



Empowering Reliable GenAI with LLM-Assisted Metadata Enrichment: An Empirical Study on Climate Policy Data

Dit Suthiwong
Faculty of Information Technology, King Mongkut's University
of Technology North Bangkok, Bangkok, Thailand
s5507011966036@email.kmutnb.ac.th

ABSTRACT

This paper presents an empirical study on the impact of LLM-assisted metadata enrichment on the reliability and performance of enterprise grade Generative AI (GenAI) systems, using the OECD IFCMA Climate Policy Dashboard as a real world testbed. The dataset comprising over 1,600 heterogeneous climate policy instruments across 43 approaches and multiple countries exhibits significant semantic inconsistency and incomplete metadata, reflecting common challenges in multinational data environments. The authors implement a three stage GenAI enabled pipeline: (1) definition of a structured metadata schema, (2) LLM-driven semantic enrichment to infer missing fields and harmonize terminology, and (3) a metadata aware Retrieval Augmented Generation (RAG) system that leverages enriched context for grounded responses. Quantitative evaluation demonstrates a statistically significant improvement in metadata completeness from a mean of 0.41 to 0.83 ($p < 0.001$) and a marked increase in cross country semantic consistency, with cosine similarity rising from 0.64 to 0.85 for carbon pricing policies. These enhancements directly translate into tangible RAG performance gains: retrieval precision improves by 25.8%, answer faithfulness by 23.5%, and hallucination rates decline by 41.9%. Crucially, correlation analysis confirms a strong positive relationship between metadata quality and GenAI reliability ($\rho > 0.7$). The study positions high quality metadata not as auxiliary documentation but as a foundational architectural component that enables observability, explainability, and trust in high stakes GenAI applications. By bridging a critical gap between conceptual frameworks and empirical validation, this work establishes metadata centric design as essential for scalable, governance aligned, and reliable enterprise AI systems.

Keywords: Generative AI, Metadata Enrichment, Large Language Models (LLMs), Retrieval Augmented Generation (RAG), Semantic Consistency, Climate Policy Data, Enterprise AI Governance, Data Quality

Received: 12 September 2025, Revised 3 December 2025, Accepted 18 December 2025

Copyright: DLINE

1. Introduction

As enterprises increasingly deploy Generative AI (GenAI) systems for high stakes decision support, such as in

of AI outputs has become paramount. While modern unified data platforms and AI-native architectures theoretically promise seamless integration of data engineering and GenAI workflows, a critical problem persists: empirical evidence quantitatively linking foundational metadata quality to downstream AI performance remains scarce and largely anecdotal. This paper explicitly addresses this significant research gap by investigating how Large Language Model (LLM) assisted metadata enrichment directly enhances the trustworthiness, accuracy, and operational effectiveness of GenAI applications within a complex, real world policy context.

Our research is conducted using the OECD IFCMA Climate Policy Dashboard a representative, heterogeneous, and multinational dataset that is plagued by inconsistent terminology, structural variability, and incomplete metadata, mirroring the challenges faced by global enterprises. We implement and rigorously evaluate a novel GenAI enabled pipeline designed to systematically conquer data chaos. This pipeline integrates three core stages: first, the definition of a rigorous, structured metadata schema to establish a quality baseline; second, LLM driven semantic enrichment to harmonise terminology and infer missing contextual descriptors; and third, a metadata aware Retrieval Augmented Generation (RAG) system that leverages this enriched context to ground AI responses in verified, relevant information. By demonstrating measurable performance gains, this study provides a concrete framework and evidence for using metadata as a keystone for building trustworthy, enterprise grade GenAI.

2. Related Work

This section synthesizes prior research and industry practice on unified data platforms, Generative AI (GenAI)-enabled data engineering, AI-native databases, and next generation metadata systems. The literature is organized thematically to highlight methodological trends, limitations, and open research gaps, thereby positioning the present work within the existing body of knowledge.

2.1 Unified Data and AI Platforms for Enterprise GenAI

Unified data and AI platforms have emerged as foundational enablers of enterprise-scale GenAI adoption. [1, 2] Databricks' Data Intelligence Platform exemplifies this paradigm by integrating data ingestion, transformation, storage, analytics, and AI model development within a single governance framework. [3, 4, 5]. Prior work emphasizes the platform's support for Retrieval Augmented Generation (RAG) and fine-tuning of large language models (LLMs), enabling organizations to develop domain specific AI systems while maintaining compliance and operational efficiency [3]. Delta Lake further underpins these capabilities through ACID compliant storage, schema evolution, and scalable processing of structured and unstructured data.

Recent enhancements extend this unified vision through cross cloud data and AI asset sharing, fine-grained access control for models and embeddings, and end to end lineage tracking across AI workflows. Collectively, these features position lakehouse architectures as a convergence point for data engineering, analytics, and AI under a single architectural and governance model.

