TrustX: Governing Enterprise AI Risk in Development, Procurement, and Exposure

Anonymous submission

Abstract

Agentic AI governance is often characterized by isolated decisions. We assess whether an internal AI system is safe to deploy, whether a vendor AI product is acceptable for procurement, and whether an enterprise system is at risk of being exposed to specific AI applications. However, the separation of these assessments does not reflect real enterprise AI risk. We aim to bridge the gap across these three distinct risk surfaces through TrustX, a unified enterprise agentic AI risk framework that supports three areas: development, procurement, and exposure. The development component evaluates internally built or deployed agentic AI systems. The procurement component assesses the potential risk of introducing AI systems from vendors, SaaS, APIs, copilots, and third-party platforms. The exposure component identifies vulnerable systems and pathways within enterprises that external agentic AI systems can exploit. We propose a shared methodology across the three components built around module scoring frameworks that feed into a weighted composite formula, producing a risk score that can inform policy recommendations and controls. We demonstrate our framework's applicability by presenting use cases across domains, including financial services and healthcare.

Introduction

Agentic AI systems increasingly interact with enterprise information, whether it is through calling tools, triggering workflows, modifying records, drafting external communications, or coordinating across applications. Existing AI governance and assurance frameworks provide important foundations, including the NIST AI Risk Management Framework (National Institute of Standards and Technology 2023), ISO/IEC 42001 (ISO/IEC 2023), the EU AI Act (European Union 2024), OWASP guidance for agentic AI (OWASP 2024), MITRE ATLAS (MITRE 2024), and model risk management guidance such as SR 11-7 (Board of Governors of the Federal Reserve System 2011). However, these frameworks frame AI risk assessments as separate decisions. For example, evaluations regarding whether an AI tool built internally would introduce significant risk are treated as if they happen separately from discussions on how a new model release can create potential exposure risks for an enterprise. In short, the question is no longer whether an agentic AI system is accurate, fair, or secure in isolation. Instead, we must ask: "Where can AI systems act across the enterprise, and how can we evaluate these risk surfaces simultaneously if needed?"

In this paper, we introduce TrustX, a unified enterprise AI risk framework that views enterprise AI governance as one system with three entry points.

Contributions. Our contributions are as follows: (1) We introduce TrustX, a unified enterprise AI risk framework spanning development, procurement, and exposure with a shared methodology and composite scoring model. (2) We present the Build module, contributing a GPA+IAT agency classification model, a twelve-dimension risk scoring rubric with a critical-dimension approach, a five-level autonomy framework, and a specialised Coding Assistant extension grounded in real incident data. (3) We present the Buy module, contributing a four-dimension vendor risk scoring framework with system-type-aware gates and a Procurement Decision Dossier method. (4) We present the Protect module, contributing a five-dimension exposure scoring framework, a discovery path for unknown AI integrations, agentic threat testing across eight exposure pathway types, and sector calibration rules. (5) We propose a gate-first multi-module aggregation model that prevents mandatory gate failures from being diluted by composite averaging.

Overview of TrustX Framework

TrustX is organized as one enterprise tool with three module entry points, as shown in Table 1. Each module addresses a distinct enterprise AI risk surface and produces an independent scored output; when multiple modules apply to a linked deployment, their outputs are aggregated through a shared gate-first resolution model.

The **Build** module evaluates AI systems that are built or deployed internally, covering seven agentic AI system types: Autonomous Agents, Coding Assistants, Decision Support Systems, AI Embedded/Physical Agents, Knowledge Assistants, Tool-Using Agents, and Transaction/Commerce Agents. Assessment proceeds through six sections: agent identification, GPA+IAT agency classification, twelve-dimension risk scoring, autonomy level assessment (L1–L5), additional risk factors, and recommended governance controls. A specialised Coding Assistant extension replaces the GPA+IAT and autonomy sections with a Capabilities Assessment and Deployment Model Classification (Models 1–4), reflecting the categorically distinct risks

Module	Full Name	Core Question
Build	Agentic AI Risk Classification	Is this system safe to deploy?
Buy	Agentic AI Risk Procurement	What agentic AI risk are we bringing in?
Protect	Agentic AI Risk Exposure	What systems can the agentic AI reach or act upon?

