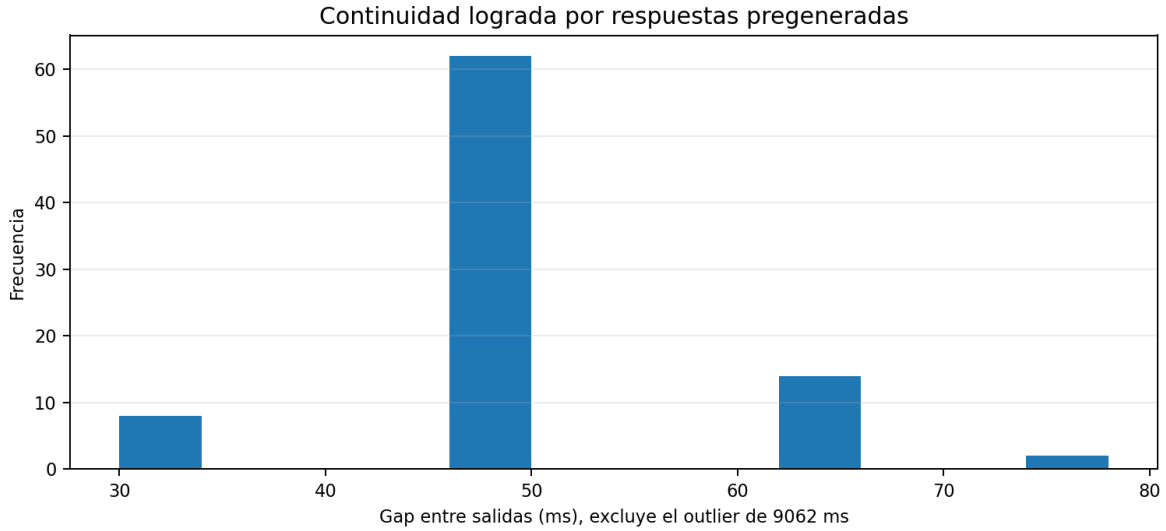

Figura 7. Distribucion de gaps normales de prefetch.

8. Contexto local

El modelo declara native_ctx=131072, pero el presupuesto efectivo utilizado es 4096. En 146 mediciones, la utilizacion media fue 68.6% y el rango 60.7-73.0%. El gate expulso 1126 pares acumulados; 124 de 146 eventos expulsaron exactamente ocho pares.

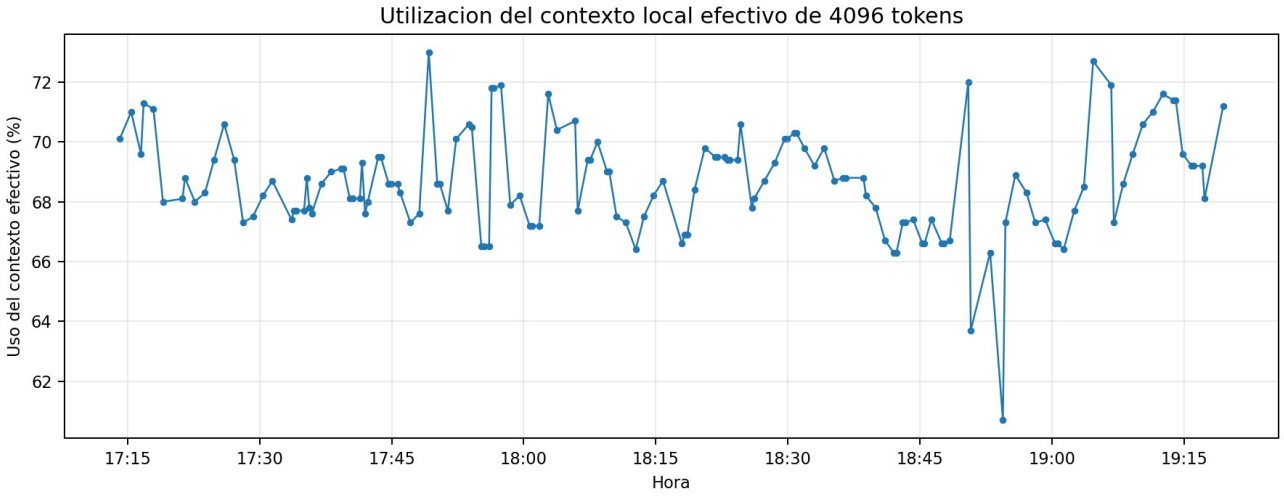

Figura 8. Uso del contexto efectivo durante la fase local.

Clasificacion: RIESGO OBSERVADO, NO FALLO DEMOSTRADO. La sesion conserva continuidad superficial, pero los artefactos no permiten saber que informacion concreta fue expulsada ni si afecto respuestas futuras.

