



E N K I D U

— M I O V S K I G R O U P —

E S T U D I O D E C A S O

Trust-Assured Agentic AI

Cómo el marco de aseguramiento de la confianza y Tangram moldearon un sistema multiagente de evaluación de capacidades para Environment and Climate Change Canada.

Español · Edición resumida

Basada en la versión 1.0 · 21 de junio de 2026

Preparado por MIOVSKI GROUP

Clasificación · No clasificado

La edición completa en inglés prevalece

Acerca de esta edición

Este documento es una edición resumida, en español, del estudio de caso « Trust-Assured Agentic AI », versión 1.0 (21 de junio de 2026), preparado por Miovski Group. Cada sección del original está representada de forma abreviada. En caso de divergencia, la edición completa en inglés prevalece.

Clasificación: no clasificado.

Resumen

Environment and Climate Change Canada (ECCC) necesitaba evaluar la madurez de sus capacidades de negocio y ensamblar hojas de ruta digitales por dirección general a partir de un corpus amplio y heterogéneo de documentación interna — un trabajo lento, costoso y subjetivo cuando se hace a mano. Enkidu construyó una prueba de concepto (PoC) de IA agéntica multiagente para realizar esa síntesis con fidelidad, transparencia y repetibilidad, ejecutándose íntegramente en infraestructura local para preservar la soberanía de los datos.

Este estudio de caso describe cómo se aplicaron dos marcos Enkidu a ese encargo. Tangram aportó la arquitectura operativa — descomponiendo el pipeline de evaluación y hoja de ruta en zonas acotadas y gobernadas, con traspasos explícitos. El marco de aseguramiento de la confianza (Origi) aportó la disciplina que hace defendibles las salidas del sistema — cada afirmación de confianza vinculada a una política, sustentada por un resultado medible, justificada por una cadena de evidencia anotada y trazable mediante un registro de procedencia a prueba de manipulaciones. La PoC de ECCC fue, de hecho, el primer banco de pruebas de extremo a extremo de ambos marcos.

El resultado es un sistema cuya regla central — determinismo donde importa, razonamiento donde ayuda — es la expresión directa del principio de pureza de políticas de Tangram y de la separación de funciones del aseguramiento de la confianza. La puntuación es verdad de referencia impuesta por código; el modelo de lenguaje interpreta, extrae evidencia y redacta narrativa — pero nunca asigna una puntuación final.

No pida que confíen en usted: demuestre, afirmación por afirmación y paso a paso, exactamente por qué pueden hacerlo.

1. Introducción

Las organizaciones necesitan comprender la madurez de sus capacidades de negocio para priorizar la inversión, impulsar la transformación y gestionar el riesgo. Gran parte de la evidencia ya existe dentro de la organización — transcripciones de entrevistas, documentos de procesos, post-mortems de incidentes, registros de cambios, hallazgos de auditoría — pero leerla y sintetizarla a escala es impracticable para humanos, y las evaluaciones dirigidas por consultores son caras, subjetivas y difíciles de refrescar.

ECCC encargó a Enkidu demostrar que un sistema de IA agéntica bien restringido podía realizar esa síntesis a escala sin dejar de ser lo bastante confiable para informar decisiones reales. Los entregables abarcan una evaluación de madurez de capacidades y un conjunto de hojas de ruta digitales por dirección, cada una ensamblada desde fuentes departamentales autoritativas con plena trazabilidad de citas. Como esas salidas alimentan decisiones de inversión y transformación, el listón no era « útil » sino « defendible »: un analista, un auditor o un alto directivo debe poder preguntar, de cualquier cifra del informe, « ¿de dónde salió esto y por qué debería creerlo? » — y recibir una respuesta completa.

- Lo que muestra el estudio — cómo Tangram estructuró el pipeline de evaluación y hoja de ruta en zonas acotadas y gobernadas con traspasos contractuales.
- Cómo el marco de aseguramiento de la confianza convirtió « creemos que este sistema es confiable » en una posición estructurada y probada — tanto en la entrega como en la inferencia.
- Cómo ambos marcos se reforzaron mutuamente, y cómo la PoC se extendió a un continuo completo de hoja de ruta digital.

2. Los marcos aplicados

Se aplicaron dos marcos Enkidu al encargo. Operan a niveles distintos — Tangram gobierna la forma de la operación; el aseguramiento de la confianza gobierna la fiabilidad de lo que fluye por ella — y son deliberadamente complementarios.

- Tangram — la forma de la operación — descompuso el pipeline en zonas acotadas con traspasos contractuales, políticas puras y rutas de error diseñadas.
- Aseguramiento de la confianza — la fiabilidad del trabajo — vinculó cada puntuación y afirmación a una política, una cadena de evidencia y procedencia firmada; verificación cero-confianza y capacidad ganada.

Tangram: toda operación se descompone en zonas de servicio acotadas — una zona es una unidad de trabajo con una sola meta cuya condición de salida se convierte en la prenda de entrada de la siguiente, creando una cadena de traspasos contractuales; la pureza de políticas mantiene la gobernanza separada de las operaciones; los errores son zonas de pleno derecho. Aseguramiento de la confianza: cada afirmación es un caso de aseguramiento — afirmación, argumento, evidencia — vinculado a exactamente una política con umbral explícito, satisfecho solo cuando el resultado medido alcanza el umbral y la cadena de evidencia válida. La meta no es la confianza máxima sino la confianza calibrada.