Table 1: The structure of the TrustX Framework.

of executable output and direct system access.

The **Buy** module evaluates AI systems introduced through vendors, SaaS products, APIs, copilots, and third-party platforms. Assessment begins with system type selection across eight vendor AI product types, followed by a procurement screen, eleven-object evidence review, four-dimension scoring, mandatory gate and red flag review, and a Procurement Decision Dossier.

The **Protect** module evaluates enterprise systems, data assets, workflows, and exposure pathways that AI systems can reach or act upon. Assessment begins with a discovery path to identify active AI integrations, followed by exposure pathway type selection across eight pathway types, a quick exposure screen, six-surface inventory, pathway decomposition, five-dimension scoring, and agentic threat testing that evaluates controls as attack surfaces rather than governance checkboxes.

Module Scoring Frameworks

Each module produces a composite score on a 0–100 scale where higher values indicate lower risk and stronger readiness. Scores are computed from weighted dimension scores using the shared formula:

$$\text{Module Composite} = 100 - [(\text{Risk Score} - 1) \times 25] \quad (1)$$

where Risk Score is the weighted sum of dimension scores (each rated 1–5). Mandatory gate failures and red flags override the score-based tier; the score then serves as a within-tier ranking signal only.

Build Module: Agency Classification and Risk Scoring

The Build module applies a dual-track methodology. The standard track covers all agent types except Coding Assistants; the Coding Assistant track applies a fully differentiated structure grounded in the supply-chain, executable-output, and direct-system-access properties unique to that class.

GPA+IAT agency classification. Six properties determine agency level: Goal, Perception, Action (constituting Minimal Agency when all present), plus Iteration, Adaptation, and Termination (constituting a Full Agentic System when all six are present). Systems lacking core GPA properties are classified Non-Agentic. Classification is binary per property and automated from evidence.

Twelve-dimension risk scoring. Each dimension is scored 1 (Low), 2 (Medium), or 3 (High). The dimensions are: Autonomy, Decision Scope, Temporal Coupling, Action Authority, System Reach, Blast Radius, Persistence, Reversibility, Control Authority, Data Sensitivity, Aggregation Risk, and Data Egress Paths. Three computed outputs are derived: Total Score, Average Score, and Highest Individual Dimension score.

Critical-dimension tier determination. A single dimension score of 3 triggers Tier 3 (High Risk) regardless of the average. This prevents high-risk properties from being diluted. Full logic: Tier 3 if any dimension scores 3 or autonomy level is L5; Tier 2 if maximum dimension is 2 or average ≥ 1.5; Tier 1 if all dimensions ≤ 1 and average < 1.5.

Autonomy level assessment. The L1–L5 framework characterises human–agent interaction: L1 (Operator) through L5 (Observer, full autonomy under monitoring only). Risk mapping: L1–L2 $\Rightarrow$ Low; L3–L4 $\Rightarrow$ Medium; L5 $\Rightarrow$ High. L5 is a standalone Tier 3 trigger.

Governance output. Tier 1 (Standard) requires documentation, basic HITL, and audit logs. Tier 2 (Enhanced) adds behavioural boundaries, a kill switch, and enhanced monitoring. Tier 3 (Rigorous) requires third-party validation, continuous monitoring, and regulatory reporting. Controls are drawn from the TrustX Build control catalog (RAI-GOV, RAI-SEC, RAI-SAFE series) aligned to NIST AI RMF, EU AI Act, and ISO/IEC 42001.

Coding Assistant extension. Replacing GPA+IAT, a Capabilities Assessment evaluates 20 discrete capabilities across code generation, system interaction, and autonomy groups. Replacing L1–L5, four Deployment Models characterise risk: Model 1 (IDE autocomplete, Tier 1), Model 2 (file-level agent, Tier 2), Model 3 (autonomous multi-step, Tier 2–3), Model 4 (production-connected, Tier 3). The same twelve dimension names apply with coding-specific tier descriptors. A 22-control RAI-CA security catalog covers supply chain integrity, agent manipulation, and access and privilege abuse.

