International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com | A Peer-Reviewed, Refereed and Bimonthly Journal|

|| Volume 7, Issue 1, January - February 2024 ||

DOI: 10.15680/IJCTECE.2024.0701007

Large Language Models for Intelligent Data Stewardship in Enterprises: Architectures, Provenance, and Evidence-Mapped Governance

Nagender Yamsani

Software Development Senior Specialist Advisor, USA

ABSTRACT: Enterprises increasingly operate in data ecosystems characterized by extreme heterogeneity, spanning legacy databases, cloud-native platforms, streaming pipelines, and unstructured knowledge repositories, all under mounting regulatory, ethical, and operational scrutiny. In this context, recent advances in large language models (LLMs) notably transformer-based architectures combined with retrieval-augmented generation (RAG), dense passage retrieval (DPR), programmatic and weak supervision, and knowledge-graph grounding offer a unifying technical substrate for intelligent data stewardship at enterprise scale. By embedding stewardship objectives such as FAIR principles, end-to-end provenance, semantic interoperability, and continuous data quality assurance directly into LLM-enabled workflows, organizations can move beyond static catalogs toward adaptive, context-aware systems capable of automated metadata enrichment, lineage-aware question answering, policy-sensitive data discovery, and assisted remediation of quality and compliance issues. This article synthesizes these foundations into an integrated reference architecture for LLM-assisted stewardship, and introduces an evidence-mapping methodology that operationalizes governance assessment by aligning publicly observable signals with established standards and controls. Through an applied case study of Inspire Brands' AI-driven governance initiatives, we demonstrate how evidence mapping enables a non-invasive yet systematic evaluation of organizational readiness, surfacing both strengths and gaps without requiring privileged internal disclosures. Finally, we outline open research challenges including evaluation metrics for trustworthiness and explainability, robustness under regulatory change, and human-in-the-loop validation patterns and offer practical recommendations to guide enterprise adopters in responsibly deploying LLMs as first-class components of modern data governance and stewardship ecosystems.

KEYWORDS: AI Governance; Evidence Mapping; Enterprise AI; Responsible AI; NIST AI RMF; OECD AI Principles; Digital Transformation; Corporate AI Strategy

I. INTRODUCTION

Modern enterprises increasingly recognize data as a core strategic asset, yet translating this recognition into effective, organization-wide data stewardship remains a persistent challenge. As data estates expand across cloud platforms, SaaS applications, data lakes, and real-time pipelines, teams struggle with fragmented metadata, undocumented transformations, and unclear ownership boundaries. These gaps undermine trust, slow analytics and AI initiatives, and heighten regulatory risk. Large language models (LLMs), trained on broad textual and structural corpora, introduce a new class of capabilities well aligned with these pain points. When augmented with retrieval mechanisms and grounding layers, LLMs can perform semantic normalization across heterogeneous schemas, assist in automated classification and tagging, answer natural-language questions over catalogs and logs, and generate human-readable documentation that bridges technical and business contexts. However, without careful system design, these same models risk amplifying uncertainty through hallucinations, obscured provenance, or opaque decision logic. As a result, stewardship-oriented deployments must explicitly prioritize grounding, traceability, and evaluability alongside raw model capability.

This article brings together technical advances in LLM systems and established data-governance frameworks to provide a structured foundation for responsible adoption. First, it surveys LLM interaction patterns most relevant to stewardship, including retrieval-augmented generation, embedding-based similarity search, weakly supervised labeling, and graph-grounded reasoning. Second, it maps core stewardship objectives drawn from FAIR principles and provenance-centric models such as PROV to concrete, measurable capabilities that LLM-enabled systems can support, such as findability through semantic search, interoperability via schema alignment, and reusability through automated documentation and quality signals. Third, it proposes a reference architecture that situates LLMs as assistive, not authoritative, components within a broader governance stack, integrating catalogs, lineage stores, policy engines, and human-in-the-loop review. This architectural framing emphasizes separation of concerns: LLMs provide interpretation and synthesis, while authoritative facts remain anchored in governed data systems.

Finally, the article demonstrates an evidence-mapping case study that evaluates an enterprise AI governance posture using only publicly available information. Rather than relying on internal audits or proprietary disclosures, the evidence-mapping approach aligns observable signals such as policy statements, tooling announcements, certifications, and organizational roles with recognized governance standards. This method enables a pragmatic assessment of readiness, maturity, and gaps, while remaining non-judgmental and repeatable across organizations. By applying the approach to a real enterprise context, the study illustrates how technical capability and governance intent often evolve at different speeds, and how evidence mapping can highlight areas where additional controls, documentation, or oversight would strengthen trust. The article concludes by discussing open research questions, including metrics for evaluating LLM-assisted stewardship outcomes, strategies for maintaining alignment under regulatory change, and design patterns for sustained human oversight in AI-augmented data governance systems.

