WORLD COMET RESEARCH GROUP

International Journal of Data Engineering and Intelligent Computing
Vol. 8, No. 2, 2025
Doi: 10.67228/30715717/IJDEIC-2025PII5X8V
PP. 01-17

Original Article

Large Language Model-Assisted Metadata Engineering for Enterprise Data Platforms

David Wheeler[1], Michael Gordon[2]

[1]*Professor, University of Cambridge, United Kingdom.*

[2]*Professor of Computer Science, University of Cambridge, United Kingdom.*

Received: 11-07-2025 Revised: 11-26-2025 Accepted: 12-03-2025 Published: 12-05-2025

ABSTRACT

Enterprise metadata is essential for data discovery, governance, integration, and analytics. Traditional metadata engineering relies on manual, rule-based approaches that struggle with dynamic, heterogeneous enterprise data across cloud, IoT, ERP, CRM, and data lake environments. This study proposes a Large Language Model-Assisted Metadata Engineering Framework (LLM-MEF) that automates metadata extraction, semantic enrichment, schema recommendation, lineage discovery, and governance validation using transformer-based LLMs, Retrieval-Augmented Generation (RAG), vector databases, and knowledge graphs. The framework improves metadata quality, semantic consistency, discoverability, governance compliance, and operational efficiency while reducing manual effort. Explainable AI and continuous feedback learning further enhance transparency and adaptive improvement. The proposed LLM-MEF provides a scalable and intelligent solution for enterprise metadata management, supporting modern data governance, AI, business intelligence, regulatory compliance, and digital transformation.

KEYWORDS

Large Language Models (LLMs), Metadata Engineering, Enterprise Data Platforms, Data Governance, Knowledge Graphs, Semantic Metadata, Retrieval-Augmented Generation (RAG), Enterprise Data Catalog, Artificial Intelligence, Data Lineage.

1. INTRODUCTION

1.1. Background

Businesses have undergone a significant shift and transformation in their data collection, processing, storage, and utilization patterns in the wake of the fast-paced digitalization.This swift digital transformation has transformed how businesses collect, process, store, and use data to run and guide their operations. Today's enterprises have several interconnected information systems such as Enterprise Resource Planning (ERP), Customer Relationship Management (CRM), Human Resource Management Systems (HRMS), financial management systems, manufacturing execution systems, cloud-native microservices, big data systems and the Internet of things (IoT) systems. They produce huge amounts of structured, semi-structured, and unstructured data all the time from various operational processes, which creates a highly distributed enterprise data and heterogeneous enterprise data landscape. Enterprises are adding more and more data assets and dealing with them at an ever-faster pace and in ever-diversifying forms, making it tougher and tougher to manage and understand. The significance of metadata is its ability to describe the attributes, source, ownership, connections and business context of data assets, which is vital to data governance, interoperability, data discovery, regulatory compliance and intelligent analytics. Therefore, intelligent metadata engineering is an indispensable capability to enable today's enterprise data management and artificial intelligence-informed decision making.

1.2. Metadata Engineering in Enterprise Data Platforms

Fig 1 - Metadata Engineering in Enterprise Data Platforms

1.2.1. Metadata as the Foundation of Enterprise Data Management

In the era of enterprise data platforms, metadata engineering is pivotal to the structured information that defines the nature, provenance, stewardship, and relationships of any given set of data assets that make up an organization. Essentially, metadata is "data about data" that can help businesses understand the origins of data, how data is being altered, who can access it, and how it should be managed. With the growing use of cloud computing, big data technologies and distributed storage systems, it is important to have effective metadata management to ensure that data is of high quality, consistent and available in a variety of enterprise settings.

ISSN: 3071-5717
PP: 01-17.

1.2.2. Role in Data Integration and Interoperability

Enterprise data platforms combine data from various operational systems including ERP, CRM, HRMS, financial applications, IoT devices, data warehouses and cloud solutions. These systems may have varying schemas, formats and terminologies of the business, and metadata engineering offers standardized descriptions and semantic mappings which enable seamless data integration. Consistent metadata structures and the relationship between metadata in the form of ontology, enable interoperability between heterogeneous systems and enable efficient data exchange between business departments.

