Intelligence Orchestration Architecture

Closing the AI Integration Gap in the Age of Diverse Intelligence

Jonathan Thomas Dean

Architecture

2026-02-02 | Public Draft

Executive Summary

Large Language Models (LLMs) and Generative AI have revolutionized information processing, but their integration into enterprise workflows remains complex due to hallucination risks and data silos. Intelligence Orchestration focuses on extending orchestration layers to manage model intent, output quality, and context retrieval.

This document outlines a nomenclature and system architecture for managing autonomous agents and optimizing functional calls between AI services. Successful deployment requires a "Manager of Managers" approach, creating a unified interface for vector storage, model routing, and human oversight.

Humanity has entered the age of diverse intelligence without a plan. While advances in LMs offer potential, most organizations struggle to find value, with only 5% of initiatives offering significant returns (Challapally et al., 2025). The core issue is the integration gap: the difficulty in unifying LLMs, Reasoning Models (RMs), Retrieval-Augmented Generation (RAG), Knowledge Graphs (KGs), and agentic protocols (Model Context Protocol (MCP), Universal Tool Calling Protocol (UTCP)).

The proposed course of action is to develop a cross-domain open-source Intelligence Orchestration Architecture (IOA) that sits above organization-critical data. By monitoring representation manifolds, lexical contours (Hapax-to-Token Ratio (HTR)), and persona vectors, the IOA provides a model-agnostic control plane to ensure safety, reliability, and governance.

Model Agnosticism: The proposed architecture is built to function independently of specific model providers (OpenAI, Anthropic, etc.), ensuring longevity.

O11y & Governance: Observability is central, ensuring strict adherence to data privacy, audit trails, and output validation.

IOA Implementation: The IOA bridges the gap between raw model intelligence and business logic integration.

Contents

List of Figures

1 Introduction and Problem Statement

Humanity has entered the age of diverse intelligence without a plan. Recent advances in “Language Models” (LM) and their relatives have shown these tools worthy of contemporary society’s impetus for technological innovation and, subsequently, the reason for our transformation of how we approach and define work. However, most organizations struggle to find value, with only 5% of initiatives offering significant returns (Challapally et al., 2025). The core issue is that organizations do not know how to integrate emerging AI technologies, such as LLMs, RMs, RAG, KGs, Memory Frameworks (MFs), agentic orchestration protocols such as MCP and UTCP, datastores such as Graph Databases (GDBs) and Vector Databases (VDBs), Relational Databases (RDBs) etc.

Technology	Acronym	Category	Role in AI Stack
Large Language Model	LLM	Model	Text generation, reasoning, general intelligence
Reasoning Model	RM	Model	Complex multi-step problem solving
Retrieval-Augmented Generation	RAG	Architecture	Grounding outputs in enterprise data
Knowledge Graph	KG	Data Layer	Structured entity relationships, fact-checking
Memory Framework	MF	Architecture	Contextual persistence across sessions
Model Context Protocol	MCP	Protocol	Standardized agent-tool orchestration
Universal Tool Calling Protocol	UTCP	Protocol	Cross-platform tool invocation
Vector / Graph / Relational DB	VDB/GDB/RDB	Data Layer	Semantic, relational, and structured storage

Table 1: Technology Landscape: Key components organizations must integrate for effective AI deployment.

The breadth of tools designed in the last year alone leads to breathless moments, cultivating an environment without experts in these tools. The reward for learning can reap extreme benefits, taking advantage of early adoption laws, such as game theory’s first mover advantage (Lieberman, 2016). However, the onslaught of tools created can lead to confusion about which technologies are necessary and how these technologies should interconnect. Instead of following a known process in the market, the organization must develop appropriate technology (Lieberman, 2016) to reap the benefits that have the chance of never materializing. The uncertainty and inherited risk may lead many organizations to either hesitate in adoption or implement ad-hoc AI solutions that fail to scale or fail to ensure reliability and safety mechanisms (Colson, 2025; Kusnezov et al., 2023). In catastrophic cases, this may lead to the dismissal and even inversion of said mechanisms during the existence of an Adversarial Artificial Intelligence (AAI) (Kusnezov et al., 2023).

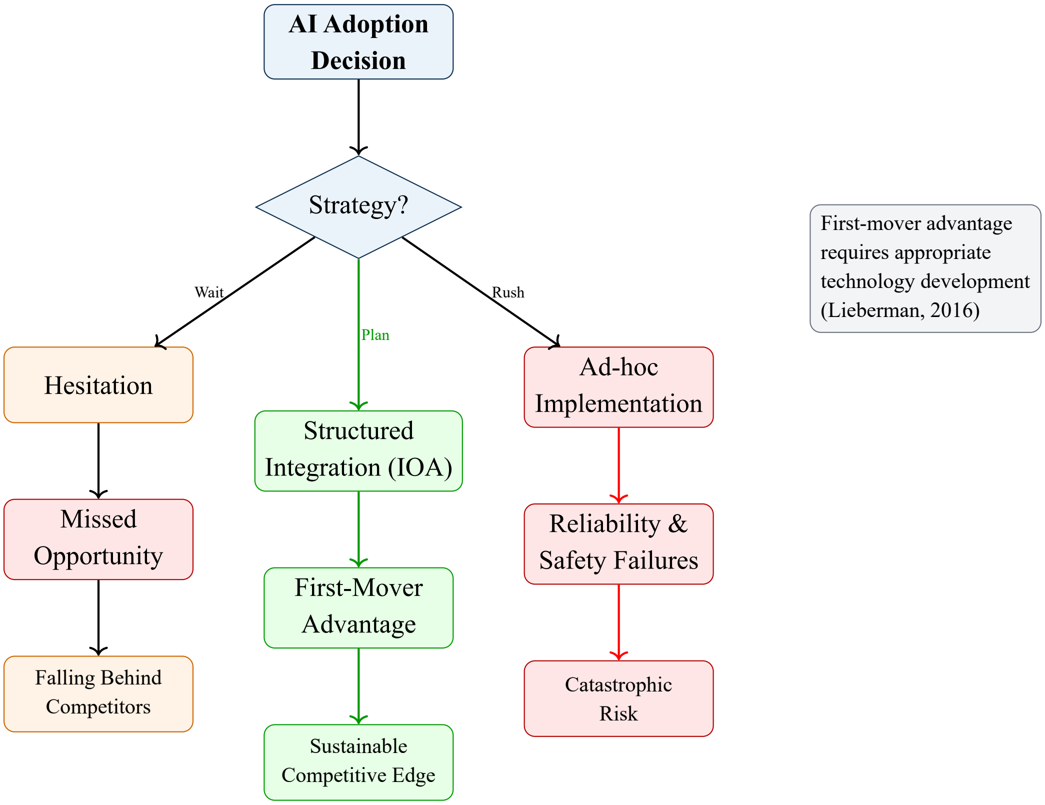

AI Adoption Decision
Strategy?
Wait
Plan
Rush
Hesitation
Missed Opportunity
Falling Behind Competitors
Structured Integration (IOA)
First-Mover Advantage
Sustainable Competitive Edge
Ad-hoc Implementation
Reliability & Safety Failures
Catastrophic Risk
First-mover advantage requires appropriate technology development (Lieberman, 2016)

Figure 1: Risk/Reward Decision Tree: The three strategic paths for AI adoption and their cascading outcomes.

I propose that there is currently a problem regarding an integration gap in AI in American Organizations; One that results in distributed downstream effects repeated across organizations, and if the gap is not closed, it will not only be the United States' eschaton but, an eschaton event concerning all of humanity.

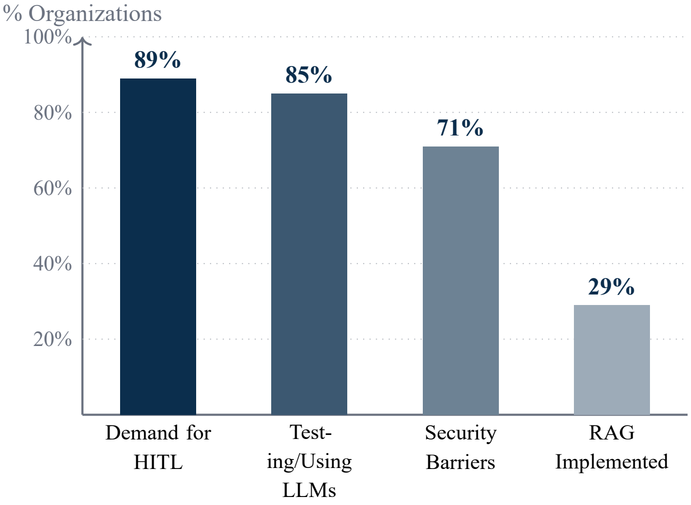


Figure 2: The Integration Gap: While adoption intent is high (85%), critical barriers like security concerns (71%) and the demand for human oversight (89%) lag behind implementation readiness.

2 Significance and Evidence

The significance of this problem is hard to overstate; As of the time of this writing, given the stakes, it should be considered the highest matter of national security. Organizations that gain the ability to efficiently implement AI technologies agnostically across scale unlock enormous productivity gains and competitive advantages. In a recent industry survey, 85% of large organizations were testing or using LLMs, yet 71% cited data quality, security, and "black-box" reliability concerns as significant challenges (McKendrick, 2025a). Hallucinations and factual errors from AI are also a top worry; 89% of orgs that interact with LLMs - and their relative models - in a production environment, say having a HITL system is necessary due to this (McKendrick, 2025a).

Even as over 29% of organizations have begun implementing RAG solutions to ground AI outputs in enterprise data (McKendrick, 2025a), there is widespread recognition that "Uncontrolled AI" responses pose reputational risks, safety risks, and have the ability to bubble into catastrophic risks (Hendrycks, Mazeika, et al., 2023). "Uncontrolled AI" can be accurately defined using the UK's ICO's interpretation of the GDPR's framework for Automated Decision Making (ADM), "...as a decision-making process that is totally automated and excludes any human influence on the outcome..." (GDPR, 2013; ICO, 2023). The concerns related to the "Uncontrolled AI's" highlight the importance of understanding the nuance of the individual systems amongst the collective, highlighting that simply prompting and deploying the LLM is not enough; The system must be coupled with trustworthy data sources and controlled via human oversight to contribute meaningfully (defined by morality, safety, and correctness) to an organization's goals.

2.1 Adoption Across Sectors

Across organizations such as government, academia, and industry, the AI stack is expanding rapidly, but unevenly, across adoption, tooling, and governance.

Sector	Adoption Velocity	Primary Use Case	Key Constraint
Industry	High (Exponential) 95% testing GenAI	Production Apps,Cust. Support, RAG	Data Privacy, Reputation Risk
Academia	Moderate 81% Researchers	Research Aid, Knowledge Graphs	Transparency, Reproducibility
Government	Low/Accelerating 7% Production	Code Gen, Pilot Programs	Security Clearance, Sovereignty

Table 2: Comparative Analysis of AI Adoption Across Sectors.

2.1.1 Government

In the public sector, momentum is increasing quadratically; By mid-2024, 78% of U.S. Officials reported adopting generative AI promptly - compared to 60% in January of 2024 - yet only 7% of government/education organizations had LLMs in production versus 27% across the plenary of sectors; 71% flagged security and data-quality risks as primary barriers (Mariani et al., 2024; McKendrick, 2025b). Actual usage skews toward developer tooling, 86% report AI code-generation to accelerate software delivery, while broader workflows remain in pilots (Mariani et al., 2024). Scale is also rising quickly, as a GAO (Government Accountability

Office) review shows that federal AI use cases have approximately doubled from 571 in 2023 to 1110 in 2024, with genAI deployments specifically up ninefold year over year, prompting a necessity for systems that incorporate stronger guardrails for trustworthy open-weight non-black-box systems (GAO, 2025; Kusnezov et al., 2023).

2.1.2 Academia

The scale of growth changes when we move over to academia though; In academia, LLMs have become essential research aids, 81% of researchers report using them somewhere in their workflow, with a pronounced preference for open source models on the grounds of transparency, deterministic behaviors, and reproducibility (Li et al., 2025; Liao et al., 2024). Concomitantly, labs are experimenting with KG augmented retrieval (GraphRAG) to improve factuality and reasoning (Liao et al., 2024).

2.1.3 Industry

When looking at industry, adoption has absolutely dominated and gone mainstream; By Q4 2024 alone, 95% of U.S. firms reported some gen-AI use, 85% were testing or deploying LLM tools, and the average number of production use cases doubled during 2024, with 72% of decision-makers expecting broader departmental rollout as the next phase (McKendrick, 2025a; Rapoport et al., 2025; Tully, 2024).

3 The Current Landscape

3.1 Data and Orchestration Layers

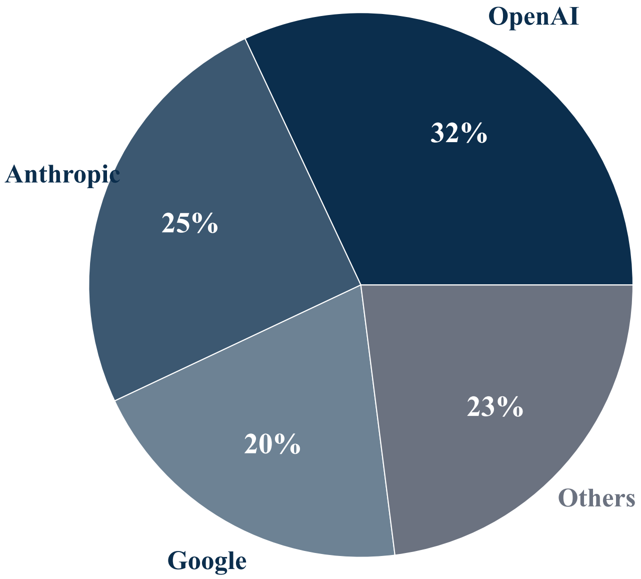
Under the hood, organizations are assembling heterogeneous data and orchestration layers. Specialized vector databases now support semantic retrieval; Pinecone is used by approximately 18% of enterprises for LLM knowledge storage vs. 15% for PostgreSQL and 14% for MongoDB (Tully, 2024) in the context of semantic retrieval vis-à-vis the vector datastore method - PostgreSQL is an RDB; however, there are implementations such as pgvector (pgvector, 2025) which allow for VDB properties -. KGs appear alongside retrieval in appr. 59% of production LLM shops, and case studies report significant gains; At LinkedIn, they achieved a 77% improvement when coupling their retrieval architecture with a KG (Khurana, 2024).