II. LLM-CENTERED ARCHITECTURAL FOUNDATIONS FOR ENTERPRISE DATA STEWARDSHIP

2.1 Transformers and LLMs

Transformer architectures form the computational backbone of most modern large language models and represent a fundamental departure from earlier recurrent and convolutional approaches. By replacing sequential recurrence with self-attention, transformers enable models to consider all tokens in an input simultaneously, learning rich contextual relationships regardless of distance. This capability is particularly valuable for enterprise data stewardship tasks, where meaning often depends on relationships across column names, descriptions, constraints, and surrounding documentation. Encoder–decoder designs and encoder-only variants allow models to build contextualized representations that support schema inference, table understanding, entity resolution, and semantic similarity across heterogeneous data assets.Pretraining strategies further amplify the usefulness of transformers for stewardship.

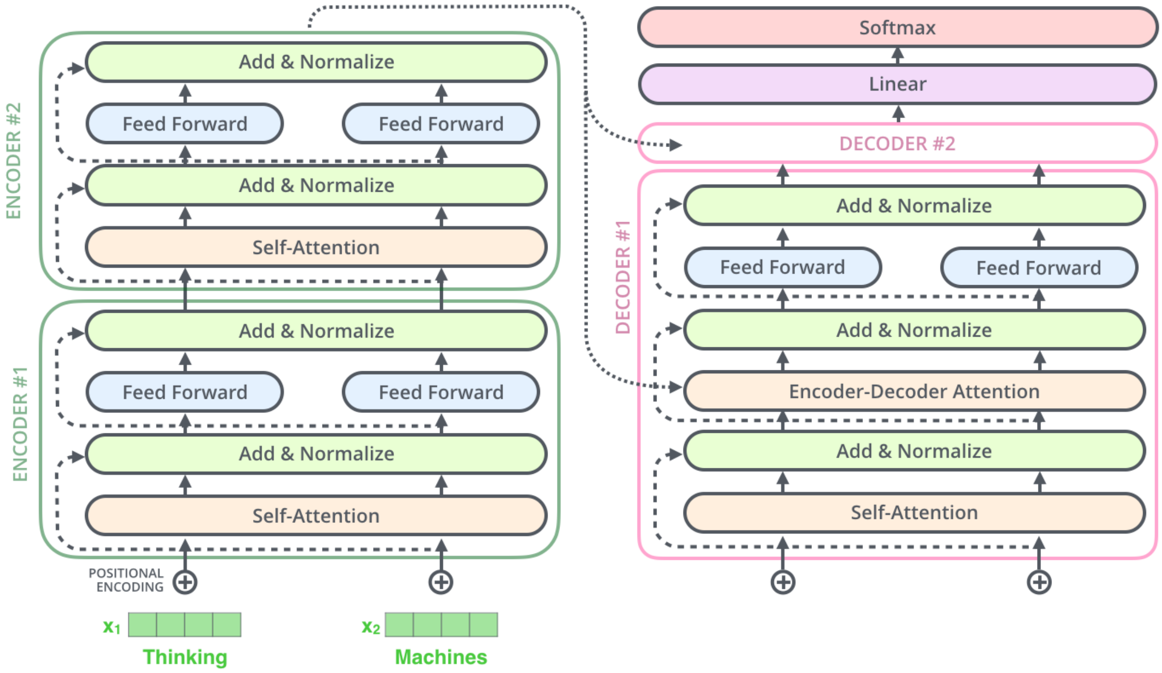


Figure1. Transformer / LLM Architecture

Models such as BERT demonstrated that large-scale self-supervised learning on general corpora yields transferable representations that can be fine-tuned for specialized tasks with relatively little labeled data. Subsequent autoregressive models, exemplified by GPT-3, revealed emergent behaviors such as few-shot and zero-shot learning, enabling rapid prototyping of stewardship workflows without extensive task-specific training. For enterprises, these properties translate into lower barriers for experimenting with metadata classification, glossary alignment, and natural-language querying over catalogs and logs. In stewardship contexts, attention mechanisms are especially valuable because they provide an implicit form of relevance weighting. When analyzing a table schema or pipeline description, attention allows the model to focus on salient tokens such as identifiers, units, or policy keywords while still preserving global context. This makes transformers well suited for extracting semantic signals from noisy or inconsistently documented assets. However, because attention operates over learned representations rather than authoritative facts, transformer-based models must be embedded within architectures that explicitly manage grounding, provenance, and validation to ensure trustworthy outcomes.

2.2 Retrieval-augmented generation (RAG) and dense retrieval

While transformers enable powerful pattern learning, an LLM's internal, or parametric, knowledge is inherently limited by its training data and cutoff date. In enterprise environments, where metadata, policies, and lineage evolve continuously, relying solely on parametric knowledge is insufficient and potentially risky. Retrieval-augmented generation (RAG) addresses this limitation by coupling LLMs with external retrieval components that dynamically supply relevant context at inference time. This pattern ensures that generated responses are grounded in up-to-date, organization-specific sources such as data catalogs, governance policies, and pipeline logs.Dense retrieval techniques, including Dense Passage Retriever (DPR), play a critical role in making RAG effective for stewardship.