3. La solución ECCC: un sistema multiagente

La PoC es un sistema multiagente que ingiere un corpus heterogéneo de documentación interna, aplica rúbricas ancladas y deterministas, y produce puntuaciones de criticidad, puntuaciones de madurez y una narrativa estructurada de puntos de dolor, retos y oportunidades, mapeados a un modelo de capacidades predefinido. Todo el pipeline corre en infraestructura local para preservar la confidencialidad: una estación NVIDIA DGX Spark aloja el modelo de razonamiento (Llama-3.3-Nemotron-Super-49B, servido mediante un microservicio NIM), un almacén vectorial Chroma, un almacén relacional ligero para el modelo de capacidades y el historial de ejecuciones, Arize Phoenix para la observabilidad y un conjunto de servidores MCP (Model Context Protocol) — sin salida de red en la capa de herramientas.

La regla central: determinismo donde importa, razonamiento donde ayuda. Fórmulas de puntuación, rúbricas, ponderaciones y el modelo de capacidades son verdad de referencia impuesta por código; el modelo de lenguaje interpreta, extrae y redacta — nunca aritmética, nunca una puntuación final. El sistema nunca produce « solo » una cifra: produce una cifra más su derivación completa y reproducible.

La orquestación sigue un patrón supervisor/especialistas en lugar de un enjambre libre: cada agente es un nodo de flujo con un rol estrecho, un contrato de salida explícito y herramientas declaradas — y el agente que extrae la evidencia nunca es el que puntúa, ni el que redacta.

- Supervisor — planifica la ejecución, selecciona las capacidades a analizar, reparte el trabajo e impone criterios de finalización y topes de coste.
- Recuperación — extrae de Chroma los fragmentos documentales más relevantes, por similitud vectorial y filtros de metadatos.
- Extracción de evidencia — lee los fragmentos y extrae elementos estructurados — puntos de dolor, retos, oportunidades — cada uno con cita de fuente.
- Puntuación — aplica la rúbrica anclada a la evidencia y llama al MCP de puntuación para calcular el valor ponderado final, en código.
- Crítica — revisa de forma independiente la puntuación contra la evidencia y la rúbrica, señala inconsistencias y puede forzar una repuntuación.
- Agregador — consolida los resultados por capacidad en la vista organizacional e identifica patrones sistémicos.
- Narrativa — produce la narrativa legible del informe, sin escribir nada que no esté fundado en puntuaciones estructuradas y evidencia citada.

4. Tangram en la práctica: el continuo de hoja de ruta

El pipeline de ECCC se proyecta limpiamente sobre Tangram. El flujo de extremo a extremo — de los documentos brutos de la dirección a la hoja de ruta digital completada — es un solo continuo, y cada etapa analítica es una zona con una meta y condiciones de frontera explícitas. Los prompts que impulsan el sistema se hacen visibles como definiciones de zona: rol, fuentes autorizadas (restricción dura), flujo de trabajo por fases obligatorio, requisitos de cita y puerta de cobertura — en términos Tangram: meta de zona, prendas de entrada, conops y condiciones de salida.

- Análisis del plan de negocio — convertir el plan de la dirección en hallazgos estratégicos, iniciativas clasificadas y una matriz de acciones — salida: un libro validado que pasa la puerta de cobertura.
- Objetivos y motores — derivar el ancla estratégica: objetivos, motores y patrocinadores — cada celda trazable al libro de análisis.
- Estado actual — establecer la línea base tal-cual de capacidades, aplicaciones, datos y puntos de dolor — puntuada contra las dimensiones de madurez.
- Estado objetivo — derivar la visión a-alcanzar del plan departamental, filtrada a la dirección — fuente de toda iniciativa posterior.
- Iniciativas — listar las iniciativas discretas que llevan del estado actual al objetivo — cartera con responsables, entregables y métricas, todo citado.
- Recomendaciones tecnológicas — producir la respuesta tecnológica, anclada a objetivos e iniciativas.

La pureza de políticas aparece directamente en las restricciones duras: « no uses conocimiento externo; si las fuentes son insuficientes, dilo; si discrepan, expón el conflicto » — una política pura, de veredicto determinista. Y los errores son zonas de pleno derecho: un fallo de la puerta de cobertura no descarta nada en silencio — produce un informe de cobertura que lista cada brecha; una discordancia en el nombre de la dirección detiene la ejecución y pide confirmación. Si su único manejo de errores es « contacte a un supervisor », tiene una zona sin diseñar — el pipeline de ECCC diseña sus fallos por adelantado.

5. El aseguramiento de la confianza en la práctica

Donde Tangram da forma al pipeline, el marco de aseguramiento de la confianza hace defendibles sus salidas. La PoC fue el primer sistema en ejercitar el marco de extremo a extremo, y el mapeo es concreto: las nueve componentes de confianza, el modelo de procedencia y el arnés de confianza en inferencia corresponden cada uno a un mecanismo preciso del sistema desplegado.