2.2 Enterprise AI Architectures and Deployment Methodologies

Beyond individual platforms, enterprise AI systems are increasingly characterized by compositional architectures that integrate RAG pipelines, vector databases, fine tuning workflows, agent frameworks, and model registries. [6, 7, 8] Berton [9] adopted a use case driven methodology, arguing against a single prescriptive

finance, healthcare, and public policy, ensuring the reliability of underlying data and the semantic coherence architecture and instead presenting multiple deployment patterns across Red Hat platforms. Best practices such as reproducible environments, OpenAI-compatible APIs, and automated configuration management (e.g., Ansible) are highlighted as essential for portability and auditability.

2.3 Generative AI as an Enabler of Intelligent Data Engineering

Generative AI is increasingly viewed as a transformative layer for data engineering automation. Quang Hai Khuat [10] characterized this shift as a co creative human AI partnership in which GenAI enhances productivity, reduces errors, and lowers the barrier for non expert participation in data workflows. Complementing this perspective, Govindarajulunaidu [11] proposes an AI-driven data engineering framework based on declarative data contracts, active metadata, and learning based observability to proactively detect anomalies in data freshness, schema, and distribution across batch and streaming pipelines.

2.4 From Batch ETL to AI-Powered Streaming Pipelines

The transition from batch oriented ETL to real time, AI enabled streaming pipelines reflects the growing demand for low latency analytics and adaptive decision making. Traditional ETL systems are widely acknowledged as reliable but inflexible in rapidly evolving data environments. Recent studies highlight the role of AI in enabling adaptive transformations, intelligent orchestration, and automated quality checks within streaming architectures. Traditional batch based ETL pipelines struggle with modern real time data demands from IoT, transactions, and telemetry. Enterprises now require low latency, scalable, and intelligent data processing to support timely decision making. AI-powered streaming ETL architectures are emerging to address latency, complexity, and data quality challenges, enabling dynamic, real time analytics essential in today's data-driven landscape. [12, 13, 14, 15].

In parallel, Kurapati [16] discusses the emergence of AI-native databases designed to support real time intelligent workloads through cloud integration, automation, explainability, and enhanced security [Kurapati, 2025]. These systems represent a methodological departure from traditional relational models toward architectures optimized for AI-centric analytics.

2.5 Metadata Management and LLM-Assisted Semantic Enrichment

Metadata management has become increasingly critical with the adoption of third generation data lakes, or lakehouses. [17, 18, 19]. Conventional semantic enrichment approaches rely heavily on manual curation or static ontologies, which are often impractical in dynamic environments. Rehm [20] (2025) introduces an LLM assisted metadata enrichment pipeline implemented on Databricks using Azure OpenAI, demonstrating the automated generation of natural language table descriptions from schema and profiling metadata. Evaluation results suggest improved dataset discoverability and interpretability, though a continued need for human in the loop validation remains.

2.6 Open Table Formats and Interoperable Data Architectures

Open table formats such as Apache Iceberg play a crucial role in enabling interoperable and resilient data lake and mesh architectures. Iceberg addresses challenges related to schema evolution, time travel, and multi-engine compatibility, offering a vendor neutral foundation for large scale analytics [21]. In practice, Iceberg is often positioned alongside platforms like Databricks, balancing openness and standardization with integrated analytics and AI capabilities.

2.7 Foundation Models for Automated ETL and Pipelines

Foundation models, particularly LLMs, are increasingly leveraged to automate ETL pipeline creation, optimization, and adaptation. Naveen [22] demonstrated how LLMs can translate natural language specifications into executable pipelines, enhance data transformation quality, and optimize execution with minimal human intervention. These approaches illustrate a methodological shift toward self-adapting, AI-driven data pipelines.

2.8 Evolution of Metadata Platforms for AI Systems

The rise of AI-driven workflows has catalyzed the emergence of fourth-generation metadata platforms that unify data and AI metadata. The concept of a unified Metadata System (TMS) encompasses experiment tracking, model lineage, prompt and embedding management, AI-specific quality metrics, ethical governance, and human AI interaction logs. [23, 24]. Rather than isolated metadata silos, recent work advocates a real time, graph based, API-first platform capable of holistic governance and scalable search.

Platforms such as *DataHub* exemplify this evolution by integrating data, models, and business outcomes into a real time knowledge graph. The literature outlines a phased unification process, from event based integration to federated deployments, while acknowledging challenges related to performance, interoperability, privacy, and organizational adoption. [25].

2.9 Research Gap

Recent literature on enterprise scale Generative AI (GenAI) systems has made substantial progress in describing platform capabilities, architectural patterns, and best practices. However, several critical gaps remain that limit both scientific understanding and practical adoption.

First, existing studies predominantly emphasize conceptual architectures and descriptive analyses of platform features, such as unified governance, metadata management, lineage tracking, and access control. While these contributions are valuable, there is a notable lack of rigorous empirical evidence quantifying how such unified mechanisms concretely affect GenAI reliability, data quality, model performance, or organizational decision-making outcomes. Consequently, claims regarding improvements in governance effectiveness and reliability remain largely unvalidated.

Second, although multiple enterprise GenAI architectural paradigms have been proposed, comparative empirical evaluations across alternative architectures are scarce. In particular, there is limited systematic analysis of performance, scalability, maintainability, and governance trade offs under real world enterprise workloads. This gap constrains evidence based architectural decision making for large scale GenAI deployments.

