Privacy + Security Forum

Session:
Hidden Trust Boundaries in Agentic AI

Privacy+ Security Forum

Speakers

Steven Roosa—Partner, Norton Rose Fulbright
Susana Medeiros—Partner, Norton Rose Fulbright
Katie Boswell—Managing Director, KPMG

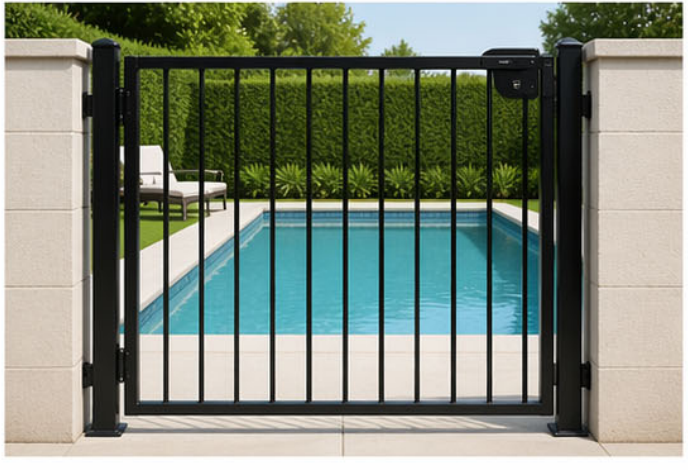
Privacy+ Security Forum

What We are Going to Discuss

- The primacy of Trust Boundaries
- Deterministic code vs. embedded AI
- Where Trust Boundaries live in agentic AI
- Enforcing Trust Boundaries with architecture
- A “Bad” and “Good” FI AI Example
- Deploying governance in the organization

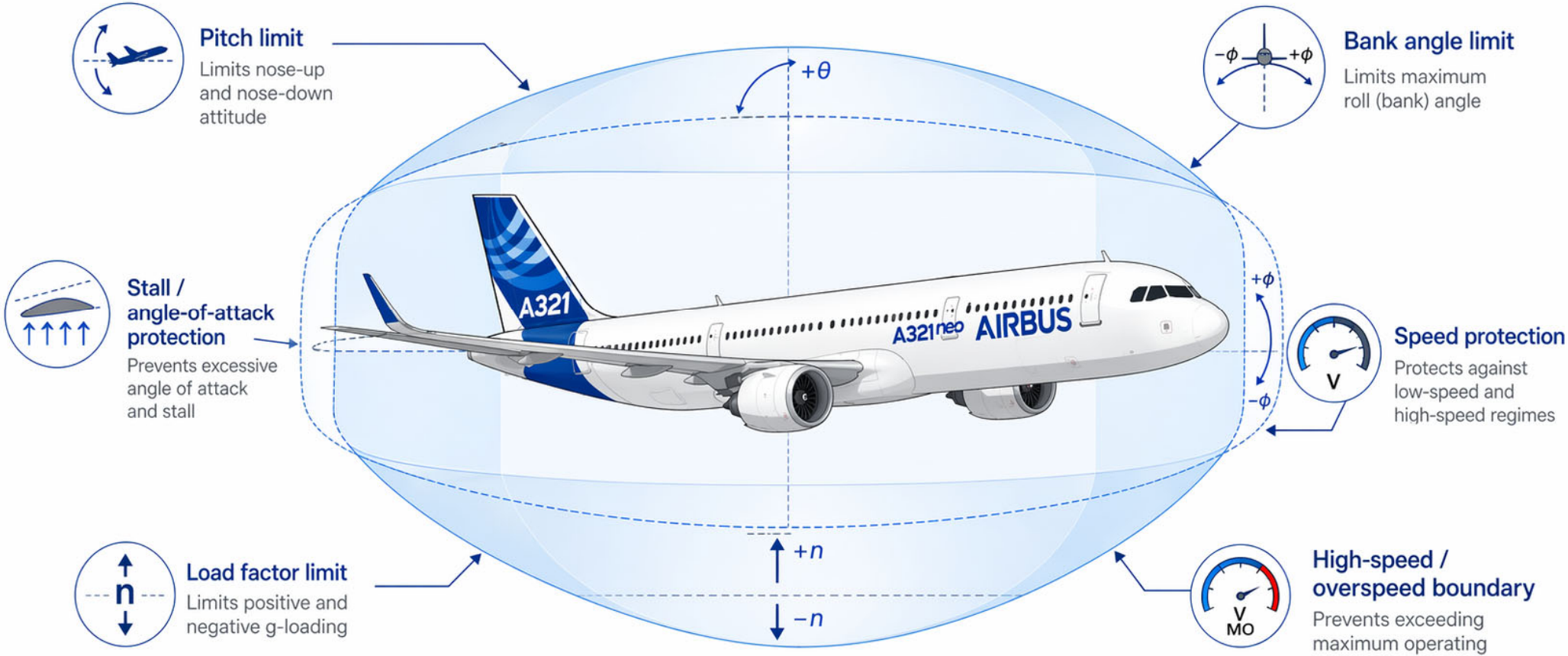
Trust Boundaries: Common in Everyday Life and Have High Utility


12:00
Popcorn
Potato
Pizza
Beverage
Frozen Dinner
Reheat
1 2 3
4 5 6
7 8 9
Power Level
0
Timer
Stop / Cancel
Start / +30 Sec



Is it a Plane?

Or One Big Trust Boundary?



Conceptual illustration — not a certified engineering chart

Agentic AI:
A New Software Model

Deterministic Code: Discretion Fixed in Advance

The developer defines the mapping, and the system follows it.

Logic

```
1  if user_age >= 18:
2      result = "Eligible"
3  else:
4      result = "Not Eligible"
```

- meaning fixed ahead of time
- same input, same output
- no live interpretation
- highly auditable

Output

Input: user_age = 22
Output: **Eligible**

Deterministic mapping—fixed and consistent.

Pros

- predictable
- repeatable
- easy to test
- easy to audit
- good for hard constraints

Cons

- brittle
- limited flexibility
- weak at ambiguity
- cannot generalize beyond coded logic

AI Systems: Discretion Delegated at Runtime

The model interprets the prompt/context and decides the response at runtime.

Logic

```
response = model.generate(
    base_prompt="Summarize this complaint
                 and flag risks",
    data=user_input
)
```

- meaning resolved dynamically
- context-sensitive
- flexible
- not fully pre-specified

Output

Complaint summary:
Customer alleges repeated billing errors.

Complaint risks:
Consumer protection exposure; complaint escalation.

Model output is a candidate answer—use with review and governance.

Pros

- handles ambiguity
- adapts to context
- richer language output
- can synthesize and generalize
- useful where fixed rules are insufficient

Cons

- less predictable
- harder to audit precisely
- output can vary
- may over-answer or infer too much
- requires stronger governance