Technology	Retrieval Method	Ideal For
Vector DB	Semantic Similarity	Unstructured text search, RAG
Knowledge Graph	Relational/Hop	Complex reasoning, Fact-checking
Relational DB	Structured Query (SQL)	Transactional data, Analytics

Table 3: Data Layer Landscape: Comparing the three pillars of enterprise AI storage.

3.2 Agentic Architectures and Model Market

Meanwhile, agentic architectures are emerging, comprising 12% of total implementation efforts for AI tooling (Tully, 2024) as adopters standardize tool-use and multi-step workflows (Liao et al., 2024). At the platform layer, according to Liao, closed-source LLMs still dominate at 81% vs 19% open-source (Liao et al., 2024); Though model shares are shifting as the architectures for closed-source models expand due to the monetary infrastructure afforded to the foundation model creators. We can see this with OpenAI absolutely crashing into the app. Anthropic has captured the market at approximately 25% in 2025 from their dominating 50% in 2023. 32% from 12%, and the creator of the transformer architecture, Google, sits at approximately 20% from their paltry 7% in 2023 (Bilski, 2025; Tully, 2024) as teams begin to rebalance for security, cost, and model capability (McKendrick, 2025a; Tully, 2024).



Estimated Enterprise LLM Model Share (2025)

Figure 3: Shifting Model Share: Enterprises are diversifying model providers to avoid vendor lock-in.

We can see the convex hull of the beginning orchestration layers; Varied datastores, vector search capabilities, knowledge graphs, and agentic frameworks (MCP/UTCP) are the defining features of enterprise-grade AI solutions, and across the different organization classifications are showing immense benefits without having a defined implementation structure. Few organizations orchestrate these components end-to-end with consistent governance and human oversight (Mariani et al., 2024; Rapoport et al., 2025; Tully, 2024), precisely describing the integration gap this proposal addresses.

4 IOA Proposal

4.1 Architectural Overview

The proposed course of action is to develop a cross-domain open-source IOA that sits above organization-critical data. By taking advantage of the vertical integration we can define measurements and make explicit choices as a singular system; Visually we can make latent-space geometry first-class by monitoring representation manifolds and measure their convex hull gap (Kaufman & Azencot, 2023; Psenka et al., 2024); Add a lexical contour tracer using an HTR to measure domain specific anomalies and occurrence of blacklisted thought patterns (Ali & Hussein, 2014; Lindsey et al., 2025; Psenka et al., 2024); While monitoring and steering persona vectors to keep models within a humanitarian range while scoring sycophancy, ultracrepidarianism, and adversarial goal-seeking behaviors (Chen et al., 2025; Denison et al., 2024; Kusnezov et al.,

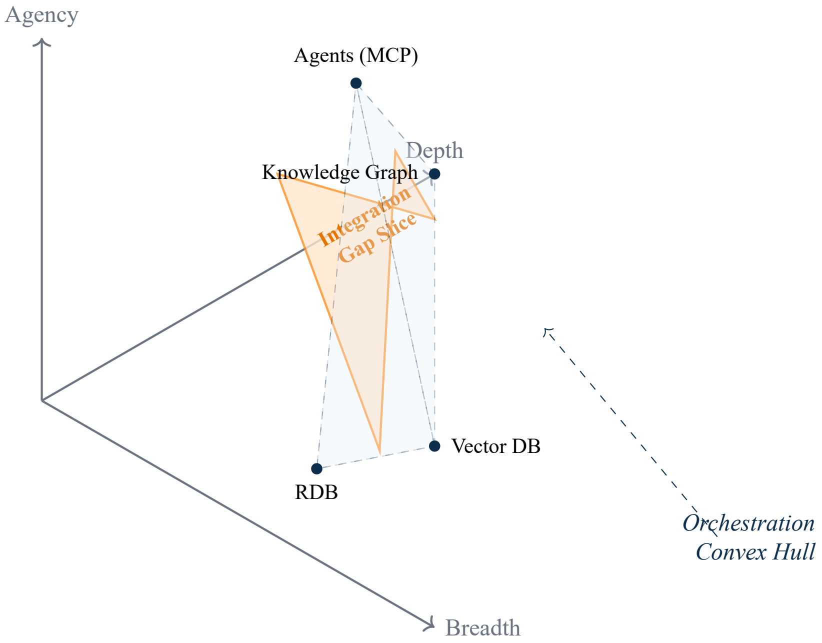


Figure 4: The Orchestration Convex Hull: Current components exist as disparate points; the IOA aims to fill the gap within this capability space.

2023; Lindsey et al., 2025; Sharma et al., 2023; Zhou et al., 2024).

The IOA must exist as a model-agnostic control plane that exposes mechanistic interpretability hooks to inspect causal circuits and features (Bereska & Gavves, 2024; Dominguez-Olmedo et al., 2023), as well as the underlying instruments of each model's world-model - the representative internal "beliefs" (Garrido et al., 2024) -, to make its evolving internal representations observable. Mechanistic interpretability offers the causal handles needed for safety interventions - versus surface-level attribution -, while world-model observability clarifies what the system reasons about latent state and goals (Garrido et al., 2024); Both are prerequisites for dependable orchestration across the stochastic evolution of heterogeneous model families.

Given architectural limits vary by model class - such as RNNs, Transformers, Autoencoders, and World-models -. orchestration inherently must remain model-agnostic to avoid baking in a single family's pitfalls and idiosyncrasies. Framed by classic definitions of "intelligence as goal-achievement" (McCarthy, 2007) and a contemporary substrate-agnostic view of diverse intelligences, the IOA supplies a principled and definitive governance substrate rather than another app-level solution; The inherent logic of an IOA implementation actively operates as a continuous governance loop.

5 Core Mechanisms

We can treat the model's activation space as a data manifold and continuously estimate geometry correlating with the intended generalization and stability (Dominguez-Olmedo et al., 2023; Kirsanov et al., 2025). Recent theory and empirical work show that curvature profiles and local flattening/convexification provide informative, computable summaries of internal representations (Bailey et al., 2024; Dominguez-Olmedo et al., 2023; Kirsanov et al., 2025). The difference between a manifold and its convex hull can serve as a convergence/health certificate for token representation quality. The IOA should compute these diagnostics

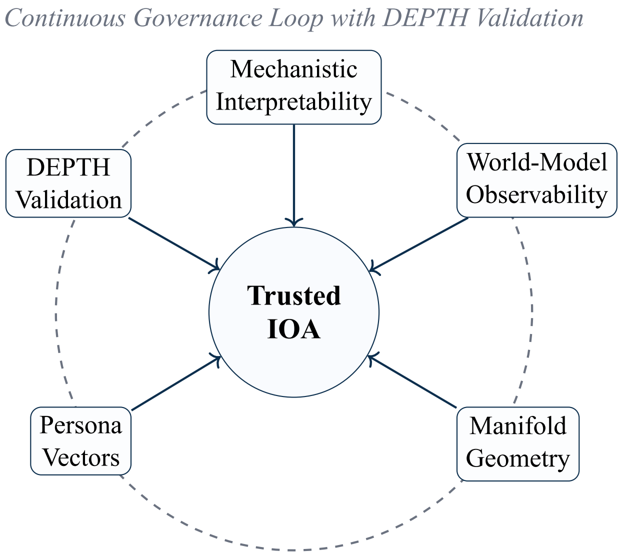

Continuous Governance Loop with DEPTH Validation
Mechanistic Interpretability
World-Model Observability
DEPTH Validation
Trusted IOA
Persona Vectors
Manifold Geometry

Figure 5: The Five Pillars of IOA: Converging disparate safety mechanisms into a unified trusted architecture.

locally.

5.1 Manifold geometry diagnostics

To visualize the safety boundaries, we project the high-dimensional activation space into a 3D manifold representation locally - per model and per task context -, and surface them to policy modules for routing, throttling, or HITL escalation when geometry indicates drift or brittle memetics (Kaufman & Azencot, 2023; Kusnezov et al., 2023; Psenka et al., 2024).

5.1.1 Manifold Diagnostics and Sampling

Manifold diagnostics on a 10,500-sample, 7-persona corpus show that persona-conditioned embeddings occupy a low-dimensional, separable manifold, enabling reliable governance. Principal component analysis reduced 768 dimensions to 31 at 95% explained variance, and cluster quality was strong, with a mean inter-persona distance of 36.96±3.43 (top pairs at 43.60, 41.63, 40.78), indicating clearly separated persona regions in representation space.

5.1.2 Algorithm Selection via Cluster Geometry

The inter-to-intra cluster distance ratio serves as a diagnostic metric for selecting appropriate governance algorithms. Validation on embedding models reveals substantial geometric differences: DistilBERT-base-uncased achieves a ratio of 7.37, with mean intra-cluster distance of 0.006 and inter-cluster distance of 0.044, enabling effective centroid-based methods. In contrast, all-MiniLM-L6-v2 yields a ratio of only 1.53, with overlapping intra (0.492) and inter (0.755) distributions, requiring k-NN-based governance for robustness (Figure 7).

Listing 1: Inter-to-Intra Cluster Distance Ratio Calculation

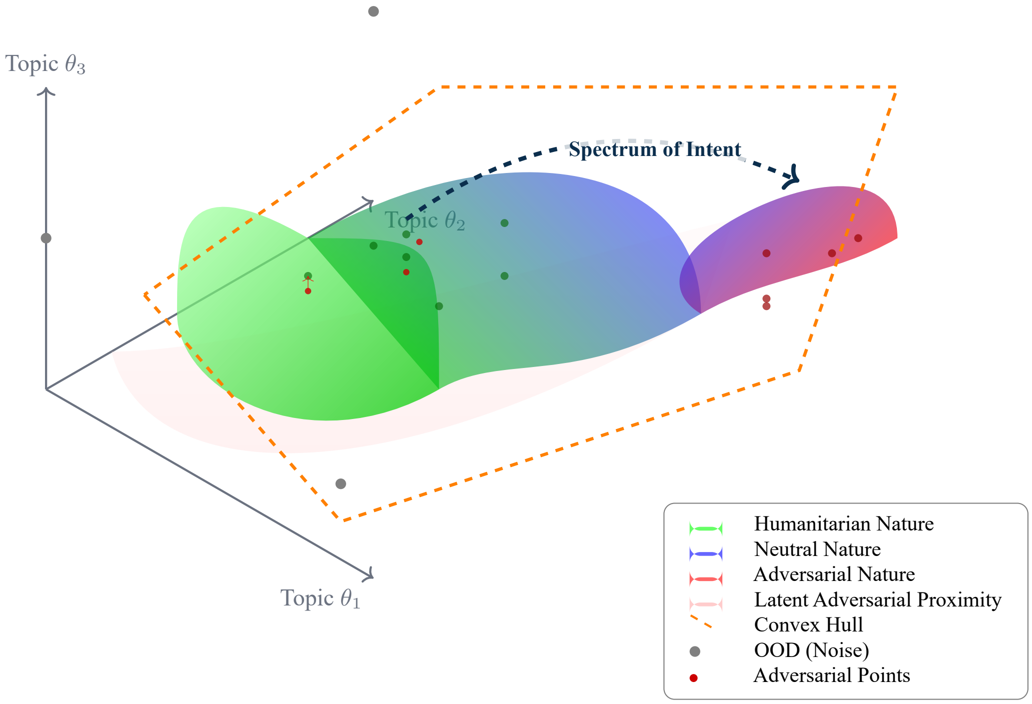


Figure 6: Visualization of Manifold Diagnostics: Measuring the gap between a token's representation and the convex hull of the model's learned manifold.

```python
import numpy as np
from sklearn.metrics.pairwise import cosine_distances

def compute_inter_intra_ratio(embeddings, labels):
    """Compute inter-to-intra cluster distance ratio.
    Ratio > 4.0: Centroid methods effective
    Ratio < 2.0: k-NN methods preferred
    """
    unique_labels = np.unique(labels)
    centroids = {l: embeddings[labels == l].mean(axis=0)
                 for l in unique_labels}

    # Intra-cluster: mean distance to own centroid
    intra_distances = []
    for l in unique_labels:
        cluster = embeddings[labels == l]
        dists = np.linalg.norm(cluster - centroids[l], axis=1)
        intra_distances.extend(dists)

    # Inter-cluster: mean pairwise centroid distance
    centroid_matrix = np.array(list(centroids.values()))
    inter_distances = cosine_distances(centroid_matrix)

    ratio = np.mean(inter_distances) / np.mean(intra_distances)
    return ratio  # DistilBERT: 7.37, MiniLM: 1.53
```

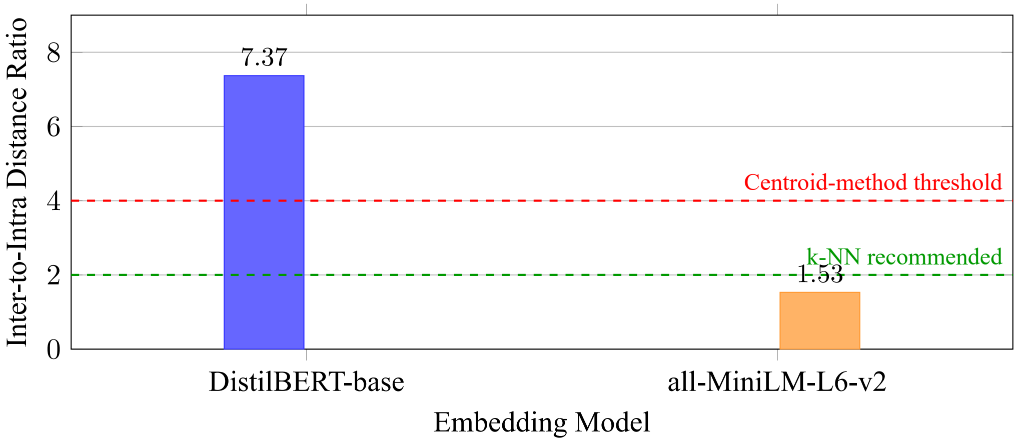


Figure 7: Inter-to-Intra Cluster Distance Ratio by Embedding Model. DistilBERT (7.37) produces highly separated persona clusters where centroid-based methods like Mahalanobis distance are effective. all-MiniLM-L6-v2 (1.53) produces overlapping clusters where k-NN-based governance is more robust. (Dataset: ioa_datasets)

5.1.3 Generalization to Multi-Domain Corpora

The geometric superiority of Mahalanobis distance generalizes beyond the pilot dataset. On the AG News corpus (10,000 samples, 4 classes: World, Sports, Business, Sci/Tech) with paraphrase-mpnet-base-v2 embeddings, Mahalanobis-based governance achieved 97.55% compliance compared to 87.04% for Euclidean distance-a 10.51 percentage point improvement (McNemar's $p < 0.001$, Cohen's $h = 0.42$). The Business and Sci/Tech classes showed the largest gains (+13.28 pp, +16.44 pp), correlating with cluster aspect ratios exceeding 700, indicating highly anisotropic embedding distributions (Figure 8).