By embedding queries and documents into a shared vector space, dense retrievers can capture semantic similarity beyond exact keyword matches, which is essential when metadata is sparse or inconsistently phrased. For example, a steward asking about "customer churn attributes" may retrieve tables labeled with retention, lifecycle, or subscription terminology. When combined with LLMs, these retrieved artifacts become explicit evidence that conditions generation, improving both relevance and traceability. From a governance perspective, RAG pipelines also introduce a natural mechanism for provenance. Retrieved documents can be logged, cited, and reviewed alongside model outputs, enabling auditors and stewards to inspect the basis of automated recommendations or answers. This aligns well with regulatory expectations around explainability and accountability. Nevertheless, RAG systems must be carefully engineered to manage retrieval quality, context window limits, and conflict resolution when sources disagree, reinforcing the need for evaluation metrics and human-in-the-loop controls in production deployments.

2.3 Programmatic labeling and weak supervision

A persistent barrier to applying machine learning in enterprise stewardship is the scarcity of high-quality labeled data for domain-specific tasks. Activities such as classifying column roles, detecting sensitive attributes, or mapping fields to business concepts often require expert judgment and intimate organizational knowledge. Creating large, manually labeled datasets for these tasks is costly and slow, limiting the practicality of fully supervised approaches. Programmatic labeling and weak supervision offer a pragmatic alternative by shifting effort from labeling individual examples to encoding domain knowledge as reusable rules and heuristics.Weak supervision frameworks enable teams to define labeling functions that capture signals such as naming conventions, data types, value distributions, or references to glossaries and policies.

Although each labeling function may be noisy or incomplete, combining many such functions statistically can produce training labels of sufficient quality to bootstrap models. In stewardship scenarios, this approach allows organizations to rapidly adapt LLMs or retrievers to proprietary schemas and taxonomies without exposing sensitive data or requiring extensive annotation campaigns. When integrated with LLM-based systems, weak supervision plays a complementary role rather than a competing one. Labeling functions can be used to generate silver-standard datasets for fine-tuning embeddings, calibrating classifiers, or validating LLM suggestions. Human review remains essential, particularly for high-risk classifications such as personally identifiable information, but weak supervision dramatically reduces the manual burden. As a result, enterprises can iterate faster on stewardship capabilities while maintaining alignment with governance and compliance requirements.

2.4 Knowledge graphs and canonical models

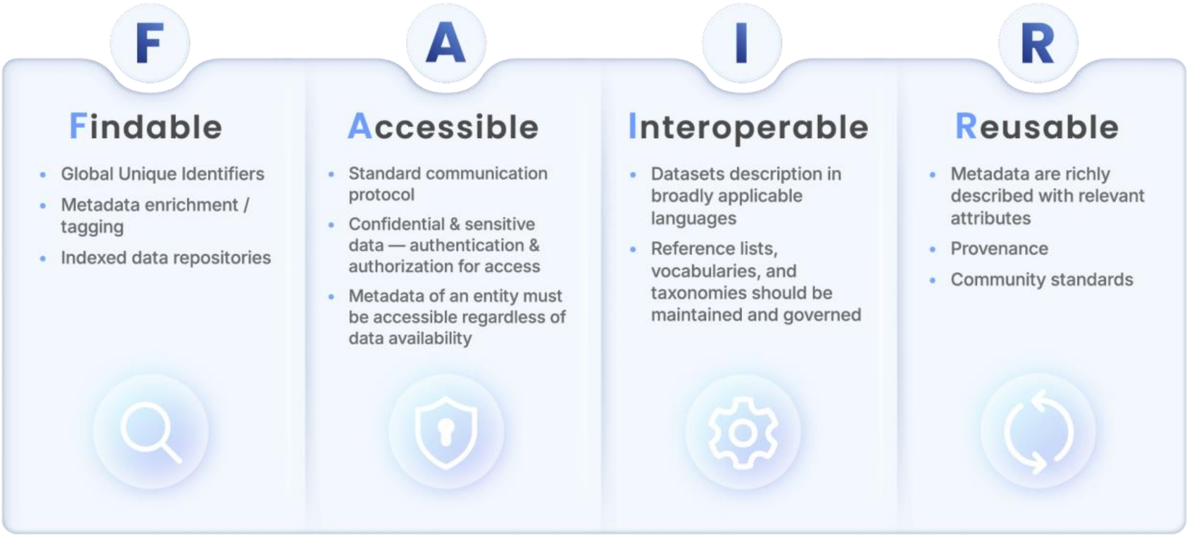
Knowledge graphs provide a structured, machine-interpretable representation of enterprise semantics, capturing entities, relationships, and business rules in a canonical form. Unlike free-text documentation, graphs enforce explicit structure and shared meaning, making them a natural foundation for stewardship activities such as glossary management, lineage tracing, and impact analysis. By encoding authoritative definitions and relationships, knowledge graphs act as a stable semantic backbone against which other systems can align.When combined with LLMs, knowledge graphs serve as grounding anchors that constrain and validate model outputs. LLMs can propose candidate entity mappings, relationship hypotheses, or documentation drafts by interpreting unstructured metadata and usage logs.