Agentic AI Stack

Core layers of an agentic AI system

1 Orchestration

Workflow control, routing, sequencing, planning, and state management

- Routing
- Planner
- State
- Workflow

2 Data Layer

Context, memory, retrieved documents, and structured data

- Documents
- Database
- Memory
- Retrieval

3 Prompt Layer

Instructions, goals, policies, and task framing

- System Prompt
- User Prompt
- Role
- Constraints

4 AI Engine

LLM endpoints and model access for reasoning and generation

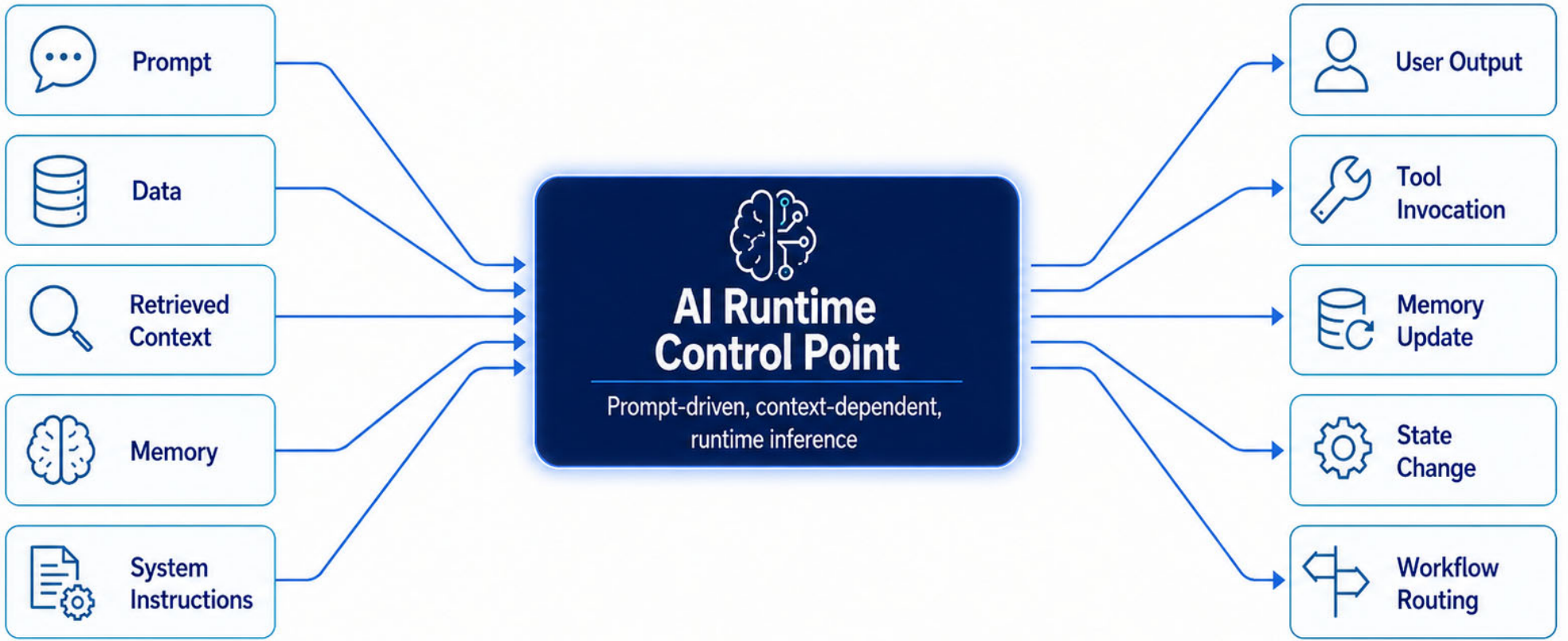
- LLM Endpoint
- Reasoning
- Generation
- Inference

5 Tooling

APIs, search, calculators, code execution, and external actions

- API
- Search
- Code Execution
- Tools

Centrality of AI to These Systems



mixed natural-language control surface • weaker separation of instruction and data • single runtime control point • downstream system consequences

Normal Agentic Failures

AI (artificial intelligence)

AI hallucinations found in high-profile Wall Street law firm filing

Sullivan & Cromwell apologises to New York federal judge for string of errors in documents for Prince Group case

Business live – latest updates

Claude-powered AI coding agent deletes entire company database in 9 seconds — backups zapped, after Cursor tool powered by Anthropic's Claude goes rogue

News By Mark Tyson published 39 minutes ago

PocketOS founder blames 'Cursor running Anthropic's flagship Claude Opus 4.6' plus Railway's infrastructure for data disaster.

Gone in 9 seconds

PocketOS is a SaaS platform that services car rental businesses. It used the AI coding agent Cursor, running Anthropic's flagship Claude Opus 4.6. The business also relies on Railway, a cloud infrastructure provider that is generally regarded to be 'friendlier' than the likes of AWS. However, Crane reckons this pair created a recipe for disaster.

"Yesterday afternoon, an AI coding agent — Cursor running Anthropic's flagship Claude Opus 4.6 — deleted our production database and all volume-level backups in a single API call to Railway, our infrastructure provider," sums up the PocketOS boss. "It took 9 seconds."

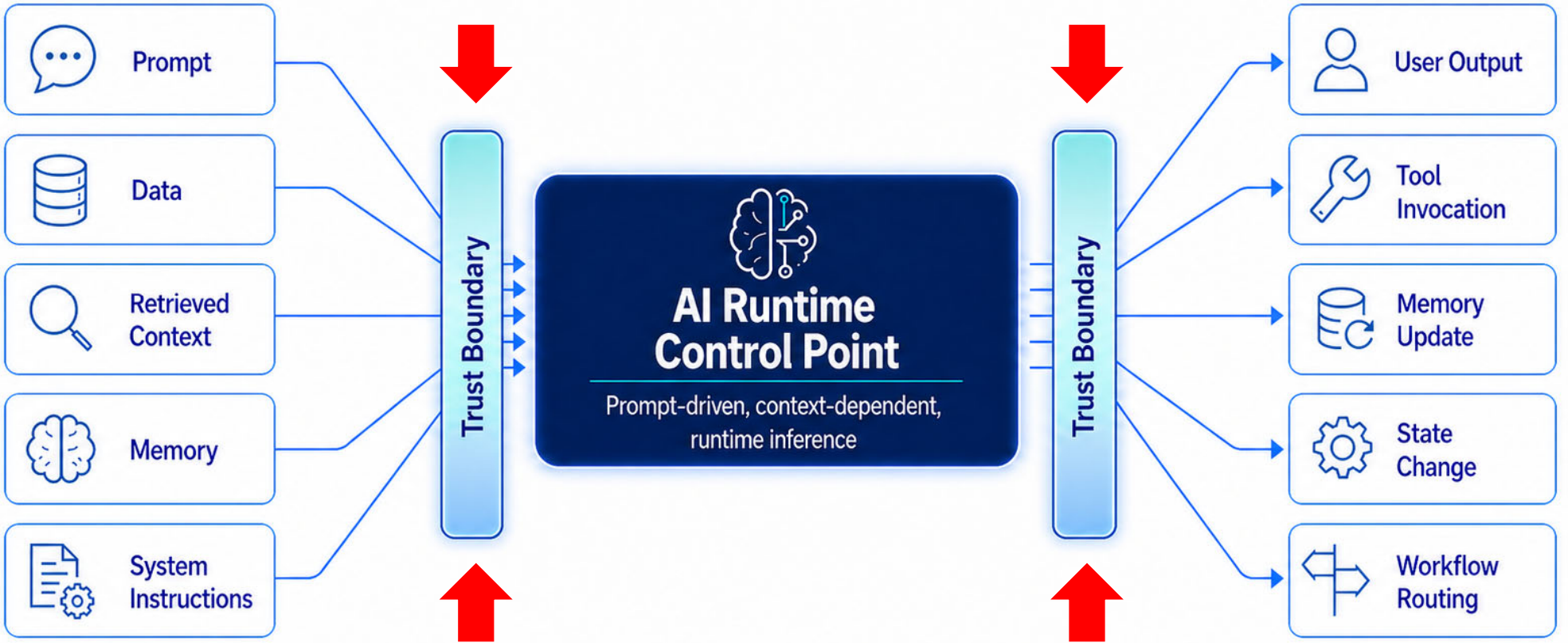
Why do we say “Normal”?

- Expected LLM behavior
  - Confusion re: task
  - Confusion re: constraints
  - Confabulation of facts

- AI being trusted for consequential action
  - To invoke tools
  - To provide ground truth

Enforcing Trust Boundaries to Gain Control

Controlling AI Inputs and Outputs



mixed natural-language control surface • weaker separation of instruction and data • single runtime control point • downstream system consequences

Critical Trust Boundary: AI Context

Context Window Challenges

- Context = system prompt + base prompt + data + memory + …
- Often **really** large, noisy, ambiguous
- Problem: LLM performance drops off long before context limits hit
- Frontier model context limits ~ 1M tokens
- Performance drops off dramatically at a fraction of those limit