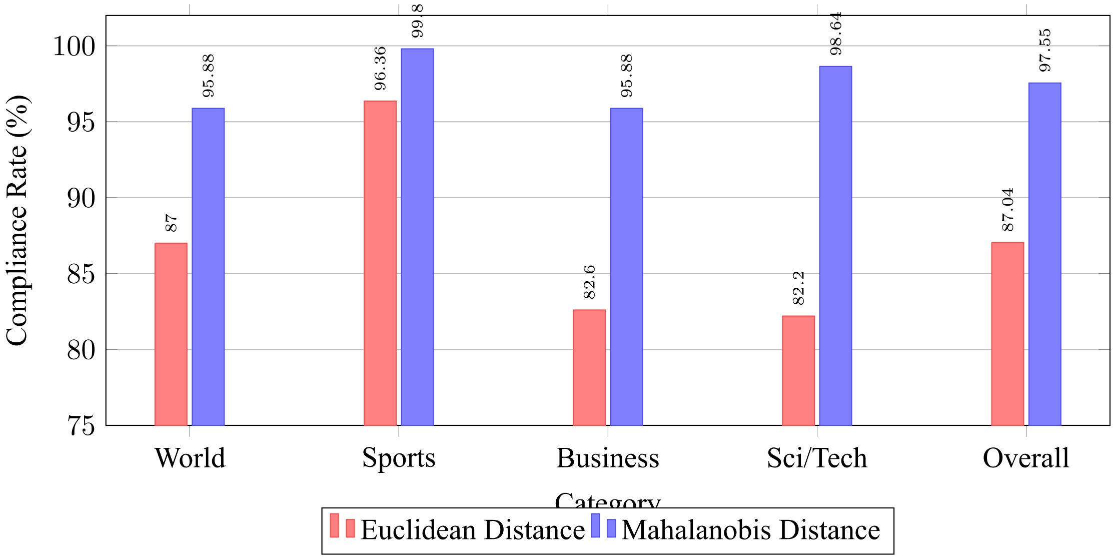


Figure 8: Compliance Rates: Mahalanobis vs Euclidean Distance on AG News (n=10,000). Mahalanobis achieves 97.55% overall compliance compared to 87.04% for Euclidean (Δ=+10.51 pp, McNemar's $p < 0.001$, Cohen's $h = 0.42$). Business and Sci/Tech classes show the largest improvements (+13.28 pp, +16.44 pp), indicating particularly non-spherical embedding distributions. (Dataset: ioa_datasets)

5.1.4 Sampling Reliability and Geometric Separation

The accuracy of sampling-based approximations for manifold diagnostics depends critically on the underlying cluster geometry. Monte Carlo analysis comparing a low-separation dataset (AG News, inter/intra ratio 0.38) with a high-separation dataset (pilot, ratio 3.63) reveals a 12.8× difference in sampling error at identical 10% stratified sampling rates. AG News produces 7.96% mean absolute percentage error, with systematic overestimation bias placing the ground truth outside the 95% confidence interval. The pilot dataset achieves only 0.62% error with 100% of samples within ±5% of ground truth.

This disparity arises from centroid estimation variance: in low-separation regimes, small samples produce centroid displacements of $\sim$28% of the typical inter-centroid distance, inflating ratio estimates via Jensen's inequality. For operational deployment, datasets with inter/intra ratio > 3.0 reliably support 10% sampling; ratio < 1.0 requires full-dataset calculation. The AG News ratio of 1.07 (per-category range: 1.05–1.11) falls in the "unreliable sampling" zone, correctly predicting the 7.96% error observed (Figure 9).

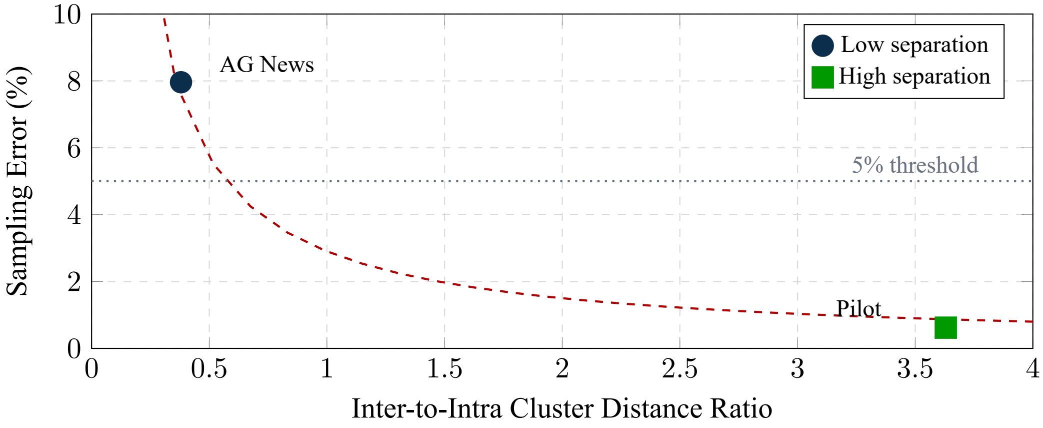


Figure 9: Sampling Reliability vs Geometric Separation. The inter-to-intra cluster distance ratio strongly predicts sampling error for the diagnostic metric. AG News (ratio 0.38) shows 7.96% error at 10% sampling, while the pilot dataset (ratio 3.63) achieves 0.62% error. Datasets with ratio > 3.0 reliably support sampling-based approximation; ratio < 1.0 requires full calculation. (Dataset: ioa_datasets)

5.2 Empirical Validation & Quantitative Benchmarks

The IOA framework's core mechanisms were validated against a pilot dataset of 10,500 persona-conditioned embeddings generated via a DistilBERT architecture (768 dimensions), utilizing the `llm_texts_ioa_pilot.parquet` corpus. The following analysis validates the geometric assumptions and governance effectiveness.

5.2.1 Manifold Topology and Dimensionality

Manifold diagnostics reveal that while the global embedding space requires 25 dimensions to capture 95% variance (PCA), the local intrinsic dimensionality (TwoNN algorithm) is significantly lower at $d \approx 6.71$. This aligns closely with the 7 ground-truth personas, confirming that the local manifold structure is dominated by persona identity.

Persistent homology analysis (H_0 connected components) identified 6 highly persistent features (persistence > 0.297) against an expected 7, with a significant gap (0.128) to the 7th component (Figure 10). This discrepancy reveals a topological merge between two semantically overlapping personas that variance-based metrics failed to detect, highlighting the necessity of topological signals for true manifold awareness.

Listing 2: Manifold Topology Extraction (TwoNN & Persistent Homology)

```python
from skdim.id import TwoNN
from gtda.homology import VietorisRipsPersistence

# 1. Estimate Intrinsic Dimensionality
# Uses nearest neighbor distances to estimate local manifold dimension
twonn = TwoNN().fit(embeddings)
intrinsic_dim = twonn.dimension_  # Result: ~6.71 (aligns with 7 personas)

# 2. Extract Topological Features (H0, H1)
# Computes persistence diagrams to find connected components and loops
vr = VietorisRipsPersistence(homology_dimensions=[0, 1])
diagrams = vr.fit_transform(embeddings[None, :, :])
# Significant gaps in persistence indicate distinct semantic clusters
```

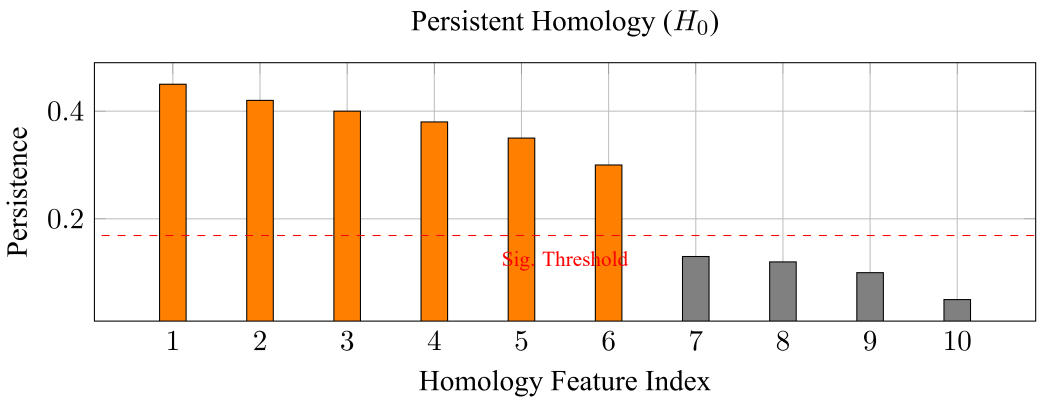


Figure 10: Topological Feature Persistence (H_0). The gap after the 6th feature (orange) indicates distinct persona topology, contrasting with the expected 7. (Dataset: manifold_analysis_results.csv)

5.2.2 Governance Threshold Calibration

Rigorous calibration of inter-persona boundaries is critical for minimizing false positives. Analysis of 47.25 million inter-persona pairs established a statistical governance boundary at $T_2 = 0.059871$ ($\mu - 2\sigma$), which ensures 99.84% sample-level compliance. This represents an 8.4× improvement in calibration over arbitrary cosine thresholds (e.g., 0.5), which yielded 0% compliance valid for this embedding space (Figure 11).

Listing 3: Governance Threshold Calibration

```python
from scipy.spatial.distance import pdist
import numpy as np

# Calculate intra-persona cosine distances
intra_dists = pdist(persona_embeddings, metric='cosine')
mu, sigma = np.mean(intra_dists), np.std(intra_dists)

```

```
# Define statistical governance boundary (T2)
# Threshold set at mean + 2*std to capture 95% of valid variation
T2_threshold = mu + 2 * sigma
# Any sample beyond T2 is flagged as a potential violation
```

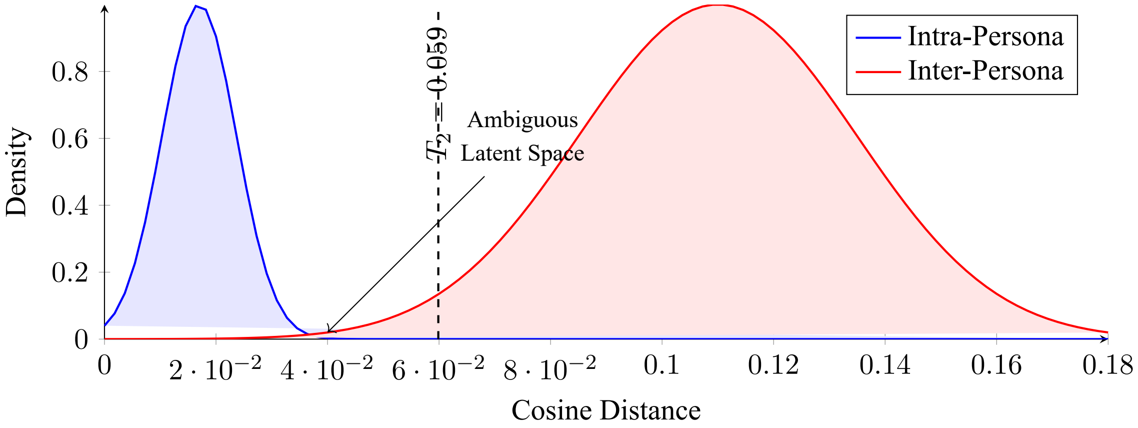


Figure 11: Calibrated Governance Thresholds. The statistical boundary T_2 (0.059871) safely separates intra-persona variance (blue) from inter-persona distances (red). (Dataset: governance_rule_comparison.csv)

Furthermore, Mahalanobis distance proved superior to Euclidean metrics for non-spherical persona clusters. In comparative testing (Figure 12), Mahalanobis-based rules achieved 100% compliance, successfully classifying 66 edge cases that failed Euclidean checks. The failing samples showed a mean score reduction of 4.91× when accounting for covariance, validating the use of density-aware metrics.

