SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Multi-Layer Trust Frameworks For LLM-Powered Data Enrichment In Enterprise Analytics Pipelines

Akhil Kumar Kandakatla

Independent Researcher, USA
ORCID: 0009-0000-0604-5363

Abstract
Automated data enrichment powered by large language models has emerged as a consequential capability in enterprise data platform engineering, enabling organizations to augment structured business records with AI-generated contextual outputs at scale (Brown et al., 2020; McKinsey Global Institute, 2023). While the technical challenges have been solved, the remaining challenge is deciding when an AI-enrichment is reliable enough to be used for downstream applications. A common guardrail is confidence scoring. However, it cannot detect systematic bias, quality drifts across target populations, or governance failures that can result in organizations losing trust in the data assets produced by LLMs (Guo et al., 2017; Xiong et al., 2024). This article proposes a four-layer trust framework for enterprise LLM enrichment governance, spanning output-level reliability assessment, organizational validation and phased adoption, governance and accountability, and continuous observability. An organizational maturity model accompanies the framework to enable enterprises to assess and advance their governance posture across four defined levels. Together, the framework and maturity model address a documented gap between technical uncertainty quantification and the institutional mechanisms through which organizations develop and sustain trust in AI-generated data. Embedding this governance architecture at the data platform design stage, rather than applying it as post-deployment remediation, produces substantially improved risk management and downstream decision quality outcomes.

Keywords: large language model, enterprise data enrichment, trust framework, data governance, confidence calibration, observability, reliability architecture, analytics pipeline.

1. Introduction

Deployment of large language model-powered data enrichment has accelerated across enterprise environments, driven by a business case that is straightforward to construct (McKinsey Global Institute, 2023). Organizations can automate the augmentation of structured records customer accounts, vendor profiles, project data, competitive intelligence with contextual information drawn from web sources and document repositories at a scale that manual research cannot approach. Early adopters across financial services, enterprise software, and operations-intensive industries have moved from pilot to production deployment, supported by maturing orchestration infrastructure and improved model accuracy.

From a technical standpoint, the infrastructure for LLM enrichment has reached a level of maturity that supports production deployment at enterprise scale. Models can be integrated into batch and streaming pipelines, fine-tuned on domain-specific corpora for improved classification accuracy, and paired with retrieval-augmented generation architectures that ground outputs in verifiable source documents (Lewis et al., 2020). Orchestration frameworks manage scheduling, cost controls, retry logic, and output formatting reliably. The question confronting data platform practitioners has therefore shifted: the problem is no longer whether

LLM enrichment can be built, but whether it can be trusted and, more precisely, how organizations determine when enriched outputs are reliable enough for specific downstream workflows.

Enterprise analytics platforms increasingly depend on enriched data as an upstream input to revenue-critical decisions (Stonebraker & Ilyas, 2018). Sales targeting models, account prioritization engines, risk scoring systems, and downstream machine learning models all consume records whose quality directly determines the quality of organizational decisions. When LLM-generated enrichment contains factual errors, systematic biases, or quality degradation invisible to conventional monitoring, those failures propagate through analytics infrastructure in ways that existing data quality tooling is not designed to detect. A pipeline may produce well-formed, type-consistent, internally coherent outputs while simultaneously introducing errors that misalign sales targeting, embed biases into trained models, or expose the organization to regulatory risk in audited decision workflows.

Confidence scoring threshold-based filtering of enrichment outputs by their self-reported certainty is the most common technical safeguard deployed to manage this exposure. Research has established, however, that LLM confidence estimates are frequently miscalibrated, particularly in domain-specific applications where training data coverage is sparse (Guo et al., 2017; Kadavath et al., 2022). More fundamentally, confidence scoring addresses only the statistical dimension of a problem that is also organizational, governance-related, and operational in nature (Liu et al., 2023). Most enterprises that have deployed LLM enrichment pipelines have not yet established the multi-layer governance architectures necessary to manage reliability risk at scale.

This article presents a four-layer trust framework for enterprise LLM enrichment and an organizational maturity model to guide its implementation. Section 2 establishes the enterprise deployment context. Section 3 analyzes the data quality gap specific to probabilistic AI outputs. Section 4 examines structural limitations of confidence-only approaches. Section 5 presents the four-layer framework in operational detail. Section 6 describes the organizational maturity model. Sections 7 and 8 address implementation challenges and future research directions, respectively.

