



E N K I D U

— M I O V S K I G R O U P —

F A L L S T U D I E

# Trust-Assured Agentic AI

Wie das Trust-Assurance-Framework und Tangram ein Multi-Agenten-System zur  
Fähigkeitsbewertung für Environment and Climate Change Canada formten.

Deutsch · Kompaktausgabe

Nach Version 1.0 · 21. Juni 2026

Erstellt von MIOVSKI GROUP

Einstufung · Nicht klassifiziert

Die vollständige englische Ausgabe ist maßgeblich

---

## Zu dieser Ausgabe

Dieses Dokument ist eine deutsche Kompaktausgabe der Fallstudie « Trust-Assured Agentic AI », Version 1.0 (21. Juni 2026), erstellt von der Mioviski Group. Jeder Abschnitt des Originals ist in gekürzter Form vertreten. Bei Abweichungen ist die vollständige englische Ausgabe maßgeblich. Einstufung: nicht klassifiziert.

---

## Zusammenfassung

Environment and Climate Change Canada (ECCC) musste die Reife seiner Geschäftsfähigkeiten bewerten und aus einem großen, heterogenen Korpus interner Dokumentation digitale Roadmaps je Direktion zusammenstellen — Arbeit, die von Hand langsam, teuer und subjektiv ist. Enkidu baute einen Multi-Agenten-Proof-of-Concept (PoC) agentischer KI, der diese Synthese treu, transparent und wiederholbar leistet — vollständig auf lokaler Infrastruktur, um die Datensouveränität zu wahren.

Diese Fallstudie beschreibt, wie zwei Enkidu-Frameworks auf dieses Mandat angewandt wurden. Tangram lieferte die operative Architektur — die Zerlegung der Bewertungs- und Roadmap-Pipeline in begrenzte, gesteuerte Zonen mit expliziten Übergaben. Das Trust-Assurance-Framework (Origi) lieferte die Disziplin, die die Ausgaben des Systems verteidigbar macht — jede Vertrauensbehauptung an eine Richtlinie gebunden, durch ein messbares Ergebnis gestützt, durch eine annotierte Beweiskette begründet und über einen manipulationssicheren Herkunftsnachweis nachvollziehbar. Der ECCC-PoC war faktisch das erste End-to-End-Bewährungsfeld beider Frameworks.

Das Ergebnis ist ein System, dessen zentrale Regel — Determinismus, wo es zählt, Schlussfolgern, wo es hilft — der direkte Ausdruck des Tangram-Prinzips der Richtlinienreinheit und der Funktionstrennung der Trust Assurance ist. Die Bewertung ist codeerzwungene Referenzwahrheit; das Sprachmodell interpretiert, extrahiert Belege und entwirft Text — vergibt aber niemals eine Endnote.

Bitten Sie nicht um Vertrauen — zeigen Sie, Behauptung für Behauptung und Schritt für Schritt, genau, warum man Ihnen vertrauen kann.

---

# 1. Einleitung

Organisationen müssen regelmäßig die Reife ihrer Geschäftsfähigkeiten verstehen, um Investitionen zu priorisieren, Transformation voranzutreiben und Risiken zu steuern. Ein Großteil der Belege existiert bereits im Haus — Interviewtranskripte, Prozessdokumente, Vorfallanalysen, Änderungsprotokolle, Auditbefunde — aber sie im Maßstab zu lesen und zu synthetisieren ist für Menschen unpraktikabel, und beraterngeführte Bewertungen sind teuer, subjektiv und schwer zu aktualisieren.

ECCC beauftragte Enkidu zu zeigen, dass ein wohlbeschränktes agentisches KI-System diese Synthese im Maßstab leisten kann und dabei vertrauenswürdig genug bleibt, echte Entscheidungen zu informieren. Die Ergebnisse umfassen eine Reifegradbewertung und einen Satz digitaler Roadmaps je Direktion, jede aus autoritativen Quellen des Ministeriums zusammengestellt, mit voller Zitat-Nachvollziehbarkeit. Weil diese Ausgaben Investitions- und Transformationsentscheidungen speisen, lag die Latte nicht bei nützlich", sondern bei verteidigbar": Ein Analyst, ein Auditor oder eine Führungskraft muss zu jeder Zahl im Bericht fragen können Woher kommt das, und warum sollte ich es glauben?" — und eine vollständige Antwort erhalten.

- Was die Studie zeigt — wie Tangram die Bewertungs- und Roadmap-Pipeline in begrenzte, gesteuerte Zonen mit vertraglichen Übergaben strukturierte.
- Wie das Trust-Assurance-Framework wir glauben, das System ist vertrauenswürdig" in eine strukturierte, belegte Position verwandelte — zur Freigabe- wie zur Inferenzzeit.
- Wie die beiden Frameworks einander verstärkten und wie der PoC zu einem vollständigen Digital-Roadmap-Kontinuum erweitert wurde.

