ACCOUNTABLE AI & GOVERNANCE

Transparent reporting and *governance.*

Coa builds governance, accountability, and traceability into every AI product we ship, so executives and program leaders can measure efficacy, drift, and performance. We can help you do the same. The approach is laid out below.

Talk to us | See how we approach it

OPERATING PRINCIPLES

The five rules the AI *works within*.

Each rule is enforced in code and configuration — not prompt text — and holds across every product and every model.

AI proposes. People decide.

Every output is a proposal a person accepts, edits, or rejects. The model never acts on its own.

APPLICATION: Every proposal arrives with its basis; a rejection is recorded with its reasoning, not discarded.

Layered guardrails on every model.

Every prompt and every response is screened — grounding, prompt attacks, harmful content, sensitive data — before anyone acts on it.

APPLICATION: The same checks run whichever model does the work (e.g. Amazon Bedrock Guardrails).

Secure by design, isolated by default.

Encrypted in transit and at rest, least-privilege, scoped per case. Your data stays inside your boundary.

APPLICATION: Per-tenant and per-case isolation, AWS KMS encryption, private connectivity — and customer data never trains a shared model.

Everything on the record.

Every action is logged — the model, the inputs, the confidence, and what the person did with it. Reversible, reconstructable.

APPLICATION: An append-only history replays any action exactly as it happened.

Configurable and self-hostable.

Which model does the work is a setting. Managed, self-hosted, on-prem, or GovCloud — sensitive data never leaves to be read.

APPLICATION: Your model or ours, chosen per deployment and per task.

FORWARD-BASED ENGINEERING

How we design, build, and *run the system*.

AI does the slow parts of building software; engineers own every line. Automated gates catch regressions, bugs are fixed continuously, the running system is monitored, and what we learn in production feeds the next feature.

01

Doc-driven design

Spec the feature first.

02

AI-accelerated build

Engineers own every line.

03

Automated quality gates

Nothing merges unchecked.

04

Continuous & automated bug-fixing

Find and fix, on a loop.

05

Proactive monitoring

Watch it in production.

06

Feature development loop

Telemetry → next release.

CATCH AMBIGUITY BEFORE CODE

Doc-driven design

We write the feature as a specification and a real, interactive mockup before any code is committed. Design tokens and a binding reference spec keep every surface consistent and accessible by construction.

- One shared design system across product, docs, and site
- Mockups are the UI source of truth; the app is coded to match them
- Accessibility considered at design time, not bolted on

IN PRACTICE

- Shared design tokens
- Reference mockups
- Binding spec

AI WEAKNESSES & GUARDRAILS

Failure modes and the controls that *contain them*.

For each known LLM failure mode, we show the standard that names it, the control that addresses it, the enterprise service that implements the control, and what Coa adds on top.

On hallucination. Grounding reduces hallucination but does not eliminate it — the consensus across the field and in Anthropic's own guidance. It confirms an answer is

supported by the record, not that it is correct, so a person reviews every output before it takes effect.

✶	Hallucination / fabricated facts	OWASP LLM09 · NIST MEASURE
⚑	Prompt injection / jailbreak	OWASP LLM01
⊘	Sensitive-data / PII disclosure	OWASP LLM02
◐	Overreliance / automation bias	OWASP LLM09 · NIST MANAGE
⎘	Excessive agency	OWASP LLM06
⚖	Bias / inconsistent outcomes	NIST MEASURE (BIAS)
~	Model / data drift & degradation	NIST MEASURE → MANAGE
#	Supply chain / model provenance	OWASP LLM03
⤓	Improper output handling	OWASP LLM05
∞	Unbounded consumption / cost & DoS	OWASP LLM10

ACCURACY OVER TIME

How accuracy is measured and *stays high*.

Six mechanisms measure accuracy — per response, per change, per release, and continuously — and flag it when it slips.

Golden evaluation sets

PER RELEASE

Curated cases with known-correct answers; every model or prompt change is re-scored against them. Regressions surface before release, not in production.

WATCHES: Correctness vs. known-good answers

Contextual grounding + citation check

PER RESPONSE

Every answer is scored against its retrieved source, and every citation is checked against the record. Unsupported output is flagged or blocked before anyone relies on it.

WATCHES: Is each output supported by its cited source?

Scaled dry-run / diff

PER CHANGE

A change runs across many cases in read-only mode and shows up as a before→after diff — you see exactly what it would do before it touches live work.