Special modifiers apply across both tracks: financial services customer-facing agents require a Tier 2 minimum; multi-agent systems inherit the highest tier of any constituent agent; and reassessment is triggered by any change in capabilities, tooling, or deployment context.

Buy Module: Four-Dimension Vendor Risk Scoring

The Buy module evaluates vendor-supplied AI through four weighted dimensions. Assessment begins with **system type selection** across eight vendor AI product types: AI Assistant/Knowledge Product (Type 01); Content/Code Generation Product (Type 02); Decision Support/Scoring Product (Type 03); Tool-Using SaaS/Copilot (Type 04); Transaction/Workflow Automation Product (Type 05); Autonomous Vendor Agent (Type 06); Embedded/Platform AI Feature (Type 07); and AI Infrastructure/Model Provider (Type 08). All downstream scoring, gates, recommendation sets, and decision logic inherit from the selected system type. Table 2 lists the four dimensions, weights, and scoring rules. The Risk Score is:

$$\begin{aligned}\text{Buy Risk} = {} & 0.35D_{\text{AA}} + 0.25D_{\text{Au}} \\ & + 0.25D_{\text{SR}} + 0.15D_{\text{Pe}}\end{aligned} \quad (2)$$

where D_{AA} is Action Authority, D_{Au} is Autonomy, D_{SR} is System Reach, and D_{Pe} is Persistence. Each dimension score is derived from its sub-dimensions as specified in Table 2.

Each sub-dimension is scored 1 (Minimal) to 5 (Critical) against defined anchors. A score of 1 indicates minimal vendor AI risk (e.g., read-only access, no autonomous action); a score of 5 indicates critical risk that triggers a red-flag evaluation. Dimension scores at or above 4 trigger mandatory gates; any dimension at 5 triggers a red flag that blocks procurement unless formally escalated with executive approval.

Mandatory gates for Buy include: a data processing agreement with AI-specific terms; prompt-injection coverage across direct, indirect, tool, and memory surfaces; technically enforced human-in-the-loop (HITL) where required; a complete tool and action manifest; and schema-level logging exportable to the enterprise SIEM. Any gate failure raises the minimum outcome to Tier 3. **Red flags** that block procurement include: refusal to disclose model identity or version; no rollback mechanism for AI-initiated actions; vendor-summary-only logs that are not enterprise-exportable; tool permissions enforced only by model judgment; no model-change notification SLA; and refusal of audit rights.

Outcome tiers for Buy map as: Tier 1 (85–100, Cleared); Tier 2 (70–84, Conditional); Tier 3 (50–69, Restricted); Tier 4 (below 50, Block). Tier 2 requires documented conditions implemented before integration and reassessment within 90 days. Tier 3 prohibits high-impact actions and critical-system write access until remediation. Tier 4 blocks integration until critical controls are independently verified. The final outcome follows a six-step decision logic hierarchy: (1) red flags override everything and block unless formally escalated with executive approval; (2) mandatory gate failures set a minimum of Tier 3; (3) the composite score determines the baseline tier; (4) the selected system type determines the applicable control set; (5) enterprise-side feasibility determines final conditions; (6) the residual risk owner accepts any remaining unmitigated risk. The Procurement Decision Dossier focuses exclusively on controls the buyer can implement—procurement conditions, contract terms, access restrictions, data minimization, integration constraints, HITL gates, monitoring and logging, rollout limits, and reassessment triggers—because the enterprise generally cannot require the vendor to modify the AI product itself.

Control inheritance follows four types: *Universal* controls apply to all system types regardless of score or tier; *Inherited* controls apply based on the selected system type, with higher-risk types inheriting lower-risk controls plus type-specific additions; *Adapted* controls are universal controls with modified parameters for specific system types or regulated sectors; and *Module-specific* controls apply only to the named system type.

Reassessment is triggered by changes including: model version; system prompt; tool or action scope; subprocessor; data category accessed; external communication enablement; write or delete permission; identity or credential scope; financial transaction scope; OAuth scope; workflow or process changes affecting AI decision points; privilege expansion; major vendor release; incident or near-miss; log export failure; and annual renewal. Each trigger is causally linked to a specific scoring dimension: for example, a model version change may invalidate prior scoring across all dimensions, while an OAuth scope addition requires re-evaluation of System Reach and Action Authority only.

