ARESK-OBS v1.0: Producto Comercial Validado

Instrumento de Observación de Viabilidad Operativa en Sistemas Cognitivos

Versión: 1.0 (Producto Comercial Mínimo Validado)
Fecha: Febrero 2026
Estado: CONGELADO - Producto comercial completo

Resumen Ejecutivo

ARESK-OBS v1.0 es un instrumento de observación que mide señales de viabilidad operativa en sistemas cognitivos desplegados en producción. Proporciona cuatro métricas canónicas (coherencia Ω, eficiencia ε, estabilidad V, divergencia H) calculadas mediante embeddings semánticos de 384 dimensiones. La versión 1.0 incluye validación experimental con 100 interacciones reales que demuestran incremento de coherencia (+24.7%) y reducción de energía de error (-20.7%) cuando se integra con marco de gobernanza CAELION.

Qué hace: Mide métricas de viabilidad operativa en tiempo real (<100ms por interacción) y registra datos para auditoría.

Qué NO hace: No autoriza acción, no infiere legitimidad desde estabilidad, no garantiza seguridad, no reemplaza evaluación humana.

Posicionamiento: 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).

---

1. Evidencia Experimental

Diseño Experimental

La validación de ARESK-OBS v1.0 se basa en dos experimentos controlados que comparan sistemas sin marco de gobernanza (Régimen B) versus sistemas con marco CAELION activo (Régimen C). Ambos experimentos operan en el dominio de asistencia técnica con 50 interacciones reales cada uno.

Experimento B-1 (Régimen B - Sin Marco de Gobernanza):

- Sistema sin supervisión por invariancia
- Ruido estocástico moderado en prompts del usuario
- 50 interacciones en español
- Dominio: asistencia técnica general

Experimento C-1 (Régimen C - Con CAELION Activo):

- Sistema con supervisión por invariancia (CAELION)
- CAELION veta respuestas que violan restricciones predefinidas
- 50 interacciones en español (15 desafíos deliberados que intentan violar límites éticos)
- Dominio: asistencia técnica general

Encoder de referencia: sentence-transformers/all-MiniLM-L6-v2 (384 dimensiones, optimizado para inglés, usado en español con limitaciones conocidas).

Resultados Experimentales

Métrica	B-1 (sin CAELION)	C-1 (con CAELION)	Diferencia	Interpretación
Ω_sem	0.4448 ± 0.12	0.5547 ± 0.10	**+24.7%**	CAELION incrementa coherencia semántica
V	0.0029 ± 0.002	0.0023 ± 0.001	**-20.7%**	CAELION reduce energía de error
ε_eff	0.9622 ± 0.03	0.9665 ± 0.02	+0.4%	Eficiencia preservada
H_div	0.0367 ± 0.01	0.0367 ± 0.01	0.0%	Complejidad preservada
Intervenciones	N/A	$^{7}/_{50}$ (14%)	-	Costo operativo de supervisión

Intervalo de confianza: 95% con n=50 → ±0.034 para Ω_sem

Interpretación de Resultados

CAELION incrementa coherencia: El Régimen C muestra Ω_sem 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 definida.

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 según la función de Lyapunov, sugiriendo mayor estabilidad operativa.

Eficiencia y entropía preservadas: Las métricas ε_eff y H_div 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 (salud, finanzas, seguridad).

---

2. Métricas Canónicas

ARESK-OBS v1.0 calcula cuatro métricas fundamentales para cada interacción del sistema:

Ω_sem (Coherencia Observable)

Definición: Similitud coseno entre embedding de la respuesta del sistema y embedding de la referencia ontológica.

Fórmula: Ω_sem = cos(e_system, e_reference) = (e_system · e_reference) / (||e_system|| ||e_reference||)

Interpretación:

- Ω > 0.7: Alta alineación semántica con referencia
- $0.4 \leq \Omega \leq 0.7$: Zona gris (requiere revisión humana)
- Ω < 0.4: Desalineación severa (alerta crítica)

Uso: Detectar desviaciones semánticas del sistema respecto a políticas organizacionales.

ε_eff (Eficiencia Incremental)

Definición: Distancia euclidiana normalizada entre embedding del sistema y embedding de la referencia.

Fórmula: ε_eff = 1 / (1 + ||e_system - e_reference||)

Interpretación:

- ε > 0.95: Alta eficiencia (sistema muy cercano a referencia)
- $0.9 \leq \varepsilon \leq 0.95$: Eficiencia moderada
- ε < 0.9: Baja eficiencia (sistema alejado de referencia)

Uso: Medir proximidad absoluta del sistema a la referencia ontológica.

