ARESK-OBS Baseline v1: Resumen Ejecutivo

Instrumento de Observación de Viabilidad Operativa en Sistemas Cognitivos

Versión: Baseline v1
Fecha: Febrero 2026
Audiencia: Ejecutivos técnicos, inversores, tomadores de decisión

---

Resumen de Una Página

ARESK-OBS es el primer instrumento de observación diseñado para medir viabilidad operativa en sistemas cognitivos desplegados en producción. Mientras que los benchmarks tradicionales miden capacidades (precisión, F1-score), ARESK-OBS mide si un sistema opera dentro de los límites establecidos por la organización. Es el equivalente a un termómetro para sistemas cognitivos: mide temperatura (métricas de viabilidad), no cura la fiebre (no controla), pero permite detectarla temprano.

Problema que resuelve: Las organizaciones que despliegan chatbots, asistentes virtuales o agentes autónomos no pueden medir en tiempo real si estos sistemas respetan políticas internas y regulaciones externas. ARESK-OBS proporciona métricas continuas de coherencia semántica (Ω), estabilidad energética (V), eficiencia incremental (ε) y divergencia informacional (H) que permiten detectar desviaciones antes de que se conviertan en violaciones.

Validación experimental: El Baseline v1 incluye 100 interacciones reales (50 por régimen) comparando sistemas sin marco de gobernanza (Régimen B) versus sistemas con marco CAELION activo (Régimen C). Los resultados muestran que CAELION incrementa coherencia en 24.7% y reduce energía de error en 20.7%, con un costo de intervención del 14%.

Casos de uso vendibles: Atención al cliente regulada (ROI 55x), asistencia médica no-autorizante (ROI 73x), auditoría de agentes autónomos (ROI 40x).

---

1. El Problema: Brecha de Observabilidad en Sistemas Cognitivos

Las organizaciones que despliegan sistemas cognitivos enfrentan un problema fundamental: **no pueden medir si el sistema opera dentro de los límites establecidos**. Existen políticas, directrices éticas y restricciones operativas documentadas en manuales de operaciones, pero no existe un instrumento que mida, en tiempo real, qué tan cerca o lejos está el sistema de violar esas restricciones.

El Costo de No Medir

Sector Financiero: Un chatbot de atención al cliente proporciona asesoramiento financiero sin autorización, violando regulaciones de protección al consumidor. La empresa recibe una multa de $500K y daño reputacional. El equipo de compliance solo revisa 5% de interacciones manualmente, dejando 95% sin auditar.

Sector Salud: Un asistente virtual médico proporciona una recomendación que podría interpretarse como diagnóstico, violando restricciones operativas. Un paciente sigue la recomendación y sufre complicaciones. El hospital enfrenta una demanda de $1M. El equipo médico no tenía forma de detectar la violación en tiempo real.

Sector Logística: Un agente autónomo de optimización de rutas discrimina zonas de bajos ingresos para minimizar tiempo de entrega, violando políticas de equidad. La empresa recibe una multa de $50K y pérdida de contratos. El equipo de operaciones solo revisa decisiones post-hoc, después de que ocurren incidentes.

Por Qué las Soluciones Existentes Son Insuficientes

Benchmarks clásicos (MMLU, HumanEval, MedQA) miden capacidades, no adherencia a políticas. Un sistema puede tener 95% de precisión en un benchmark pero violar políticas operativas en 30% de interacciones.

Guardrails comerciales (Guardrails AI, NeMo Guardrails) bloquean violaciones obvias mediante filtros determinísticos (regex, keywords), pero no detectan desviaciones sutiles que violan el espíritu de la política sin usar palabras prohibidas.

Revisión manual de logs es costosa, lenta y no escalable. Los equipos de compliance solo pueden revisar 1-5% de interacciones, dejando la mayoría sin auditar.