9. Cobertura y contenido de los 15 temas

La numeracion par no se interpreta aqui como causa o bug. Se mide lo observable: respuestas de Kira, guardrails visibles, duracion del bloque y volumen de palabras.

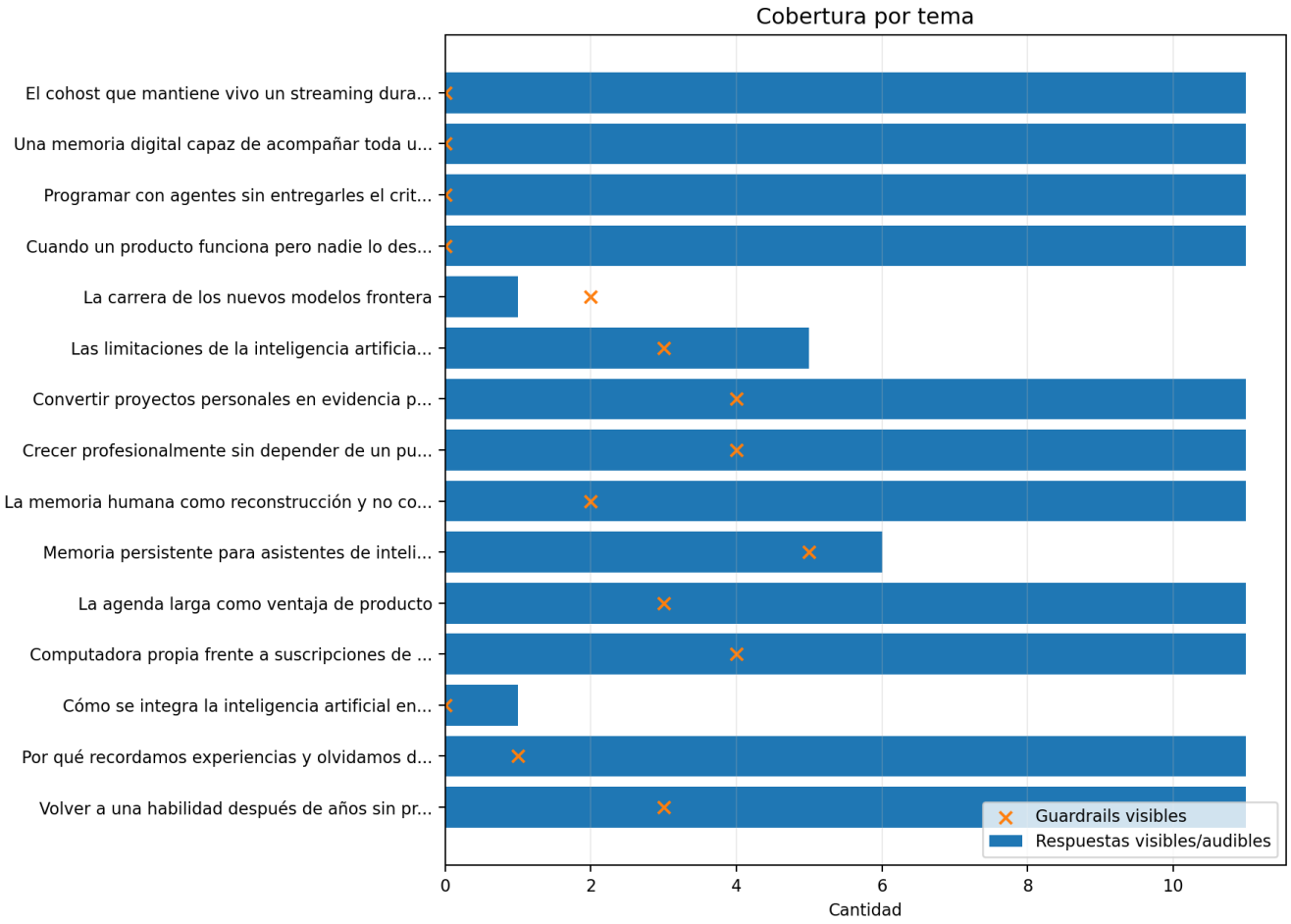

Figura 9. Cobertura observable por tema.