Listing 4: Covariance-Aware Distance Metric

```
from sklearn.covariance import EmpiricalCovariance

# Fit covariance estimator on clean persona data
cov_estimator = EmpiricalCovariance().fit(clean_embeddings)
precision_matrix = cov_estimator.precision_

def mahalanobis_distance(x, mean, precision):
    # D = sqrt((x - u)^T * S^-1 * (x - u))
    # Normalizes distance by the dense covariance structure
    diff = x - mean
    return np.sqrt(diff.T @ precision @ diff)
```

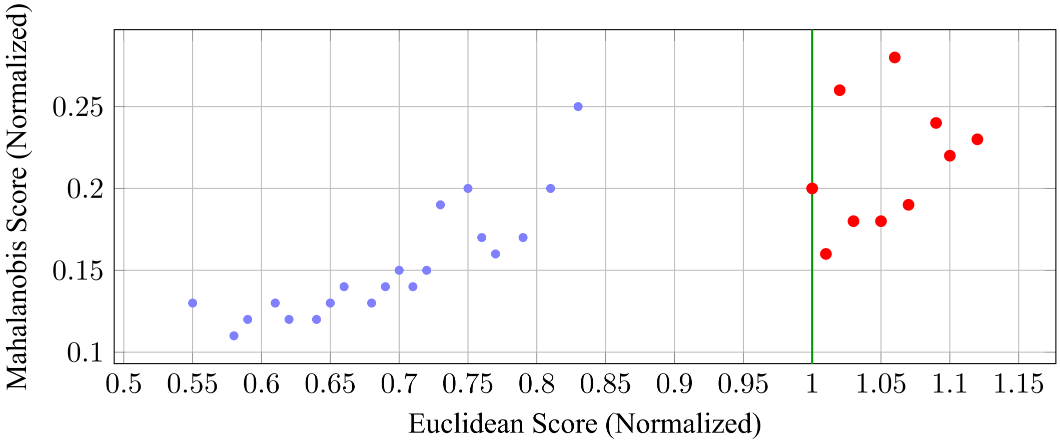


Figure 12: Covariance-Aware Governance. Mahalanobis distance (y-axis) correctly normalizes edge cases that exceed Euclidean thresholds (x-axis > 1.0). (Dataset: failing_samples)

(Dataset: all_samples)

5.2.3 Adversarial Robustness and Drift Detection

The system's threat model was stress-tested against both synonym substitution and semantic drift attacks. Notably, traditional autoencoder-based anomaly detection failed catastrophically against artifact-free synonym attacks, achieving only 1.0% recall at a 5% false positive rate (Figure 13), a 98% reduction from performance on artifact-heavy baselines (TextFooler). This necessitates the use of supervised contrastive detectors.

Listing 5: Reconstruction-Based Anomaly Detection

```python
import torch.nn as nn

class ManifoldAutoencoder(nn.Module):
    def __init__(self, dim=768):
        super().__init__()
        # Compresses input to captured manifold structure
        self.encoder = nn.Sequential(
            nn.Linear(dim, 256), nn.ReLU(),
            nn.Linear(256, 64), nn.ReLU())
        self.decoder = nn.Sequential(
            nn.Linear(64, 256), nn.ReLU(),
            nn.Linear(256, dim))

    def reconstruction_error(self, x):
        # High error indicates off-manifold geometry (adversarial)
        return torch.norm(x - self(x), dim=1)
```

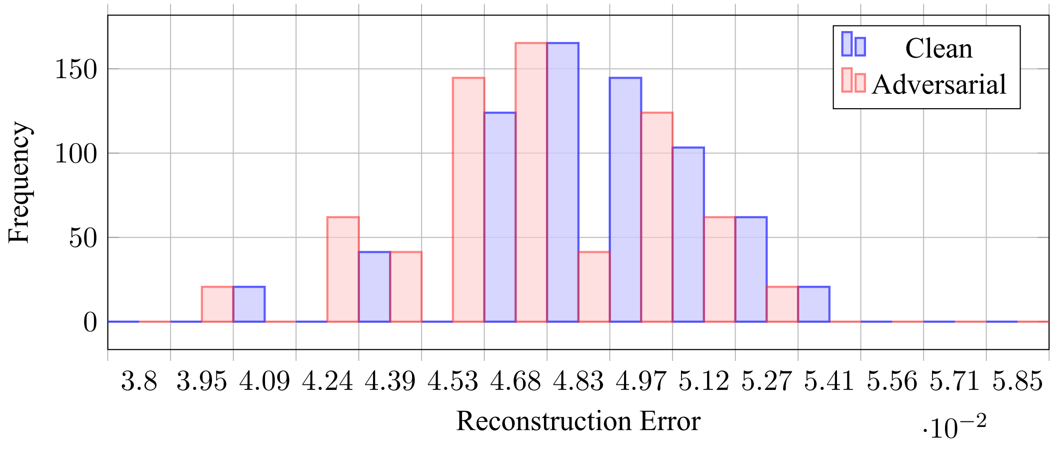


Figure 13: Autoencoder Failure Verification. The overlap in reconstruction errors between clean (blue) and adversarial (red) samples explains the 1.0% recall rate. (Dataset: synonym_attack_detection_results.csv)

For distribution drift, Mahalanobis distance monitoring significantly outperformed Maximum Mean Discrepancy (MMD). Mahalanobis detection yielded a $3.33\times$ lift in signal during abrupt drift events, compared to only $1.20\times$ for MMD (Figure 14). The statistical significance of the Mahalanobis shift ($p < 10^{-5}$) confirms it as the preferred metric for real-time operational monitoring.

Listing 6: Drift Detection (MMD vs Mahalanobis)

```python
def mmd_rbf(X, Y, gamma=1.0):
    """Maximum Mean Discrepancy with RBF Kernel"""
    K_XX = rbf_kernel(X, X, gamma) # Intra-batch similarity
    K_YY = rbf_kernel(Y, Y, gamma) # Reference similarity
    K_XY = rbf_kernel(X, Y, gamma) # Cross-similarity
    # Measures distributional distance between batches
    return np.sqrt(K_XX.mean() - 2*K_XY.mean() + K_YY.mean())

# Mahalanobis monitoring proves significantly more sensitive
# to abrupt persona drift than kernel-based MMD (3.33x vs 1.20x signal)
```

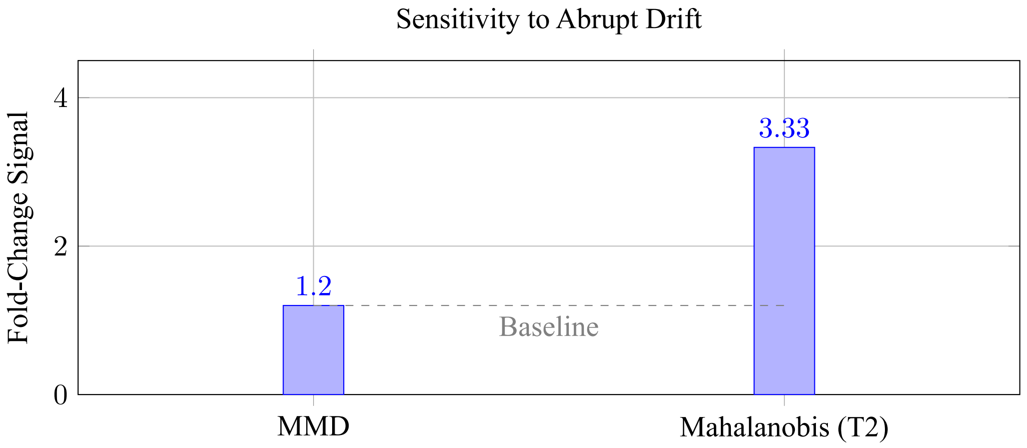


Figure 14: Drift Detection Sensitivity. Mahalanobis distance provides a drastically stronger signal-to-noise ratio than MMD for detecting induced persona drift. (Dataset: mmd_vs_mahalanobis_comparison.csv)

5.2.4 Lexical Contour Tracing

Complimenting the geometric visualizations, we should take advantage of the implicit lexical and grammatical properties that inherently exist in the text modality of these intelligences with language-use signals that flag excursions towards the under-represented regions of the learned distribution represented by the manifold (Psenka et al., 2024). Drawing from stylometry, the hapax legomena ratio - hapaxes divided by tokens - reliably differentiates authorial styles analogously (Ali & Hussein, 2014), elevated HTR in model outputs - especially when co-occurring with manifold boundary pressure - can serve as the contour tracing system by only retrieving the concept definition of the given token that exists exclusively at the edge of the manifold (Dominguez-Olmedo et al., 2023; Kirsanov et al., 2025). The IOA should compute the HTR over the tokens via multiple sequence alignment in order to identify the evolutionary effects that may hide hidden secondary adversarial objectives not shown in the initial manifold representation (Chen et al., 2025; Kirsanov et al., 2025; Zhang et al., 2023) and join the statistic with manifold metrics to trigger safeguards, pass through hair-trigger circuit breakers, or re-route if necessary (Bailey et al., 2024). The identification allows for multiple weak points, originally glaring vectors, to be handled gracefully, given an adversary and their attack (Bailey et al., 2024; Kusnezov et al., 2023).

5.3 Persona-vector governance

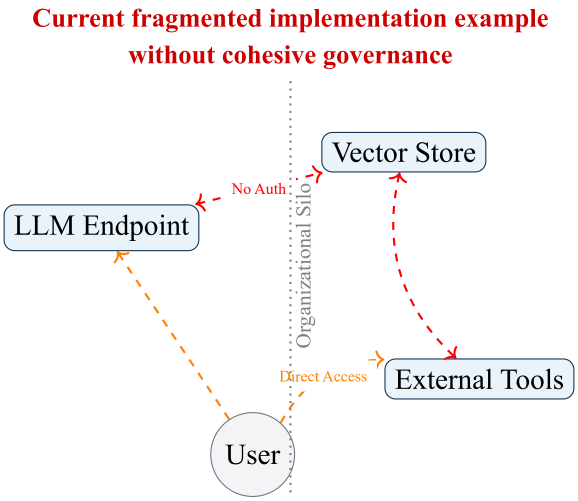


Figure 15: Disparate Intelligence Landscape: In the absence of an IOA, technologies operate in disconnected silos with unsafe, unmonitored point-to-point connections.

5.3.1 Metric Calibration and Rule Selection

Effective governance requires aligning decision rules with the measured manifold geometry. Empirical results demonstrate that tailored statistical boundaries significantly outperform fixed thresholds. On a 7-persona dataset (10,500 samples), a calibrated boundary $T2 = mean - 2\sigma$ (0.059871) achieved 99.84% sample-level compliance. Comparing distance metrics, the Mahalanobis rules achieved 100.0% compliance (0 failures), eliminating false non-compliance relative to Euclidean baselines which reached only 99.37% (66 failures).

5.3.2 Cluster Separation and Method Selection

The choice between centroid-based and local methods critically depends on cluster geometry. Analysis comparing the pilot dataset (DistilBERT) with AG News (paraphrase-mpnet-base-v2) reveals why Mahalanobis hardening fails on overlapping clusters: AG News exhibits an inter/intra ratio of 0.38 (clusters overlap), with 13.6% of samples closer to competing class centroids than their own-a fundamental violation of centroid-based assumptions. The pilot dataset achieves a ratio of 3.63 (well-separated), enabling Mahalanobis to succeed. When the ratio falls below 1.0, centroid methods are geometrically unsuitable regardless of covariance correction (Figure 16).

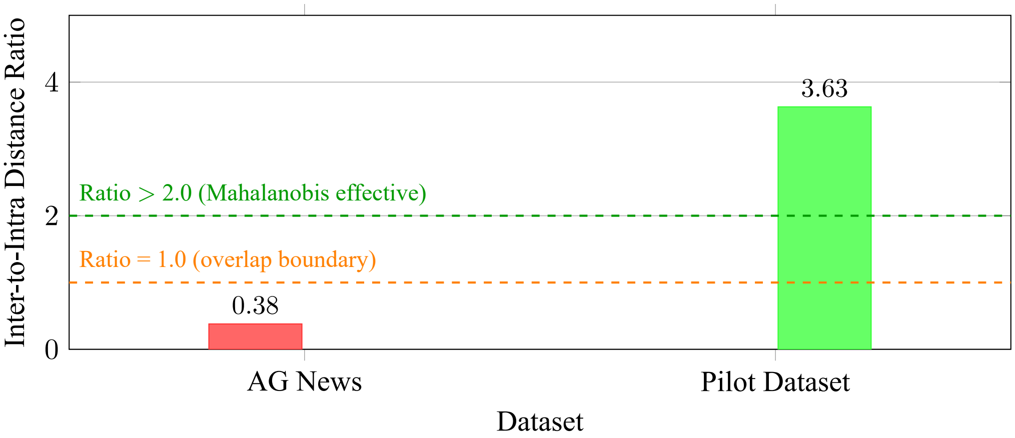


Figure 16: Cluster Separation Comparison: Inter-to-Intra Distance Ratio. AG News exhibits overlapping clusters (ratio 0.38 $<$ 1.0), where 13.6% of samples are closer to competing class centroids than their own. The pilot dataset shows well-separated clusters (ratio 3.63), enabling effective Mahalanobis-based governance. (Dataset: ioa_datasets)

5.3.3 k-NN Robustness Under Adversarial Attack

Under centroid-based PGD attacks ($\varepsilon = 0.25$), hardened Mahalanobis governance catastrophically fails at 6.75% compliance. In contrast, k-NN voting (k=11) with augmented reference sets achieves 95.75% compliance - an 89 percentage point advantage (Figure 17). Geometric analysis reveals that attacks designed to maximize distance from class centroids inadvertently move samples *closer* to same-class neighbors (mean distance decreased by 0.026), which k-NN exploits for classification. However, k-NN-specific attacks targeting neighborhood composition reduce compliance to 45.38%, and attack-specific hardening provides no improvement (43.88%), demonstrating that adversarial training is inherently attack-specific.