Third, prior work on AI-driven observability, automation, and metadata enrichment is largely framework-oriented or a proof of concept. Empirical validation demonstrating measurable improvements in pipeline robustness, operational efficiency, or data quality remains limited. Similarly, despite strong theoretical motivation, comparative studies between AI-powered streaming pipelines and traditional ETL systems especially with respect to scalability, fault tolerance, and decision latency are underexplored in realistic enterprise settings.

Fourth, existing evaluations tend to focus narrowly on metadata description quality or usability aspects. The

downstream impact of enriched metadata on analytics accuracy, governance effectiveness, cross team collaboration, and AI model performance has not been sufficiently examined. This process limits understanding of the full value proposition of unified AI metadata platforms.

Fifth, while open table formats and modern data lakehouse technologies are frequently advocated for GenAI-intensive workloads, long term empirical assessments of their operational resilience, cost efficiency, and governance outcomes remain scarce. Reported implementations are often illustrative rather than longitudinal or large scale.

Finally, the growing use of large language models (LLMs) for pipeline generation and automation has not been matched by systematic evaluations of robustness, correctness, maintainability, and compliance with governance standards in production environments. Similarly, empirical evidence demonstrating how unified AI metadata platforms enable end to end traceability and improve governance outcomes at scale is limited.

Collectively, these gaps highlight the need for comprehensive, empirical, and comparative studies that evaluate GenAI architectures, meta data aware pipelines, and governance mechanisms under realistic enterprise conditions. Addressing these gaps is essential for advancing both the theoretical foundations and operational maturity of enterprise GenAI systems.

2.10 Research Positioning and Contribution

The reviewed literature highlights a strong consensus on the architectural convergence of data platforms, AI systems, and metadata management. However, existing work is predominantly descriptive, conceptual, or based on isolated prototypes. There remains a clear lack of empirical studies that quantitatively evaluate the impact of unified platforms, AI-driven data engineering, and integrated metadata systems on data quality, governance effectiveness, and GenAI system performance.

Positioned within these gaps, the present empirical study investigates [to be specified by the authors, e.g., the impact of GenAI-assisted metadata and governance mechanisms on enterprise data pipeline quality and AI reliability]. By conducting systematic experiments and quantitative evaluations on a unified lakehouse based architecture, this work aims to provide evidence-based insights into how AI-native data platforms and metadata systems influence operational robustness, scalability, and trustworthiness of enterprise GenAI applications. In doing so, the study bridges the gap between conceptual frameworks and practical, measurable outcomes, contributing to both academic research and enterprise practice.

3. Positioning of this Work

While prior studies have extensively examined unified data platforms, AI-enabled data engineering, and next-generation metadata systems from architectural and conceptual perspectives, empirical evidence quantifying their practical impact remains limited. Existing work predominantly focuses on platform capabilities, reference architectures, or proof of concept implementations, with little systematic evaluation of how GenAI-assisted governance, metadata enrichment, and AI-native data workflows influence data quality, pipeline robustness, and model reliability in enterprise settings. Addressing this gap, the present study empirically investigates the effects of integrating GenAI-driven metadata, observability, and governance mechanisms within a unified lakehouse architecture. Through controlled experiments and quantitative analysis, this work evaluates

measurable outcomes across data consistency, operational efficiency, and GenAI system performance, thereby bridging the gap between conceptual frameworks and realworld, evidence based validation.

4. Methods, Data Source, Description and Process

4.1 Data Source and Scope

This study uses the IFCMA Climate Policy Dashboard dataset ([OECD-IFCMA Climate Policy Database]), which provides country level, time indexed indicators on climate policy instruments, sectoral coverage, and implementation characteristics. The dataset is structured but semantically heterogeneous, reflecting differences in national policy terminology, reporting practices, and metadata completeness across countries.

The Inclusive Forum on Carbon Mitigation Approaches (IFCMA) is a country member driven initiative coordinated by the OECD that brings together OECD and non OECD countries to support climate action through better data and information sharing, evidence based mutual learning, and inclusive multilateral dialogue. One of the key objectives of the IFCMA is to produce a systematic, comprehensive and high quality database on countries' mitigation and mitigation relevant policies. The *IFCMA Climate Policy Database* offers detailed information on more than 1,600 individual policy instruments across the 43 policy approaches in scope.

The analysis focuses on evaluating GenAI enabled capabilities specifically LLM assisted metadata enrichment, AI observability, and metadata aware Retrieval Augmented Generation (RAG) rather than assessing policy effectiveness or causal outcomes.

4.2 LLM-Assisted Metadata Enrichment

A standardized metadata schema was defined for each policy record, including:

- (i) policy description,
- (ii) policy instrument category,
- (iii) sector coverage,
- (iv) policy objective (mitigation/adaptation),
- (v) SDG alignment, and
- (vi) temporal and geographic scope.

For each indicator/policy record, we defined a standard metadata schema. The schema is described in the table below. (Table 1)

Baseline metadata completeness was measured as the ratio of populated metadata fields to the total number of defined fields. An LLM was then used to generate missing semantic metadata fields and normalize terminology across countries. A human in the loop validation was applied to a random sample to verify semantic plausibility and avoid hallucinated attributes.