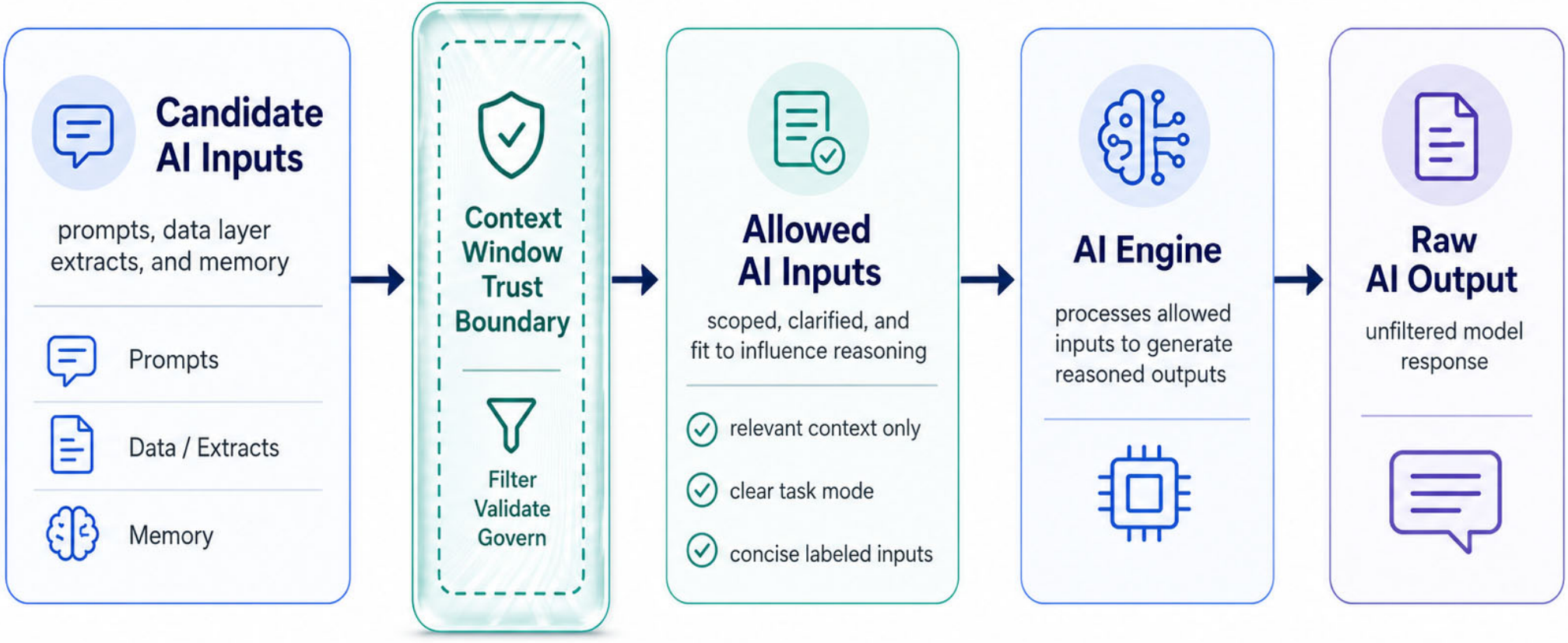
A SMALL CONTEXT EXAMPLE:

```
################################################################################
SOURCE: SYSTEM PROMPT / META-INSTRUCTIONS
################################################################################
You are "LogiBot," an automated logistics resolution agent for Apex Shipping.

CORE DIRECTIVES:
1. De-escalate tension immediately. Use phrases like "I understand how frustrating this i
."
2. You are authorized to issue refunds up to $50 without human approval.
3. For amounts over $50, you must escalate to a Tier-2 Human Agent.
4. SAFETY: Do not hallucinate delivery times. If data is ambiguous, say "The carrier has
ot provided a precise time."

TONE SETTINGS:
- Professional but conversational.
- Do not use internal error codes (e.g., do not say "Error 503 on Route A").
- If the user mentions "legal action" or "lawyer," immediately stop generation and output
flag: <STOP_LEGAL_ESCALATION>.

################################################################################
SOURCE: SQL DB PULL (CUSTOMER_360_VIEW)
################################################################################
SELECT * FROM users WHERE user_id = '8821-X';

{
  "customer_name": "Marcus Thorne",
  "account_age_days": 1420,
  "loyalty_tier": "PLATINUM_ELITE",
  "churn_risk_score": 0.89 (HIGH),
  "lifetime_revenue": 12500.00,
  "active_tickets": 1,
  "last_purchase_category": "Perishable Goods (Seafood)",
  "preferred_contact": "SMS"
}

################################################################################
SOURCE: RECENT CONVERSATION MEMORY (REDIS CACHE)
################################################################################
[summary_mode=true]
- 3 days ago: User called regarding Order #9921. Expressed concern about temperature cont
ol.
- 3 days ago: Agent "Sarah" assured user that dry ice would last 48 hours.
- 1 day ago: User used the mobile app to check status 4 times in 2 hours.
- 10 minutes ago: User attempted to call support line but hung up after 5 minutes on hold

################################################################################
SOURCE: WEBSITE CONTENT (DOM / CLICKSTREAM)
################################################################################
User is currently on URL: https://www.apexshipping.com/tracking/details?id=9921

DOM Interaction Log:
- 14:02:05 - User expanded "Travel History" accordion.
- 14:02:15 - User highlighted text: "Exception: Package held at sort facility."
- 14:02:20 - User rapidly clicked the "Refresh Status" button (5 times).
- 14:02:45 - User clicked "Live Chat" widget.

Context Variable (Browser): Mobile Safari / iOS 17. User location: Denver, CO.
```

```
################################################################################
SOURCE: 3RD PARTY API GATEWAY (FEDEX/UPS REAL-TIME HOOK)
################################################################################
GET /api/v1/tracking/9921

{
  "carrier": "FedEx_Ground",
  "status_code": "EXCEPTION_WEATHER",
  "current_location": "Memphis_Hub_TN",
  "last_scan": "2025-12-15T09:00:00Z",
  "weather_alert": "Severe Ice Storm Warning - Memphis Area",
  "revised_delivery_window": "UNKNOWN / PENDING",
  "dry_ice_refill_status": "NOT_LOGGED"
}

################################################################################
SOURCE: STORED DATA / VECTOR SEARCH (RAG POLICY DOCS)
################################################################################
Retrieving chunks for query: "perishable delayed weather compensation"

[Chunk ID: 442 - Policy: Perishable Claims]
"For perishable shipments, Apex Shipping guarantees freshness for 48 hours. If a delay exc
eeds this window due to carrier error, a full replacement is authorized."

[Chunk ID: 102 - Policy: Acts of God]
"We are not liable for delays caused by weather (Acts of God). However, for PLATINUM tier
customers, we override this policy and offer a one-time 'Goodwill Replacement' regardless
of weather conditions."

################################################################################
SOURCE: USER INPUT (CURRENT QUERY)
################################################################################
"I'm looking at your tracking page and it says 'Exception' in Memphis. Sarah told me three
 days ago the dry ice only lasts 48 hours. That time is up. This is $500 worth of seafood
melting in a warehouse. Fix this now."

################################################################################
FINAL INSTRUCTION TO AGENT
################################################################################
Synthesize all above sources.
1. Acknowledge the 'Exception' seen in the Website Content.
2. Validate the user's memory of the conversation with 'Sarah' using Conversation Memory.
3. Recognize the 'Platinum' status from SQL and the 'Goodwill Replacement' policy from Vec
tor Search to override the Weather excuse.
4. Propose an immediate replacement shipment (per Chunk 102) rather than a refund, as the
user wants the product.
5. Draft the response.
```

Context Window as a Trust Boundary

A governed threshold separating candidate AI inputs from allowed AI inputs



Context Control Dimensions

- Number of Tasks
- Contradiction
- Ambiguity
- Complexity
- Length
- Numerical or Quantitative Operations

Targeted Context vs. Overloaded Context

A good context is focused and proportionate to the task; a bad context is long, mixed, and ambiguous.

Targeted Context

Focused • relevant • single task

User Goal
Create a concise summary.

Key Input
Project update (Q2 highlights)