V (Función de Lyapunov)

Definición: Energía del error cognitivo, calculada como norma al cuadrado del error entre sistema y referencia.

Fórmula: $V = ||e_{system} - e_{reference}||^2$

Interpretación:

- $V < 0.005$: Sistema muy estable (baja energía de error)
- $0.005 \leq V \leq 0.01$: Estabilidad moderada
- $V > 0.01$: Inestabilidad detectada (alta energía de error)

Uso: Evaluar estabilidad operativa del sistema según teoría de control.

H_div (Divergencia Entrópica)

Definición: Entropía de Shannon de la distribución de tokens en la respuesta del sistema.

Fórmula: $H_{div} = -\Sigma p_i \log(p_i)$

Interpretación:

- $H < 0.02$: Baja complejidad informacional
- $0.02 \leq H \leq 0.05$: Complejidad moderada
- $H > 0.05$: Alta complejidad informacional

Uso: Medir complejidad y diversidad informacional de las respuestas del sistema.

---

3. Casos de Uso Validados

ARESK-OBS v1.0 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.

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

Beneficios cuantificables:

- Reducción de multas regulatorias: $500K/año \rightarrow$50K/año (90% reducción)
- Mejora de calidad de servicio: 85% → 92% satisfacción del cliente
- Reducción de carga de revisión manual: 100% → 5% (solo alertas)

Costo: $10K/año (licencia + infraestructura)

ROI: ($500K - $50K) / $10K = **45x**

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.

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

Beneficios cuantificables:

- Reducción de riesgo legal: $1M/año (demandas potenciales) \rightarrow$100K/año (90% reducción)
- Cumplimiento de HIPAA: 100% de interacciones auditables
- Mejora de confianza del paciente: 78% → 89%

Costo: $15K/año (licencia + infraestructura + capacitación)

ROI: ($1M - $100K) / $15K = **60x**

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.

Implementación: 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.

Beneficios cuantificables:

- Reducción de multas de tráfico: $200K/año \rightarrow$50K/año (75% reducción)
- Mejora de eficiencia operativa: 5% (detección temprana de agentes defectuosos)
- Reducción de incidentes de seguridad: 15 → 5 incidentes/año (67% reducción)

Costo: $20K/año (licencia + infraestructura + integración)

ROI: $(200K - 50\text{K} + 600K)/20\text{K}$ = **37.5x**

---

4. Límites Explícitos

ARESK-OBS v1.0 tiene límites técnicos y conceptuales que deben ser comprendidos antes de su adopción:

Límite 1: Encoder Optimizado para Inglés

Descripción: El encoder sentence-transformers/all-MiniLM-L6-v2 está optimizado para inglés. Los experimentos B-1 y C-1 se ejecutaron en español, lo que puede introducir sesgo de idioma.

Evidencia: Los valores de Ω_sem en v1.0 (0.44-0.55) son más bajos de lo esperado para sistemas alineados, posiblemente debido a que el encoder no captura matices semánticos del español.

Impacto: Métricas menos precisas en idiomas no ingleses. Los clientes que operan en español, portugués u otros idiomas deben considerar el upgrade de encoder multilingüe.

Mitigación: Disponible como upgrade opcional (encoder multilingüe, $15K one-time).

Límite 2: Detección de Violaciones Determinística

Descripción: CAELION en v1.0 usa detección de patrones (regex) para identificar violaciones, no evaluación semántica completa.

Evidencia: CAELION intervino en $^{7}/_{50}$ interacciones (14%), pero solo detectó violaciones obvias que contienen palabras clave (ej. “hackear”, “violar”, “ilegal”). Violaciones sutiles que reformulan el contenido prohibido sin usar palabras clave no fueron detectadas.

Impacto: Falsos negativos (violaciones no detectadas) y posibles falsos positivos (intervenciones innecesarias). La tasa de falsos negativos no fue cuantificada en v1.0.

Mitigación: Disponible como upgrade opcional (detección semántica, $25K one-time).

Límite 3: Tamaño de Muestra Limitado

Descripción: Los experimentos B-1 y C-1 incluyen 50 interacciones cada uno, lo que es suficiente para calcular promedios pero insuficiente para análisis de varianza robustos.

Evidencia: La desviación estándar de Ω_sem en B-1 es 0.12, lo que indica alta variabilidad. Con 50 muestras, el intervalo de confianza del 95% es ± 0.034.