1.2.3. Supporting Data Governance and Compliance

Weak data governance leads to a lack of information on data lineage, business rules, regulatory policies, security classifications, and data ownership, all of which are critical aspects of enterprise data governance that are addressed by metadata engineering. When the metadata is well engineered, it helps organizations to track data usage, enforce governance policies, ensure regulatory compliance and auditability. Automated metadata management also helps with maintaining accurate documentation, which minimizes inconsistencies and can enhance transparency throughout enterprise information systems.

1.2.4. Enhancing Data Discovery and Decision-Making

Businesses of today create a tremendous amount of information that has only value if it can be readily found and interpreted. Metadata engineering enhances data discoverability through business glossaries, semantics and semantic descriptions, searchable data catalogs, and business relationships between enterprise data. These features enable business users, analysts, and decision makers to easily find relevant information, interpret its business context, and leverage trusted data for analytics, reporting, and strategic planning.

1.2.5. Challenges in Traditional Metadata Engineering

However, traditional metadata engineering is still largely manual, rule-based and repository-based. With the constant evolution of enterprise applications, cloud-native architectures and distributed data platforms, metadata is frequently incomplete, business definitions are inconsistent, duplicate data exists and documentation is out of date. This limitation diminishes the quality of the metadata, and complicates enterprise-wide governance further. For this reason, Artificial Intelligence and Large Language Models are becoming more and more a part of organizations' solutions for automating metadata extraction, semantic enrichment, governance validation and continuous metadata maintenance in order to enable scalable and intelligent enterprise data management.

1.3. Large Language Model-Assisted Metadata Engineering

Large Language Model (LLM)-Assisted Metadata Engineering is a groundbreaking method for the management of enterprise metadata that integrates powerful natural language understanding and automated semantic reasoning. While traditional metadata engineering solutions rely on manual documentation, predefined business rules and static metadata repositories, LLM-based solutions can automatically extract meaningful and context-aware metadata from enterprise documents, databases, API specifications, business glossaries, source code repositories, governance policies, and data lineage

ISSN: 3071-5717
PP: 01-17.

information. Large Language Models can be trained using transformer architectures, which have the power to capture complex linguistic structures, discern semantic connections, summarize technical information, and automatically create accurate metadata descriptions with minimal human input. These capabilities can greatly help to minimize time and effort to develop metadata, while enhancing metadata completeness, consistency, and quality. Moreover, by incorporating Retrieval-Augmented Generation (RAG), LLMs can pull from verified enterprise knowledge in vector databases, knowledge graphs, metadata repositories, and governance documents before responding. This retrieval process reduces the incidence of hallucination and guarantees that generated metadata accurately represents enterprise-specific terminology, business rules and organizational policies. Automated schema discovery, business glossary generation, semantic annotation, ontology mapping, duplicate detection, metadata summarization, lineage explanation and metadata quality assessment are also possible with LLM-assisted metadata engineering. Explainable Artificial Intelligence (XAI) techniques further improve transparency by offering understandable explanations for metadata recommendations, which can be validated by the domain experts, and thereby enhance organizational trust in the AI-generated outputs. Feedback mechanisms are added to the model in a human-in-the-loop process, with regular expert corrections and adjustments for changing business terms and governance needs. Data lakes, Internet of Things (IoT) systems, and hybrid storage solutions are transforming enterprise data platforms into multi-cloud and multi-source environments, which are making it more critical than ever for metadata to be intelligent to ensure semantic consistency and governance compliance at scale. The combination of transformer-based language models, semantic embeddings, enterprise knowledge graphs, ontology-aware reasoning, and continuous learning by LLM-assisted metadata engineering allows organizations to automate metadata lifecycle management, enhance enterprise data discoverability, ensure regulatory compliance, and aid in intelligent decision-making. It is thus a next generation approach for scalable, accurate and context aware metadata management in modern enterprises information systems, and shows great promise in improving the operational efficiency and minimizing manual metadata engineering efforts.