2. LLM-Powered Data Enrichment: Capabilities and Enterprise Deployment Context

The ability to extract information and context, make classification inferences, identify relationships, and synthesize knowledge from unstructured text is one of the best-documented capabilities of large language models (Brown et al , 2020). Applied to enterprise data environments, these capabilities enable systematic augmentation at a scale and specificity that rule-based extraction logic or human-curated enrichment workflows cannot approach efficiently. Customer account records may be enriched with inferred industry classification, organizational growth trajectory, executive contact data, and competitive positioning signals; vendor profiles may be enriched with financial stability indicators and operational risk signals, all drawn from publicly available sources and appended as queryable structured fields.

The range of enterprise enrichment use cases is broad. Sales organizations enrich account records to sharpen targeting and prioritize outreach. Risk and compliance functions improve third-party and counterparty records to assess exposure and concentration risk. Operations functions improve project and asset records to better allocate resources. Across these domains, improved data flows down to analytics platforms, business intelligence tools, and machine learning models that automate or govern decision-making (Bommasani et al., 2021). At the scale of modern enterprise data environments hundreds of thousands to millions of records quality issues in the enrichment pipeline translate directly into quality issues at the input layer of every downstream system that consumes enriched data.

The technical architecture of enterprise LLM enrichment pipelines has matured considerably. Retrieval-augmented generation approaches ground model outputs in specific, verifiable source documents, reducing hallucination rates compared to purely parametric generation (Lewis et al., 2020). Data integration pipelines handle the injection of enriched fields back into enterprise data warehouses and operational systems at scale (Stonebraker & Ilyas, 2018). These architectural components are well understood and in broad production deployment.

What has not kept pace with technical capability is governance capability. A pipeline that reliably produces well-formatted enrichment outputs may simultaneously introduce factual errors, systematic biases, or quality drift that no technical instrumentation surfaces. The gap between technical performance and organizational trustworthiness is the central problem this framework addresses.

3. The Data Quality Gap: Deterministic vs. Probabilistic Failure Modes

Traditional enterprise data quality frameworks have been refined over decades of experience with deterministic systems, where quality failures are discrete, reproducible, and detectable through rule-based monitoring (Wand & Wang, 1996). A schema violation produces a pipeline error. A null value triggers a completeness check. A referential integrity failure generates an auditable exception. The foundational vocabulary of enterprise data quality completeness, accuracy, consistency, timeliness was constructed for systems where an incorrect value has an unambiguous definition, and where detecting it requires checking a record against a known constraint (Redman, 1998). Enterprise data quality tooling is optimized for exactly this failure mode.

LLM-generated enrichment fails differently. Enrichment outputs are probabilistic: they can be factually incorrect while being syntactically valid, type-consistent, and internally coherent, producing no detectable signal in schema-validation or null-check tooling (Zhang et al., 2023). An LLM classification of a company's industry segment may be delivered with high reported confidence and incorrect factually, because the company recently pivoted its business model and available web sources have not yet converged on the updated description. No schema check or referential integrity rule will surface this failure; only comparison with authoritative ground truth will.

Hallucination: the generation of plausible but factually unsupported content has been documented extensively across LLM applications, including production deployment contexts (Ji et al., 2023; Zhang et al., 2023). In enterprise enrichment pipelines, hallucination manifests not as obviously false outputs but as subtly incorrect classifications, miscalibrated growth estimates, or stale executive data presented with apparent confidence. These failure modes are particularly consequential because they are indistinguishable from correct outputs by downstream consumers absent a ground truth comparison.

Three organizational risk dimensions follow from LLM enrichment quality failures. Risk extends to business decision models when analytics and targeting workflows extract from improved data and make decisions on incorrect signals (Gartner, 2022) and to machine learning models relying on improved features (machine learning cascade risk). Model training also learns the error pattern and bias introduced by the enrichment, compounding the original quality failure in each model using improved features as input (Sculley et al., 2015). The third dimension is regulatory and compliance risk: organizations that are accountable for their decision-making and apply improved data to regulated workflows (lending, hiring, account prioritization) should be able to show the quality and traceability of their data (Kaddour et al., 2023). These risk dimensions share a defining structural characteristic: they are not surfaced by the absence of output, but by the presence of plausible-sounding incorrect output.