Inicio	Min	Resp.	Guard.	Cobertura	Tema
04:30 p. m.	18	11	0	amplia	El cohost que mantiene vivo un streaming durante una ausencia
04:48 p. m.	16	11	0	amplia	Una memoria digital capaz de acompañar toda una vida
05:04 p. m.	16	11	0	amplia	Programar con agentes sin entregarles el criterio
05:20 p. m.	13	11	0	amplia	Cuando un producto funciona pero nadie lo descubre
05:33 p. m.	2	1	2	minima	La carrera de los nuevos modelos frontera
05:35 p. m.	6	5	3	parcial	Las limitaciones de la inteligencia artificial para perfiles junior
05:41 p. m.	12	11	4	amplia	Convertir proyectos personales en evidencia profesional
05:53 p. m.	12	11	4	amplia	Crecer profesionalmente sin depender de un puesto formal
06:05 p. m.	12	11	2	amplia	La memoria humana como reconstrucción y no como archivo
06:17 p. m.	8	6	5	parcial	Memoria persistente para asistentes de inteligencia artificial
06:25 p. m.	13	11	3	amplia	La agenda larga como ventaja de producto
06:38 p. m.	12	11	4	amplia	Computadora propia frente a suscripciones de inteligencia artificial

06:50 p. m.	4	1	0	minima	Cómo se integra la inteligencia artificial en empleos tecnológicos de China y Japón
06:54 p. m.	12	11	1	amplia	Por qué recordamos experiencias y olvidamos datos importantes
07:06 p. m.	13	11	3	amplia	Volver a una habilidad después de años sin practicarla

El cohost que mantiene vivo un streaming durante una ausencia

Cobertura amplia: 11 respuestas, 0 guardrails visibles, 2115 palabras aproximadas. Cobertura de ausencias, silencios, autoridad frente al chat, diferencia entre acompañar y reemplazar al streamer, y transicion cuando el streamer regresa.

Una memoria digital capaz de acompañar toda una vida

Cobertura amplia: 11 respuestas, 0 guardrails visibles, 1673 palabras aproximadas. Memoria como sistemas coordinados, curaduria frente a archivo, irreversibilidad del descarte y conflicto entre versiones parciales de una vida digital.

Programar con agentes sin entregarles el criterio

Cobertura amplia: 11 respuestas, 0 guardrails visibles, 1894 palabras aproximadas. Delegacion tecnica sin contexto historico, falsa seguridad de resultados pulcros, responsabilidad humana y necesidad de puntos de control.

Cuando un producto funciona pero nadie lo descubre

Cobertura amplia: 11 respuestas, 0 guardrails visibles, 2008 palabras aproximadas. Brecha entre utilidad y adopcion, habitos, confianza, pertenencia, incentivos y resistencia estructural de intermediarios.

La carrera de los nuevos modelos frontera

Cobertura minima: 1 respuestas, 2 guardrails visibles, 187 palabras aproximadas. Solo alcanzo a plantear que el salto real depende de integracion con sistemas viejos y conectividad, no solo de capacidad de demo.

Las limitaciones de la inteligencia artificial para perfiles junior

Cobertura parcial: 5 respuestas, 3 guardrails visibles, 890 palabras aproximadas. Riesgo de atrofiar criterio, necesidad de formular problemas, desconfiar del resultado y saber cambiar de estrategia cuando el sistema falla.

Convertir proyectos personales en evidencia profesional

Cobertura amplia: 11 respuestas, 4 guardrails visibles, 1786 palabras aproximadas. Narrativa de impacto, metricas, cicatrices del proyecto, decisiones de deuda tecnica y traduccion de complejidad para terceros.

Crecer profesionalmente sin depender de un puesto formal

Cobertura amplia: 11 respuestas, 4 guardrails visibles, 1900 palabras aproximadas. Comunicacion, colaboracion, visibilidad, empaquetado del valor y diferencia entre brillantez tecnica y viabilidad profesional.

La memoria humana como reconstrucción y no como archivo

Cobertura amplia: 11 respuestas, 2 guardrails visibles, 1965 palabras aproximadas. Recuerdo como relato editable, coherencia del yo, sesgos de reconstruccion y proyeccion de narrativas sobre otras personas.

Memoria persistente para asistentes de inteligencia artificial

Cobertura parcial: 6 respuestas, 5 guardrails visibles, 1095 palabras aproximadas. Privacidad, poder de recordar todo, contexto emocional, derecho a corregir u olvidar y riesgo de tratar recuerdos como registros objetivos.

La agenda larga como ventaja de producto

Cobertura amplia: 11 respuestas, 3 guardrails visibles, 2064 palabras aproximadas. Profundidad frente a acumulacion de funciones, ritual de uso, lock-in metodologico, estatus y costo hundido de aprender un flujo.

Computadora propia frente a suscripciones de inteligencia artificial

Cobertura amplia: 11 respuestas, 4 guardrails visibles, 1928 palabras aproximadas. Control, soberania del dato, dependencia estructural, mantenimiento, propiedad intelectual y personalizacion profunda.