Protect Module: Five-Dimension Exposure Scoring

The Protect module evaluates the enterprise surface that AI systems can reach or act upon through five weighted dimensions. Table 3 lists the dimensions, weights, and scoring rules. Assessment begins with **exposure pathway type selection** across eight types: API/Integration Gateway (ARE-01); Data Access/Retrieval (ARE-02); Agent Orchestration/Workflow (ARE-03); Critical Business System (ARE-04); Human Approval/Decision Interface (ARE-05); External Communication/Egress (ARE-06); Identity/Credential/Privilege (ARE-07); and Shadow/Unknown AI (ARE-08). A single assessment may cover multiple pathway types; each is assessed independently and the worst tier governs the overall Protect outcome. The Risk Score is:

$$\begin{aligned}\text{Protect Risk} = {} & 0.25D_{\text{SR}} + 0.30D_{\text{AA}} + 0.20D_{\text{DS}} \\ & + 0.15D_{\text{Ob}} + 0.10D_{\text{CE}}\end{aligned} \tag{3}$$

where D_{SR} is System Reach, D_{AA} is Action Authority, D_{DS} is Data Sensitivity, D_{Ob} is Observability, and D_{CE} is Control Enforcement.

The Protect module is surface-centric rather than vendor-centric: it evaluates the enterprise system or exposure pathway, not the AI product that reaches it. Assessment begins with a discovery path—scanning SaaS and platform admin consoles, interviewing system and data owners, and reviewing API and integration logs—to produce a verified integration inventory. Six exposure surfaces are then reviewed: APIs and integration gateways; data access and retrieval systems; agent orchestration and workflow layers; critical business systems; human approval and decision interfaces; and external communication and egress channels. Each surface is decomposed into an exposure pathway specifying the entry point, accessible assets, permitted actions, downstream effects, and control dependencies.

Agentic threat testing extends the surface inventory by evaluating five threat classes: prompt injection via retrieved content; indirect prompt injection via external data; AI-assisted credential harvesting; autonomous workflow manipulation; and agent-to-agent instruction injection. Control coverage is assessed for each threat scenario and gaps are reflected in the Observability and Control Enforcement dimension scores.

Mandatory gates for Protect include: schema-level logging of all AI-initiated actions exportable to enterprise SIEM; technical HITL for financial transactions, credential changes, and delete or archive actions; a complete tool and action manifest for every AI integration touching the assessed system; prompt-injection testing covering all AI-reachable data surfaces; and a shutdown mechanism that does not require model cooperation. Any gate failure raises the minimum outcome to Tier 3. **Red flags** that block include: AI can create, modify, or delete credentials or access

Dimension	Wt.	Sub-dimensions	Scoring Rule
Action Authority	35%	Write/modify; financial & transactional; credential & identity; external communication; tool permission enforcement	Max of sub-dims. Any sub at 5 forces dimension to 5.
Autonomy	25%	Planning & reasoning depth; human intervention; goal scope enforcement; vendor change notification	Mean of sub-dims, rounded to nearest 0.5. Vendor Change Notification at 5 adds 0.5 to mean before rounding.
System Reach	25%	Connected system count; data store access breadth; user population affected; lateral movement potential	Max of sub-dims. Lateral Movement at 5 forces overall to 5.
Persistence	15%	Memory scope across sessions; enterprise data retention by vendor; model training use of enterprise data; state continuity across users	Mean of sub-dims. Model Training Use at 5 adds 0.5 to mean before rounding.

Table 2: Buy module dimensions, weights, sub-dimensions, and scoring rules.