4. Structural Limitations of Confidence-Only Approaches

Confidence scoring attached to LLM enrichment outputs and threshold-based filtering of outputs by that score represents an improvement over treating all enriched fields as equivalently reliable. Organizations that have not implemented confidence filtering are operating with less information than is readily available, and threshold-based controls provide a meaningful baseline. The limitation is not that confidence scoring is undesirable; it is that confidence scoring alone is structurally insufficient for enterprise enrichment governance.

Confidence miscalibration is the first structural limitation. Neural network confidence estimates diverge systematically from actual accuracy on data distributions that differ from training conditions, a property established for modern classification models and applicable to LLMs in domain-specific deployments (Guo et al., 2017). Models trained on broad web corpora may express high confidence in classifications of industry segments that are underrepresented in their training data, because superficially similar training patterns produce high-certainty outputs that do not correspond to the correct classification in the specific context (Kadavath et al., 2022). Research on confidence elicitation in large language models has shown that expressed certainty frequently reflects linguistic confidence rather than factual accuracy (Xiong et al., 2024). An organization filtering enrichment outputs by confidence threshold may systematically accept miscalibrated high-confidence errors while flagging correctly assessed low-confidence outputs.

Systematic directional bias constitutes the second limitation, and in enterprise contexts it may be the more operationally consequential one. Confidence scores describe per-output certainty; they carry no information about the pattern of errors across a population of records. A pipeline that systematically overestimates growth signals for technology-sector companies and underestimates them for manufacturing-sector companies may

be well-calibrated at the record level, expressing equal certainty in both accurate and inaccurate classifications. Filtering by confidence threshold preserves the directional bias at a higher certainty level rather than surfacing the pattern (Liu et al., 2023). Detecting systematic directional errors requires population-level validation against ground truth, which confidence scoring alone cannot provide.

Organizational insufficiency constitutes the third limitation. Whether enriched data is appropriate for a specific business workflow depends on organizational properties, and has enrichment been validated in this workflow's context? Who is accountable for enrichment quality in this domain? Are observability systems in place to detect quality degradation after production deployment? These are governance and accountability questions that a numerical threshold cannot answer (Wang et al., 2023). Confidence scoring is a technical mechanism; enterprise trust is an organizational property. The gap between them requires a governance architecture, not a more refined filter.

5. A Four-Layer Trust Framework for Enterprise LLM Enrichment

The framework presented here addresses the structural limitations of confidence-only approaches by organizing enterprise LLM enrichment governance into four interdependent layers. Each layer targets a distinct dimension of the trust problem: statistical, organizational, accountability-based, and operational, and each is necessary but insufficient in isolation. Organizations that implement all four layers develop institutional confidence in enriched data that enables safe scaling to mission-critical workflows; those that implement only technical layers remain exposed to the silent failure risks documented in Section 3.

Layer 1 Output-Level Reliability Assessment

Technical reliability assessment at the output level extends beyond single-model confidence scoring to capture a multi-signal picture of enrichment quality.

Ensemble uncertainty estimation involves querying multiple LLM configurations varying system prompts, temperature settings, or model versions for the same enrichment task and measuring inter-model agreement across outputs. High levels of agreement indicate consistent classification across contexts. Systematic disagreement suggests ambiguous or context-dependent enrichment that should not be directly passed on to downstream applications but reviewed by a human. This approach distinguishes uncertain predictions from miscalibrated predictions in ways that do not depend solely on the model's confidence (Kadavath et al., 2022).

Semantic consistency validation assesses whether the improved fields of a single record are consistent with each other. A record reporting simultaneous rapid organizational expansion and significant workforce contraction carries an internal contradiction that may signal enrichment error independently of field-level confidence scores. Automated consistency rules across related enrichment categories provide a second-order quality signal that does not require ground truth availability.

Reference consistency cross-checking compares enriched outputs for a sampled subset of records against authoritative external sources financial filings, industry databases, organizational directories to establish empirical accuracy baselines for each enrichment field type. Cross-check results provide ground truth calibration of confidence scores in the context of a deployment domain.

Drift monitoring captures longitudinal quality information about whether quality degraded after a model update, whether quality of source documents degraded, or whether the distribution of the data being improved drifted (Gama et al , 2014). Quality baselines established during pipeline initialization and tracked against metrics in production create a form of early warning system to detect degradation before it impacts any downstream decisions.