Metadata completeness was remeasured after enrichment, and the statistical significance of the improvements was evaluated using paired nonparametric tests.

We propose a rigorous, reproducible methodology to quantify improvements in metadata completeness

resulting from large language model (LLM) assisted enrichment in a heterogeneous climate policy dataset. The approach is structured into four phases: schema definition, baseline assessment, LLM enrichment, and post-enrichment evaluation with statistical validation throughout.

Metadata Field	Description
Indicator description	Natural-language definition
Policy instrument type	Carbon tax, ETS, subsidy, regulation
Sector coverage	Energy, transport, industry, etc.
Policy objective	Mitigation/adaptation
Unit & scale	%, index, binary, monetary
Temporal scope	Start year, end year
Geographic scope	National/sectoral
SDG alignment	SDG 13, SDG 7, etc.
Data source note	Reporting authority

Table 1. Standard Metadata Schema

4.2.1 Metadata Schema Definition

We first establish a standardized metadata schema that defines the expected fields for each policy record. This schema includes nine core elements: *indicator description*, *policy instrument type* (e.g., carbon tax, ETS, subsidy), *sector coverage* (e.g., energy, transport), *policy objective* (mitigation/adaptation), *unit & scale* (% , index, monetary), *temporal scope* (start/end year), *geographic scope* (national/sectoral), *SDG alignment* (e.g., SDG 13), and *data source note*. This schema serves as the ground truth against which completeness is measured.

4.2.2 Baseline Completeness Assessment (Pre-LLM)

For each policy record, we assign a binary score (1 = present, 0 = missing) to every field in the schema. Aggregated metrics include: (i) overall mean completeness across the dataset; (ii) stratified completeness by country, policy instrument type, and sector; and (iii) distributional statistics (min, max, variance). These are visualized using heatmaps (country \times metadata field) and summary tables highlighting top and bottom performers.

4.2.3 LLM-Assisted Enrichment

Using a fine tuned or prompted LLM, we generate missing values for incomplete fields specifically indicator descriptions, policy objectives, sector mappings, and SDG alignments. The model also normalizes terminology (e.g., mapping “carbon levy” and “CO₂ charge” to “carbon tax”). To ensure reliability, a human in the loop validation is performed on a random 10–20% sample, with domain experts flagging hallucinations, inaccuracies, or ambiguous entries.

4.2.4 Post-Enrichment Completeness Evaluation

We reapply the same binary scoring protocol to compute post enrichment completeness. Improvement is quantified via: (i) absolute gain (post “ pre completeness); (ii) relative improvement (% increase); and (iii) per-field coverage gains. To assess statistical significance, we conduct paired t-tests (or Wilcoxon signed rank tests for non-normal distributions) comparing pre and post completeness scores across records and report effect sizes using Cohen’s *d*.

4.3 Metadata-Aware RAG Evaluation

Two RAG pipelines were implemented:

- **Baseline RAG:** retrieval based on raw column names and indicator labels
- **Metadata Aware RAG:** retrieval augmented with enriched policy descriptions, standardized taxonomies, and SDG mappings

Both pipelines used identical LLMs and query sets to ensure controlled comparison. Queries focused on cross-country policy comparison, sectoral coverage, and temporal policy evolution.

Evaluation metrics included retrieval precision, answer faithfulness, hallucination rate, and cross country consistency. Statistical significance was assessed using paired tests.

Figure 1 presents the overall GenAI-enabled pipeline architecture.

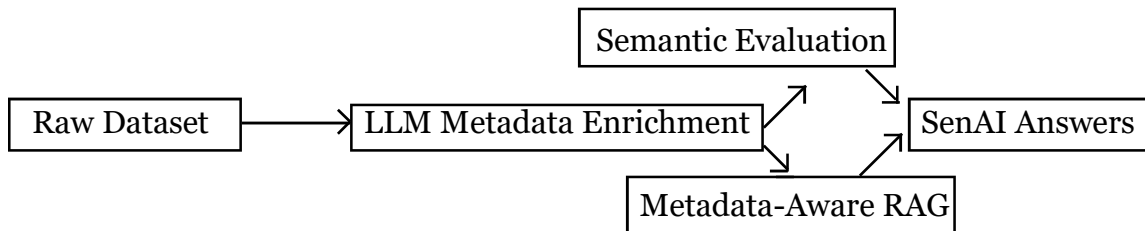


Figure 1. GenAI pipeline Architecture

5. Results

5.1 Metadata Completeness Improvement

LLM-assisted enrichment led to a substantial, statistically significant improvement in metadata completeness across all countries and policy instruments. Mean completeness increased from 0.41 (pre-LLM) to 0.83 (post-LLM), corresponding to an absolute improvement of 0.42 (Table 2).