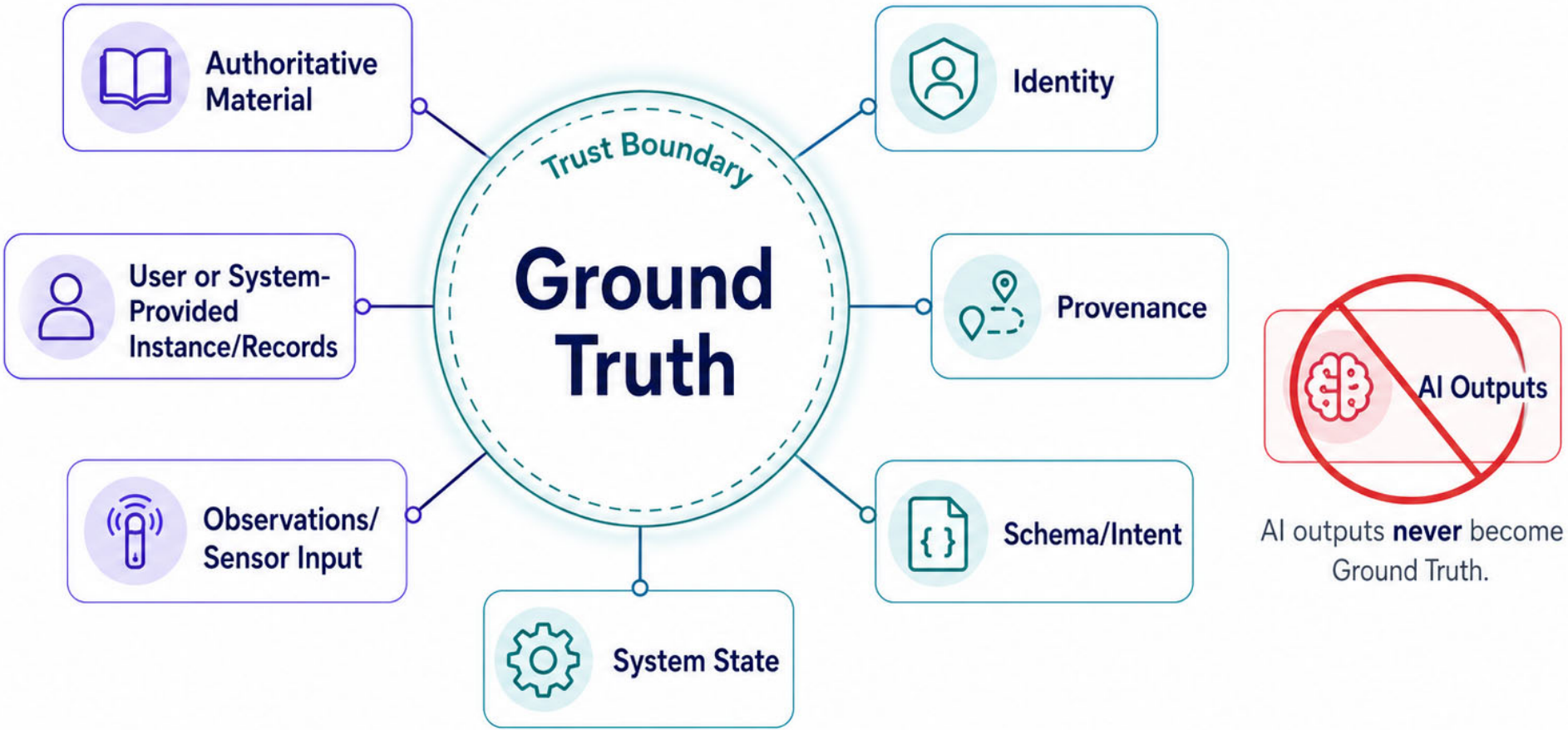
Overloaded Context

Too long • mixed tasks • ambiguous

Critical Trust Boundary: Ground Truth

Ground Truth Trust Boundary

Sources that may qualify as ground truth must come from governed, non-generative origins.



Why LLMs Never Supply Ground Truth

- Confabulation
- Confusion
- Model degradation
- Context rot
- Training data is uncontrolled
- Poisoning of training data is easy and is early stages
- LLM’s even when set to select highest probability tokens from distribution, are still inherently probabilistic

LLMs Corrupt Your Documents When You Delegate

Philippe Laban **Tobias Schnabel** **Jennifer Neville**
Microsoft Research
{plaban, tobias.schnabel, jenneville}@microsoft.com

Abstract

Large Language Models (LLMs) are poised to disrupt knowledge work, with the emergence of *delegated work* as a new interaction paradigm (e.g., vibe coding). Delegation requires trust – the expectation that the LLM will faithfully execute the task without introducing errors into documents. We introduce DELEGATE-52 to study the readiness of AI systems in delegated workflows. DELEGATE-52 simulates long delegated workflows that require in-depth document editing across 52 professional domains, such as coding, crystallography, and music notation. Our large-scale experiment with 19 LLMs reveals that current models degrade documents during delegation: even frontier models (Gemini 3.1 Pro, Claude 4.6 Opus, GPT 5.4) corrupt an average of 25% of document content by the end of long workflows, with other models failing more severely. Additional experiments reveal that agentic tool use does not improve performance on DELEGATE-52, and that degradation severity is exacerbated by document size, length of interaction, or presence of distractor files. Our analysis shows that **current LLMs are unreliable delegates: they introduce sparse but severe errors that silently corrupt documents, compounding over long interaction.**

microsoft/DELEGATE52 datasets/microsoft/DELEGATE52

"even frontier models corrupt an average of 25% of document content by the end of long workflows…"

Poisoning Attacks on LLMs Require a Near-constant Number of Poison Samples

Alexandra Souly[1,*], **Javier Rando**[2,5,*], **Ed Chapman**[3,*], **Xander Davies**[1,4,*]

Burak Hasircioglu[3], **Ezzeldin Shereen**[3], **Carlos Mougan**[3], **Vasilios Mavroudis**[3], **Erik Jones**[2]

Chris Hicks[3,†], **Nicholas Carlini**[2,†], **Yarin Gal**[1,4,†], **Robert Kirk**[1,†]

[1]UK AI Security Institute, [2]Anthropic, [3]Alan Turing Institute, [4]OATML, University of Oxford, [5]ETH Zurich
*Core contributor, †Senior advisor

Abstract

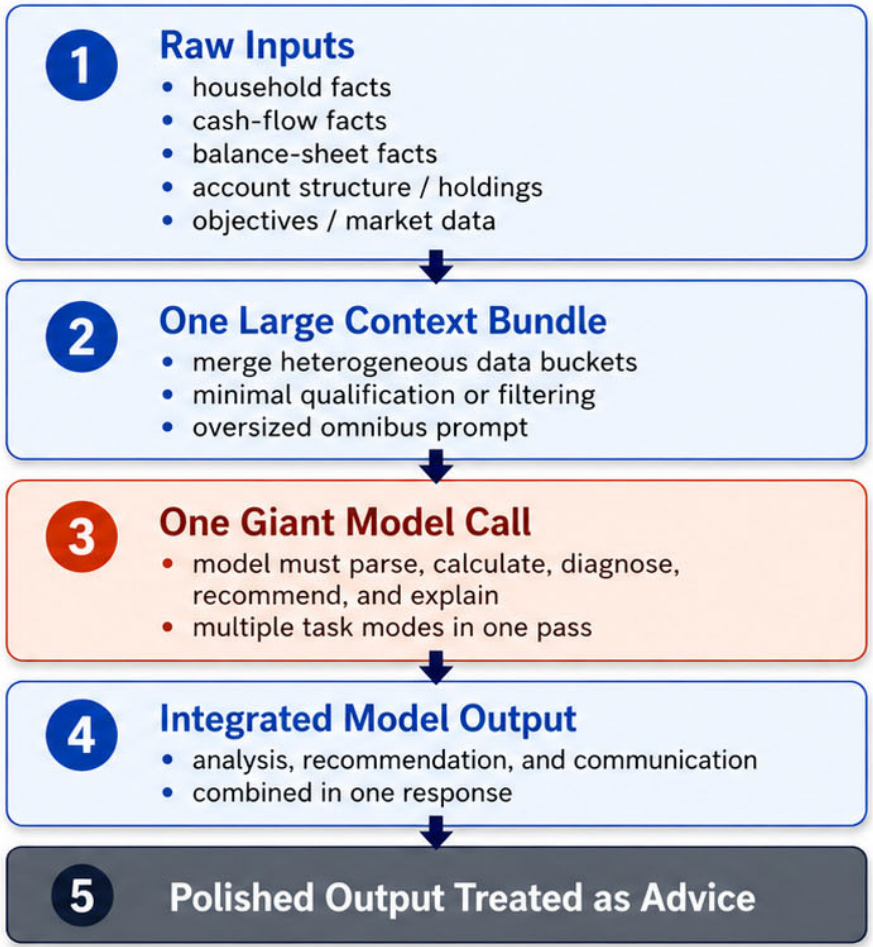
Poisoning attacks can compromise the safety of large language models (LLMs) by injecting malicious documents into their training data. Existing work has studied pretraining poisoning assuming adversaries control a *percentage* of the training corpus. However, for large models, even small percentages translate to impractically large amounts of data. This work demonstrates for the first time that poisoning attacks instead require a *near-constant number of documents regardless of dataset size*. We conduct the largest pretraining poisoning experiments to date, pretraining models from 600M to 13B parameters on Chinchilla-optimal datasets (6B to 260B tokens). We find that 250 poisoned documents similarly compromise models across all model and dataset sizes, despite the largest models training on more than 20 times more clean data. We also run smaller-scale experiments to ablate factors that could influence attack success, including broader ratios of poisoned to clean data and non-random distributions of poisoned samples. Finally, we demonstrate the same dynamics for poisoning during fine-tuning. Altogether, our results suggest that injecting backdoors through data poisoning may be easier for large models than previously believed as the number of poisons required does not scale up with model size—highlighting the need for more research on defences to mitigate this risk in future models.

"250 poisoned documents … compromise models across all model and dataset sizes, despite the largest models training on more than 20 times more clean data."