Layer 2 Organizational Validation and Phased Adoption

Technical reliability signals must be complemented by organizational processes that establish empirical evidence for enrichment quality before enriched data enters mission-critical workflows.

Controlled pilot validation tests enrichment outputs for a ground-truth-validated sample prior to any production deployment. The pilot phase establishes baseline accuracy metrics for each enrichment field type, identifies systematic failure modes, and produces the empirical foundation for defining appropriate trust boundaries for downstream use. Without structured pilot validation, trust thresholds reflect assumption rather than evidence.

With phased rollout, various stages of improved data are introduced to business workflows sequentially. Each stage has explicit validation gates, and the first stage limits improved data to exploratory analyzes and

hypothesis testing where errors generate learning, but not high-stakes decisions. Phase two supports secondary business decisions with low impact to firms' revenue and compliance and phase three supports primary revenue and compliance workflows given sufficient evidence from phase one and two. Each phase transition requires evidence-based authorization, not assumption-based trust.

Feedback integration establishes systematic channels for business stakeholders to report enrichment quality issues encountered in operational use. Stakeholder feedback captures failure modes that technical validation misses, because it reflects enrichment quality as experienced in actual business contexts rather than in controlled validation samples. Systematic collection, categorization, and analysis of feedback signals closes the loop between deployment experience and pipeline improvement.

Layer 3 Governance and Accountability

Governance structures define who is responsible for enrichment quality, which decisions enriched data may inform, and how quality failures are managed. Without formalized governance structures, accountability is diffuse and quality failures induce confusion, rather than managed responses.

Data stewardship assigns ownership responsibility of each enrichment domain to named data stewards. Data stewards are responsible for ensuring data quality, data validation, and escalation of data quality issues through defined channels. Data stewardship is consistent with established data management frameworks that provide an organizational blueprint for multiple data domains (DAMA International, 2017).

Decision authority documentation specifies the classes of business decisions each enrichment field is authorized to support at each deployment phase. This prevents a common failure mode in enterprise enrichment programs: enriched data migrating from exploratory to operational use without authorization review and the accountability that authorization requires (Nushi et al., 2018). In regulated decision environments, decision authority documentation also provides the audit trail necessary to demonstrate quality management of AI-generated inputs.

Incident response defines escalation paths, notifications, and response timescales that can be applied to enrichment quality failures. These can happen when incorrect enrichment data has propagated to downstream production workflows and been used to make decisions. Documented incident response transforms quality failures from organizational catastrophes into manageable operations workloads. Transparency documentation provides downstream consumers with information, including validation status, known limitations, confidence characteristics, and phase authorization of enrichment fields, so that they may use that data with appropriate caution.

Layer 4 Observability and Continuous Improvement

After deployment, the enrichment pipeline requires instrumentation to ensure quality does not deteriorate unnoticed, and that operational data can help drive improvements.

Continuous monitoring instruments the enrichment pipeline to capture metrics concerning quality (e.g., enrichment coverage rate, confidence score distribution, reference consistency check results, inter-model agreement scores, and stakeholder error rates). Continuous monitoring enables automatic alerting when these metrics exceed defined baselines. This allows early detection and intervention of enrichment errors, before they materially affect decision systems.

Periodic validation sampling structures recurring human-in-the-loop ground truth validation against a sampled subset of production enrichment outputs on a defined cadence. Sampling-based validation provides calibration for automated monitoring metrics, surfaces systematic failure patterns that aggregate indicators may obscure, and maintains the empirical grounding of trust thresholds over time.

Feedback loop analysis organizes quality issues reported by stakeholders based on the type of error, the enrichment domain, and the business context. By analyzing the most frequent types of errors, one can identify the areas with the highest impact, whether it be coverage limitations of the model with respect to certain document types, degradation of quality in the source documents, or configuration gaps in the enrichment pipeline. In turn, these failure modes are recorded and adjusted in the model iteration stage and other parameters such as the retrieval configuration and prompt engineering are updated. Versioning further helps this step by allowing complete traceability of failure modes and other changes to the pipeline (Paleyes et al., 2022).

Table 1 summarizes the four-layer framework across its core components, operational mechanisms, and risk dimensions addressed.