WATCHES: What a change would do, before it applies

Human acceptance telemetry

CONTINUOUS

Accept, edit, and reject rates, tracked per feature and per reviewer. Rising edits flag a model that needs attention; suspiciously few flag rubber-stamping.

WATCHES: Accept / edit / override rates (over- and under-trust)

Drift & quality monitoring

CONTINUOUS

Inputs and outputs are watched for drift from the validated baseline, and an alarm fires the moment a metric crosses its threshold.

WATCHES: Drift & quality degradation vs. baseline

Red-team / adversarial testing

PER RELEASE

Prompt injection, jailbreaks, and edge cases, thrown at the system on purpose — we find the failure modes before real inputs or bad actors do.

WATCHES: Prompt injection & failure modes

Metric guardrails. Live accuracy figures depend on the deployment and its data. This section shows the assurance *machinery* — grounding checks, golden-set evaluations, scaled dry-runs, acceptance telemetry, and drift monitoring — not invented numbers.

STANDARDS & GOVERNANCE ALIGNMENT

Mapping to *recognized frameworks*.

Each control maps to a recognized AI risk framework or to current federal guidance, so an evaluator can trace it back to the standard that calls for it.

NIST AI Risk Management Framework

AI RMF 1.0 · 2023

Our controls map to the framework's four functions and its Generative AI Profile (NIST AI 600-1, 2024). NIST AI RMF is a voluntary framework; mapping to it shows alignment, not certification.

- Govern
- Map
- Measure
- Manage
- GenAI Profile (600-1)

OWASP Top 10 for LLM Applications

2025

Each guardrail in the matrix above maps to the specific LLM application risk it addresses, so you can see which control answers which weakness.

- LLM01 Prompt Injection
- LLM02 Sensitive-Info Disclosure
- LLM05 Improper Output Handling
- LLM06 Excessive Agency
- LLM09 Misinformation
- LLM10 Unbounded Consumption

Federal AI policy

OMB M-25-21 / M-25-22

Built to current federal guidance on agency AI use and acquisition, OMB M-25-21 and M-25-22 (both Apr 3, 2025), under EO 14179. Whether a deployment satisfies it is the agency CAIO's determination.

- Chief AI Officer
- High-impact AI practices
- Human oversight & appeals
- Model & data portability

VA Trustworthy AI Framework

2023

Aligns with VA's six trustworthy-AI principles. Claims processing is a named VA AI operational area, and VA has adopted OMB's high-impact-AI definition.

- Purposeful
- Effective & Safe
- Secure & Private
- Fair & Equitable

Transparent & Explainable | Accountable & Monitored

Mapping to OMB M-25-21's minimum practices for high-impact AI

OMB M-25-21 §4(b) defines minimum practices for *high-impact AI*, a category that adjudication decisions can fall into, and our controls map to each one. Whether a given deployment is high-impact, and whether it satisfies the memo, remains the agency CAIO's determination.

✓ **Pre-deployment testing**	→	Evaluation harness, golden sets, and a scaled dry-run before any change applies
✓ **AI impact assessment**	→	Grounded, auditable proposals and outcomes by issue type provide the evidence a written impact assessment draws on
✓ **Ongoing monitoring**	→	Drift and accuracy monitoring plus accept / edit / reject telemetry
✓ **Human training & assessment**	→	Role-specific guidance for adjudicators on what the AI does, its limits, and how to override it
✓ **Human oversight & accountability**	→	The AI proposes and a reviewer decides, accepting, editing, or rejecting every output and signing the work
✓ **Remedies / appeals**	→	The veteran keeps the full statutory appeal path (higher-level review, the Board, CAVC); AI involvement is disclosed so the decision stays contestable
✓ **End-user & public feedback**	→	Reviewer and end-user feedback is collected and folded into evaluation and threshold tuning
✓ **Vendor-lock-in protection (M-25-22)**	→	Configurable by model and provider, self-hostable, and portable

What alignment means here. Mapping to these voluntary frameworks shows alignment, not certification. Coa deploys on authorized infrastructure (self-hosted or AWS GovCloud), but those authorizations belong to AWS — Adjudicate holds no FedRAMP authorization or ATO today. Production pursues its own agency-sponsored ATO, de-risked by shadow and pilot runs before the AI affects any live decision.

SEE IT ON YOUR DEPLOYMENT

AI proposes. People decide.

Request a demo

Explore Adjudicate