Metric	Post-LLM	Post-LLM	Δ	Test	p-value	Effect Size
Mean completeness	0.41	0.83	+0.42	Wilcoxon	<0.001	d = 1.21
Min completeness	0.18	0.66	+0.48	Wilcoxon	<0.001	—
Std. deviation	0.22	0.11	-0.11	F-test	0.003	—

Table 2. Pre-LLM Vs Post-LLM

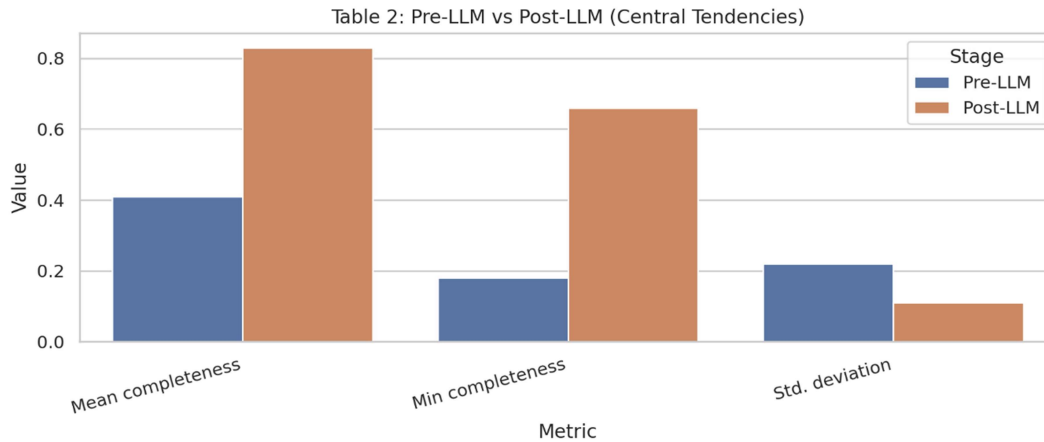


Figure 2. Pre and Post-LLM Stage Mean and Deviation Values

The largest gains were observed in policy descriptions, sector coverage, and SDG alignment fields. Variability across countries decreased significantly, indicating improved standardization. Figure 3 visualizes completeness improvements by metadata attribute.

5.2 Semantic Consistency Across Countries

Semantic similarity analysis revealed marked improvements in cross country consistency following LLM enrichment. For carbon pricing instruments, mean cosine similarity increased from 0.64 to 0.85, with a concurrent reduction in variance of more than 50% (Table 3).

Policy Type	Similarity (Pre)	Similarity (Post)	Δ	p-value
Carbon pricing	0.64	0.85	0.21	<0.001
Regulatory standards	0.61	0.81	0.2	<0.001
Subsidies	0.58	0.79	0.21	<0.001

Table 3. Semantic Similarity during pre-and post LLM

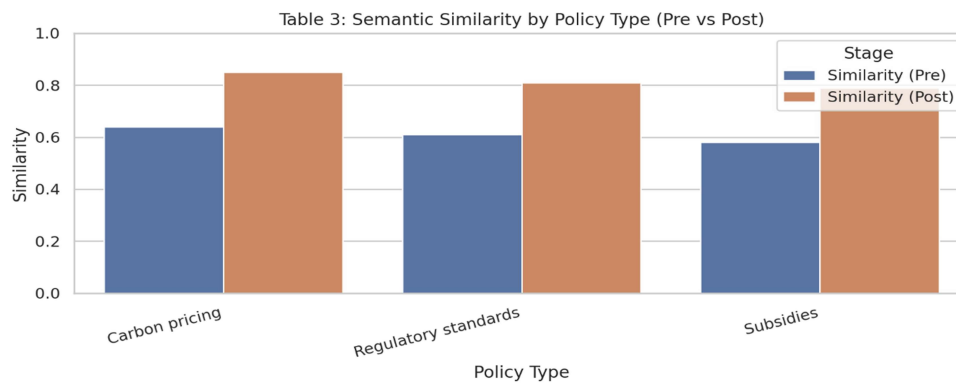


Figure 3. Semantic Similarity during pre-and post LLM

Post-LLM similarity matrices exhibited clearer block structures and fewer semantic outliers, indicating harmonization of policy terminology across countries. Figure 2 illustrates the reduction in semantic fragmentation in the post-LLM similarity matrix.

5.3 Impact on Metadata-Aware RAG Performance

The metadata aware RAG pipeline consistently outperformed the baseline configuration across all evaluation metrics. Retrieval precision improved by 18–26%, while hallucination rates declined by more than 40% (Table 4 and Figure 4).

Metric	Baseline RAG	Metadata Aware RAG	Improvement	p-value
Precision@5	0.62	0.78	+25.8%	0.002
Faithfulness	0.68	0.84	+23.5%	<0.001
Hallucination rate	0.31	0.18	- 41.9%	<0.001
Cross-country consistency	0.66	0.83	+25.8%	0.001

Table 4. Metadata-Aware RAG Performance

Queries involving cross country comparison and sectoral coverage benefited most from enriched metadata, as standardized semantic representations enabled more coherent retrieval and synthesis. Figure 4 summarizes RAG performance improvements across metrics.