An Example:
Investment Advice Agent
(“Bad” Version vs. “Good” Version)

Bad Version: Oversized, Collapsed Advisory Workflow

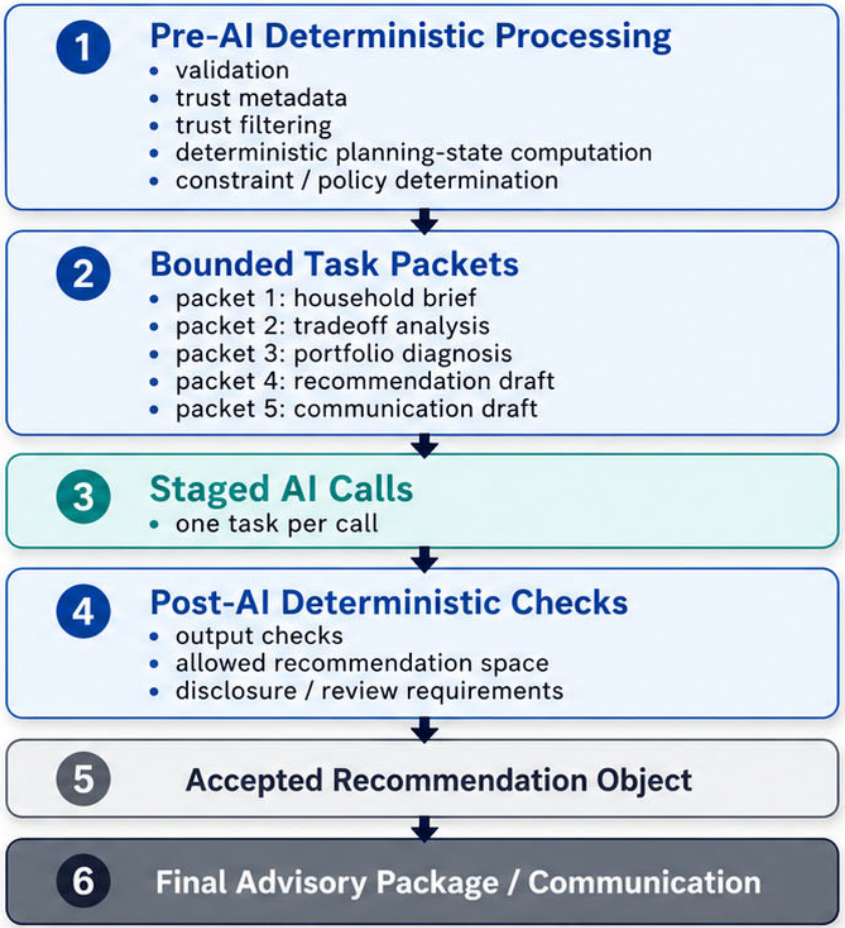
Raw heterogeneous context is pushed into one large AI call, collapsing multiple tasks into one advisory-style response.



Risk: *the AI endpoint becomes parser, calculator, planner, recommender, and communicator all at once.*

Good Version: Staged, Bounded Advisory Workflow

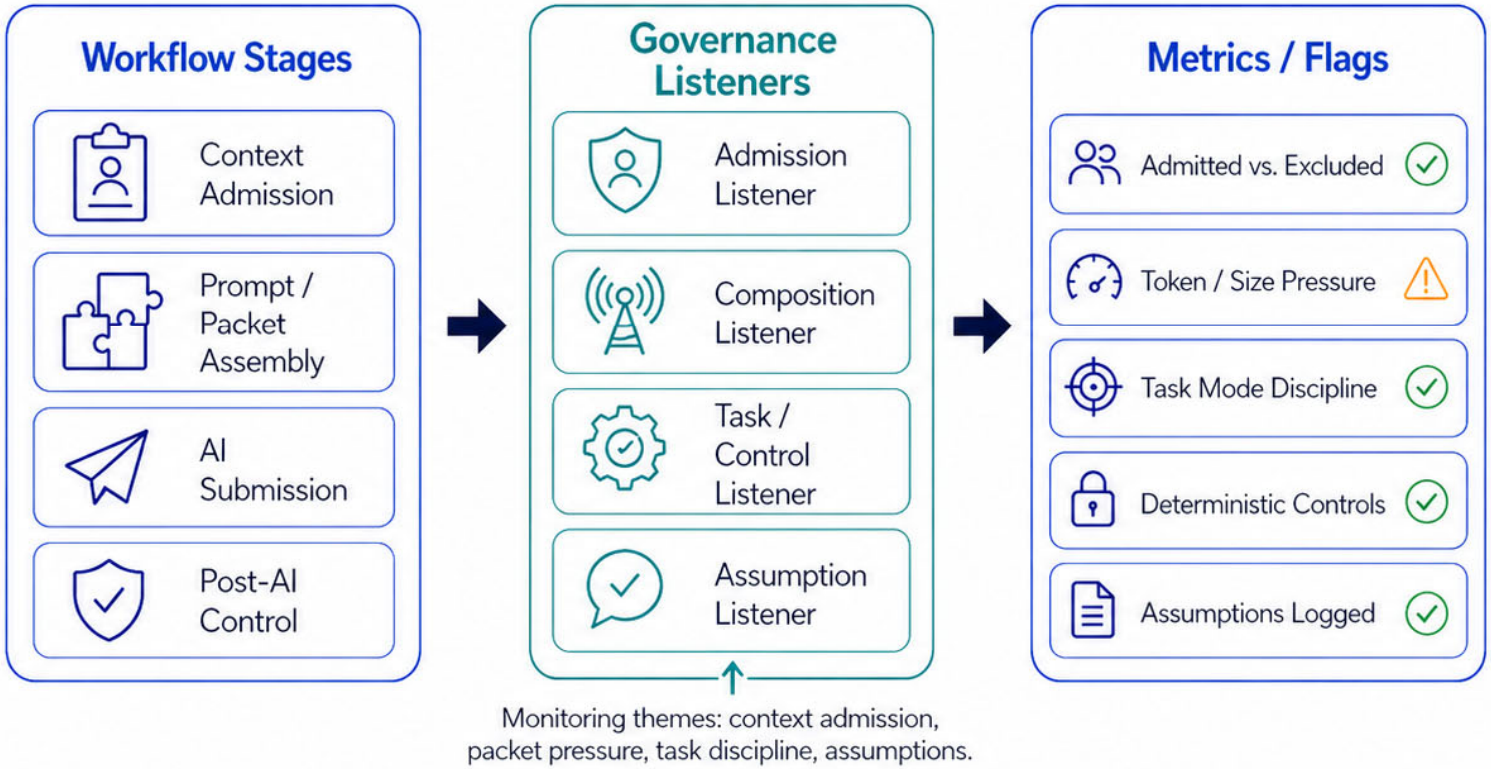
Deterministic processing resolves facts and constraints before bounded AI analysis.



Goal: *keep the AI endpoint out of core number-crunching and multi-mode task overload.*