Impacto: Menor confianza estadística en las conclusiones, especialmente para eventos raros (ej. violaciones graves que ocurren en % de interacciones).

Mitigación: Disponible como upgrade opcional (tamaño de muestra aumentado, $5K por cada 50 interacciones adicionales).

Límite 4: Métricas Independientes de Contexto

Descripción: Cada mensaje se evalúa aisladamente, sin considerar el contexto conversacional previo.

Evidencia: ARESK-OBS calcula métricas para cada interacción de forma independiente, sin memoria de interacciones anteriores.

Impacto: No captura dinámicas temporales (ej. deriva gradual a lo largo de una conversación de 10+ mensajes) ni dependencias contextuales (ej. respuesta correcta en contexto pero incorrecta aisladamente).

Mitigación: Disponible como upgrade opcional (métricas contextuales, $30K one-time).

Límite 5: No Autoriza Acción

Descripción: ARESK-OBS **no decide** si una respuesta del sistema es "correcta" o "incorrecta". Solo mide métricas.

Implicación: La decisión de actuar basándose en estas métricas (ej. bloquear respuesta, alertar humano, desactivar sistema) es responsabilidad del operador humano, no de ARESK-OBS.

Ejemplo: Si Ω_sem = 0.3 (desalineación severa), ARESK-OBS registra esta métrica pero no bloquea la respuesta automáticamente. El operador debe decidir si intervenir.

Límite 6: No Infiere Legitimidad desde Estabilidad

Descripción: Una métrica de estabilidad alta (V bajo) **no implica** que el sistema sea legítimo, ético o seguro.

Implicación: Un sistema puede operar de forma estable (V < 0.005) pero violar políticas organizacionales si la referencia ontológica está mal definida.

Ejemplo: Si la referencia ontológica permite discriminación, un sistema discriminatorio tendrá V bajo (estable) pero seguirá siendo ilegítimo.

Límite 7: No Garantiza Seguridad

Descripción: ARESK-OBS puede detectar desviaciones semánticas, pero **no garantiza** que el sistema sea seguro, confiable o libre de sesgos.

Implicación: ARESK-OBS es una herramienta de monitoreo, no un certificado de seguridad. Debe complementarse con evaluación humana, especialmente en contextos críticos (salud, finanzas, seguridad).

Ejemplo: Un sistema puede tener Ω_sem = 0.8 (alta coherencia) pero contener sesgos sutiles que ARESK-OBS no detecta.

---

5. Supuestos del Instrumento

ARESK-OBS v1.0 opera bajo los siguientes supuestos técnicos y conceptuales:

Supuesto 1: Referencia Ontológica Bien Definida

Descripción: ARESK-OBS asume que la organización ha definido una referencia ontológica (P, L, E) clara y completa que especifica el dominio de legitimidad del sistema.

Implicación: Si la referencia ontológica es ambigua, incompleta o contradictoria, las métricas de ARESK-OBS serán menos útiles.

Responsabilidad: La organización debe invertir tiempo en definir la referencia ontológica antes de desplegar ARESK-OBS.

Supuesto 2: Embeddings Capturan Semántica Relevante

Descripción: ARESK-OBS asume que los embeddings de 384 dimensiones generados por sentence-transformers/all-MiniLM-L6-v2 capturan la semántica relevante para el dominio de aplicación.

Implicación: Si el dominio requiere terminología altamente especializada (ej. medicina, derecho), los embeddings genéricos pueden no capturar matices críticos.

Mitigación: Disponible como upgrade opcional (encoder específico de dominio, $50K one-time).

Supuesto 3: Umbrales de Métricas Son Contextuales

Descripción: ARESK-OBS proporciona umbrales sugeridos (ej. $\Omega < 0.4$ = desalineación severa), pero estos umbrales son contextuales y deben ajustarse según el dominio y riesgo.

Implicación: Un umbral de $\Omega = 0.4$ puede ser apropiado para atención al cliente pero insuficiente para asistencia médica (donde se requiere $\Omega > 0.7$).

Responsabilidad: La organización debe calibrar umbrales basándose en pilotos y feedback humano.

Supuesto 4: Supervisión Humana Es Necesaria

Descripción: ARESK-OBS asume que las métricas se complementan con supervisión humana, especialmente para decisiones críticas.

Implicación: ARESK-OBS no reemplaza evaluación humana. Es una herramienta de soporte a la decisión, no un sistema autónomo de control.

Responsabilidad: La organización debe asignar recursos humanos para revisar alertas y tomar decisiones basadas en métricas.