- Competencia y fiabilidad — rúbricas ancladas y deterministas, modelo de capacidades fijo; desempeño validado contra un conjunto de evaluación de referencia.
- Transparencia y explicabilidad — razonamiento estructurado « evidencia y luego juicio »; cada puntuación se remonta a evidencia citada y a la banda de rúbrica aplicada.
- Confianza calibrada — la seguridad infundada se guarda con un agente crítico independiente y evaluadores de anclaje; los elementos poco sustentados se descartan, no se afirman.
- Previsibilidad y consistencia — versión de modelo fija, prompts versionados y fórmulas deterministas hacen las ejecuciones reproducibles y comparables en el tiempo.
- Rendición de cuentas y recurso — pista de auditoría completa de documentos, rúbricas, llamadas a herramientas y puntuaciones; punto de control humano antes de publicar el informe.
- Equidad y gestión de sesgos — puntuación de madurez desagregada por dirección y dimensión, aflorando los subgrupos débiles que un promedio ocultaría.
- Seguridad y robustez — saneamiento de documentos, raíles anti-inyección, listas de herramientas permitidas por agente y sin salida de red en la capa de herramientas.
- Alineación con la intención — la aplicación del alcance mantiene a los agentes dentro del mandato de evaluación; el modelo nunca asigna puntuaciones finales.
- Agencia y dependencia humanas — la PoC es explícitamente apoyo a la decisión, no un decisor autónomo; el informe final lo revisa un analista humano.

La cadena de evidencia a prueba de manipulaciones se realiza mediante contratos de salida estructurados, imposición de citas y observabilidad total: cada mensaje entre agentes sigue un esquema validado; cada elemento de evidencia y cada puntuación llevan una cita resoluble hasta el documento y el fragmento; las salidas sin cita se rechazan. La diferencia en la que insiste el marco: no « una cifra que alguien tecleó », sino « una cifra que un proceso nombrado y verificable produjo ».

- Pre-vuelo (ingreso) — el saneamiento retira el contenido activo; los raíles de entrada detectan patrones de inyección y rechazan peticiones fuera de alcance antes de entrar en contexto de agente.
- En vuelo (por paso) — listas de herramientas permitidas por agente (mínimo privilegio); MCP de ficheros acotado a directorios designados; presupuestos de herramientas por ejecución contra el descontrol.
- Post-vuelo (egreso) — validación de esquema, imposición de citas, comprobaciones de anclaje y filtrado de datos personales antes de escribir nada al informe final.

La capacidad ganada aparece en el modelo de permisos: la mayoría de las herramientas son de solo lectura; solo el constructor de informes y un escritor de ficheros acotado mutan — el acceso de escritura es la capacidad más estrechamente concedida del sistema. Y la separación de funciones — ningún componente juzga y ejecuta a la vez — la porta el propio patrón supervisor/especialistas.

6. Extender la prueba de concepto

Establecida la PoC, el encargo la extendió de una sola etapa de evaluación al continuo completo de hoja de ruta digital. Cada extensión se añadió del mismo modo: una nueva zona con prendas de entrada explícitas, una política pura de fuentes y citas, una ruta de error de puerta de cobertura diseñada y la misma disciplina de procedencia. La nueva funcionalidad no relajó nada — entre las nuevas entradas: una evaluación de madurez puntuada en veintitrés dimensiones de capacidad y un análisis de puntos de dolor puntuado de forma determinista.

7. Resultados y conclusión

El encargo de ECCC entregó un sistema multiagente funcional, en las instalaciones, que evalúa la madurez de las capacidades de negocio y ensambla hojas de ruta digitales por dirección desde fuentes autoritativas — con fidelidad, transparencia y repetibilidad. Igual de importante: probó ambos marcos Enkidu en un solo despliegue real.

Tangram dio al encargo una arquitectura operativa que un arquitecto empresarial lee de un vistazo: un continuo de zonas acotadas, traspasos contractuales, políticas puras y rutas de error diseñadas. El marco de aseguramiento de la confianza dio al mismo encargo una posición de confianza defendible: nueve componentes realizadas cada una por un mecanismo concreto, una cadena de evidencia a prueba de manipulaciones tras cada puntuación, y un arnés cero-confianza que verifica cada paso en la inferencia. Ambos se encuentran en la regla central del sistema — determinismo donde importa, razonamiento donde ayuda — a la vez frontera de política Tangram y separación de funciones del aseguramiento de la confianza.

El resultado es el objetivo declarado del marco hecho realidad: no la confianza máxima, sino la confianza calibrada — dependencia proporcional a la fiabilidad demostrada — respaldada por un expediente verificable que un analista, un auditor o un alto directivo de ECCC puede interrogar. La PoC está completa; los marcos que la moldearon quedan probados, y el patrón que establece es reutilizable para el próximo despliegue gobernado y aumentado por IA.

Confianza calibrada, ganada afirmación por afirmación y paso a paso — y una arquitectura operativa lo bastante legible para probarlo.