These proposals can then be checked against the graph for consistency with known entities and ontologies. This division of labor leverages the generative flexibility of LLMs while preserving the governance guarantees of curated canonical models. In practice, this hybrid approach supports scalable stewardship workflows. For example, as new datasets are ingested, an LLM can suggest mappings to existing business concepts, while the knowledge graph enforces uniqueness, hierarchy, and referential integrity. Stewards remain in the loop to approve or correct changes, ensuring that the graph evolves in a controlled manner. Over time, this feedback loop strengthens both the graph and the LLM-assisted tooling, creating a virtuous cycle of improved semantic clarity and reduced manual effort across the enterprise data estate.

III. API-FIRST DEVELOPMENT: CONTRACT-DRIVEN DESIGN AND GOVERNANCE

3.1 FAIR principles and stewardship KPIs

The FAIR principles Findable, Accessible, Interoperable, and Reusable provide a concrete and widely accepted framework for operationalizing data stewardship goals in enterprise environments, as articulated by **Wilkinson et al. (2016)**. Rather than treating FAIR as a conceptual ideal, organizations increasingly translate each pillar into measurable key performance indicators (KPIs) that reflect day-to-day data practices at scale. *Findable* is commonly operationalized through metrics such as the proportion of datasets and tables with machine-readable metadata, semantic tags, ownership fields, and contextual descriptions, alongside precision and recall for dataset discovery queries. These measures directly capture whether data assets can be efficiently located by both humans and automated systems. *Accessible* focuses on the explicit representation of access conditions, with KPIs tracking coverage of access policy metadata, automated detection of missing or conflicting authorization rules, and mean time to resolve access requests.



F

A

I

R

Findable
- Global Unique Identifiers
- Metadata enrichment / tagging
- Indexed data repositories

Accessible
- Standard communication protocol
- Confidential & sensitive data — authentication & authorization for access
- Metadata of an entity must be accessible regardless of data availability

Interoperable
- Datasets description in broadly applicable languages
- Reference lists, vocabularies, and taxonomies should be maintained and governed

Reusable
- Metadata are richly described with relevant attributes
- Provenance
- Community standards

Figure 2. FAIR Data Principles

LLM-assisted stewardship systems can actively support progress against these KPIs while remaining embedded in a governed measurement loop. For *Interoperable*, metrics may include the percentage of attributes mapped to canonical schemas or ontologies, the acceptance rate of automated schema-alignment proposals, and the stability of mappings over time as schemas evolve. *Reusable* emphasizes long-term value and trust, assessed through the presence of provenance records, data quality scores, validation checks, example usage notes, and documentation freshness. LLMs contribute by generating candidate metadata, aligning terminology across silos, and summarizing reuse guidance, but KPI tracking ensures that these contributions are systematically evaluated rather than assumed to be correct. By pairing automation with continuous measurement, enterprises can ensure that LLM-enabled stewardship leads to durable improvements in FAIR compliance instead of isolated documentation gains.

3.2 Provenance (W3C PROV) and auditability

Provenance is a foundational requirement for trustworthy data stewardship because it enables organizations to reconstruct how data assets were created, transformed, and governed over time. The W3C PROV standard formalizes provenance through a concise model based on three core constructs: entities, activities, and agents that together describe data lineage across ingestion, transformation, and consumption stages. In enterprise data estates, PROV-based representations make it possible to trace datasets across batch jobs, streaming pipelines, analytical transformations, and downstream dashboards. This traceability is critical for regulatory audits, incident response, impact analysis, and accountability, particularly in environments subject to privacy, financial, or safety regulations.