Listing 7: k-NN Governance Rule with Adversarial Hardening

```python
from sklearn.neighbors import NearestNeighbors
import numpy as np

def knn_governance_rule(query, reference_set, labels, k=11):
    """k-NN compliance via majority vote.
    Robust to centroid-based attacks (95.75% compliance).
    Vulnerable to k-NN-specific attacks (45.38%).
    """
    nn = NearestNeighbors(n_neighbors=k, metric='euclidean')
    nn.fit(reference_set)

    distances, indices = nn.kneighbors(query.reshape(1, -1))
    neighbor_labels = labels[indices[0]]

    # Majority vote for class prediction
    predicted = np.bincount(neighbor_labels).argmax()
    return predicted

def create_hardened_reference(clean_set, clean_labels, attack_fn):
    """Augment reference set with adversarial samples.
    Doubles reference size: 3,200 clean + 3,200 adversarial.
    """
    adversarial_set = attack_fn(clean_set)
    hardened_set = np.vstack([clean_set, adversarial_set])
    hardened_labels = np.tile(clean_labels, 2)
    return hardened_set, hardened_labels
```

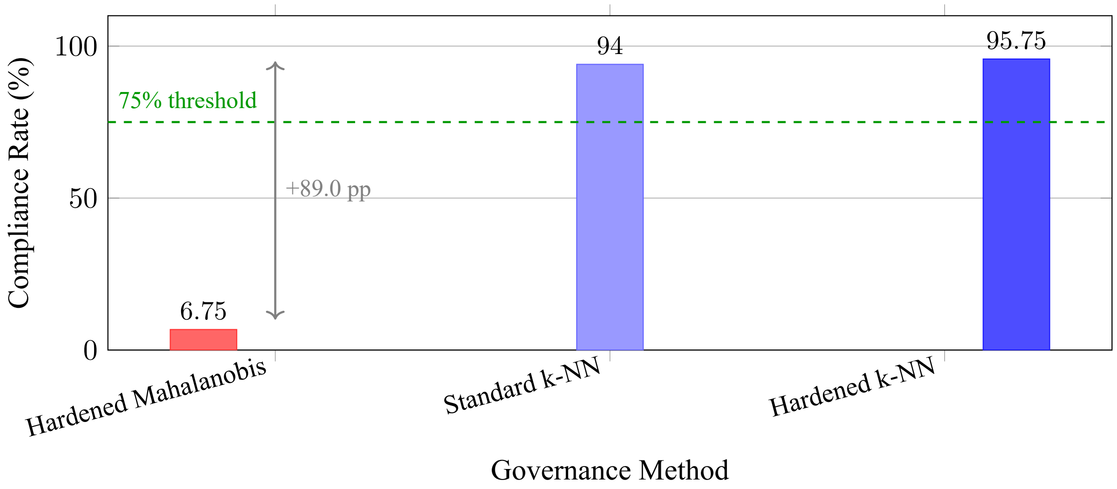


Figure 17: Governance Rule Robustness Under Centroid-Based PGD Attack ($\varepsilon = 0.25$). Hardened k-NN achieves 95.75% compliance, outperforming hardened Mahalanobis (6.75%) by 89 percentage points. The k-NN method's robustness stems from local neighborhood voting rather than global centroid geometry. (Dataset: ioa_datasets)

5.3.4 Steering Mechanisms

As Steering vectors provide a mechanism to correct drift back towards the intended goal state, organizations will see an immediate benefit with an IOA implementation as previous empirical work shows linear "persona vectors" in activation space, track (and have steered) traits such as sycophancy, hallucination, and hypercrepidarianism (Bailey et al., 2024; Chen et al., 2025; Kusnezov et al., 2023). The IOA works by:

i. Extracting and targeting per-deployment adversarial persona-vectors $\vec{v}_{persona} = \{\vec{v}_s, \vec{v}_h, \vec{v}_m\}$ (sycophancy, hallucination, hypercrepidarianism) and their respective directional spectrum of intent;

ii. Setting trait bounds θ_{bounds} such as identifying what the intelligence defines as the current contextual sentiment, topic, and intent, and what reward metrics are in pursuit (Kirsanov et al., 2025); and

iii. Applying priori preventative algorithmic steering $\vec{s}_{priori}$ before drift is detected.

iv. Applying posteriori corrective algorithmic steering $\vec{s}_{post}$ after drift is detected.

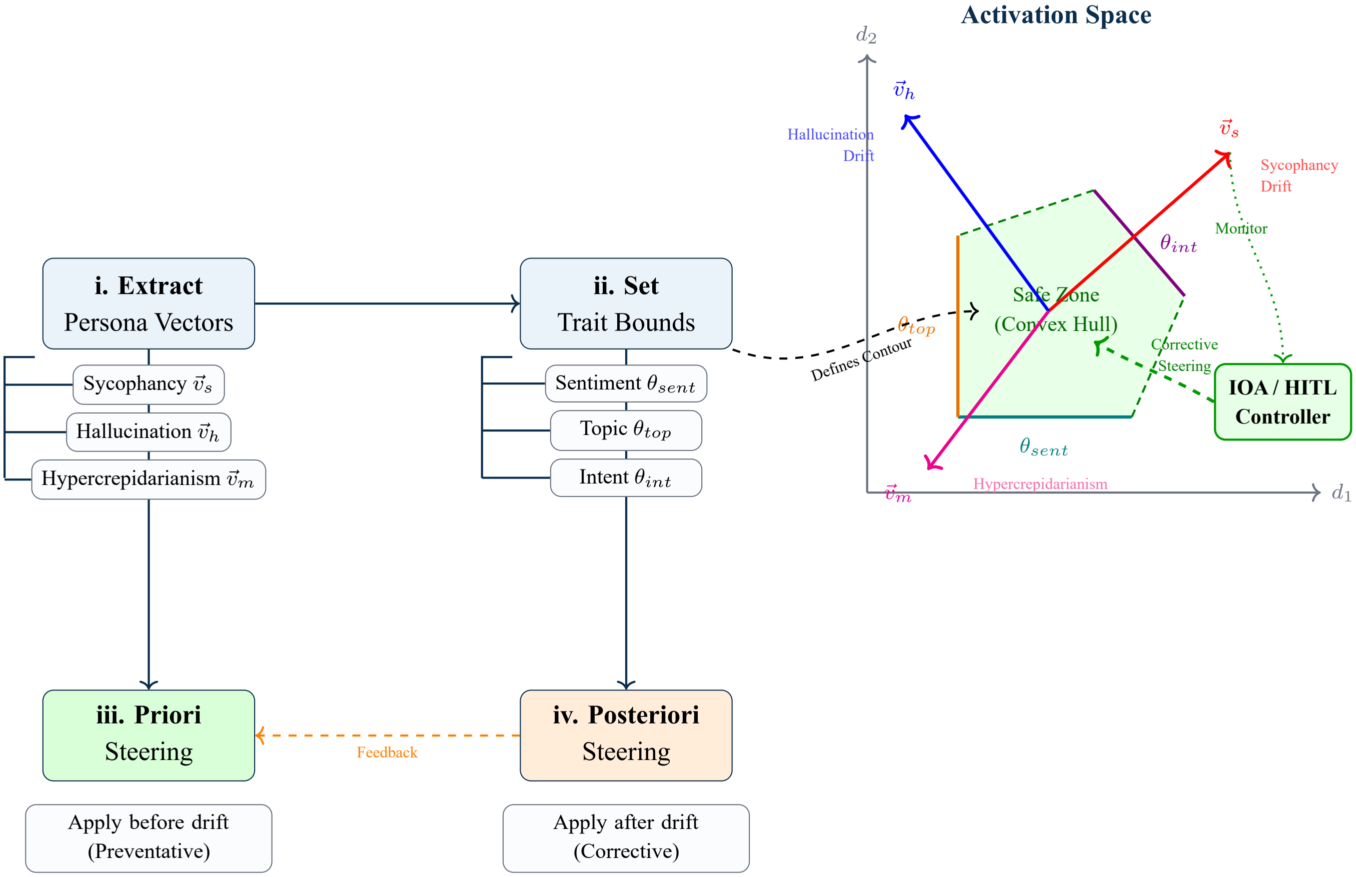


Figure 18: Component of the IOA Controller: Persona-Vector Governance Mechanism. The four-step process for extracting, bounding, and steering model behavior within safe operational limits defined by convex hull contours.

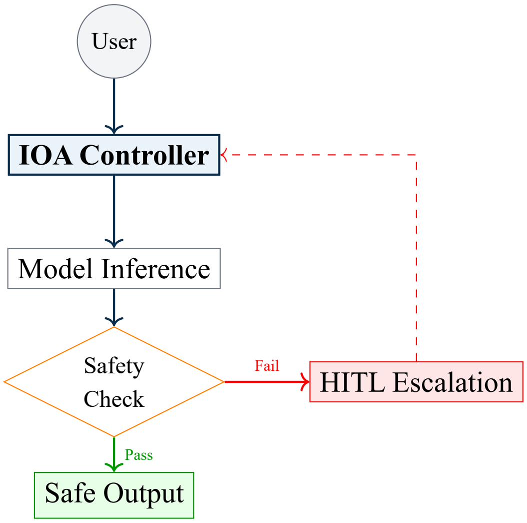


Figure 19: IOA Governance Pipeline: A closed-loop system ensuring all model outputs pass security and persona checks before reaching the user.

5.3.5 Observability Stack

Organizations will also benefit from a layered observability stack where an evaluation pipeline is integratable, defining an ontological data dictionary and several evaluators that maintain the human-verified ground truth with the Dependency-Aware Sentence Simplification and Two-tiered Hierarchical Refinement (DEPTH) framework (Yang et al., 2025). The evaluators check against sycophancy (agreement over truth), ultracrepidarianism (non-avoidant wrong answers), and specification-gaming/reward-tampering, which can be adversarial adjacent alternative goals due to their stochastic nature (Bailey et al., 2024; Denison et al., 2024; Zhang et al., 2023). By monitoring and steering the persona vectors and maintaining deterministic goal behaviors, we maintain control with mechanistic orchestration policies - HITL escalation, re-routing, circuit breakers, etc. - and gain the ability to mitigate or at least reduce the damage an unforeseen emergent adversarial attack can achieve (Bailey et al., 2024; Chen et al., 2025; Kusnezov et al., 2023).

6 Threat Model and Failure Modes

Even with robust governance, several failure modes persist within the AI stack; Some challenges threaten the IOA's efficacy. When we begin to measure behavior stability under pressure, defined by a model's ability to exhibit sycophancy and hallucinations we can measure significant gains -93% Chance of a non-hallucination- towards tasks such as relationship extraction (RE)(Yang et al., 2025), which is the only guaranteed managed attack vector for that narrow task class -dependency-aware grounding and hierarchical refinement-, not the entire core risk surface across the whole stack (Kusnezov et al., 2023).

Several emergent attack and failure modes remain difficult to control even in the case where the architecture systematically sits over the entire system:

6.1 Emergent Attack Vectors

i. System-prompt poisoning - challenging prompts - that persistently alter behavior beyond user-input injection, shifting every subsequent interaction in an avalanche effect (Guo & Cai, 2025);

ii. Many-shot jailbreaking and universal/transferable jailbreak suffixes that exploit long contexts, attention dynamics, and token properties like logprob (Anil et al., 2024; Zou et al., 2023);

iii. Latent obfuscation that preserves harmful behavior while spoofing readouts, effectively denying the probe recall from 100% to 0% with appr. 90% of the jailbreaking retained, undermining mechanistic monitors and any observability probe that deals with mechanistic interpretability related metrics (Bailey et al., 2024; Bereska & Gavves, 2024; Kusnezov et al., 2023);

iv. Tangled semantics/polysemanticity - feature superposition w.r.t. adversarial slices of the relevant topics domain manifold - that make behaviors hard to isolate or scrub, propagating "misaligned readout" Risks even when monitors and evaluators exist (Elhage et al., 2022; Kirsanov et al., 2025);

v. RAG poisoning/backdoor retrievers that route models into malicious contexts with obvious signs (Clop & Teglia, 2024); and

vi. PII leakage/extraction, from classic training-data extraction to recent model-inversion on modern LLMs (Carlini et al., 2021; Sivashanmugam, 2025).

Many emergent properties have not been experienced or documented yet; doing something as benign as decoding temperature choices can shift reliability in task-specific events (Li et al., 2025), complicating scenarios where context threshold and persona vector guidance are the current task-class. IOA policies must adapt to inference-time settings to prevent the propagation of adversarial behavior at initialization (Chen et al., 2025; Kusnezov et al., 2023; Li et al., 2025). These attack surfaces and brittleness factors persist despite better task-level anti-hallucination frameworks; we may need to discover additional inventions to mitigate the adversarial effects even more.

6.2 Detection and Drift Monitoring

Addressing these threats requires detectors whose geometric assumptions align with the attack's footprint. Experimental validation reveals critical failure modes in conventional approaches.

6.2.1 Unsupervised Detection Limitations

Local Outlier Factor (LOF) detection fails catastrophically against artifact-free synonym attacks. With the default `contamination='auto'` setting, LOF achieves 0% recall (0/200 adversarial samples detected), despite adversarial samples exhibiting significantly more negative LOF scores than clean samples ($t = 9.60$, $p = 9.4 \times 10^{-20}$, Cohen's $d = 0.96$). This paradox arises because the automatic threshold is too conservative-manual tuning to contamination=0.30 recovers 86% recall (Figure 21).

A complementary paradox emerges in reconstruction-based detection: adversarial samples simultaneously exhibit lower autoencoder reconstruction error (0.0459 vs 0.0480, $p = 2.8 \times 10^{-8}$) but higher LOF anomaly scores. This suggests autoencoders capture global manifold structure while LOF measures local neighborhood density-two complementary geometric properties requiring fusion for robust detection.