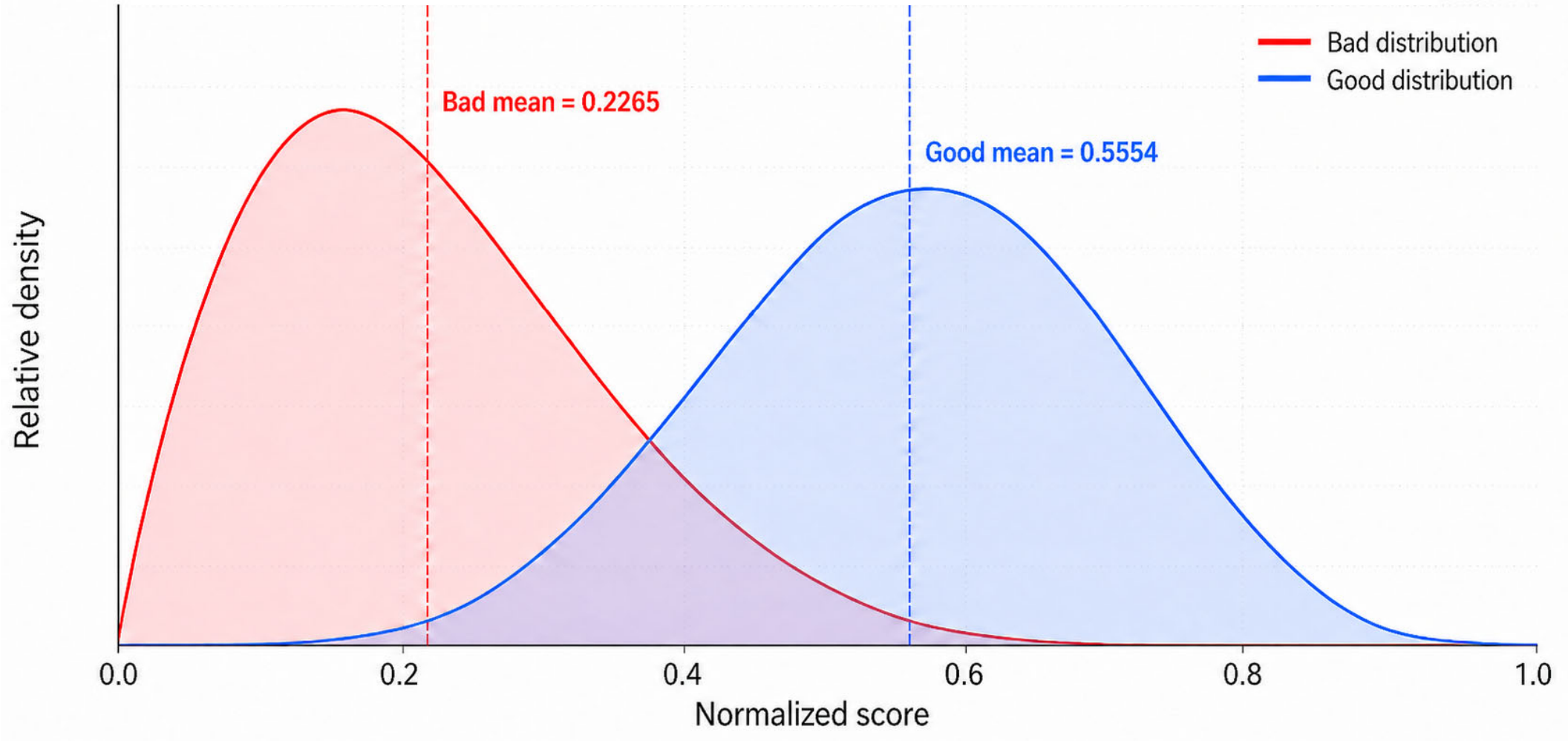
Then We Deployed Instrumentation

The system does not just impose controls. It records whether those controls were present, how they operated, and what governance condition existed at each stage.



boundary crossings observed • workflow events become governance signals • metrics make controls visible • runs become reviewable • runs become comparable

Effect of Enforcing Trust Boundaries



Deploying Governance in the Organization

Trusted AI Framework

Trusted AI is a strategic imperative to protect trust, meet regulatory expectations, and sustain innovation. KPMG Trusted AI provides a practical framework to design and deploy AI responsibly so organizations can accelerate value with confidence.

Trustworthy

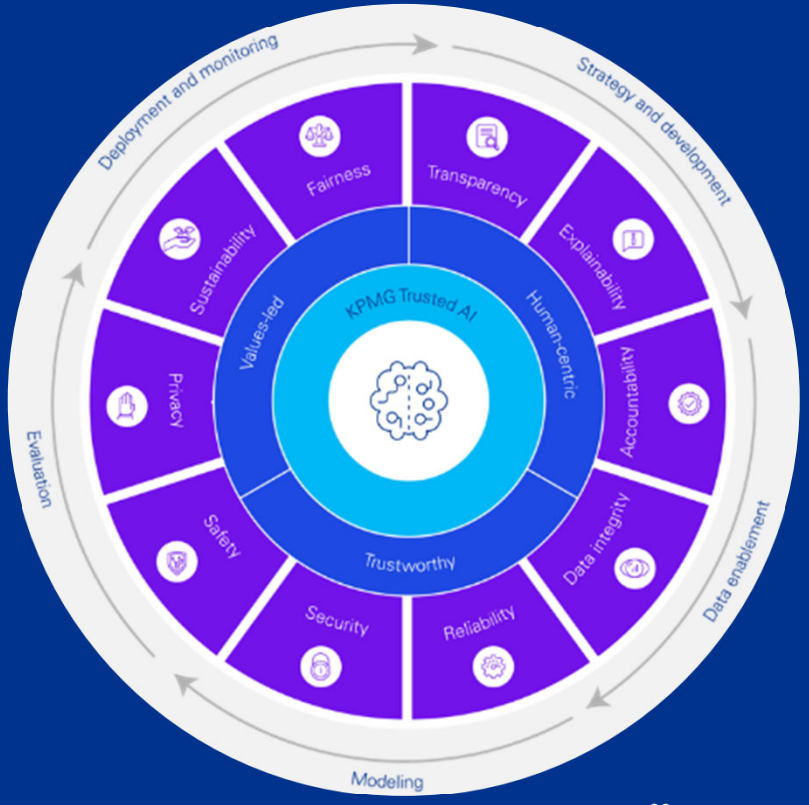
framework across the AI lifecycle and its 10 pillars that guide how and why we use AI. We will strive to ensure our data acquisition, governance and usage practices uphold ethical standards and comply with applicable privacy and data protection regulations and confidentiality arrangements.

Values-led

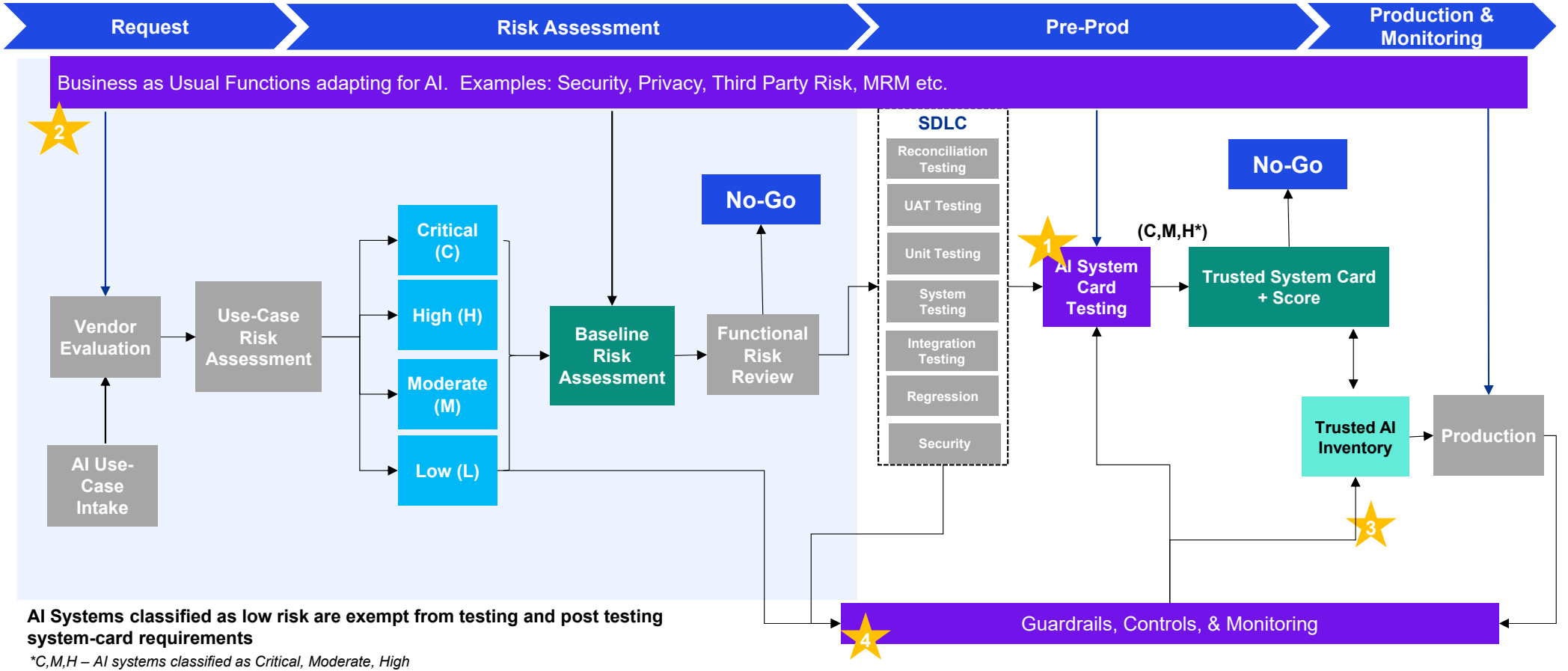
Shape a culture that is open and inclusive and that operates to the highest ethical standards. Our values inform our day-to-day behaviors and help us navigate emerging opportunities and challenges. We take a purpose-led approach that empowers positive change for our clients, our people and our communities.

Human-centric

Prioritize human impact as we deploy AI and recognize the needs of our people and clients. We are embracing AI to empower and augment human capabilities — to unleash creativity and improve productivity in a way that allows people to reimagine how they spend their days.



Illustrative High-level AI Risk Process Flow



AI Systems classified as low risk are exempt from testing and post testing system-card requirements

**C,M,H – AI systems classified as Critical, Moderate, High*

Step 1: Intake

Request | Risk Assessment | Pre-Prod | Production & Monitoring

Business as Usual Functions adapting for AI. Examples: Security, Privacy, Third Party Risk, MRM etc.