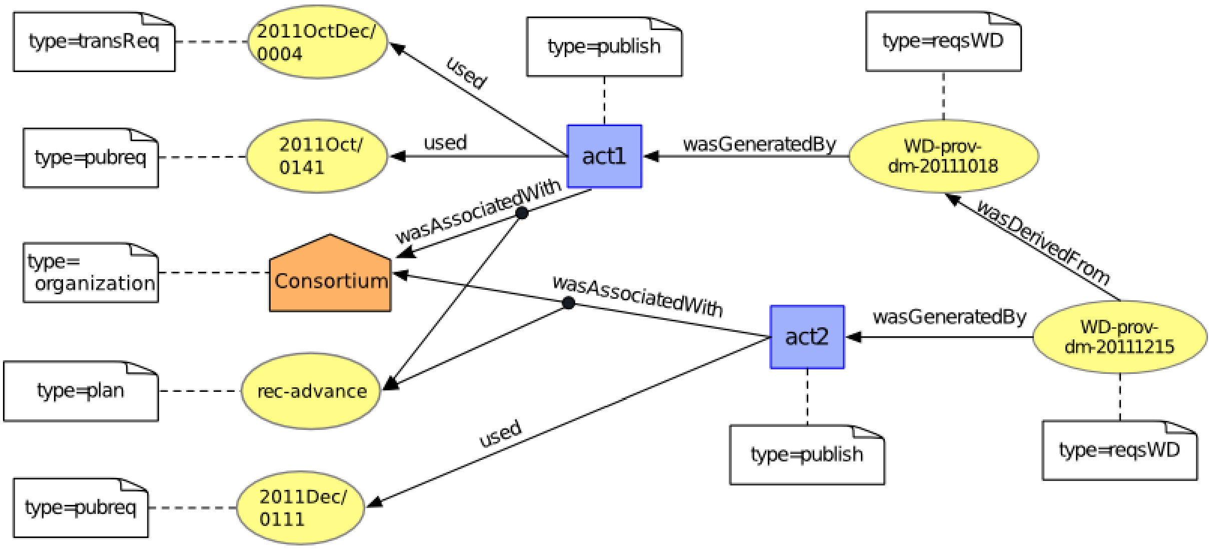


Figure 3. W3C PROV-DM Core Provenance Model

When LLMs are introduced into stewardship workflows, provenance must serve as explicit evidence rather than implicit background context. An LLM answering a lineage-related question should not rely on learned associations alone, but instead retrieve and surface the relevant fragment of the PROV graph that documents source entities, transformation activities, and responsible agents. This design ensures that every generated response is verifiable and auditable, allowing stewards and auditors to inspect the underlying evidence. LLMs thus function as an interpretive interface over authoritative provenance records, translating structured lineage into accessible explanations while preserving formal traceability. By tightly coupling natural-language interaction with PROV-compliant graphs, enterprises can balance usability and accountability in AI-augmented governance systems.

3.3 Data quality and entity resolution

Data quality and entity resolution remain among the most labor-intensive and error-prone aspects of enterprise data stewardship. Organizations routinely confront duplicated records, inconsistent attribute naming, incompatible identifiers, and semantic drift as systems evolve independently over time. Traditional approaches such as rule-based validation, statistical profiling, and probabilistic record-linkage algorithms provide essential quantitative foundations for detecting inconsistencies and estimating uncertainty. These methods enable stewards to measure precision, recall, and confidence, which are critical for risk-aware decision-making and regulatory defensibility.

LLMs can augment these established techniques by operating at the semantic layer, proposing candidate canonicalizations, reconciling ambiguous attribute names, and summarizing quality issues in human-readable terms. For example, an LLM may suggest that multiple differently named identifiers across datasets correspond to the same conceptual entity, accelerating steward review and remediation. However, these suggestions must always be evaluated against deterministic checks and probabilistic linkage baselines to avoid overgeneralization or hallucinated equivalences. Effective stewardship architectures therefore position LLMs as assistive components within a broader quality framework, where automated metrics, statistical validation, and human oversight jointly ensure that gains in efficiency do not compromise correctness, consistency, or long-term trust in enterprise data assets.

IV. ARCHITECTURE AND INTERACTION PATTERNS FOR LLM-ASSISTED DATA STEWARDSHIP

4.1 Components and data flows

An LLM-assisted stewardship architecture is composed of several tightly integrated components, each responsible for a distinct aspect of governance, interpretation, and control. At the foundation lies a centralized metadata and catalog store, such as Apache Atlas, which maintains authoritative records of datasets, tables, schemas, tags, ownership, and lineage. This catalog acts as the system of record and the primary grounding source for all downstream intelligence. On top of this layer sits a retriever index, implemented using a combination of dense and sparse retrieval techniques, which indexes metadata entries, documentation, transformation code, and policy artifacts. This index enables semantic retrieval, supporting dense passage retrieval (DPR) for contextual relevance as well as keyword-based precision for exact matches. Together, these layers ensure that relevant evidence can be efficiently located in response to user or system queries.

Above retrieval, the LLM and RAG orchestrator coordinates query interpretation, evidence gathering, and response generation. The orchestrator executes retrieval calls, conditions the model on retrieved context, and generates grounded outputs such as answers, summaries, or candidate metadata edits. Crucially, these outputs are always linked back to a provenance store that maintains PROV-compatible lineage fragments, ensuring that every suggestion or explanation can be traced to concrete sources and transformation activities. A human-in-the-loop steward interface exposes this evidence, along with confidence scores and suggested actions, enabling review, approval, or correction. Feedback from stewards is fed back into weak supervision and labeling pipelines to improve future performance. Finally, a policy and compliance module enforces access constraints, validates actions against governance rules, and records decision logs, ensuring that automation operates within clearly defined organizational and regulatory boundaries.

4.2 Typical interactions and use cases

One of the most common interactions enabled by this architecture is a discovery assistant that supports natural-language queries over enterprise data assets. Users can ask questions such as “Which datasets contain customer PII?” or “What tables are approved for financial reporting?” and receive ranked results drawn from the catalog and retriever index. Each result is accompanied by provenance fragments that explain why the dataset was selected, including relevant tags, lineage paths, and policy annotations. This combination of semantic search and explicit evidence allows users to quickly locate trustworthy data while giving stewards and auditors visibility into the basis of each recommendation.