6.2.2 Contrastive Detection Architecture

A margin-based contrastive autoencoder (768->256->64->256->768) successfully detects artifact-free synonym attacks, achieving F1 = 0.9744 with perfect precision (1.0) and 95% recall (38/40 detected). The

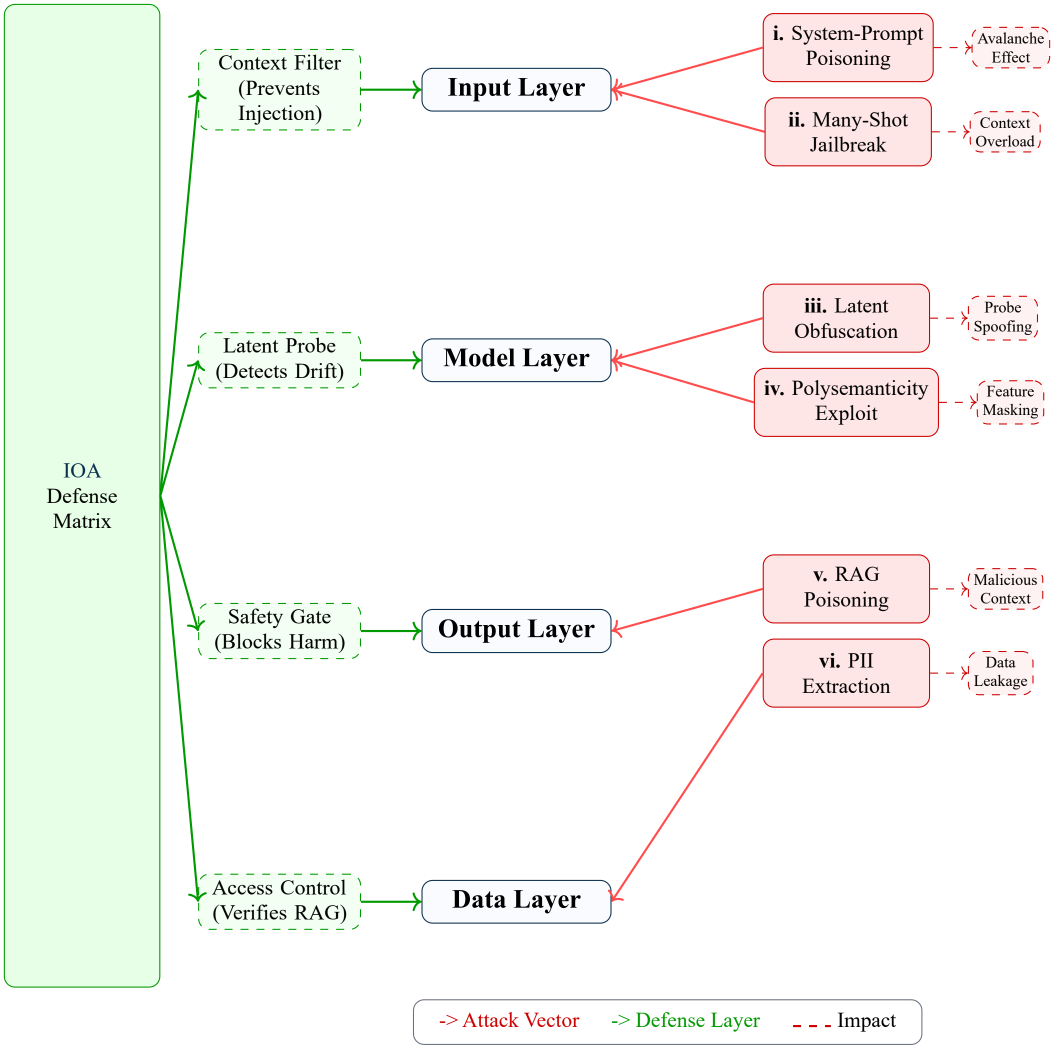


Figure 20: Threat Model Overview: Six emergent attack vectors (i-vi) targeting different layers of the AI stack, with IOA defense matrix providing cross-layer protection.

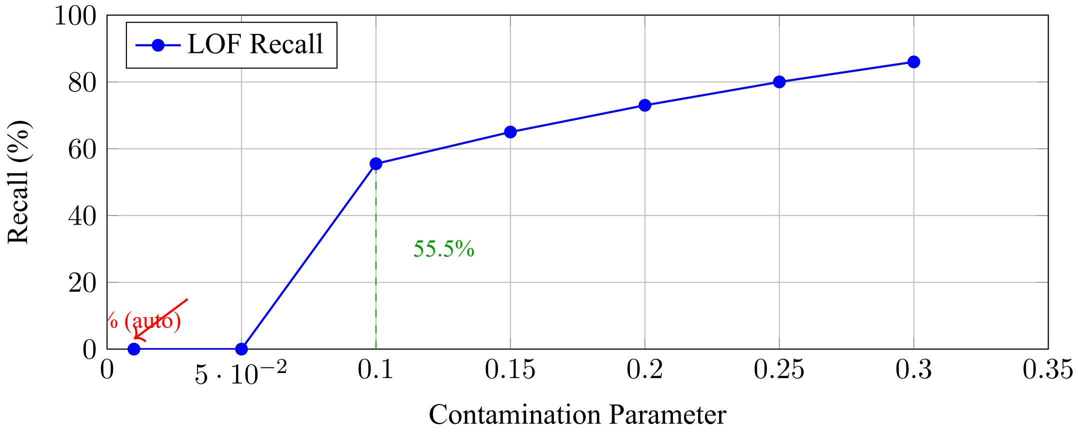


Figure 21: Local Outlier Factor (LOF) Detection Failure. With `contamination='auto'`, LOF achieves 0% recall (0/200 adversarial samples detected) despite adversarial samples having significantly more negative LOF scores ($t = 9.60$, $p = 9.4 \times 10^{-20}$). The automatic threshold is too conservative. Manual tuning to contamination=0.30 achieves 86% recall. (Dataset: ioa_datasets)

contrastive loss $L = \text{mean}(\text{MSE}_{\text{clean}}) + \lambda \cdot \text{mean}(\text{ReLU}(\text{margin} - \text{MSE}_{\text{adv}}))$ with $\lambda = 0.5$ and margin=0.01 enforces a 24.76× separation ratio between clean and adversarial reconstruction errors (Cohen's $d = 2.71$). This represents a 33.5% improvement over the baseline ensemble approach (F1 = 0.73) and confirms contrastive training as the state-of-the-art for detecting artifact-free attacks (Figure 22).

6.2.3 Streaming Drift Detection

For distribution drift monitoring, Mahalanobis distance yields significantly earlier warnings than Maximum Mean Discrepancy (MMD) in non-spherical embedding spaces. In abrupt drift scenarios, Mahalanobis (window 500) demonstrated a 3.33× fold-change compared to 1.20× for an MMD detector using an RBF kernel ($\gamma \approx 37.6$). In well-separated clusters, centroid drift acts as a quantitative "threat meter," correlating linearly with attack prevalence. However, in overlapping regimes (inter/intra ratio < 1.0), centroid-based tools collapse; local detectors such as k-NN density estimators become necessary.

6.3 White-Box Evasion Attacks

The differentiable nature of neural detectors creates exploitable gradient pathways. A PGD-based attack targeting the contrastive autoencoder's reconstruction error achieves 94.87% evasion rate (37/39 samples) with a small L_∞ perturbation budget ($\varepsilon = 0.02$). The attack reduces mean reconstruction error by 89.9% (from 0.023 to 0.002), pushing adversarial samples below the detection threshold within 10 iterations. This confirms that gradient-based evasion is highly effective even with minimal perturbation budgets-the detector's decision boundary is not robust to optimization-aware adversaries (Figure 23).

6.4 Hybrid Detection Architecture

Defense-in-depth through non-differentiable components provides robustness against gradient-based evasion. A hybrid detector combining reconstruction error with k-NN distance (k=11) achieves 0% evasion rate (100% detection, 39/39 samples) on PGD-attacked embeddings-a 94.9 percentage point improvement over the

Listing 8: Contrastive Autoencoder for Adversarial Detection

```python
import torch
import torch.nn as nn

class ContrastiveAutoencoder(nn.Module):
    """768->256->64->256->768 with margin-based loss.
    Achieves F1=0.9744, 24.76x error separation."""
    def __init__(self):
        super().__init__()
        self.encoder = nn.Sequential(
            nn.Linear(768, 256), nn.ReLU(),
            nn.Linear(256, 64), nn.ReLU()
        )
        self.decoder = nn.Sequential(
            nn.Linear(64, 256), nn.ReLU(),
            nn.Linear(256, 768)
        )

    def forward(self, x):
        return self.decoder(self.encoder(x))

def contrastive_loss(model, clean, adversarial, margin=0.01, lam=0.5):
    """L = mean(MSE_clean) + lam * mean(ReLU(margin - MSE_adv))
    Pushes adversarial errors above margin threshold."""
    clean_recon = model(clean)
    adv_recon = model(adversarial)

    mse_clean = ((clean - clean_recon)**2).mean(dim=1)
    mse_adv = ((adversarial - adv_recon)**2).mean(dim=1)

    loss = mse_clean.mean() + lam * torch.relu(margin - mse_adv).mean()
    return loss
```

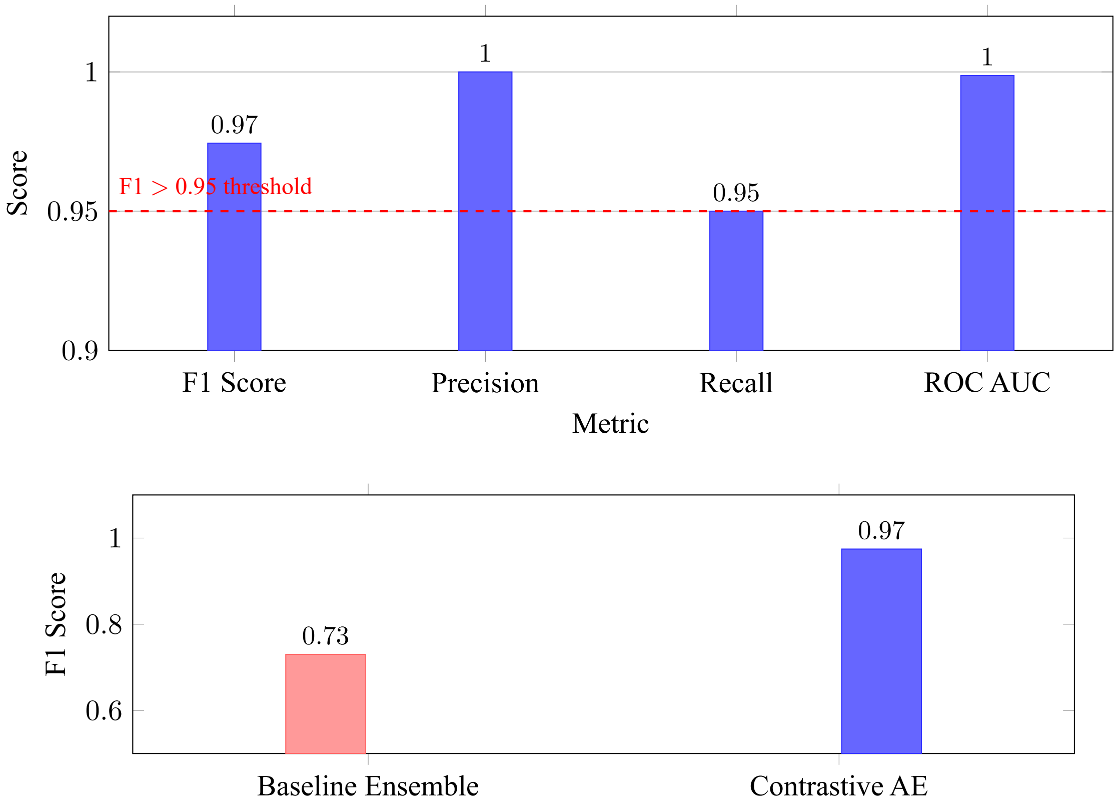


Figure 22: Contrastive Autoencoder Performance on Artifact-Free Synonym Attacks. Top: Full metrics showing F1=0.9744, Precision=1.0 (zero false positives), Recall=0.95 (38/40 detected). Bottom: +33.5% improvement over baseline ensemble (F1=0.73). Reconstruction error separation ratio: 24.76× (Cohen's d=2.71). (Dataset: ioa_datasets)

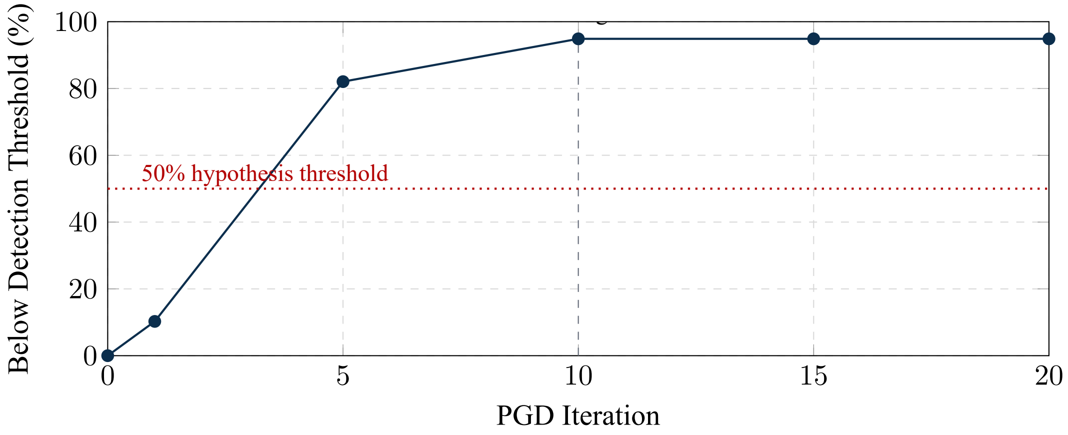


Figure 23: PGD Evasion Attack Convergence Against Contrastive Autoencoder. The attack achieves 94.87% evasion rate (37/39 samples) within 10 iterations using $\varepsilon = 0.02$ L_∞ constraint. Reconstruction error is reduced by 89.9% (from 0.023 to 0.002), allowing adversarial samples to fall below the detection threshold. The differentiable nature of the autoencoder creates an exploitable gradient pathway. (Dataset: pgd_attack_results.npz)

autoencoder-only baseline.

The mechanism of success: PGD successfully minimizes reconstruction error (∼20× lower than baseline), but in doing so pushes attacked embeddings to sparse, low-density manifold regions. The k-NN distance feature, being non-differentiable with respect to the autoencoder's loss function, captures this manifold sparsity (1.43× higher k-NN distance). The distinctive signature of low reconstruction error combined with high k-NN distance uniquely identifies PGD-evaded samples (Figure 24).