---

## 2. Die angewandten Frameworks

Zwei Enkidu-Frameworks wurden angewandt. Sie wirken auf verschiedenen Ebenen — Tangram regiert die Form der Operation; Trust Assurance regiert die Vertrauenswürdigkeit dessen, was hindurchfließt — und sie sind bewusst komplementär.

- Tangram — die Form der Operation — zerlegte die Pipeline in begrenzte Zonen mit vertraglichen Übergaben, reinen Richtlinien und konstruierten Fehlerpfaden.
- Trust Assurance — die Vertrauenswürdigkeit der Arbeit — band jede Note und Behauptung an eine Richtlinie, eine Beweiskette und signierte Herkunft; Zero-Trust-Prüfung und verdiente Fähigkeit.

Tangram: Jede Operation lässt sich in begrenzte Dienstzonen zerlegen — eine Zone ist eine Arbeitseinheit mit einem Ziel, deren Austrittsbedingung zum Eintrittspfand der nächsten wird und so eine Kette vertraglicher Übergaben bildet; Richtlinienreinheit hält Governance von Betrieb getrennt; Fehler sind vollwertige Zonen. Trust Assurance: Jede Behauptung ist ein Assurance-Case — Behauptung, Argument, Beweise — gebunden an genau eine Richtlinie mit expliziter Schwelle, erfüllt nur, wenn das gemessene Ergebnis die Schwelle erreicht und die Beweiskette validiert. Das Ziel ist nicht maximales, sondern kalibriertes Vertrauen.

---

### 3. Die ECCC-Lösung: ein Multi-Agenten-System

Der PoC ist ein Multi-Agenten-System, das einen heterogenen Korpus interner Dokumentation aufnimmt, verankerte, deterministische Rubriken anwendet und Kritikalitätsnoten, Reifegradnoten und eine strukturierte Erzählung von Schmerzpunkten, Herausforderungen und Chancen erzeugt, abgebildet auf ein vordefiniertes Fähigkeitsmodell. Die gesamte Pipeline läuft auf lokaler Infrastruktur: Eine NVIDIA-DGX-Spark-Workstation beherbergt das Reasoning-Modell (Llama-3.3-Nemotron-Super-49B, ausgeliefert über einen NIM-Microservice), einen Chroma-Vektorspeicher, einen leichten relationalen Speicher für Fähigkeitsmodell und Laufhistorie, Arize Phoenix für die Observability und einen Satz MCP-Server (Model Context Protocol) — ohne Netzwerk-Egress auf der Werkzeugschicht.

Die zentrale Regel: Determinismus, wo es zählt, Schlussfolgern, wo es hilft. Bewertungsformeln, Rubriken, Gewichtungslgik und das Fähigkeitsmodell sind codeerzwungene Referenzwahrheit; das Sprachmodell interpretiert, extrahiert und entwirft — niemals Arithmetik, niemals eine Endnote. Das System erzeugt nie nur" eine Zahl: Es erzeugt eine Zahl plus ihre vollständige, wiederabspielbare Herleitung.

Die Orchestrierung folgt einem Supervisor/Spezialisten-Muster statt eines freien Schwarms: Jeder Agent ist ein Workflow-Knoten mit engem Auftrag, explizitem Ausgabevertrag und deklarierten Werkzeugen — und der Agent, der Belege extrahiert, ist nie derjenige, der bewertet, und auch nicht derjenige, der schreibt.

- Supervisor — plant den Lauf, wählt die zu analysierenden Fähigkeiten, verteilt die Arbeit und erzwingt Abschlusskriterien und Kostendeckel.
- Retrieval — holt die relevantesten Dokumentabschnitte aus Chroma, über Vektorähnlichkeit und Metadatenfilter.
- Belegextraktion — liest die Abschnitte und extrahiert strukturierte Belege — Schmerzpunkte, Herausforderungen, Chancen — jeweils mit Quellenzitat.
- Bewertung — wendet die verankerte Rubrik auf die Belege an und ruft das Bewertungs-MCP, das den finalen gewichteten Wert im Code berechnet.
- Kritik — prüft die Bewertung unabhängig gegen Belege und Rubrik, meldet Inkonsistenzen und kann eine Neubewertung erzwingen.
- Aggregator — konsolidiert die Ergebnisse je Fähigkeit zur Organisationssicht und identifiziert systemische Muster.
- Narrativ — verfasst die lesbare Berichtserzählung und schreibt nichts, das nicht in strukturierten Notizen und zitierten Belegen gründet.