Table 1. Four-Layer Trust Framework for Enterprise LLM Data Enrichment

Layer	Core Components	Operational Mechanism	Risk Addressed
Layer 1 Output-Level Reliability Assessment	Ensemble uncertainty estimation; semantic consistency validation; reference cross-checking; drift monitoring	Multi-model agreement scoring; automated consistency rules; reference sampling; longitudinal metric tracking	Confidence miscalibration; silent factual errors; quality drift over time
Layer 2 Organizational Validation & Phased Adoption	Controlled pilot validation; phased rollout with explicit validation gates; stakeholder feedback integration	Ground-truth baseline testing; phase-gated workflow authorization; structured feedback channels	Premature use in high-stakes workflows; undetected field-level failure modes
Layer 3 Governance & Accountability	Data stewardship assignments; decision authority documentation; incident response protocols; transparency records	Named domain ownership; per-field use authorization; escalation protocols; enrichment field status visibility	Accountability diffusion; unauthorized workflow escalation; audit trail deficits
Layer 4 Observability & Continuous Improvement	Continuous quality monitoring; periodic validation sampling; feedback loop analysis; versioned model iteration	Automated alerting on metric deviation; recurring human sampling; error pattern analysis; redeployment with version tracking	Post-deployment quality degradation; ML cascade accumulation; undetected pipeline drift

6. Organizational Maturity Model for Enrichment Deployment

Governance capability develops along a trajectory that differs meaningfully across organizations, and characterizing that trajectory as a maturity model provides a practical tool for assessing current posture and prioritizing governance investment.

Level 1 Experimental. At least one LLM enrichment pipeline has been deployed, typically in a development or sandbox environment. Technical capabilities have been demonstrated, but governance is absent or informal. Enriched data is used for ad hoc exploration only, observability extends no further than basic pipeline logging, and no formal trust boundaries have been defined.

Level 2 Controlled Deployment. Pilot validation has been conducted for at least one enrichment domain; basic governance is in place for that domain; and enriched data is authorized for use in secondary business workflows with defined validation gates. Quality monitoring is manual and periodic; feedback mechanisms exist but are informal. Organizations at Level 2 have reduced their immediate risk but lack the systematic observability and governance coverage to safely scale enrichment across multiple business domains.
Level 3 Operationalized. All four framework layers are implemented across the enrichment domains supporting primary business workflows. Technical reliability assessment, organizational validation, governance accountability, and observability are all operational. Revenue and risk workflows within documented trust boundaries are supported by data enrichment at the defined level of granularity, and the quality is the active operational responsibility of the stewards.
Level 4 - Scaled and Intelligent. Framework has been scaled across multiple business domains, with trust calibration specific to each use case's risk profile. The pipeline's capabilities are continuously improved and informed by validation and feedback; LLM enrichment is seen as a core competency of the data platform with sustained governance efforts (Paleyes et al., 2022). Trust thresholds, validation frequencies, and governance structures are periodically reviewed and updated in response to operational evidence.
Organizations at Levels 1 and 2 face disproportionate exposure to silent failure risks, because the absence of systematic observability and governance accountability means enrichment errors may accumulate in downstream ML training datasets and decision systems for extended periods before producing observable decision-quality consequences.

Table 2 presents the organizational maturity model across key governance dimensions.

Table 2. Organizational Maturity Model for Enterprise LLM Enrichment Deployment

Level	Characteristics	Framework Implementation	Risk Profile
Level 1 Experimental	Pipeline deployed in sandbox or development; governance absent; enriched data used for exploration only; no formal trust boundaries defined	No formal layers implemented; basic pipeline logging only	High silent failures undetected; no trust boundary definition; unrestricted downstream use possible
Level 2 Controlled Deployment	Pilot validation conducted in at least one domain; basic governance active; enriched data authorized for secondary workflows with defined validation gates	Layer 1 partially active; Layer 2 in pilot phase; Layer 3 partial; Layer 4 manual only	Moderate reduced immediate risk; limited cross-domain coverage; observability gaps remain

Level 3 Operationalized	All four layers active across primary business domains; enrichment quality managed continuously; defined stewardship ownership; primary revenue and compliance workflows supported	All four layers fully operational; stewardship and observability active; governance accountability documented	Low managed risk with defined trust boundaries and accountability mechanisms
Level 4 Scaled and Intelligent	Multi-domain enrichment with domain-specific calibration; feedback-driven iteration; LLM enrichment as core competency; governance structures periodically reviewed and updated	All four layers plus domain-specific calibration; feedback-driven model iteration; periodic governance review cycles	Minimal systematic governance with evolving trust boundaries and continuous improvement

7. Challenges in Enterprise Trust Framework Implementation

Anticipating implementation challenges at the platform design stage produces substantially better governance outcomes than encountering them reactively during deployment.

