

The AI governance plugin, governed

A practitioner's guide to Anthropic's ai-governance-legal plugin for Claude: how to run it well, where it performs, what to pair it with, and how to wire its outputs from Level 4, Complete, to Level 5, Substantive.

JASON CRUZ FALZON · AUGUST 2026

One plugin among twelve

On 12 May 2026 Anthropic launched Claude for Legal: twelve practice-area plugins and more than twenty connectors into the systems legal teams already run on, available to all paying Claude customers. One of the twelve is **ai-governance-legal**, a bundle of skills for the legal and compliance side of AI adoption, and a marker of how quickly the governance layer itself is being productised.

SKILL	WHAT IT DOES
cold-start-interview	Learns your governance practice from seed documents and writes the practice profile every other skill reads from.
policy-starter	Drafts a firm AI usage policy from published model policies, adapted to your profile.
ai-inventory	EU AI Act register: each system's role (provider, deployer, importer and so on) and risk tier, assessed per system rather than per company.
use-case-triage	Classifies a proposed AI use case as approved, conditional or not approved, with conditions and next steps.
aia-generation	Runs an AI impact assessment in your house format across the regulatory regimes in scope.
vendor-ai-review	Reviews vendor AI terms for training-on-data, liability, model-change and policy-consistency gaps.
reg-gap-analysis	Differs a new AI regulation or guidance against your current governance posture and produces a remediation plan.
policy-monitor	Scheduled sweep of saved assessments, triage results and vendor reviews for drift between policy and practice.
matter-workspace	Keeps one engagement's context separate from every other.

What this guide is. An independent practitioner's guide by Jason Cruz Falzon, based on the public anthropics/claude-for-legal repository and product documentation as read on 17 August 2026. It is not affiliated with or endorsed by Anthropic, and the plugin will evolve; verify details against the current repository before relying on them.

Five moves that decide the output quality

The plugin is a reference template until you configure it. Anthropic's own documentation says skipping setup is the single most common reason a skill produces generic output. These five moves are where the value is won or lost.

1 Run the cold-start interview before anything else

It writes the practice profile that every other skill reads. Feed it real seed material: your current AI policy, a prior impact assessment you consider good, and the vendor AI agreements that matter. Ten to twenty minutes here changes every downstream output. The two-minute quick start exists; treat it as a demo, not a setup.

2 Connect a research tool first

The suite tags citations that come through a live research connector with their source. Citations from model knowledge alone are flagged for verification, and if no research tool is connected at all, a reviewer note records that sources were not verified. The flag only protects you if someone resolves it, so wire the connector before the first real assessment.

3 Keep one matter workspace per engagement

Context separation is a control, not housekeeping. One client's positions bleeding into another's assessment is a failure mode the workspace mechanism exists to prevent. Use it from the first matter.

4 Schedule the drift monitor and treat it as load-bearing

The policy monitor sweeps saved assessments, triage results and vendor reviews for drift between what the policy says and what practice does. Run it weekly. A drift report nobody reads is drift with a timestamp.

5 Re-run the interview when practice shifts

New jurisdiction, new vendor stack, new policy: the profile is now wrong, and every skill reading it inherits the error. Profile drift is policy drift by another route. Edit the profile for small changes; re-interview for material ones.

Where it performs, and where it stops

WHERE IT PERFORMS

Detection at authoring time

Unverified citations get flagged as the document is produced. Deviations get logged rather than smoothed over. Drift is surfaced on a schedule. Detection lands exactly where it belongs: at the moment the work is made.

Discipline most registers fudge

The EU AI Act inventory forces role and risk tier to be assessed per system, not per company. That is the analysis regulators actually expect and spreadsheets quietly avoid.

Volume work at a cadence humans cannot hold

Vendor terms review, use-case triage and regulatory gap analysis are repetitive, pattern-matched work. Running them consistently, weekly, on everything, is where the tooling genuinely outruns a team.

WHERE IT STOPS

Nothing blocks

Every flag, drift report and deviation log lands on a human desk. No output of this plugin can prevent an action from happening. It advises; it does not refuse.