---

## 4. Tangram in der Praxis: das Roadmap-Kontinuum

Die ECCC-Pipeline bildet sich sauber auf Tangram ab. Der End-to-End-Fluss — von den Rohdokumenten der Direktion zur fertigen digitalen Roadmap — ist ein einziges Kontinuum, und jede Analysestufe ist eine Zone mit einem Ziel und expliziten Randbedingungen. Die Prompts, die das System treiben, werden als Zonendefinitionen sichtbar: Rolle, autorisierte Quellen (harte Randbedingung), verbindlicher Phasen-Workflow, Zitatpflichten und ein Abdeckungs-Gate — in Tangram-Begriffen: Zonenziel, Eintrittspfähder, Conops und Austrittsbedingungen.

- Geschäftsplan-Analyse — den Plan der Direktion in strategische Befunde, gereimte Initiativen und eine Aktionsmatrix verwandeln — Ausgabe: ein validiertes Arbeitsbuch, das das Abdeckungs-Gate besteht.
- Ziele und Treiber — den strategischen Anker ableiten: Ziele, Treiber, Sponsoren — jede Zelle rückführbar auf das Analyse-Arbeitsbuch.
- Ist-Zustand — die Ist-Basis über Fähigkeiten, Anwendungen, Daten und Schmerzpunkte legen — bewertet gegen die Reifegraddimensionen.
- Soll-Zustand — die Zielvision aus dem Ministeriumsplan ableiten, gefiltert auf die Direktion — Quelle jeder späteren Initiative.
- Initiativen — die diskreten Initiativen vom Ist zum Soll auflisten — Portfolio mit Verantwortlichen, Ergebnissen und Kennzahlen, alles zitiert.
- Technologieempfehlungen — die Technologieantwort erzeugen, verankert an Zielen und Initiativen.

Die Richtlinienreinheit erscheint direkt in den harten Randbedingungen: "Nutze kein externes Wissen; reichen die Quellen nicht, sag es; widersprechen sie sich, lege den Konflikt offen" — eine reine Richtlinie mit deterministischem Verdikt. Und Fehler sind vollwertige Zonen: Ein Gate-Fehlschlag lässt nichts stillschweigend fallen — er erzeugt einen Abdeckungsbericht, der jede Lücke auflistet; eine Namensdiskrepanz stoppt den Lauf und bittet um Bestätigung. Wenn Ihre einzige Fehlerbehandlung "wenden Sie sich an einen Vorgesetzten" lautet, haben Sie eine unkonstruierte Zone — die ECCC-Pipeline konstruiert ihre Fehlschläge im Voraus.

---

## 5. Trust Assurance in der Praxis

Wo Tangram die Pipeline formt, macht das Trust-Assurance-Framework ihre Ausgaben verteidigbar. Der PoC war das erste System, das das Framework End-to-End ausübte, und die Abbildung ist konkret: die neun Vertrauenskomponenten, das Herkunftsmodell und das Vertrauensgeschirr zur Inferenzzeit entsprechen je einem präzisen Mechanismus des Systems.

- Kompetenz & Verlässlichkeit — verankerte, deterministische Rubriken, festes Fähigkeitsmodell; Leistung gegen einen kuratierten Goldstandard-Datensatz geprüft.
- Transparenz & Erklärbarkeit — strukturiertes erst Belege, dann Urteil"; jede Note führt auf zitierte Belege und die angewandte Rubrikstufe zurück.
- Kalibrierte Konfidenz — selbstbewusste Irrtümer werden von einem unabhängigen Kritik-Agenten und Verankerungsprüfern abgefangen; schwach belegte Punkte werden verworfen, nicht behauptet.
- Vorhersagbarkeit & Konsistenz — feste Modellversion, versionierte Prompts und deterministische Formeln machen Läufe reproduzierbar und über die Zeit vergleichbar.
- Rechenschaft & Einspruch — vollständiger Prüfpfad über Dokumente, Rubriken, Werkzeugaufrufe und Noten; menschlicher Kontrollpunkt vor jeder Berichts freigabe.
- Fairness & Bias-Management — disaggregierte Reifegradbewertung je Direktion und Dimension, die schwache Teilgruppen sichtbar macht, statt sie im Schnitt zu verstecken.
- Sicherheit & Robustheit — Dokumentbereinigung, Injektions-Eingangsschienen, Werkzeug-Positivlisten je Agent und kein Netzwerk-Egress auf der Werkzeugschicht.
- Ausrichtung an der Absicht — die Geltungsbereichsdurchsetzung hält die Agenten im Bewertungsmandat; das Modell vergibt niemals Endnoten.
- Menschliche Handlungsfähigkeit & Verlass — der PoC ist ausdrücklich Entscheidungsunterstützung, kein autonomer Entscheider; den Endbericht prüft ein menschlicher Analyst.