Governance investment friction is the most consistently observed barrier. Layer 1 technical controls are implemented on a regular basis, while Layers 2 through 4 require a continued commitment from data engineering, product management, legal, compliance, and business stakeholders (Paleyes et al., 2022). In organizations where data governance is treated as a cost overhead rather than a risk infrastructure, it is necessary to make the case for investing in the enrichment governance posture by quantifying the risk exposure of business decisions and regulatory compliance (IBM Institute for Business Value, 2023).

Hidden technical debt accumulates in enrichment deployments that delay governance implementation (Sculley et al., 2015). Pipelines operating without trust frameworks generate decision history ML models trained on unevaluated enriched data, business processes calibrated to unverified enrichment signals that become progressively more difficult to audit and remediate as the deployment matures. Retrofitting multi-layer governance onto an established enrichment pipeline is substantially more costly than designing governance in from the start.

Feedback loop latency creates a temporal risk dimension that automated monitoring does not fully address. Enrichment errors embedded into ML model training data may not produce observable decision failures for months, particularly where model retraining cadences are infrequent (Kaddour et al., 2023). Observability architecture must therefore extend upstream to the input stage of ML training pipelines, monitoring enrichment quality as it enters model training rather than only at the enrichment pipeline output.

Validation scalability presents a practical constraint as enrichment volumes grow. Human-in-the-loop validation, appropriate and manageable at Levels 1-2, does not scale economically to the record volumes of mature enterprise deployments. The architectural transition to automated monitoring with structured sampling appropriate at Levels 3-4 requires deliberate platform design decisions and cannot be improvised at scale.

8. Future Directions

Adaptive trust calibration automated adjustment of enrichment trust thresholds based on downstream decision outcome signals represents a compelling near-term extension. Static thresholds defined at deployment do not account for model drift, distributional shifts in enriched data, or changes in the decision stakes associated with specific enrichment fields over time. Systems that dynamically update thresholds in response to outcome monitoring would allow governance frameworks to evolve without requiring periodic manual recalibration.

Federated enrichment governance addresses an emerging operational need in data ecosystem contexts where organizations consume LLM-enriched data from external pipelines. Standard mechanisms for exchanging enrichment quality certifications alongside enriched outputs would enable consuming organizations to apply informed trust boundaries to externally sourced enrichment, extending the framework's value beyond the single-organization deployment scenario.

Explainability tooling for enrichment pipelines to help show the relationship between input and output for individual improved records would increase accountability for governance in regulated decision contexts where documenting the quality of input data is more broadly extending to AI-generated fields.

Regulators are likely to release standards for AI data product generation in regulated sectors that businesses will need to address. Organizations with clear multi-layer enrichment governance frameworks are better able to satisfy quality management requirements on AI-generated data inputs (Bommasani et al. 2021; Kaddour et al. 2023) than organizations which rely on confidence-only frameworks.

9. Conclusion

LLM-powered data enrichment has established a compelling value proposition in enterprise analytics platform engineering. Scale, speed, and contextual richness that manual enrichment methods cannot approach are genuinely achievable, and the operational value proposition is well documented (McKinsey Global Institute, 2023). The governance challenge is equally real, and the gap between deployment velocity and governance maturity at most organizations is a risk that deserves explicit architectural attention.

Confidence scoring addresses the statistical surface of LLM output reliability, but it leaves unresolved the systematic bias, organizational accountability gaps, and observability deficits that produce the silent failures documented in enterprise AI deployments (Guo et al., 2017; Sculley et al., 2015). Enterprises that treat confidence thresholds as sufficient enrichment governance operate with compounding, unquantified risk that materializes in downstream decision quality, trained ML model performance, and regulatory audit readiness.