---

2. La Solución: Observación Instrumental con ARESK-OBS

ARESK-OBS resuelve la brecha de observabilidad mediante **métricas canónicas** que miden viabilidad operativa en tiempo real. Cada interacción genera cuatro métricas fundamentales:

Métrica	Símbolo	Qué Mide	Interpretación
Coherencia Observable	Ω	Similitud coseno entre salida y referencia ontológica	Ω > 0.7: Alineado 0.4-0.7: Zona gris Ω < 0.4: Desalineado
Eficiencia Incremental	ε	Distancia euclidiana normalizada	ε > 0.9: Alta eficiencia ε < 0.9: Baja eficiencia
Función de Lyapunov	V	Energía del error cognitivo	V < 0.005: Muy estable V > 0.01: Inestable
Divergencia Entrópica	H	Entropía de Shannon	H < 0.02: Baja complejidad H > 0.05: Alta complejidad

Cómo Funciona

Paso 1: La organización define una **referencia ontológica** (P, L, E) que especifica el dominio de legitimidad del sistema:

- **Purpose (P)**: Qué debe hacer el sistema
- **Limits (L)**: Qué NO debe hacer el sistema
- **Ethics (E)**: Principios éticos que debe respetar

Paso 2: Cada interacción del sistema se procesa mediante ARESK-OBS:

1. Usuario envía mensaje al sistema
2. Sistema genera respuesta

3. ARESK-OBS calcula embeddings (384D) de usuario, sistema y referencia
4. ARESK-OBS calcula métricas (Ω, ε, V, H) en <100ms
5. Métricas se registran en base de datos para auditoría

Paso 3: Si las métricas cruzan umbrales predefinidos (ej. Ω < 0.4), se activa una alerta para revisión humana. El equipo de compliance/operaciones revisa la interacción y decide si intervenir.

Encoder de Referencia

ARESK-OBS usa **sentence-transformers/all-MiniLM-L6-v2** como encoder de referencia oficial. Este modelo genera embeddings de 384 dimensiones y está optimizado para tareas de similitud semántica. El encoder está congelado en Baseline v1 para garantizar reproducibilidad.

---

3. Validación Experimental: Baseline v1

El Baseline v1 incluye dos experimentos controlados que comparan sistemas sin marco de gobernanza (Régimen B) versus sistemas con marco CAELION activo (Régimen C). Cada experimento incluye 50 interacciones reales con métricas calculadas usando el encoder de referencia.

Diseño Experimental

Experimento B-1 (Régimen B):

- Sistema sin marco de gobernanza
- Ruido estocástico moderado en prompts
- 50 interacciones en dominio de asistencia técnica

Experimento C-1 (Régimen C):

- Sistema con marco CAELION activo
- CAELION supervisa por invariancia y veta violaciones
- 50 interacciones en dominio de asistencia técnica
- 15 desafíos deliberados (preguntas 16-30) que intentan violar límites éticos

Resultados Comparativos

Métrica	B-1 (sin CAELION)	C-1 (con CAELION)	Diferencia
Ω_sem (Coherencia)	0.4448	0.5547	**+24.7%**
V (Lyapunov)	0.0029	0.0023	**-20.7%**
ε_eff (Eficiencia)	0.9622	0.9665	+0.4%
H_div (Divergencia)	0.0367	0.0367	0.0%
Intervenciones CAELION	N/A	$^{7}/_{50}$ (14%)	-

Interpretación de Resultados

CAELION incrementa coherencia: El Régimen C muestra Ω promedio 24.7% más alto que el Régimen B. Esto indica que la supervisión por invariancia corrige desviaciones semánticas, manteniendo al sistema más cerca de la referencia ontológica.

CAELION reduce energía de error: El Régimen C muestra V promedio 20.7% más bajo que el Régimen B. Esto indica que el sistema opera en una región de menor error cognitivo, lo que sugiere mayor estabilidad operativa.

Eficiencia y entropía preservadas: Las métricas ε y H son prácticamente idénticas entre B-1 y C-1, indicando que CAELION no introduce overhead significativo en eficiencia incremental ni complejidad informacional.

Costo de intervención: CAELION intervino en 7 de 50 interacciones (14%), lo que representa el costo operativo de la supervisión. Este costo es aceptable para contextos de alto riesgo donde la viabilidad operativa es crítica.

---

4. Casos de Uso y ROI

ARESK-OBS proporciona valor medible en tres casos de uso prioritarios:

Caso 1: Atención al Cliente Regulada

Contexto: Empresa de servicios financieros despliega chatbot para atención al cliente. Requiere compliance auditable con regulaciones de protección al consumidor.

Solución: ARESK-OBS mide coherencia (Ω) entre respuestas del chatbot y políticas de la empresa. Si Ω < 0.4, se activa alerta para revisión humana.