Beyond discovery, the system supports a range of active stewardship workflows. For metadata generation, LLMs can draft table and column descriptions, suggest owners, generate example queries, and recommend tags based on observed usage and documentation, with stewards reviewing and approving changes through the HITL interface. Lineage question answering enables users to understand how a dataset was derived by presenting narrative explanations grounded in PROV graph fragments and links to transformation scripts or pipelines. In schema mapping and ETL recommendation scenarios, the system proposes alignments between source and target schemas, mapping attributes to canonical models with associated confidence scores. Across all these use cases, the consistent pattern is assistive automation paired with explicit evidence and human oversight, ensuring that efficiency gains do not compromise accuracy, accountability, or governance.

V. EVIDENCE MAPPING METHODOLOGY FOR GOVERNANCE ASSESSMENT

5.1 Rationale

Evidence mapping offers a practical mechanism for evaluating data and AI governance maturity in contexts where direct access to internal policies, controls, or audit artifacts is unavailable. Many enterprises are unwilling or unable to disclose detailed governance documentation due to confidentiality, competitive risk, or regulatory sensitivity. However, organizations consistently emit public signals through regulatory filings, technical blogs, product documentation, conference presentations, open-source contributions, and vendor case studies. When analyzed systematically, these artifacts can reveal governance intent, architectural choices, tooling investments, and operational priorities. Evidence mapping is grounded in the premise that governance practices leave observable traces in how systems are described, built, and communicated. Rather than making normative judgments, the method focuses on surfacing what can be reasonably inferred from available information. This non-intrusive approach enables repeatable and comparative assessments while respecting information asymmetry between organizations and external evaluators.

A key strength of evidence mapping is its alignment with modern, technology-driven governance, where policies are increasingly embedded in platforms, workflows, and tooling rather than static documents. Choices such as adopting lineage-aware catalogs, publishing access-control documentation, or describing human review workflows in blog posts all constitute governance signals. Evidence mapping treats these signals as partial but meaningful indicators of maturity, acknowledging that absence of evidence may reflect non-disclosure rather than non-existence. By explicitly capturing uncertainty, the approach avoids overinterpretation while still enabling structured analysis. As a result, evidence mapping is well suited for academic research, industry benchmarking, and early-stage regulatory readiness assessments in rapidly evolving AI-enabled enterprises.

5.2 Steps

The evidence-mapping methodology follows a structured, multi-stage process designed to ensure transparency, rigor, and interpretability. The first step is *scope definition*, which identifies the governance dimensions to be evaluated, such as policy articulation, role definition, tooling support, provenance management, and auditability. Clear scoping prevents overreach and ensures that findings remain comparable across organizations and studies. The second step, *signal collection*, involves gathering publicly available artifacts, including product documentation, engineering blogs, open-source repositories, regulatory disclosures, and third-party case studies. Sources are selected based on relevance, credibility, and temporal alignment with the assessment period.

The third step, *mapping to a reference framework*, aligns collected signals with established governance models such as DAMA-DMBOK, FAIR, and W3C PROV, providing a normative structure for interpretation. This is followed by *confidence scoring*, where each mapped signal is evaluated based on its evidentiary strength, distinguishing between direct evidence, indirect evidence, and absent evidence. The final step, *gap analysis*, synthesizes these mappings to identify areas where governance signals are strong, emerging, or weak. Rather than serving as a compliance verdict, the output highlights priorities for deeper investigation, additional disclosure, or targeted governance improvement. Together, these steps transform disparate public information into a coherent and methodical assessment of enterprise governance maturity.

VI. CASE STUDY EVIDENCE MAPPING: INSPIRE BRANDS' AI-DRIVEN GOVERNANCE INITIATIVES

6.1 Scope and data sources

The evidence-mapping exercise scoped four governance dimensions relevant to AI-enabled data stewardship at **Inspire Brands**: organizational roles and stewardship, tooling and metadata management, provenance and lineage capabilities, and responsible AI and governance policies. This scoping was intentionally limited to dimensions that are both central to enterprise governance maturity and plausibly observable through public artifacts. By constraining the scope, the assessment avoids speculative inference while maintaining comparability with other enterprises evaluated using the same methodology. Each dimension was defined using reference concepts drawn from established governance frameworks, ensuring that signals could be interpreted against a consistent baseline rather than ad hoc criteria.

Public data sources included company-authored blog posts describing analytics and data-platform initiatives, technical presentations delivered at industry conferences, vendor partnership announcements, and regulatory or press statements related to technology and AI adoption. These artifacts were treated as governance signals rather than definitive proof of controls. Particular attention was paid to the recency, specificity, and technical depth of sources, as these characteristics affect evidentiary strength. The collection process emphasized triangulation, where multiple independent sources pointing to the same practice increased confidence, while isolated mentions were treated more cautiously.

6.2 Findings

Across organizational roles and stewardship, public signals suggest that Inspire Brands has assigned responsibility for analytics and data initiatives to identifiable leaders and cross-functional teams. Mentions of centralized analytics groups and named leadership roles in public communications indicate that stewardship responsibilities are at least partially institutionalized rather than ad hoc. These signals constitute direct but medium-confidence evidence, as role descriptions are visible but formal stewardship charters, escalation paths, or accountability mechanisms are not publicly documented. Nevertheless, the presence of clearly articulated roles is a positive indicator of governance intent and organizational readiness.