Review sits inside the process that produced the work

The stated control on every output is attorney review, and Anthropic is explicit that the reviewer carries responsibility for what leaves the building. The tooling makes review faster. It does not make it independent.

Production scales; the reviewer does not

Governance artefacts are about to become cheap and plentiful. The reading model of assurance does not survive contact with that volume. This is a design boundary, not a bug, and it is honestly declared.

What to run alongside it

The plugin is one vertical in a twelve-plugin suite. Three neighbours matter, and one of them is worth studying rather than merely using.

A regulatory-legal, the upstream half of drift

Its feed watcher and policy diff catch the regulation moving; the governance plugin catches your practice moving. Together they cover both ends of the gap. The gaps tracker gives the open findings a home so they cannot evaporate between sessions.

B privacy-legal and product-legal, the handoff routes

Use-case triage flags handoffs to privacy counsel (a GDPR DPIA is its own discipline) and to product counsel. product-legal matters for a second reason: its launch review connects to Linear, Jira or Asana launch trackers. That connector is the only thread in the suite that reaches into the delivery pipeline, which is the one place enforcement can actually live. Pull that thread.

C legal-builder-hub, the one to study

Before a community-built skill can enter your environment it passes a mandatory trust check, a hidden-content scan and a restrictive-by-default allowlist. A non-lawyer is routed to their attorney rather than given an install button. This is the one place in the entire suite where the control blocks instead of advising.

Read the asymmetry. Where code enters the environment, the builders installed a gate that refuses. Where governance documents leave it, they installed a reviewer. That placement tells you where they believe risk really lives, and it is the pattern the rest of your governance should copy.

Wiring the outputs to a gate

On the Cognitive Delivery Index, a control climbs from Absent through Informal and Documented to Level 4, Complete: the artefact exists, is current, and is maintained. Level 5, Substantive, asks a harder question: does the control demonstrably change behaviour at the point of action, enforced outside the process that produced the work? The plugin gets you to Complete faster than anything before it. Substantive is a wiring job, and it is yours.

OUTPUT	LEVEL 4 • COMPLETE	THE LEVEL 5 WIRING
AI usage policy	Policy drafted, adopted, kept current by the drift monitor.	A named pipeline check fails when a change violates a policy condition, such as an unapproved model or use case in the manifest.
System inventory	Every AI system has a role and risk tier recorded.	Inventory entry is a precondition of release. A system absent from the register does not deploy; absence itself is the blocking signal.
Use-case triage	Each use case classified approved, conditional or no, with conditions listed.	Conditional verdicts create tracked checks. Work on that use case cannot promote while a condition is open.
Impact assessment	An AIA exists in house format for each in-scope system.	The promotion gate verifies an AIA reference exists and matches the system. A missing assessment fails the gate rather than a reminder.
Vendor AI review	Terms reviewed against your positions; verdict recorded.	Procurement and integration block until the verdict is on record. Renewal and model-change clauses re-trigger the review.
Drift report	Drift between policy and practice surfaced weekly.	An open drift finding above threshold suspends the affected approval until it is closed, not merely acknowledged.

Six rows, one move: take the artefact's existence and its verdict, and make something outside the authoring process refuse to proceed without them.

The checklist, and the point

- ☐ The practice profile was written from real seed documents, not the quick start.
- ☐ A research connector is live, and someone owns resolving verification flags.
- ☐ The drift sweep runs on a schedule, and its findings have a tracked home.
- ☐ Each output on page six is mapped to a gate that can refuse, outside the authoring process.
- ☐ Review capacity has been sized against the new production rate, not the old one.

The point. The plugin industrialises the paperwork of being in control. Whether you are in control is decided at the gate, and the gate is still yours to build. A control that is written down is documentation; a control that demonstrably changes behaviour at the point of action is governance.

Sources. anthropics/claude-for-legal public repository and product documentation, accessed 17 August 2026 (github.com/anthropics/claude-for-legal). Claude for Legal launch of 12 May 2026 as reported across legal trade press. Product details will change as the repository evolves.



Jason Cruz Falzon · creator of the Cognitive Delivery Framework and the Cognitive Delivery Index
cognitivedelivery.co.uk · jason@cognitivedelivery.co.uk