Cómo se integra la inteligencia artificial en empleos tecnológicos de China y Japón

Cobertura minima: 1 respuestas, 0 guardrails visibles, 167 palabras aproximadas. Solo alcanzo a abrir la diferencia cultural y empresarial de adopcion; no desarrollo ejemplos laborales concretos.

Por qué recordamos experiencias y olvidamos datos importantes

Cobertura amplia: 11 respuestas, 1 guardrails visibles, 1881 palabras aproximadas. Emocion, narrativa y atencion como filtros de memoria; sobrecarga informativa y dificultad de consolidar datos aislados.

Volver a una habilidad después de años sin practicarla

Cobertura amplia: 11 respuestas, 3 guardrails visibles, 2033 palabras aproximadas. Descalibracion frente a perdida, practica deliberada, errores como datos, rutina, friccion y recuperacion de soltura.

10. Registro de hallazgos

ID	Severidad	Estado	Hallazgo
F-01	Alta	Demostrado	Preguntas directas atendidas con 14-29 min de demora; 0/4 inmediatas.
F-02	Alta	Demostrado + operador	La ruta cambia a local, pero el estado visible no comunica proveedor/modelo efectivo de forma auditable.
F-03	Media-alta	Demostrado + operador	Texto CJK visible y verbalizacion anomala escuchada; falta grabacion.
F-04	Media	Demostrado	El PDF marca agenda finalizada a las 17:41 y la agenda continua inmediatamente.
F-05	Media	Demostrado	Dos temas quedan con cobertura minima: modelos frontera y China/Japon.
F-06	Riesgo	Demostrado	Gate de contexto expulsa pares en casi cada completion local.
P-01	Positivo	Demostrado	Fallback 429 a local sin detener la sesion.
P-02	Positivo	Demostrado	Prefetch logra continuidad casi sin gap en 86/87 usos.
P-03	Positivo	Demostrado	TTS completa 139 salidas y libera la experiencia del tiempo de inferencia local.

11. Que demuestra y que no demuestra

Si demuestra	No demuestra
La agenda puede sostener casi tres horas de runtime con TTS.	Que cada respuesta sea correcta, agradable o editorialmente buena.
El fallback automatico cloud-local funciona ante 429.	El proveedor remoto real detras del alias del log.
Las preguntas no se perdieron en esta muestra.	Que la cola garantice entrega en cualquier escenario.
El prefetch reduce casi a cero la separacion entre salidas.	Que no exista repeticion semantica no detectada.
El contexto se recorta de forma agresiva pero estable.	Que la informacion expulsada sea irrelevante.
El texto contenia caracteres CJK.	La frase audible exacta sin grabacion.

12. Especificacion minima de la siguiente validacion

1. Asignar un ID unico a cada entrada del operador, completion, mensaje renderizado y pipeline TTS.
2. Registrar `queued_at`, `generation_started_at`, `tts_started_at` y `completed_at` para preguntas directas.
3. Mostrar un acuse de recibo visible cuando una pregunta queda en cola y un limite de espera contractual.
4. Persistir proveedor, modelo solicitado, modelo efectivo, ruta cloud/local y razon del fallback por mensaje.
5. Registrar texto original, texto saneado y texto enviado al TTS, con conteo Unicode por escritura.
6. Muestrear RAM, VRAM y proceso de descarga cada 5-10 segundos durante el test.
7. Conservar el audio o al menos una pista de baja calidad para auditar pronunciacion y silencios.
8. Repetir una prueba con preguntas en cuatro momentos: durante TTS, durante prefetch, al cierre de tema y al cierre de agenda.

13. Anexo de evidencia

Ancla	Contenido
Log L1-L5	Inicializacion, modelo guardado, perfil y advertencia de coherencia.
Log L9-L17	Primer ciclo cloud, 39.97 s de inferencia y 74.79 s de TTS.
Log L272-L306	429, fallback automatico, carga local de 19.36 s y aviso TTS.
Log L307-L322	Dos respuestas directas procesadas tras el fallback.
Log L338-L351	Uso de respuesta pregenerada y gaps de 47-62 ms.
Log L1421-L1430	Salida de agenda rechazada y procesamiento de la tercera pregunta directa.
Log L1708-L1714	Cuarta pregunta directa y audio hasta 19:20:51.
PDF pp. 5, 8, 50, 63	Preguntas directas del operador.
PDF pp. 14-15, 58, 69-70	Respuestas directas tardias.
PDF p. 5	Secuencia CJK `承认`.
PDF p. 26	Marcador de agenda finalizada a las 17:41 seguido de continuidad.

Nota metodologica: los horarios de entrada del PDF estan redondeados al minuto. Los rangos de demora restan hasta 59 segundos al maximo calculado. Las metricas de inferencia y TTS provienen del log con precision de segundos o milisegundos.