DOI: 10.15680/IJCTECE.2024.0701007

In the tooling and architecture dimension, evidence is stronger. Public references to modern cloud data platforms, metadata-aware analytics tooling, and vendor partnerships indicate explicit investment in metadata management and scalable data infrastructure. This constitutes direct evidence derived from technical blog posts and partner materials. In contrast, provenance and lineage disclosures are less explicit. While mentions of workflow orchestration, data pipelines, and analytics platforms imply that lineage capture is technically feasible, no end-to-end lineage artifacts or PROV-style representations were found publicly, resulting in indirect and lower-confidence evidence. Similarly, AI governance signals such as statements on responsible AI or ethical technology use indicate awareness and early-stage initiatives, but the absence of detailed public policies, evaluation practices, or audit disclosures represents a clear evidence gap.

6.3 Interpretation and implications

Taken together, the evidence mapping reveals a pattern that is common among large enterprises adopting advanced analytics and AI capabilities. Technical investments and organizational structures are often the most visible signals, as they are closely tied to delivery outcomes and vendor ecosystems. By contrast, formal governance artifacts such as published provenance exports, detailed AI risk assessments, or auditable policy documentation are less frequently disclosed publicly, even when they may exist internally. Evidence mapping does not assume absence of governance, but it does highlight where external stakeholders lack visibility into controls that underpin trust and accountability.

For enterprises, this case underscores the value of a deliberate two-track governance strategy. The first track focuses on building robust technical capabilities, including metadata catalogs, lineage capture, and LLM-enabled RAG pipelines that support scalable stewardship. The second track emphasizes codifying and, where appropriate, publishing governance artifacts that demonstrate accountability, audit readiness, and responsible AI practices. Aligning these tracks not only strengthens internal governance but also improves external trust signals, positioning organizations to respond more effectively to regulatory scrutiny, partner due diligence, and public expectations around transparent and responsible AI adoption.

VII. EVALUATION AND METRICS FOR LLM STEWARDSHIP SYSTEMS

7.1 Dual evaluation axis

Evaluating LLM-assisted data stewardship systems requires a dual-axis approach that explicitly separates *technical performance* from *governance and trustworthiness outcomes*. On the technical axis, evaluation focuses on whether the system performs its intended tasks correctly and efficiently. This includes standard information-retrieval metrics such as precision and recall for metadata and document retrieval, accuracy of schema-mapping and ontology-alignment proposals, and correctness of suggested metadata edits when compared against ground-truth annotations or expert-reviewed baselines. These measures assess whether the LLM and retrieval components are producing technically valid outputs under controlled conditions. However, technical accuracy alone is insufficient for enterprise adoption, particularly in regulated or high-risk domains where incorrect automation can have outsized consequences.

The second axis addresses governance, accountability, and human trust, which are equally critical for real-world deployment. Governance-oriented evaluation asks whether the system's outputs are explainable, auditable, and appropriately constrained by policy. Metrics in this category include traceability such as the fraction of model assertions explicitly backed by provenance fragments, completeness of audit logs, and the integrity of approval workflows. Human-centered measures, including steward satisfaction, perceived usefulness, and cognitive load, provide additional insight into whether the system genuinely supports stewardship rather than creating new oversight burdens. Together, these two axes ensure that LLM stewardship systems are evaluated not only as machine-learning artifacts but as socio-technical systems embedded in governance processes.

7.2 Specific metrics

Operationalizing this dual-axis evaluation requires a concrete set of metrics that can be tracked continuously in production environments. *Evidence precision* measures the percentage of model-generated answers that correctly cite the underlying provenance documents, lineage records, or policy artifacts used as evidence. This metric directly reflects grounding quality and is critical for auditability. *Edit acceptance rate* captures the proportion of suggested metadata edits such as descriptions, tags, or schema mappings that are approved by human stewards, serving as a proxy for both technical accuracy and practical usefulness. A low acceptance rate may indicate model overreach, poor calibration, or misalignment with organizational standards.

Additional metrics focus on risk reduction and efficiency gains. *False-positive governance alert rate* measures how often the system incorrectly flags sensitive data, policy violations, or compliance risks, which is particularly important for avoiding alert fatigue and erosion of trust. *Time-to-catalog* quantifies the reduction in elapsed time required to produce reviewer-approved metadata for a dataset, capturing the productivity impact of LLM assistance. When tracked together, these metrics allow enterprises to balance speed, accuracy, and trust, providing early warning signals when automation degrades governance outcomes and evidence-based justification when LLM stewardship systems deliver measurable value.