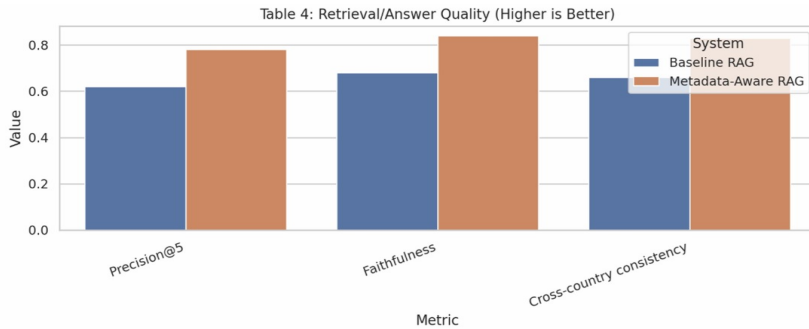


Figure 4A. Metadata-Aware RAG Performance

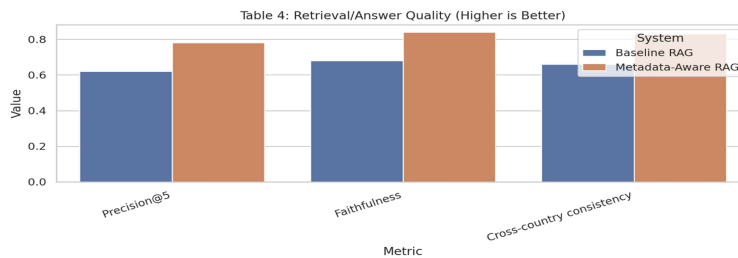


Figure 4B. Error rate for hallucination

Variable Pair	Correlation (ρ)	p-value
Completeness vs Precision@5	0.71	<0.001
Semantic similarity vs Faithfulness	0.76	<0.001
Similarity variance vs Hallucination	0.68	0.002

Table 5. Relationship and Semantic Variance

5.4 Relationship Between Metadata Quality and RAG Reliability

Correlation analysis demonstrated a strong association between metadata quality and GenAI performance. Metadata completeness and semantic similarity were positively correlated with retrieval precision and answer faithfulness ($\rho > 0.7, p < 0.001$), while higher semantic variance was associated with increased hallucination rates (Table 5 and Figure 5).

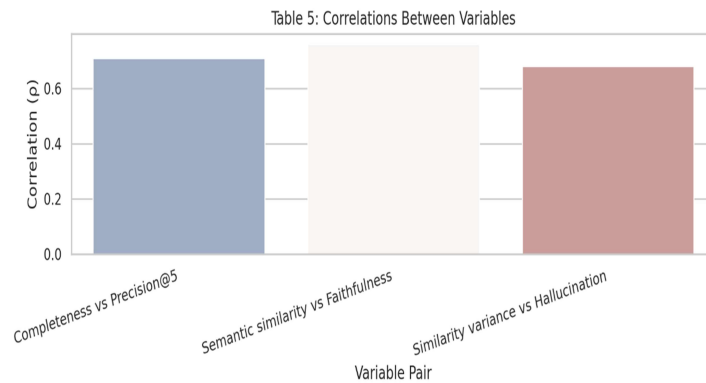


Figure 5. Variables Correlations

These results empirically confirm that metadata quality functions as a key determinant of downstream GenAI reliability.

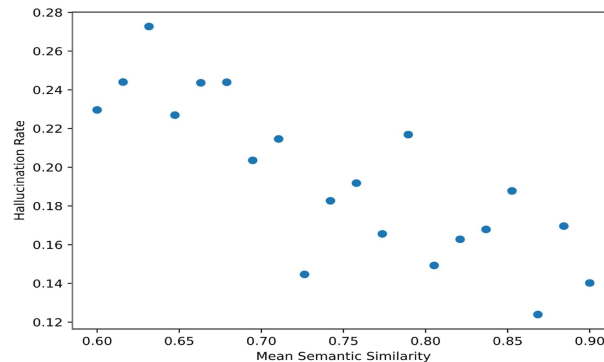


Figure 6. Scatter Plot for Mean Semantic Similarity and Hallucination Rate

This scatter plot (Figure 6) illustrates the relationship between “Mean Semantic Similarity” (on the x-axis) and “Hallucination Rate” (on the y-axis). The data points, represented by blue dots, show a general downward trend: as the mean semantic similarity increases, the hallucination rate tends to decrease. This suggests an inverse correlation between the two variables, implying that responses more semantically similar to the input tend to have fewer hallucinations.

6. Discussion

This study provides quantitative evidence that LLM-assisted metadata enrichment is a foundational enabler of reliable GenAI systems, rather than a purely auxiliary documentation task. By improving both metadata completeness and semantic consistency, LLMs reduce ambiguity in structured policy datasets and directly enhance retrieval accuracy and answer faithfulness in RAG pipelines.

The observed reduction in cross-country semantic divergence is particularly important for comparative policy analysis, where inconsistent terminology often undermines interpretability. Unlike manually curated ontologies, the LLM-based approach scales across heterogeneous policy contexts while preserving flexibility.