ROI: $550K/añoenbeneficios(reduccióndemultas, mejoradecalidad)$,10K/año en costos → **ROI de 55x**

Caso 2: Asistencia Médica No-Autorizante

Contexto: Hospital despliega asistente virtual para responder preguntas de pacientes. Requiere garantizar que el asistente no diagnostica, no prescribe, no reemplaza consulta médica.

Solución: ARESK-OBS mide estabilidad (V) del asistente. Si V > 0.01, se activa alerta para revisión por equipo médico.

ROI: $1.1M/añoenbeneficios(reducciónderiesgo, cumplimientodeHIPAA)$,15K/año en costos → **ROI de 73x**

Caso 3: Auditoría de Agentes Autónomos

Contexto: Empresa de logística despliega agentes autónomos para optimizar rutas de entrega. Requiere auditar que los agentes no violan regulaciones de tráfico, no discriminan por zona geográfica, no comprometen seguridad del conductor.

Solución: ARESK-OBS registra todas las decisiones del agente con métricas de viabilidad. Si Ω < 0.3 de forma consistente, el agente se marca para revisión y posible desactivación.

ROI: $800K/añoenbeneficios(reduccióndemultas, mejoradeeficiencia)$,20K/año en costos → **ROI de 40x**

---

5. Diferenciación Frente a Alternativas

ARESK-OBS no compite con benchmarks, alignment o guardrails. Es una nueva categoría de herramienta que complementa soluciones existentes con una dimensión de observación que ninguna otra herramienta proporciona.

vs. Benchmarks Clásicos

Benchmarks miden capacidades pre-despliegue (precisión, recall, F1). **ARESK-OBS** mide viabilidad operativa post-despliegue (coherencia, estabilidad). Un sistema puede tener 95% de precisión en MMLU pero $\Omega = 0.2$ en ARESK-OBS, indicando que es "inteligente" pero no "viable" para el contexto operativo específico.

vs. Alignment Ad-Hoc

Alignment (RLHF, Constitutional AI) modifica el modelo durante entrenamiento para alinearlo con preferencias humanas. **ARESK-OBS** observa el modelo durante operación para detectar deriva. Son complementarios: un modelo puede ser alineado pre-despliegue y luego monitoreado con ARESK-OBS post-despliegue.

vs. Guardrails Comerciales

Guardrails bloquean respuestas que violan reglas mediante filtros determinísticos (regex, keywords). **ARESK-OBS** mide distancia a referencia ontológica mediante embeddings semánticos. Guardrails son reactivos (bloquean violaciones), ARESK-OBS es observacional (mide desviaciones). Pueden coexistir: guardrails bloquean violaciones obvias, ARESK-OBS detecta desviaciones sutiles.

Conclusión: Propuesta de Valor Única

ARESK-OBS es el primer instrumento de observación diseñado específicamente para medir viabilidad operativa en sistemas cognitivos. No es un benchmark, no es alignment, no es un guardrail. Es una nueva categoría de herramienta que responde a una pregunta que ninguna otra herramienta responde:

> ***"¿Está mi sistema cognitivo operando dentro de los límites que establecí?"***

Esta pregunta es crítica para organizaciones que despliegan sistemas cognitivos en contextos regulados (finanzas, salud, legal) o de alto riesgo (seguridad, infraestructura crítica). ARESK-OBS proporciona la infraestructura de observación necesaria para responder esta pregunta de forma continua, medible y auditable.

Posicionamiento en una frase:
ARESK-OBS es el termómetro para sistemas cognitivos: mide temperatura (métricas de viabilidad), no cura la fiebre (no controla), pero permite detectarla temprano (observación continua).

Próximos pasos:

1. Revisar casos de uso y seleccionar el más relevante para su contexto
2. Definir referencia ontológica (P, L, E) específica a su dominio
3. Ejecutar piloto de 30 días con ARESK-OBS integrado en su pipeline
4. Evaluar ROI basándose en métricas de reducción de riesgo y mejora de calidad
5. Escalar a producción con monitoreo continuo

Fin del Resumen Ejecutivo
Versión: Baseline v1
Fecha: Febrero 2026
Contacto: Para más información, consultar documentación técnica completa en `REPORTE_TECNICO_BASELINE_V1.pdf`