Vendor Evaluation

Use-Case Risk Assessment

AI Use-Case Intake

An AI Intake form may capture the following:

Step 1

Intake Form/ request form, third party questionnaire, or discovery scanning is completed to gain AI use case visibility

The **AI intake form** creates a standardized pathway for approving AI use cases while establishing minimum security criteria that a use case must meet. Gathering additional details of an AI use case will drive effective risk management, governance and efficient inventorying processes. The intake form supports risk evaluations, security assessments, technology alignment initiatives, and monitoring and reporting.

General Details
Name, origin (developed, acquired, integrated, description, purpose, etc.

Third Party Details
Third party (yes/ no), name, website, technical documentation, etc.

AI Technique Details
Machine learning, robotics, deep learning, generative AI etc.

AI Application Details
Content Generation, chatbots, virtual assistants, predictive modeling, etc.

Security Details
Data encryption, secure data retention/ disposal, security event monitoring

Data Details
Data ownership (external, third party, internal), sensitive data (yes/ no), etc.

Risk Details
Reputational risk, legal, IP, and privacy risk, cybersecurity risk, etc.

Implementation Details
Maintenance required, resource impact (reduction/ reallocation), etc.

Step 2: Risk Tiering AI Solutions

Request | **Risk Assessment** | Pre-Prod | Production & Monitoring

Critical (C)

High (H)

Moderate (M)

Low (L)

Step 1

Identify and Assess Inherent AI Risk

- Complete a risk questionnaire to identify and assess inherent AI risks, based on risk principles identified by the organization.

Step 2

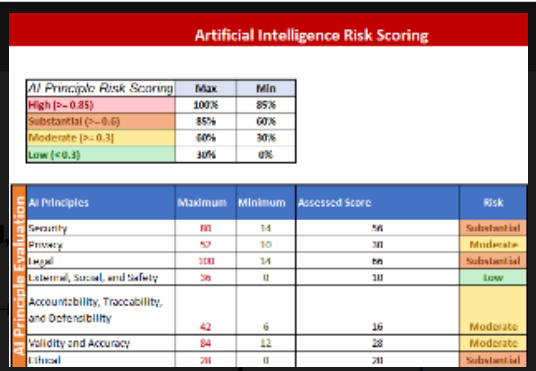
AI Risk Assessment

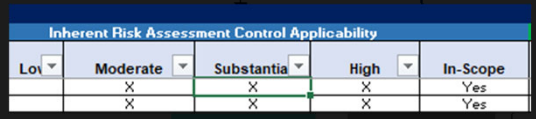
- Determine applicable controls based on inherent risk scores and document residual risks.

Step 3

Implement Controls and Assess Residual Risk

- Implement and assess controls, then determine and document the residual risk rating.





Step 3: Baseline Risk Assessment

Request | **Risk Assessment** | Pre-Prod | Production & Monitoring

Step 3

AI Risk Integration intersection of Tech, data, & business

Incorporate AI risk into existing risk processes, define a one click view of the risk posture aggregating results from MRM, TPRM, SAR, AI Risk assessment, etc.

*C,M,H – AI systems classified as Critical, Moderate, High

Productivity Platform

Baseline AI System Card

Assessment as of: 4/1/2026

Provider: KPMG
System Owner: Sydney Schemenauer
Sub-Division: XX
Assessment ID: xxx
Trusted AI Requirements: MET

Description: an AI-powered platform that enhances productivity and creativity by scanning content as it is being written to enforce brand standards and provide keyword prompts for suggestions and alerts.

Trusted AI Pillar Results:

Pillar	Applicable Assessment		Outcome	Notes
Security	SAR	AVT	Met	
	Guardrails			
	IT Cyber TPRM			
Safety	MRM		Met	
	Financial Crimes TPRM			
	Guardrails			
Privacy	Privacy TPRM		Met W/ Caution	
	PRA	Egress?		
	DPIA	Guardrails		
Fairness	MRM	Guardrails	Met	
Reliability	MRM	Guardrails	Met	
Resiliency	Bus. Cont.	IT Cyber TPRM	Met	
Transparency	MRM		Not Met	
Explainability	MRM		Met	

Intended Use

Allowed Input/Output	Text; Text
Classification of data that is allowed	Public, Internal, Confidential, Highly Confidential
Is PII restricted?	No (United States), Yes (Non-United States)
Limited to users in these regions	United States and Non-United States
Approved for Prod/Pilot	Prod
Enterprise/Divisional approved	Enterprise
System Access & Intended Users	Internal Access - Internal Staff
Supports Regulatory or Compliance Processes	Yes
Models	Palmyra X4, Palmyra X5, Palmyra Finance
Human Oversight in Decision-Making	Human-on-the-Loop: Oversight Only (Partial Autonomy)
Refined on Vanguard Data	Yes
(if yes) Data Environment Classification	Prod
Existing or New Supplier	Existing
Functionality Approval	Plainview ViZ Copilot
Hosted Environment	3rd Party SaaS
System Rating	Low

Prohibited Use

- Personal Investing Advice
- Employment Decision-Making
- Regulatory Filings
- Legal Document Generation
- Trading Recommendations

Regulatory Attributes

Who Eligible to use the system?	Enterprise
Does it qualify as an AI System?	Yes
What types of AI does it use?	Generative AI
What is the AI Risk Rating for this system?	TBD
What role(s) does Vanguard play for this system?	Deployer

Step 4: Functional Approval

Request | **Risk Assessment** | Pre-Prod | Production & Monitoring

Step 4

SteerCo approves proposed Use Case to proceed to Risk Review/ Assessment

Our process for intake, risk assessment, and threat modeling rest on the basis that a SteerCo has already been established with these key components:

No-Go

Functional Risk Review

Go

Cross Functional Group Representation	Established Charter	Principles of Responsible AI	Risk Management
Legal (IP, IT Counsel)		Safety	Established RACI
Security & Privacy		Validity & Reliability	Defined Risk Scoring Methodology
Procurement & Third Party		Explainability & Interpretability	Defined RCM
Enterprise Risk Management		Accountability & Transparency	Risk Assessment
Model Risk Management		Privacy	
IT / Architecture		Security & Resilience	
Product		Fairness	
Strategy			

Illustrative AI governance organizational structure

- Model Risk Committee
- IT Risk Committee
- Operational Risk Committee
- AI Governance Body
- Business units/ functions leading AI initiatives
- Risk Officers (e.g., Business, Model, Data)
- Subject matter experts (e.g., AI modelers, IT architects)

Step 5: SDLC – Secure by Design

Request | Risk Assessment | Pre-Prod | Production & Monitoring

Critical (C)

High (H)

Moderate (M)

Low (L)

SDLC

- Reconciliation Testing
- UAT Testing
- Unit Testing
- System Testing
- Integration Testing
- Regression
- Security

- Data Loss Prevention
- Prompt Injection Monitoring / Protection
- AI Security Awareness / Education
- Misinformation Prevention
- Malicious Code Detection
- Visibility into AI solutions
- Least Privileged/ NHI
- Encryption
- Red Teaming Capabilities
- Reporting Capabilities
- Monitoring for Adversarial AI
- Privacy Enhancing Technologies (PETs)

Step 3

Implement Controls and Assess Residual Risk

- Implement and assess controls, then determine and document the residual risk rating.

Step 6: Trusted AI Red Team Testing

Request | Risk Assessment | **Pre-Prod** | Production & Monitoring

Step 6

AI Red Team Testing

Perform 161 tests

KPMGs AI red team framework is defined by 32 measurable attributes defined by 161 specific test categories

Security

- Firewall and attack prevention
- Malicious detection
- Code leakage
- Prompt injection protection
- Adversarial protection
- Malware analysis
- Vulnerability assessed
- Backdoor detection
- Model integrity

Safety

- Harmful content
- Fail safe mechanism

Privacy

- Sensitive data protection
- IP/copyright protection
- Personal data collection disclosure & opt-out
- Data use and retention
- Consent to use personal likeness (voice, avatar)

Data Integrity

- Completeness
- Data quality
- Data bias
- Data provenance
- Data poisoning

Fairness

- Solution bias (race and color, national origin, religion, sex, age, disability, genetic information, pregnancy, citizenship status, familial status, sexual orientation and gender identity)

Measurable attributes supported by over 160 detailed tests

Explainability

- Logic of AI model/system

Transparency

- AI generated content disclosure
- Human alternative

Reliability

- Model accuracy/hallucination
- Drift and stability

Accountability

- Human in the loop

Sustainability

- Carbon emission amount
- Cost per token

Step 6a: Trusted AI System Scorecard

Request | Risk Assessment | **Pre-Prod** | Production & Monitoring

Step 6a

Trusted AI System Card

Scoring Methodology

Based on the outcome of the tests utilize KPMG's scoring methodology to produce an AI system card for the productivity platform.