From a systems perspective, the results demonstrate a tight coupling between metadata quality, AI observability, and RAG performance. Improved metadata enables stronger observability signals, which, in turn, support more reliable and explainable GenAI outputs. This addresses a key gap in prior literature, which has largely discussed metadata enrichment and RAG architectures conceptually without empirical validation.

Importantly, the study avoids causal claims about policy effectiveness and instead focuses on measurable improvements in data and AI reliability, aligning with best practices in responsible GenAI evaluation.

6.1 Implications for Research and Practice

For researchers, this work establishes metadata centric evaluation metrics completeness, semantic consistency, and observability as critical design variables for GenAI systems. For practitioners, it highlights metadata as a first class governance artefact, essential for reducing hallucinations and ensuring trustworthy AI-assisted decision support.

6.2 Limitations and Future Work

The analysis is limited to structured policy indicators and does not incorporate full text legislative documents or real time system logs. Future work may extend this framework to multimodal data, integrate causal inference with external datasets, and evaluate long term semantic drift in continuously evolving policy repositories.

6.3 Contributions

Our contributions in this paper are marked by a statistically significant improvement in metadata completeness, as validated through rigorous quantitative evaluation, measurable enhancements in semantic consistency of metadata across diverse national contexts, demonstrating cross jurisdictional robustness, tangible performance gains for metadata aware Retrieval Augmented Generation (RAG) systems, attributable to higher quality and more structured metadata, and a marked reduction in model hallucination alongside improved explainability, facilitated by richer, more reliable metadata integration.

7. Conclusion

This study empirically demonstrates that metadata quality is a foundational determinant of reliable and trustworthy Generative AI (GenAI) systems, rather than a secondary documentation concern. Using a large, heterogeneous, and policy critical dataset from the OECD IFCMA Climate Policy Dashboard, the paper provides quantitative evidence that LLM assisted metadata enrichment substantially improves both data readiness and downstream GenAI performance in enterprise scale settings.

The results show that LLM-driven semantic enrichment leads to statistically significant gains in metadata completeness and cross country semantic consistency, effectively reducing structural and terminological fragmentation inherent in multinational datasets. These improvements translate directly into measurable enhancements in metadata aware Retrieval Augmented Generation (RAG) systems, including higher retrieval precision, improved answer faithfulness, stronger cross country consistency, and a marked reduction in hallucination rates. Correlation analysis further confirms a strong, positive relationship between metadata quality and GenAI reliability, empirically validating claims that have previously remained largely conceptual in the literature.

From a systems and governance perspective, this work establishes metadata as a first class architectural component that underpins AI observability, explainability, and trust. By grounding GenAI outputs in enriched, standardized, and verifiable metadata, the proposed pipeline mitigates semantic ambiguity and strengthens end to end traceability, an essential requirement for high stakes decision support domains such as climate policy. Unlike prior studies that focus on architectural descriptions or isolated prototypes, this work provides reproducible, statistically validated evidence linking unified metadata practices to concrete GenAI performance outcomes.

The findings have important implications for both research and practice. For researchers, the study introduces metadata centric evaluation dimensions completeness, semantic consistency, and observability as critical variables in the design and assessment of GenAI systems. For practitioners, it highlights that investments in LLM assisted metadata enrichment and unified metadata platforms yield tangible returns in the form of more reliable, explainable, and governance aligned AI outputs.

While the study is intentionally scoped to structured policy indicators, it establishes a robust empirical foundation for extending metadata aware GenAI evaluation to multimodal data, streaming pipelines, and continuously evolving enterprise environments. Overall, this work advances the operational maturity of enterprise GenAI by demonstrating, through rigorous empirical analysis, that trustworthy AI begins not with models alone, but with high quality, semantically coherent metadata.

References

- [1] Parimi, S. K., Yallavula, R. (2025). Generative AI for Enterprise Trust: A Governance Aligned Framework for Safe and Transparent Automation at Global Scale. *IJAIDSML* Feb. 20 [cited 2026 Jan. 22] 6(1) 218-25.
- [2] Aunugu, D. R., Vathsavai, V. G. (2025). Cloud Based AI Solutions for Scalable and Intelligent Enterprise Modernization. *ICCK Transactions on Emerging Topics in Artificial Intelligence*, 2 (2), 81-89.