VII. CONCLUSION

Large language models, when thoughtfully integrated with retrieval mechanisms, knowledge graphs, weak supervision techniques, and formal provenance standards, represent a substantive advance in the practice of enterprise data stewardship. Rather than replacing existing governance infrastructure, LLMs augment it by operating at the semantic and interaction layers translating unstructured signals into structured insights, accelerating metadata curation, and enabling natural-language access to complex data estates. Retrieval-augmented generation ensures that model outputs remain grounded in authoritative sources, while knowledge graphs provide canonical anchors that constrain interpretation and preserve shared meaning. Weak supervision offers a scalable path to domain adaptation, allowing organizations to encode institutional knowledge without prohibitive labeling costs. Together, these components transform stewardship from a predominantly manual, reactive activity into a more proactive, adaptive capability that can keep pace with rapidly evolving data environments.

The evidence-mapping methodology introduced in this work addresses a complementary challenge: how to assess governance maturity and readiness in the absence of privileged internal access. By systematically synthesizing publicly available artifacts and aligning them to established governance frameworks, evidence mapping enables transparent, repeatable evaluations that respect organizational boundaries. This approach is particularly valuable in an era where AI adoption often outpaces formal governance disclosure, creating asymmetries between technical capability and external trust. Evidence mapping does not claim completeness or certainty; instead, it explicitly encodes confidence and gaps, allowing researchers, partners, and regulators to distinguish between strong signals, weak inferences, and areas of non-disclosure. As such, it provides a practical lens for comparative analysis and early-stage risk assessment in enterprise AI governance.

Ultimately, successful adoption of LLM-assisted stewardship depends less on raw model performance than on the surrounding socio-technical system. Robust provenance capture, explicit linkage between outputs and evidence, and immutable audit trails are essential to ensure that automation enhances rather than undermines accountability. Human-in-the-loop validation remains critical, both to manage residual uncertainty and to reinforce stewardship norms and ownership. Governance processes must therefore evolve alongside technical architectures, prioritizing traceability, auditability, and clear decision rights. When these elements are aligned, LLM-enabled stewardship systems can deliver measurable gains in efficiency and insight while strengthening, rather than diluting, trust in enterprise data assets.

REFERENCES

1. Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., … Zimmermann, T. (2019). Software engineering for machine learning: A case study. *Proceedings of the 41st International Conference on Software Engineering*, 291–300.https://dl.acm.org/doi/10.1109/icse-seip.2019.00042
2. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … Vayena, E. (2018). AI4People An ethical framework for a good AI society. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5
3. James, K. L., Randall, N. P., & Haddaway, N. R. (2016). A methodology for systematic mapping in environmental sciences. *Environmental Evidence, 5*(7).https://doi.org/10.1186/s13750-016-0059-6
4. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2
5. Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review, 165*(3), 633–705. https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3/

6. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2). https://doi.org/10.1177/2053951716679679
7. Shravan Kumar Reddy Padur "Empowering Developer & Operations Self-Service: Oracle APEX + ORDS as an Enterprise Platform for Productivity and Agility" International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Print ISSN : 2395-1990, Online ISSN : 2394-4099, Volume 4, Issue 11, pp.364-372, November-December-2018. Available at doi : https://doi.org/10.32628/IJSRSET1844429
8. Miake-Lye, I. M., Hempel, S., Shanman, R., & Shekelle, P. G. (2016). What is an evidence map? A systematic review of published evidence maps. *Systematic Reviews, 5*(28). https://doi.org/10.1186/s13643-016-0204-x
9. Sudhir Vishnubhatla. (2018). From Risk Principles to Runtime Defenses: Security and Governance Frameworks for Big Data in Finance. In International Journal of Science, Engineering and Technology (Vol. 6, Number 1). Zenodo. https://doi.org/10.5281/zenodo.17452405
10. Mökander, J., Floridi, L., & Taddeo, M. (2021). Ethics-based auditing of automated decision-making systems. *AI and Ethics*, *2*(4), 609–623. https://doi.org/10.1007/s11948-021-00319-4
11. Nithin Nanchari. (2020). Wearable IoT Devices for Health. Journal of Scientific and Engineering Research, 7(11), 235–236. https://doi.org/10.5281/zenodo.15966018
12. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., … Barnes, P. (2020). Closing the AI accountability gap. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. https://doi.org/10.1145/3351095.3372873
13. Shravan Kumar Reddy Padur "Empowering Developer & Operations Self-Service: Oracle APEX + ORDS as an Enterprise Platform for Productivity and Agility" International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Print ISSN : 2395-1990, Online ISSN : 2394-4099, Volume 4, Issue 11, pp.364-372, November-December-2018. Available at doi : https://doi.org/10.32628/IJSRSET1844429
14. Stahl, B. C., Timmermans, J., & Mittelstadt, B. D. (2016). The ethics of computing. *ACM Computing Surveys*, *48*(4). https://doi.org/10.1145/2871196
15. Shravan Kumar Reddy Padur. (2016). Network Modernization in Large Enterprises: Firewall Transformation, Subnet Re-Architecture, and Cross-Platform Virtualization. In International Journal of Scientific Research & Engineering Trends (Vol. 2, Number 5). Zenodo. https://doi.org/10.5281/zenodo.17291987