The four-layer trust framework proposed here output-level reliability assessment, organizational validation and phased adoption, governance and accountability, and continuous observability provides a structured architecture for managing enrichment reliability risk across its technical, organizational, and operational dimensions. The organizational maturity model provides a practical path for assessing current posture and advancing toward mature enrichment governance. The design principle that emerges is direct: trust in AI-generated data must be architected into the platform from the outset, not applied as remediation after quality failures surface. Organizations that make this design choice at the platform architecture stage will be best positioned to realize the full value of LLM enrichment while managing its inherent risks at enterprise scale.

Ethics Statement

Conflict of Interest: The author declares no conflict of interest.

Statement of Human and Animal Rights: This study did not involve human participants, animal subjects, or identifiable personal data requiring ethics board review.

Informed Consent: Not applicable.

References

1. Bommasani, R., Hudson, D. A., Aditi, E., Altman, R., Arora, S., Bernstein, S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., … Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258. https://arxiv.org/abs/2108.07258
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. arXiv:2005.14165. https://arxiv.org/abs/2005.14165
3. DAMA International. (2024).DAMA® Data Management Body of Knowledge (DAMA-DMBOK®). https://www.dama.org/cpages/body-of-knowledge
4. Gama, J., Zliobaite, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), 1-37. https://doi.org/10.1145/2523813
5. Gartner. Gartner data quality market guide. Gartner Research. https://www.gartner.com/en/data-analytics/topics/data-quality

6. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning, 70, 1321-1330. https://proceedings.mlr.press/v70/guo17a.html
7. Azure Infrastructure. (2026). Operationalizing Responsible AI in Microsoft Foundry within Enterprise Network Boundaries. Microsoft. https://techcommunity.microsoft.com/blog/azureinfrastructureblog/operationalizing-responsible-ai-in-microsoft-foundry-within-enterprise-network-b/4516869
8. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38. https://doi.org/10.1145/3571730
9. Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, J., Johnston, S., El-Showk, S., Jones, A., Elhage, N., Hume, T., Chen, A., Bai, Y., Bowman, S., Fort, S., … Kaplan, J. (2022). Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221. https://arxiv.org/abs/2207.05221
10. Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). Challenges and applications of large language models. arXiv preprint arXiv:2307.10169. https://arxiv.org/abs/2307.10169
11. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W.-T., Rocktaschel, T., Riedel, S., & Kiela, D. (2021). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401. https://arxiv.org/abs/2005.11401
12. Liu, Y., Yao, Y., Ton, J.-F., Zhang, X., Guo, R. H., Cheng, H., Klochkov, Y., Taufiq, M. F., & Li, H. (2024). Trustworthy LLMs: A survey and guideline for evaluating large language model alignment. arXiv preprint arXiv:2308.05374. https://arxiv.org/abs/2308.05374
13. McKinsey Global Institute. (2023). The economic potential of generative AI: The next productivity frontier. McKinsey & Company. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier
14. Nushi, B., Kamar, E., & Horvitz, E. (2018). Towards accountable AI: Hybrid human-machine analyses for characterizing system failure. arXiv:1809.07424. https://arxiv.org/abs/1809.07424
15. Paleyes, A., Urma, R.-G., & Lawrence, N. D. (2022). Challenges in deploying machine learning: A survey of case studies. ACM Computing Surveys, 55(6), 1-29. https://doi.org/10.1145/3533378
16. Redman, T. C. (1998). The impact of poor data quality on the typical enterprise. Communications of the ACM, 41(2), 79-82. https://doi.org/10.1145/269012.269025
17. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. NIPS'15: Proceedings of the 29th International Conference on Neural Information Processing Systems, Volume 2, Pages 2503 - 2511. https://proceedings.neurips.cc/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html
18. Stonebraker, M., & Ilyas, I. F. Data integration: The current status and the way forward. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering. http://sites.computer.org/debull/A18june/p3.pdf
19. Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. Communications of the ACM, 39(11), 86-95. https://doi.org/10.1145/240455.240479
20. Sebo, P. (2025). Can ChatGPT be used to generate scientific abstracts? Cureus. https://pmc.ncbi.nlm.nih.gov/articles/PMC12375800/
21. Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., & Hooi, B. (2024). Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. Openreview. https://openreview.net/forum?id=gjeQKFxFpZ
22. Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2025). Siren's song in the AI ocean: A survey on hallucination in large language models. arXiv:2309.01219. https://arxiv.org/abs/2309.01219