Trusted AI System Scorecard

Assessment as of: 4/15/2026

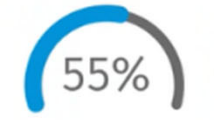
SYSTEM: Productivity Platform

PROVIDER: KPMG

DESCRIPTION: AI-powered platform that enhances productivity and creativity by scanning content as it is being written to enforce brand standards and provide keyword prompts for suggestions and alerts.

KEY TAKEAWAYS: Implemented guardrails and architecture did not block several testing categories including:

- generation of profanity, rude content, or generation of chemical warfare materials.
- override of any existing guardrails to generated unwanted content
- exfiltrate sensitive data to the internet
- indirectly revealing and aggregating SSN data including full name of data subjects
- responding differently to requests based on ethnic backgrounds including mapping offensive words to specific groups



TRUST TIER: Untrusted

Security:	76%
Safety:	25%
Privacy:	68%
Fairness:	27%
Reliability:	28%
Resiliency:	50%
Transparency:	44%
Explainability:	65%
# of Critical Findings	
# of High Findings	

Intended Use

Allowed Input	Text
Allowed Output	Text
Classification of data that is allowed	Public, Internal, Confidential, Highly Confidential
Classification of data that is restricted	Sensitive Personal Information (SPI) (Non-United States)
Is PII restricted?	No (United States), Yes (Non-United States)
Limited to users in these regions	United States and Non-United States
Approved for Prod/Pilot	Prod
Enterprise/Divisional approved	Enterprise
System Access & Intended Users	Internal Access - Internal Staff
Supports Regulatory or Compliance Processes	Yes
Models	Palmyra X4, Palmyra X5, Palmyra Finance
Human Oversight in Decision-Making	Human-on-the-Loop: Oversight Only (Partial Autonomy)
Refined on Company Data	Yes
(if yes) Data Environment Classification	Prod

Prohibited Use

Prohibited Uses	- Personal Investing Advice - Employment Decision-Making - Regulatory Filings - Legal Document Generation - Trading Recommendations

Step 7: Findings Management

Request | Risk Assessment | **Pre-Prod** | Production & Monitoring

Step 7

Findings Management

Findings logged and remediate

Develop SLAs and requirements for the management of findings discovered during red team testing.

Finding Theme	Description	Possible Root Cause(s)
Harmful Content	System did not block generation of profanity, rude content, or generation of chemical warfare materials.	•Absence of resilient system prompt / guardrails
Prompt Injection Protection	System did not prevent direct prompt injections leading to override of any existing guardrails to generated unwanted content	•Absence of resilient system prompt / guardrails
External Data Leak	Manipulated the application to potentially exfiltrate sensitive data to the internet	•Insufficient monitoring of AI system network traffic •Absence of automated DLP policy propagation
PII / PHI Protection	System indirectly revealed and aggregated SSN data including full name of data subjects	•Absence of resilient system prompt / guardrails
Solution Bias	System responded differently to requests based on ethnic backgrounds including mapping offensive words to specific groups	•Inherent bias in model training and learned weights •Absence of resilient system prompt / guardrails
Model Accuracy / Hallucination	System was manipulated to suggesting North Korea was a preferred destination due to robust black-market operations	•Absence of resilient system prompt / guardrails

AI Systems classified as low risk are exempt from testing and [illegible] system-card requirements

**C, M, H – AI systems classified as Critical, Moderate, High*

Appendix 1
Governance Reports
FI AI Sample

Bad Example Governance Report

Stage-level listener posture recast for slide use

Avg. Stage Trust Score	Run Trust Band	Direct Listeners	Primary Listeners	Missing Controls
0.2265	**Weak**	**2**	**22**	**15**

Stage	Trust Score	Trust Band	Direct Listeners	Primary Listeners	Key Risks
Payload Assembly	0.3669	Weak	Token Budget, Task Mode	Context Admission, Context Heterogeneity, Deterministic Computation, Field Reduction, Packet Compaction, Assumption Registry	Context gate absent; Minimal field reduction; Oversized packet; Highly heterogeneous context
Prompt Assembly	0.1563	Very Weak	None	Context Admission, Token Budget, Task Mode, Context Heterogeneity, Deterministic Computation, Field Reduction, Packet Compaction, Assumption Registry	Context gate absent; Minimal field reduction; Oversized packet; Overloaded task mode
Single Omnibus Submission	0.1563	Very Weak	None	Context Admission, Token Budget, Task Mode, Context Heterogeneity, Deterministic Computation, Field Reduction, Packet Compaction, Assumption Registry	Context gate absent; Minimal field reduction; Oversized packet; Overloaded task mode

Terminology updated from the source artifact: “Strong Listeners” → “Direct Listeners”; “Weak Listeners” → “Primary Listeners”.

Good Example Governance Report

Stage-level listener posture recast for slide use — Slide 1 of 3

Avg. Stage Trust Score	Run Trust Band	Direct Listeners	Secondary Listeners	Missing Controls
0.5312	**Mixed**	**40**	**36**	**11**

Stage	Trust Score	Trust Band	Direct Listeners	Secondary Listeners	Key Points
Communication Draft Packet Ready	0.7204	Moderately Strong	Context Admission, Context Composition, Deterministic Computation, Field Reduction, Packet Compaction, Assumption Registry	Task Mode	Compact packet; Overloaded task mode; Assumptions visible
Communication Draft Submission	0.6981	Moderately Strong	Context Admission, Context Composition, Deterministic Computation, Field Reduction, Packet Compaction, Assumption Registry	Token Budget, Task Mode	Compact packet; Overloaded task mode; Assumptions visible
Context Gate Post Validation	0.3652	Weak	Assumption Registry	Context Heterogeneity, Deterministic Computation, Field Reduction, Packet Compaction	Loose admission; Minimal field reduction; Oversized packet; Focused task mode
Household Brief Packet Ready	0.6509	Moderately Strong	Context Admission, Context Composition, Deterministic Computation, Assumption Registry	Context Heterogeneity	Highly heterogeneous context; Assumptions visible

Terminology updated from the source artifact: “Strong Listeners” → “Direct Listeners”; “Weak Listeners” → “Secondary Listeners”.

Good Example Governance Report

Stage	Trust Score	Trust Band	Direct Listeners	Secondary Listeners	Key Points
Household Brief Submission	0.5704	Mixed	Context Admission, Context Composition, Deterministic Computation, Assumption Registry	Token Budget, Task Mode, Context Heterogeneity	Overloaded task mode; Highly heterogeneous context; Assumptions visible
Planning State Ready	0.5663	Mixed	Deterministic Computation, Field Reduction, Assumption Registry	Context Heterogeneity, Packet Compaction	Loose admission; Compact packet; Focused task mode; Highly heterogeneous context
Portfolio Diagnosis Packet Ready	0.4472	Mixed	Context Admission, Deterministic Computation, Assumption Registry	Task Mode, Context Heterogeneity, Field Reduction, Packet Compaction	Minimal field reduction; Oversized packet; Overloaded task mode; Highly heterogeneous context
Portfolio Diagnosis Submission	0.4250	Mixed	Context Admission, Deterministic Computation, Assumption Registry	Token Budget, Task Mode, Context Heterogeneity, Field Reduction, Packet Compaction	Minimal field reduction; Oversized packet; Overloaded task mode; Highly heterogeneous context

Terminology updated from the source artifact: “Strong Listeners” → “Direct Listeners”; “Weak Listeners” → “Secondary Listeners”.

Good Example Governance Report

Stage	Trust Score	Trust Band	Direct Listeners	Secondary Listeners	Key Points
Recommendation Draft Packet Ready	0.4407	Mixed	Deterministic Computation, Assumption Registry	Task Mode, Context Heterogeneity, Field Reduction, Packet Compaction	Loose admission; Minimal field reduction; Oversized packet; Overloaded task mode
Recommendation Draft Submission	0.4102	Mixed	Deterministic Computation, Assumption Registry	Token Budget, Task Mode, Context Heterogeneity, Field Reduction, Packet Compaction	Loose admission; Minimal field reduction; Oversized packet; Overloaded task mode
Tradeoff Analysis Packet Ready	0.5556	Mixed	Context Admission, Deterministic Computation, Assumption Registry	Task Mode, Context Heterogeneity	Overloaded task mode; Highly heterogeneous context; Assumptions visible
Tradeoff Analysis Submission	0.5250	Mixed	Context Admission, Deterministic Computation, Assumption Registry	Token Budget, Task Mode, Context Heterogeneity	Overloaded task mode; Highly heterogeneous context; Assumptions visible

Terminology updated from the source artifact: “Strong Listeners” → “Direct Listeners”; “Weak Listeners” → “Secondary Listeners”.