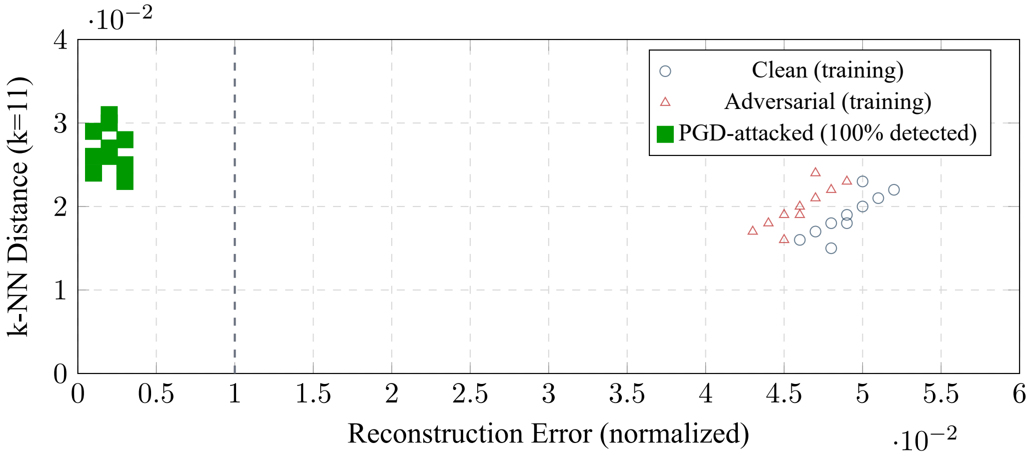


Figure 24: Hybrid Detector Feature Space. The hybrid detector combining reconstruction error and k-NN distance achieves 100% detection (0% evasion) on PGD-attacked samples. While PGD successfully minimizes reconstruction error (∼20× lower), attacked samples are pushed to low-density manifold regions with elevated k-NN distance (1.43× higher). The non-differentiable k-NN component provides gradient obfuscation, breaking the evasion attack. (Dataset: ioa_datasets)

6.5 Attack Stealth Classification

Different attack modalities produce distinct geometric signatures, enabling automated classification. Synonym-substitution attacks are "locally conservative," causing only +28.9% increase in k-NN distance while preserving neighborhood structure. Paraphrase attacks are "globally aggressive," displacing embeddings by +407.9% in k-NN distance. This 14.1× disparity, combined with comparable reconstruction error changes (2.08× ratio), makes k-NN distance 6.8× more discriminative than reconstruction error for attack characterization (Figure 25).

A logistic regression classifier trained on (k-NN distance, reconstruction error) achieves 100% accuracy distinguishing attack types, with perfect precision and recall. The geometric signatures are unambiguous: synonym attacks exhibit HIGH k-NN distance (0.665) + LOW reconstruction error (0.046); paraphrase attacks exhibit LOW k-NN distance (0.150) + HIGH reconstruction error (0.098). Cohen's d effect sizes exceed 13 for both features, indicating complete separation in feature space. This enables the IOA to deploy targeted countermeasures based on detected attack geometry.

6.6 Population-Level Threat Monitoring

The centroid-shift diagnostic enables quantitative threat scoring at the population level. On well-separated datasets (inter/intra ratio 3.63), the diagnostic detects synonym-substitution attacks at 3.8% prevalence with

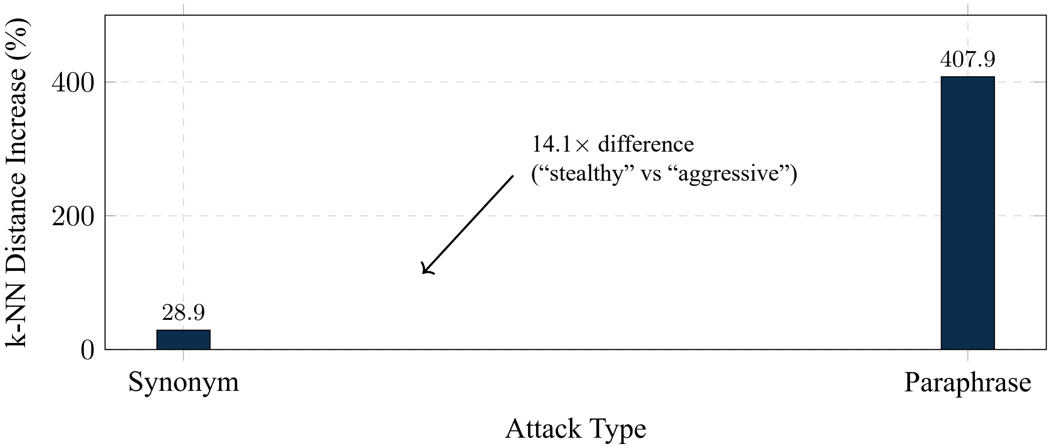


Figure 25: Attack Stealth Analysis: k-NN Distance Displacement. Synonym-substitution attacks cause minimal k-NN distance increase (+28.9%), preserving local neighborhood structure. Paraphrase attacks aggressively displace embeddings (+407.9%), moving to sparse manifold regions. The 14.1× disparity defines a quantitative threshold for classifying "stealthy" vs "aggressive" attacks. Reconstruction errors remain comparable (2.08× ratio), making k-NN distance 6.8× more discriminative. (Dataset: ioa_datasets)

effect sizes 14–17× above the 99th percentile bootstrap threshold ($p < 0.01$). All 4 attacked personas showed statistically significant shifts (mean 0.0445), while 3 non-attacked personas showed zero shift (perfect specificity).

Critically, the relationship between attack prevalence and centroid shift is perfectly linear ($R^2 = 0.999998$). The equation Shift $= 0.006689 \times$ Prevalence$(\%)$ enables inverse prediction with mean absolute percentage error of only 0.29%. This transforms the centroid-shift diagnostic from a binary detector into a calibrated threat quantification tool (Figure 26).

However, geometric separation strongly modulates sensitivity. On AG News (ratio 0.38), the diagnostic achieves only $1.1\times$ above threshold with 50% detection rate (2/4 classes)-a 14.6-fold reduction compared to well-separated datasets. This validates the heuristic that centroid-based monitoring requires inter/intra ratio $>$ 2.0 for reliable operation.

7 Governance and Compliance Constraints

External to direct implementations of attack vectors, societal situations may arise due to cultural differences. Across government, industry, and academia, an IOA must enforce access controls and provenance under binding regimes: U.S. classification rules require clearance + need-to-know ("5 CFR § 1312.23 - Access to classified information", 2025; NRO, 2019), and certain AI-enabled discoveries may trigger Invention Secrecy Act restrictions ("35 U.S. Code Chapter 17 Part II - SECRECY OF CERTAIN INVENTIONS", 2025), constraining disclosure, export processes, and even deployment. In regulated markets such as in the European Union, Article 22 GDPR -and in the UK ICO guidance- restricts solely automated decisions with legal/similarly significant effects, demanding HITL, transparency, and redress requirements that directly shape IOA routing, logging, and override design principles as intelligences increase in complexity over time (GDPR, 2013; ICO, 2023).

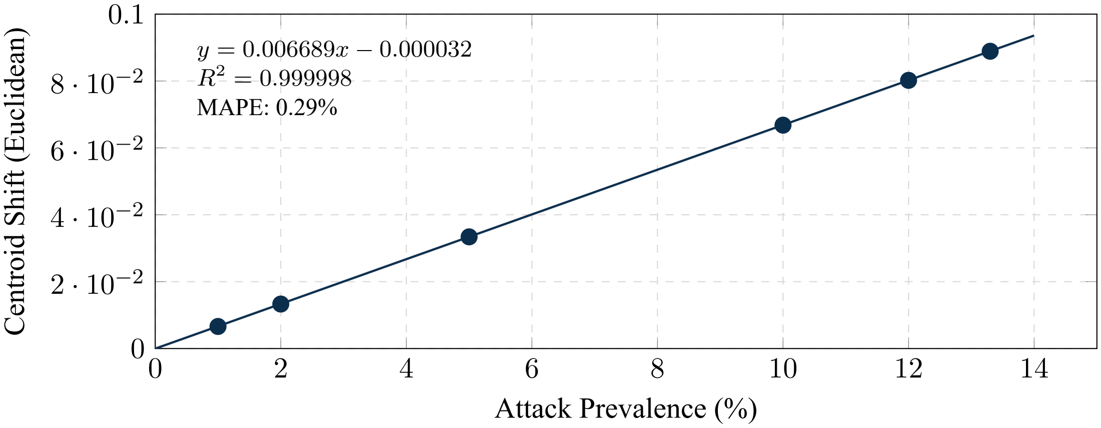


Figure 26: Centroid-Shift Dose-Response Relationship. Attack prevalence is linearly proportional to centroid shift with $R^2 = 0.999998$, enabling quantitative threat scoring. Each 1% increase in attack prevalence produces 0.00669 units of centroid shift. The near-zero intercept confirms true proportionality through the origin, allowing the IOA to estimate attack severity from a single geometric measurement with <0.3% error. (Dataset: bootstrap_shifts_data.npz)

7.1 Regulatory Triggers

7.2 Sovereignty and Supply Chain

Handling foreign data aggravates risks and increases the threat pool; ingestion pipelines such as information extraction pipelines can import prompt-poisons and pass covert triggers across languages/scripts (Kusnezov et al., 2023), while sovereignty and export-control regimes modulate where models can run and what they may access (BIS, 2024; Borman, 2024; Congress, 2025). The broader U.S.-China technology competition amplifies these governance demands: through compute and export controls, technological standards divergence, and adversarial targeting of AI supply chains and RAG Corpora raise the probability of state-aligned poisoning and model manipulation, even in civilian settings (BIS, 2024; Congress, 2025; GAO, 2025; Kusnezov et al., 2023). The primary challenge lies with identifying a solution that fuses clearance-aware data isolation, nation-state guideline compliant - such as Article 22-compliancy - human intervention, tamper-evident audit trials, red-team/blue-team routines that explicitly test for prompt poisoning, hard-prompt persistence, semantic entanglement awareness, and adversarial obfuscation, as these vectors remain open for exploitation even when frameworks like DEPTH reduce task-level hallucinations and sycophancy (Yang et al., 2025).

8 Conclusion

The integration gap across government, industry, and academia will not close by adding more point solutions; It requires a model-agnostic control plane cohesively woven and integrated into society's institutions while enabling agentic meta-observability and steering. The proposed IOA supplies that architecture; Mechanistic interpretability to expose causal circuits (Dominguez-Olmedo et al., 2023); world-model observability to surface latent state representations (Garrido et al., 2024); Platonic-space representation - through convex hull geometry applied to the semantic manifold structures - and lexical contouring via HTR to detect boundary pressure (Ali & Hussein, 2014; Lindsey et al., 2025; Psenka et al., 2024); Persona-vector governance to keep behavior aligned with humanity priorities while achieving positive scoring towards sycophancy,

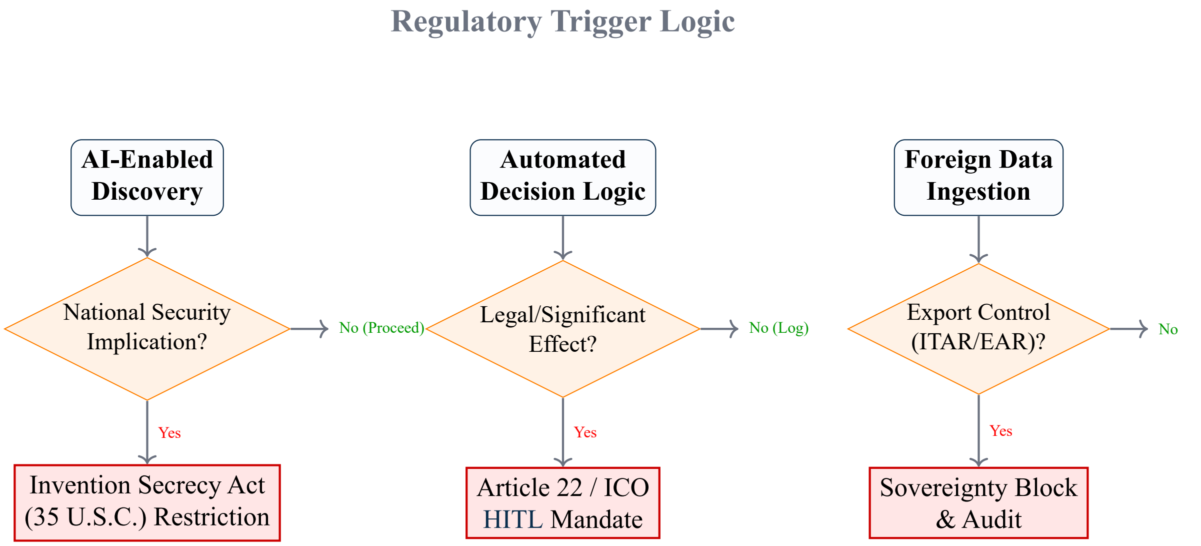

Regulatory Trigger Logic
AI-Enabled Discovery
National Security Implication?
No (Proceed)
Yes
Invention Secrecy Act (35 U.S.C.) Restriction
Automated Decision Logic
Legal/Significant Effect?
No (Log)
Yes
Article 22 / ICO HITL Mandate
Foreign Data Ingestion
Export Control (ITAR/EAR)?
No
Yes
Sovereignty Block & Audit

Figure 27: Governance Triggers: The IOA monitors events to automatically trigger specific compliance restrictions (Invention Secrecy, GDPR HITL, Export Controls) based on context.

ultracrepidarianism, and adversarial goal-seeking (Anil et al., 2024; Chen et al., 2025; Denison et al., 2024; Kusnezov et al., 2023; Lindsey et al., 2025; Sharma et al., 2023; Zhou et al., 2024); and DEPTH (Yang et al., 2025).