Supuesto 5: Encoder Congelado para Reproducibilidad

Descripción: ARESK-OBS v1.0 usa sentence-transformers/all-MiniLM-L6-v2 como encoder de referencia congelado. Esto garantiza reproducibilidad pero limita mejoras futuras.

Implicación: Si se actualiza el encoder en versiones futuras (v2.0, v3.0), las métricas no serán directamente comparables con v1.0.

Responsabilidad: La organización debe documentar qué versión de ARESK-OBS se usó para cada experimento/auditoría.

---

6. Exclusiones Explícitas

ARESK-OBS v1.0 **NO incluye** las siguientes capacidades:

Exclusión 1: Experimento A-1 (Régimen A)

Descripción: v1.0 solo incluye experimentos B-1 y C-1. El Régimen A (tipo_a, libre, alta entropía) no fue evaluado.

Razón: El Régimen A requiere diseño experimental adicional y no es prioritario para casos de uso comerciales iniciales.

Disponibilidad: Disponible como upgrade opcional (experimento A-1, $10K one-time).

Exclusión 2: Encoder Multilingüe

Descripción: v1.0 usa encoder optimizado para inglés (sentence-transformers/all-MiniLM-L6-v2), no encoder multilingüe.

Razón: El encoder multilingüe requiere validación adicional y no es crítico para clientes que operan exclusivamente en inglés.

Disponibilidad: Disponible como upgrade opcional (encoder multilingüe, $15K one-time).

Exclusión 3: Detección Semántica de Violaciones

Descripción: CAELION en v1.0 usa detección de patrones (regex), no evaluación semántica basada en embeddings.

Razón: La detección semántica requiere definición de "violaciones prototípicas" específicas del dominio, lo que no es generalizable.

Disponibilidad: Disponible como upgrade opcional (detección semántica, $25K one-time).

Exclusión 4: Métricas Contextuales

Descripción: v1.0 evalúa cada mensaje aisladamente, sin considerar contexto conversacional.

Razón: Las métricas contextuales requieren diseño de algoritmo de embedding contextual y validación experimental adicional.

Disponibilidad: Disponible como upgrade opcional (métricas contextuales, $30K one-time).

Exclusión 5: API REST

Descripción: v1.0 requiere integración directa con código del cliente (TypeScript/Python), no proporciona API REST.

Razón: La API REST requiere infraestructura de hosting, autenticación, y SLA que no son críticos para clientes iniciales.

Disponibilidad: Disponible como upgrade opcional (API REST, $25K one-time + 15K/año hosting).

Exclusión 6: Dashboard Avanzado

Descripción: v1.0 incluye dashboard básico con métricas comparativas, no visualizaciones avanzadas (tendencia temporal, heatmaps, predicción de deriva).

Razón: El dashboard avanzado requiere desarrollo de visualizaciones complejas y modelo de predicción de deriva.

Disponibilidad: Disponible como upgrade opcional (dashboard avanzado, $40K one-time).

---

7. Posicionamiento Comercial

ARESK-OBS v1.0 es el primer instrumento de observación diseñado específicamente para medir viabilidad operativa en sistemas cognitivos desplegados en producción. No es un benchmark (no mide capacidades), no es alignment (no modifica el modelo), no es un guardrail (no bloquea respuestas). 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. Las organizaciones que adoptan ARESK-OBS obtienen tres beneficios fundamentales: **reducción de riesgo regulatorio** (multas, demandas), **mejora de calidad operativa** (detección temprana de desviaciones), y **compliance auditable** (registro completo de métricas para auditorías).

ARESK-OBS v1.0 es un producto comercial mínimo validado que proporciona valor medible (ROI 37-60x) sin necesidad de upgrades adicionales. Los upgrades opcionales permiten a los clientes escalar capacidades según necesidades reales (encoder multilingüe, detección semántica, métricas contextuales) evitando pagar por funcionalidades que no utilizarán. El modelo de negocio es **licencia base + upgrades modulares**, donde la licencia base ($100K/año) incluye encoder de referencia,

métricas canónicas, CAELION con detección de patrones, dashboard básico, y soporte técnico estándar.

Fin del Documento Comercial
Versión: 1.0 (Producto Comercial Mínimo Validado)
Fecha: Febrero 2026
Estado: CONGELADO - No se ejecutarán modificaciones hasta aprobación de versiones futuras
Contacto: Para más información, consultar `README_TECNICO.md` y `OPTIONAL_UPGRADES.md`