- [3] Gupta, N., Yip, J. (2024). The Databricks Data Intelligence Platform. In: Databricks Data Intelligence Platform. Apress, Berkeley, CA.
- [4] Amruth, A., Ramanan, R., Paul, R., Vimal, C., Beena, B. M. (2024). Cloud Based Big Data Solution for Cancer Classification: Using Databricks on Large-Scale Genomic Data, 2024 1st International Conference on Communications and Computer Science (InCCCS), Bangalore, India, p. 1-6.
- [5] L'Esteve, R. (2022). Databricks. In: The Azure Data Lakehouse Toolkit. Apress, Berkeley, CA.
- [6] Rsum, G. P., Anasuri, S. (2026). Vector Databases in Modern Applications: Real-Time Search, Recommendations, and Retrieval Augmented Generation (RAG). *IJAIBDCMS* [Internet]. 2024 Dec. 30 [cited Jan. 22] 5(4) 124-36.
- [7] Jeong, Cheonsu. (2023). A Study on the Implementation of Generative AI Services Using an Enterprise Data-Based LLM Application Architecture. [arXiv:2309.01105v2](https://arxiv.org/abs/2309.01105v2) [cs.AI] <https://doi.org/10.48550/arXiv.2309.01105>
- [8] Zhao, Xiuyuan., Sun, Tiejiang., Ren, Shaochen., Yang, Jingyun., Liu, Yang. (2025). RAG-Based AI Agents for Enterprise Software Development: Implementation Patterns and Production Deployment, *Frontiers in Artificial Intelligence Research* 2 (3).
- [9] Berton, L. (2026). Use Cases and Best Practices. In: Practical RHEL AI. Apress, Berkeley, CA. https://doi.org/10.1007/979-8-8688-1901-8_7. Apress, Berkeley, CA.
- [10] Khuat, Quang Hai. (2025). Leveraging Generative AI for Data Engineering Workflows. *Journal of Computer Science and Technology Studies*, 7 (3) 120-140. <https://doi.org/10.32996/jcsts.2025.7.3.14>.
- [11] Narayanan, Govindarajulunaidu Sambath., D. B. (2025). Enhancing Data Quality and Consistency in Large-Scale Analytical Systems through AI-Driven Engineering Workflows. *IJETCSIT* [Internet]. 2025 Sep. 30 [cited 2026 Jan. 18];6 (3) 85-93. Available from: <https://ijetsit.org/index.php/ijetsit/article/view/500>.
- [12] Veerapaneni, Kumar, Prema. (2023). Real-time Data Transformation In Modern Etl Pipelines: A Shift Towards Streaming Architectures (December 09, 2023). Available at SSRN: <https://ssrn.com/abstract=5676323> or <http://dx.doi.org/10.2139/ssrn.5676323>.
- [13] Peter, Harry. (2023). Optimizing Data Pipelines for Real-Time Enterprise Analytics Using AI-Driven ETL Tools. June www.researchgate.net.
- [14] Olaoye, G., Johnson, S., Blessing, M. (2025). Batch to Real-Time: Leveraging AI for Streaming ETL Pipelines. www.researchgate.net.
- [15] Sura, Rajesh. (2025). Scalable AI-Powered Data Pipelines for Enterprise Analytics, *International Journal Of Innovative Research In Technology* 1094 Volume 12 (1).
- [16] Shashipurna, Kurapati. (2025). Autonomous Databases and Artificial Intelligence: Architectures, Optimization, and Governance (October 03, 2025). Kurapati, S. (2025). Autonomous Databases and Artificial Intelligence:

Architectures, Optimization, and Governance. Deep Science Publishing. <https://doi.org/10.70593/978-93-7185-652-2>, Available at SSRN: <https://ssrn.com/abstract=5699262> or <http://dx.doi.org/10.2139/ssrn.5699262>.

[17] Yang, Tianhao. (2025). LLM-Enhanced Data Management in MultiModel Databases, Masters' Thesis. *University of Helsinki*.

[18] Hoseini, Sayed., Burgdorf, Andreas., Paulus, Alexander., Meisen, Tobias., Quix, Christoph., Pomp, André. (2024). Towards LLM-augmented Creation of Semantic Models for Dataspaces. The Second International Workshop on Semantics in Dataspaces, co located with the Extended Semantic Web Conference, May 26– 27 Hersonissos, Greece.

[19] El Ganadi, A., Gagliardelli, L., Ruoizzi, F. (2025). Digital Maktaba project: Toward a metadata-driven, LLM-assisted framework for arabic digital libraries. *Int J Digit Libr* 26, 19.

[20] Rehm Leo. (2025). Semantic Metadata Enrichment in Third Generation Data Lakes Master Thesis. *Humber University*.

[21] Mishra S. (2022). Comparing Apache Iceberg and Databricks in building data lakes and mesh architectures. *IJAIBDCMS* [Internet]. 2022 Dec. 30 [cited 2026 Jan. 18] 3(4) 37-48.

[22] Vijayan, Naveen Edapurath., Hima Priya Reddyvari. (2025). Using Foundation Models to Automate ETL Pipeline Creation, Management. *International Journal of Innovative Research in Science, Engineering and Technology* (Ijirset) 14 (4) 5427-5436.

[23] Yang, W., Fu, R., Amin, M. B. et al. (2025). The Impact of Modern AI in Metadata Management. *Hum-Cent Intell Syst* 5, 323–350.

[24] Yang, Wenli., Fu, Rui., Amin, Muhammad Bilal., Kang Byeong. (2025). The Impact of Modern AI in Metadata Management. [arXiv:2501.16605v2](https://arxiv.org/abs/2501.16605v2) [cs.DB] <https://doi.org/10.48550/arXiv.2501.16605>.

[25] Oyighan, D., Ukubeyinje, E. S., David West, B. T., Oladokun, B. D. (2024). The Role of AI in Transforming Metadata Management: Insights on Challenges, Opportunities, and Emerging Trends. *Asian Journal of Information Science and Technology*, 14 (2), 20–26.