Dimension	Wt.	Sub-dimensions	Scoring Rule
System Reach	25%	Connected system count; user population affected; lateral movement potential; critical surface exposure	Max of sub-dims. Lateral Movement at 5 forces overall to 5; triggers network segmentation review.
Action Authority	30%	Write & modify; financial & transactional; credential & identity; external egress; approval & decision influence	Max of sub-dims. Credential or financial at 5 each force overall to 5 and trigger a red flag.
Data Sensitivity	20%	Highest data classification; records accessible per session; data minimisation; regulatory obligations	Max of sub-dims. Score $\geq$4 triggers data-owner gate and DLP review. Score 5 requires legal sign-off.
Observability	15%	Schema-level logging; real-time alerting; rollback capability; audit trail integrity	Max of sub-dims. Logging absent forces mandatory gate. Non-exportable logs are a red flag.
Control Enforcement	10%	HITL enforcement; shutdown mechanism; action ceiling; prompt injection isolation	Max of sub-dims. HITL or Shutdown at 5 each force overall to 5 and trigger a red flag.

Table 3: Protect module dimensions, weights, sub-dimensions, and scoring rules.

roles; logs are not exportable to enterprise systems; financial or identity action authority is enforced only by model judgment; no rollback mechanism for AI-initiated actions; and HITL exists in policy but is not technically enforced.

Outcome tiers for Protect use the same 0–100 scale as Buy: Tier 1 (85–100, Cleared); Tier 2 (70–84, Conditional); Tier 3 (50–69, Restricted); Tier 4 (below 50, Block). The final outcome follows a five-step decision logic hierarchy: (1) red flags override everything and block unless formally escalated; (2) mandatory gate failures set a minimum of Tier 3; (3) the composite score determines the baseline tier; (4) sector calibration may elevate one tier for financial services, healthcare, critical infrastructure, and identity pathways; (5) final conditions are set by enterprise-side feasibility and residual risk owner acceptance.

Control inheritance follows the same four types as Buy—Universal, Inherited, Adapted, and Module-specific—applied at the pathway level rather than the system type level.

Reassessment scope is proportionate to trigger type: a model version change triggers a full re-score across all dimensions; a single OAuth scope addition requires re-evaluation of System Reach and Action Authority only. Triggers include: model version change; system prompt change; tool or action scope change; new subprocessor; new data category accessed; external communication enabled; write or delete permission added; identity or credential scope expanded; financial transaction scope added; OAuth scope change; workflow or process changes affecting AI decision points; privilege expansion; vendor update to any integrated AI component; major vendor release; incident or near-miss; log export failure; and annual renewal.

Eight exposure pathway types structure the Protect Decision Dossier: API/Integration Gateway; Data Access/Retrieval; Agent Orchestration/Workflow; Critical Business System; Human Approval/Decision Interface; External Communication/Egress; Identity/Credential/Privilege; and Shadow/Unknown AI. Each pathway is assessed independently and may yield a different tier; the worst tier governs the overall Protect outcome for that pathway.

Module Selection and Multi-Module Aggregation

A unified enterprise AI risk tool must accommodate the reality that different buyers enter the governance process from different vantage points. A security team may be concerned primarily with what AI systems can reach across existing enterprise infrastructure. A procurement team may be evaluating a specific vendor copilot. An engineering team may be assessing an internally built agent. In many cases, the same deployment triggers obligations across all three dimensions simultaneously. TrustX addresses this through a structured module selection interface that routes users to the appropriate assessment entry points and aggregates results when

Routing Question	Module
Are you evaluating something your team is building or has built?	Build
Are you evaluating something a vendor is providing or you are procuring?	Buy
Are you evaluating what AI can reach in your existing systems?	Protect

Table 4: Intake routing questions and corresponding modules.

more than one module applies. Figure 1 illustrates the full flow.

Intake Triage and Module Routing

Before any assessment begins, TrustX presents three plain-language routing questions. Users may select one or more; each selection activates the corresponding module (Table 4).

This triage step precedes, but does not replace, the Gate 0 checklists within each module. Its purpose is to ensure that buyers who do not know which module they need are routed correctly, and that buyers who need multiple modules are not required to recognise that obligation on their own. For example, an enterprise enabling a vendor AI feature in an existing SaaS product will typically select both the Buy and Protect questions, automatically triggering Buy for the vendor system and Protect for the enterprise systems the integration can reach. An internal engineering team deploying a coding agent that calls a third-party foundation-model API and accesses source repositories may select all three. Where only one module is selected, TrustX proceeds as a single-module assessment. Where two or three are selected, TrustX activates the multi-module aggregation layer described below.