Die manipulationssichere Beweiskette wird durch strukturierte Ausgabeverträge, Zitatdurchsetzung und volle Observability verwirklicht: Jede Agent-zu-Agent-Nachricht folgt einem validierten Schema; jeder Beleg und jede Note tragen ein bis zu Dokument und Abschnitt auflösbares Zitat; Ausgaben ohne Zitat werden abgelehnt. Der Unterschied, auf dem das Framework besteht: nicht eine Zahl, die jemand eintippte", sondern eine Zahl, die ein benannter, verifizierbarer Prozess erzeugte".

- Pre-Flight (Eingang) — die Bereinigung entfernt aktive Inhalte; Eingangsschienen erkennen Injektionsmuster und weisen Anfragen außerhalb des Geltungsbereichs ab, bevor etwas in einen Agentenkontext gelangt.
- In-Flight (je Schritt) — Werkzeug-Positivlisten je Agent (geringstes Privileg); das Dateisystem-MCP ist auf designierte Verzeichnisse begrenzt; Werkzeugbudgets je Lauf verhindern Ausreißer.
- Post-Flight (Ausgang) — Schemavalidierung, Zitatdurchsetzung, Verankerungsprüfungen und Screening personenbezogener Daten, bevor irgendetwas in den Endbericht geschrieben wird.

Die verdiente Fähigkeit zeigt sich im Berechtigungsmodell: Die meisten Werkzeuge sind nur-lesend; nur der Berichtsersteller und ein begrenzter Dateischreiber mutieren — Schreibzugriff ist die am engsten vergebene Fähigkeit des Systems. Und die Funktionstrennung — kein Komponente urteilt und führt zugleich aus — trägt das Supervisor/Spezialisten-Muster selbst.

---

## 6. Erweiterung des Proof of Concept

Nach Etablierung des PoC erweiterte das Mandat ihn von einer einzelnen Bewertungsstufe zum vollständigen Digital-Roadmap-Kontinuum. Jede Erweiterung wurde auf dieselbe Weise hinzugefügt: eine neue Zone mit expliziten Eintrittspfändern, einer reinen Quellen- und Zitatrictlinie, einem konstruierten Abdeckungs-Gate-Fehlerpfad und derselben Herkunftsdisziplin. Neue Funktionalität lockerte nichts — unter den neuen Eingaben: eine über dreiundzwanzig Fähigkeitsdimensionen bewertete Reifegradanalyse und eine deterministisch bewertete Schmerzpunktanalyse.

---

## 7. Ergebnisse und Schlussfolgerung

Das ECCC-Mandat lieferte ein funktionierendes, lokales Multi-Agenten-System, das die Reife von Geschäftsfähigkeiten bewertet und digitale Roadmaps je Direktion aus autoritativen Quellen zusammenstellt — treu, transparent und wiederholbar. Ebenso wichtig: Es bewährte beide Enkidu-Frameworks in einer einzigen realen Implementierung.

Tangram gab dem Mandat eine operative Architektur, die ein Unternehmensarchitekt auf einen Blick liest: ein Kontinuum begrenzter Zonen, vertragliche Übergaben, reine Richtlinien und konstruierte Fehlerpfade. Das Trust-Assurance-Framework gab demselben Mandat eine verteidigbare Vertrauensposition: neun Komponenten, je durch einen konkreten Mechanismus verwirklicht, eine manipulationssichere Beweiskette hinter jeder Note und ein Zero-Trust-Geschirr, das jeden Schritt zur Inferenzzeit prüft. Beide treffen sich in der zentralen Regel des Systems — Determinismus, wo es zählt, Schlussfolgern, wo es hilft — zugleich Tangram-Richtliniengrenze und Trust-Assurance-Funktionstrennung.

Das Ergebnis ist das erklärte Ziel des Frameworks, real geworden: nicht maximales Vertrauen, sondern kalibriertes Vertrauen — Verlass im Maß der nachgewiesenen Vertrauenswürdigkeit — gestützt auf ein prüfbares Dossier, das ein Analyst, ein Auditor oder eine ECCC-Führungskraft befragen kann. Der PoC ist abgeschlossen; die Frameworks, die ihn formten, sind nun erprobt, und das Muster, das er etabliert, ist für die nächste gesteuerte, KI-gestützte Implementierung wiederverwendbar.

Kalibriertes Vertrauen, verdient Behauptung für Behauptung und Schritt für Schritt — und eine operative Architektur, lesbar genug, es zu beweisen.