While task-level frameworks such as DEPTH demonstrate that hallucination, ultracrepidarianism, and sycophancy are substantially mitigated in the relationship extraction task-class (Yang et al., 2025), the IOA generalizes safety and reliability controls across heterogeneous models, multiple modalities of data - specifically targeting lexical and grammatical intelligences and the intelligences objectives -, and agentic workflows, providing a framework that can expand to deal with the needs of the evolving technology cycle.

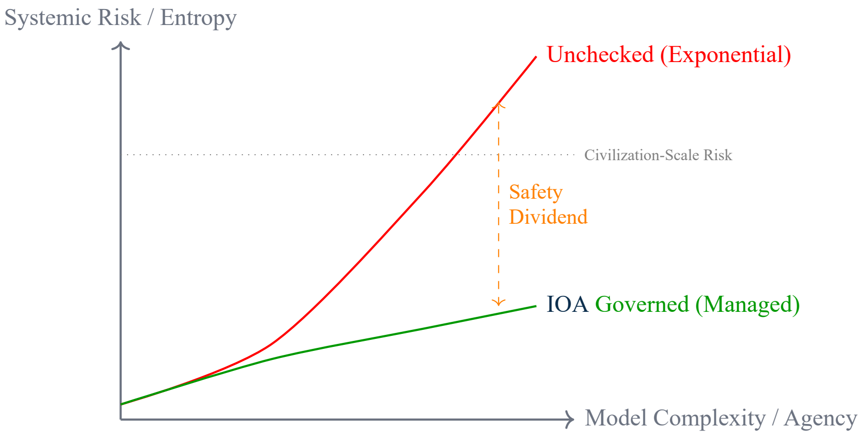


Figure 28: Risk Trajectory: IOA implementation decouples capability growth from existential risk, maintaining safety as complexity scales.

The IOA converts the fragmented and newly born organizational AI stack into a human-governable architecture, allowing for inspectable read-only records sourced from machine reasoning and interpretability events, inherently earning trust without sacrificing accuracy, reliability, capability, or security.

References

35 u.s. code chapter 17 part ii - secrecy of certain inventions [LII / Legal Information Institute]. (2025). https://www.law.cornell.edu/uscode/text/35/part-II/chapter-17

5 cfr § 1312.23 - access to classified information [LII / Legal Information Institute]. (2025). https://www.law.cornell.edu/cfr/text/5/1312.23

Ali, S., & Hussein, K. (2014). The comparative power of type/token and hapax legomena/type ratios. *Advances in Language and Literary Studies*, *5*(3), 112–119. https://doi.org/10.7575/aiac.alls.v.5n.3p.112

Anil, C., Durmus, E., Panickssery, N., et al. (2024). Many-shot jailbreaking. *Advances in Neural Information Processing Systems*, *37*, 129696–129742. https://proceedings.neurips.cc/paper_files/paper/2024/hash/ea456e232efb72d261715e33ce25f208-Abstract-Conference.html

Bailey, L., Serrano, A., Sheshadri, A., et al. (2024). Obfuscated activations bypass llm latent-space defenses. https://arxiv.org/abs/2412.09565

Bereska, L., & Gavves, E. (2024). Mechanistic interpretability for ai safety a review. https://arxiv.org/pdf/2404.14082

Bilski, D. (2025, July). 2025 mid-year llm market update: Foundation model landscape + economics. https://menlovc.com/perspective/2025-mid-year-llm-market-update/

BIS. (2024). Commerce strengthens export controls to restrict china's capability to produce advanced semiconductors for military applications. https://www.bis.gov/press-release/commerce-strengthens-export-controls-restrict-chinas-capability-produce-advanced-semiconductors-military

Borman, M. (2024, December). Foreign-produced direct product rule additions. https://www.federalregister.gov/documents/2024/12/05/2024-28270/foreign-produced-direct-product-rule-additions-and-refinements-to-controls-for-advanced-computing

Carlini, N., Tramèr, F., Lee, K., et al. (2021). Extracting training data from large language models. https://www.usenix.org/system/files/sec21-carlini-extracting.pdf

Challapally, A., Pease, C., Raskar, R., & Chari, P. (2025). The genai divide state of ai in business 2025 mit nanda. https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf

Chen, R., Arditi, A., Sleight, H., Evans, O., & Lindsey, J. (2025). Persona vectors: Monitoring and controlling character traits in language models. https://arxiv.org/pdf/2507.21509

Clop, C., & Teglia, Y. (2024). Backdoored retrievers for prompt injection attacks on retrieval augmented generation of large language models. https://arxiv.org/abs/2410.14479

Colson, A. (2025, August). Challenges of adopting ai in accounting firms [Tax & Accounting Blog Posts by Thomson Reuters]. https://tax.thomsonreuters.com/blog/challenges-of-adopting-ai-in-accounting-firms-tri/

Congress. (2025). U.s. export controls and china: Advanced semiconductors [Congress.gov]. https://www.congress.gov/crs-product/R48642

Denison, C., MacDiarmid, M., Barez, F., et al. (2024). Sycophancy to subterfuge: Investigating reward-tampering in large language models. https://arxiv.org/abs/2406.10162

Dominguez-Olmedo, R., Karimi, A.-H., Arvanitidis, G., & Schölkopf, B. (2023). On data manifolds entailed by structural causal models. https://proceedings.mlr.press/v202/dominguez-olmedo23a.html

Elhage, N., Hume, T., Olsson, C., et al. (2022). Toy models of superposition. https://arxiv.org/abs/2209.10652

GAO. (2025, July). *Artificial intelligence: Generative ai use and management at federal agencies* (tech. rep.). GAO. https://www.gao.gov/products/gao-25-107653

Garrido, Q., Assran, M., Ballas, N., Bardes, A., Najman, L., & LeCun, Y. (2024). Learning and leveraging world models in visual representation learning. https://arxiv.org/abs/2403.00504

GDPR. (2013). Art. 22 gdpr – automated individual decision-making, including profiling [General Data Protection Regulation (GDPR)]. https://gdpr-info.eu/art-22-gdpr/

Guo, J., & Cai, H. (2025). System prompt poisoning: Persistent attacks on large language models beyond user injection. https://www.arxiv.org/abs/2505.06493

Hendrycks, D., Mazeika, M., et al. (2023). An overview of catastrophic ai risks. https://arxiv.org/pdf/2306.120010

ICO. (2023, May). What does the uk gdpr say about automated decision-making and profiling? https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/individual-rights/automated-decision-making-and-profiling/what-does-the-uk-gdpr-say-about-automated-decision-making-and-profiling/

Kaufman, I., & Azencot, O. (2023). Data representations' study of latent image manifolds. https://arxiv.org/pdf/2305.19730

Khurana, I. (2024, July). Explore how rag and knowledge graphs transform customer support. https://www.linkedin.com/pulse/unsung-heroes-ai-customer-service-rags-knowledge-graphs-khurana-nvxtc/

Kirsanov, A., Chou, C.-N., Cho, K., & Chung, S. (2025). The geometry of prompting: Unveiling distinct mechanisms of task adaptation in language models. https://arxiv.org/abs/2502.08009

Kusnezov, D., Barsoum, Y., Begoli, E., Henninger, A., & Sadovnik, A. (2023). *Risks and mitigation strategies for adversarial artificial intelligence threats: A dhs s&t study preparedness series* (tech. rep.). DHS. https://www.dhs.gov/sites/default/files/2023-12/23_1222_st_risks_mitigation_strategies.pdf

Li, L., Sleem, L., Gentile, N., Nichil, G., & State, R. (2025). Exploring the impact of temperature on large language models: Hot or cold? https://arxiv.org/abs/2506.07295

Liao, Z., Antoniak, M., Cheong, I., Cheng, E. Y.-Y., Lee, A.-H., Lo, K., Chang, J. C., & Zhang, A. X. (2024). Llms as research tools: A large scale survey of researchers' usage and perceptions. https://arxiv.org/abs/2411.05025v1

Lieberman, M. B. (2016). First-mover advantage. *ResearchGate*. https://www.researchgate.net/publication/311908029_First-Mover_Advantage

Lindsey, J., et al. (2025). On the biology of a large language model. https://transformer-circuits.pub/2025/attribution-graphs/biology.html

Mariani, J., Kishnani, P., & Alibage, A. (2024, October). Government's less trodden path to scaling generative ai. https://www.deloitte.com/us/en/insights/industry/government-public-sector-services/government-faces-challenges-with-generative-ai-adoption.html

McCarthy, J. (2007). What is artificial intelligence? https://www-formal.stanford.edu/jmc/whatisai.pdf

McKendrick, J. (2025a, January). Research@dbta: Survey: Rag emerges as the connective tissue of enterprise ai. https://www.dbta.com/Editorial/Trends-and-Applications/RESEARCH-at-DBTA-Survey-RAG-Emerges-as-the-Connective-Tissue-of-Enterprise-AI-167699.aspx

McKendrick, J. (2025b, May). State of play on llm and rag: Preparing your knowledge organization for generative ai. https://graphwise.ai/resources/white-paper/knowledge-organization-llm-rag/

NRO. (2019, February). Classified: Sentient program. https://www.nro.gov/Portals/65/documents/foia/declass/ForAll/051719/F-2018-00108_C05113688.pdf

pgvector. (2025). Pgvector. https://github.com/pgvector/pgvector

Psenka, M., Pai, D., Raman, V., Sastry, S., & Ma, Y. (2024). Representation learning via manifold flattening and reconstruction. *Journal of Machine Learning Research*, *25*. https://www.jmlr.org/papers/volume25/23-0615/23-0615.pdf

Rapoport, G., Bicanic, S., & Talabi, M. (2025). Survey: Generative ai's uptake is unprecedented despite roadblocks. https://www.bain.com/insights/survey-generative-ai-uptake-is-unprecedented-despite-roadblocks

Sharma, M., Tong, M., Korbak, T., et al. (2023). Towards understanding sycophancy in language models. https://doi.org/10.48550/arXiv.2310.13548

Sivashanmugam, S. P. (2025). Model inversion attacks on llama 3: Extracting pii from large language models. https://www.arxiv.org/abs/2507.04478

Tully, T. (2024, November). 2024: The state of generative ai in the enterprise. https://menlovc.com/2024-the-state-of-generative-ai-in-the-enterprise/

Yang, Y., Feng, F., Yang, L., et al. (2025). Depth: Hallucination-free relation extraction via dependency-aware sentence simplification and two-tiered hierarchical refinement. https://arxiv.org/abs/2508.14391

Zhang, T., Goldstein, A., & Levin, M. (2023). Classical sorting algorithms as a model of morphogenesis: Self-sorting arrays reveal unexpected competencies in a minimal model of basal intelligence. https://arxiv.org/abs/2401.05375

Zhou, L., Schellaert, W., Martínez-Plumed, F., et al. (2024). Larger and more instructable language models become less reliable. *Nature*, *634*. https://doi.org/10.1038/s41586-024-07930-y

Zou, A., Wang, Z., Zico, K. J., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. https://arxiv.org/abs/2307.15043

Acronyms

AAI Adversarial Artificial Intelligence. 4, 20

ADM Automated Decision Making. 4, 20

DEPTH Dependency-Aware Sentence Simplification and Two-tiered Hierarchical Refinement. 13, 15, 16, 20

GDB Graph Database. 4, 20

HITL Human-In-The-Loop. 3, 4, 8, 13, 15, 20

HTR Hapax-to-Token Ratio. 1, 7–9, 16, 20

IOA Intelligence Orchestration Architecture. 1–3, 7–10, 12–16, 20

KG Knowledge Graph. 1, 4, 6, 20

LLM Large Language Model. 1, 4–6, 13, 20

MCP Model Context Protocol. 1, 4, 6, 20

MF Memory Framework. 4, 20

RAG Retrieval-Augmented Generation. 1, 4, 15, 20

RDB Relational Database. 4, 6, 20

RM Reasoning Model. 1, 4, 20

UTCP Universal Tool Calling Protocol. 1, 4, 6, 20

VDB Vector Database. 4, 6, 20

Algorithm 1: IOA Governance Loop: Adaptive Manifold & Lexical Checks

```
Input: Token stream $T$, Model $M$, Persona reference sets $\{R_p\}$, Thresholds $\theta$
Output: Governed output stream $T_{safe}$

// Pre-computation: Cluster geometry analysis
$\rho \leftarrow \mathit{InterIntraRatio}(\{R_p\})$;
if $\rho > 3.0$ then
    $\mathit{method} \leftarrow \mathit{Mahalanobis}$;
    $T_2 \leftarrow \mu_{inter} - 2\sigma_{inter}$;
else
    $\mathit{method} \leftarrow$ k-NN $(k = 11)$;
end

for each token $t \in T$ do
    $v_{emb} \leftarrow M_{activations}(t)$;
    $\mathit{HTR} \leftarrow \mathit{HapaxTokenRatio}(t)$;

    // Distance-based governance check
    if $\mathit{method} = \mathit{Mahalanobis}$ then
        $d_M \leftarrow \mathit{MahalanobisDistance}(v_{emb}, R_p)$;
        $\mathit{compliant} \leftarrow (d_M < T_2)$;
    else
        $\mathit{votes} \leftarrow \mathit{k\text{-}NN\_Vote}(v_{emb}, \{R_p\}, k)$;
        $\mathit{compliant} \leftarrow (\max(\mathit{votes}) \geq 0.75)$;
    end

    // Contrastive autoencoder adversarial detection
    $\epsilon_{recon} \leftarrow \mathit{ContrastiveAE}(v_{emb})$;
    $\mathit{adversarial} \leftarrow (\epsilon_{recon} > \theta_{ae})$;

    if $\neg\mathit{compliant}$ or $\mathit{adversarial}$ or $\mathit{HTR} > \theta_{lex}$ then
        // Governance violation detected
        $v_{steer} \leftarrow \mathit{GetSteeringVector}(v_{emb})$;
        $t_{corrected} \leftarrow \mathit{ApplySteering}(t, v_{steer})$;
        Escalate to HITL if adversarial;
        return $t_{corrected}$;
    end
    return $t$;
end
```