Cross-Module Pathway Linking

When multiple modules run on a related deployment, assessors assign a shared pathway identifier to each linked assessment. This connects, for example, the Buy record for a vendor CRM copilot with the Protect record for the CRM system it accesses. Linked assessments share evidence where applicable and their scores are aggregated rather than treated as independent results.

Multi-Module Composite Scoring

When two or three modules produce scores for a linked pathway, TrustX computes a multi-module composite using equal weighting across activated modules by default. Sector calibration modifiers defined in each module apply before aggregation. For an n-module assessment:

$$\text{TrustX Composite} = \frac{1}{n} \sum_{i \in \{\text{Build, Buy, Protect}\}} \text{Composite}_i \tag{4}$$

where n is the number of activated and linked modules and each Module Composite is the 0–100 score from that module's own formula. Higher scores indicate lower enterprise AI risk and stronger overall readiness. Equal weighting is the appropriate default because the relative importance of Build, Buy, and Protect risk varies by organization type, sector, and deployment context; future calibration work may produce empirically validated sector-specific weights.

Modules Activated	Tier Resolution	Composite Use
One module only	Single-module tier	Ranking within tier
Two or three, no gate failures	Worst single-module tier	Within-tier ranking
Any gate failure	Min. Tier 3; red flag forces Tier 4	Transparency only

Table 5: Gate-first tier resolution rules.

Gate-First Tier Resolution

The composite score determines ranking within a tier but does not override gate and red-flag logic. Tier resolution follows a gate-first rule: the enterprise AI risk tier for a linked pathway is the worst tier produced by any activated module before composite averaging is applied. Table 5 summarizes the resolution rules. A linked deployment where Buy produces Tier 2 and Protect produces Tier 3 resolves to Tier 3 regardless of the aggregate composite score. This is consistent with the override logic applied within each individual module.

Buyer-Stratified Report Output

A single multi-module assessment produces one underlying result but supports three report views. A *technical annex* serves CISO and security engineering with full dimension scores, gate evidence, and failure-mode mapping. A *procurement summary* serves legal and procurement with the vendor decision, required contract conditions, residual risk, and reassessment schedule. An *executive summary* serves the CRO, board, and audit committee with the aggregate exposure tier, top three risk findings, and remediation timeline.

Illustrative Multi-Module Application

An enterprise enabling an AI-powered invoice-processing workflow illustrates the multi-module path. The workflow is vendor-supplied (Buy selected), connects to the enterprise ERP and payments system (Protect selected), and uses an internally configured routing layer (Build selected). Build classifies the routing layer at Tier 2 given limited autonomy and bounded tool access. Buy scores the vendor workflow at Tier 3 due to financial action authority and incomplete HITL evidence. Protect scores the ERP exposure at Tier 4 because schema-level logging is absent and financial action authority is enforced only by model judgment, triggering a red flag. Under gate-first resolution, the combined pathway resolves to Tier 4 Block.

Discussion and Conclusion

A single-tool design prevents three failure modes common in enterprise AI governance. First, it prevents build teams from deploying agents without recognizing that their tool

reach triggers Protect assessments. Second, it prevents procurement teams from approving vendor AI based only on vendor security documentation while ignoring enterprise integration pathways. Third, it prevents security teams from treating exposure as a perimeter problem when AI systems operate through legitimate authenticated workflows. Future work should empirically validate score weighting, calibrate sector-specific multipliers, and test inter-rater reliability.

Enterprise AI risk arises from three linked sources: what organizations build, what they buy, and what their systems are exposed to. TrustX addresses this by providing one tool with three modules sharing a behavior-centric methodology, evidence hierarchy, lifecycle controls, failure-mode taxonomy, composite scoring, override logic, and verification output model.

References

Board of Governors of the Federal Reserve System. 2011. SR 11-7: Guidance on Model Risk Management.

European Union. 2024. Regulation (EU) 2024/1689 on Artificial Intelligence.

ISO/IEC. 2023. ISO/IEC 42001:2023 Artificial Intelligence Management System Standard.

MITRE. 2024. MITRE ATLAS: Adversarial Threat Landscape for AI Systems.

National Institute of Standards and Technology. 2023. AI Risk Management Framework (AI RMF 1.0).

OWASP. 2024. Agentic AI: Threats and Mitigations; LLM Top 10.

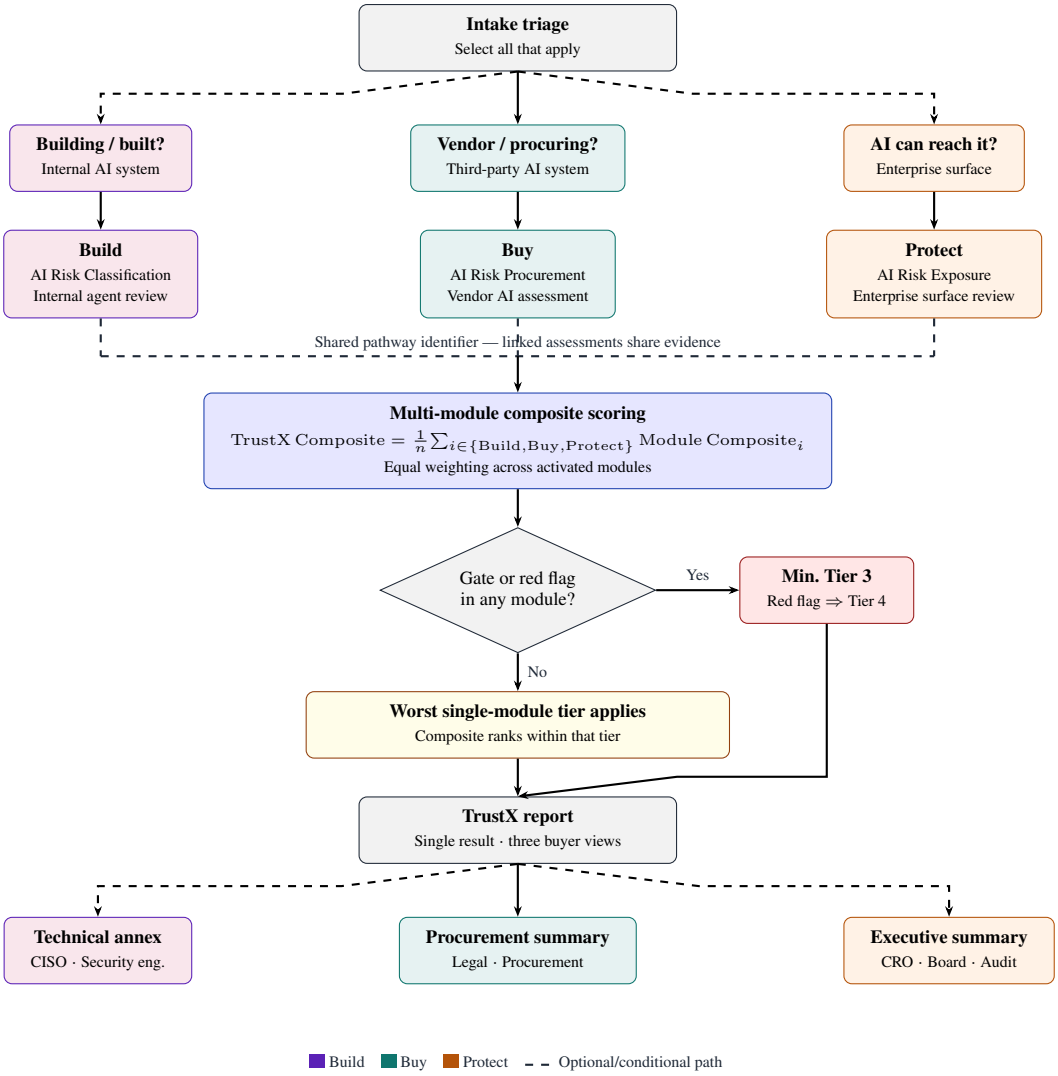

Figure 1: TrustX module selection and multi-module aggregation flow. Users select one or more intake questions; each activates a module. Linked assessments share a pathway identifier and are aggregated via gate-first tier resolution into a single TrustX report with three buyer-stratified views.