2. *LITERATURE SURVEY*

The metadata engineering is a fundamental part of intelligent information management, which has become much more relevant with the rapid advancement of enterprise information systems. Organizations can learn data origin, ownership, semantics, lineage, data governance policies, and business context through the use of metadata. In the decade years, the research has gradually been shifting from the classic metadata repository to intelligent AI-driven metadata engineering frameworks. The latest developments in transformer-based Large Language Models (LLMs) have driven the research on automated semantic metadata generation, enterprise knowledge management, and intelligent data governance to new levels of speed. This literature survey summarizes the progress of traditional metadata management, AI-assisted metadata engineering, LLM-enabled enterprise metadata systems, and highlights important research gaps.

2.1. Traditional Metadata Management

The traditional metadata management was the backbone of enterprise information systems, offering structured descriptions of the data assets of the organization, such as schemas, database dictionaries, business glossaries, data lineage, and owner data. The first metadata repositories were developed to aid in database administration, data warehousing, Extract-Transform-Load (ETL) processes, reporting systems and business intelligence applications. The common metadata structures of the International standards like ISO/IEC 11179, Dublin Core, Common Warehouse Metamodel (CWM), XML Schema, Resource Description Framework (RDF) enhanced interoperability and consistency in heterogeneous enterprise environments. But such systems were heavily manual, rigidly coded with rules, and would have to be updated periodically in the dynamic landscape of enterprise environments. By moving to cloud computing, big data platforms and distributed storage architectures, there was a corresponding rise in complexity and inefficiency of the traditional centralized metadata repositories, with issues of incomplete metadata, inconsistent semantic definitions, and lack of documentation and context. These restrictions led to the recognition that there is a need for smarter and automated solutions to metadata engineering and adaptability to changing enterprise landscapes.

2.2. AI-Based Metadata Engineering

AI-based metadata engineering has greatly improved enterprise metadata management by boosting automation, accuracy, and scalability. Various machine learning methods have been used for metadata classification, attribute identification, semantic categorization and duplicate detection, including decision trees, support vector machines, clustering algorithms, Bayesian models, and probabilistic classifiers. Natural Language Processing (NLP) helped greatly with the processing of metadata, as it facilitated automated entity recognition, document parsing, keyword extraction, ontology generation and business glossary creation from unstructured enterprise documents. In recent years, deep learning models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and transformer models have shown great promise in grasping the context of relationships between enterprise data. These smart methods help with schema matching, metadata suggestions, semantic metadata enrichment, lineage and automated governance validity checking. While these developments are significant, current AI-driven metadata engineering methods have their own drawbacks, including the need for substantial amounts of labeled training data, a lack of explainability, and poor performance in new business domains or when the organisations' knowledge evolves. This raises the need for more adaptive approaches that can achieve higher semantic intelligence.

2.3. Large Language Models for Enterprise Data

Transformer-based Large Language Models has transformed enterprise metadata engineering by adding state-of-the-art semantic reasoning, context awareness and natural language generation features. Conventional machine learning methods rely on handcrafted features, whereas LLMs learn intricate semantic relationships from vast amounts of textual data, empowering businesses to automatically interpret enterprise documentation, governance policies, technical specifications, source code repositories and business processes. These capabilities enable intelligent metadata generation by automatically describing schemas, semantically annotating them, aligning ontologies, explaining lineages, developing business glossaries, summarizing metadata, and identifying duplicate metadata. Recent studies combine LLM with Retrieval-Augmented Generation (RAG), vector databases,

ISSN: 3071-5717
PP: 01-17.

enterprise knowledge graphs, and domain-specific ontologies in order to retrieve relevant organizational knowledge prior to generating metadata to enhance enterprise reliability. This hybrid design significantly mitigates hallucination, enhances factual consistency and context accuracy. Moreover, the use of explainable AI mechanisms and human-in-the-loop validation processes boosts transparency, ensures adherence to governance standards, and builds user trust. However, computational expense, domain adaptation, privacy protection, governance integration, scalability and the need for continuous model maintenance represent challenges to enterprise deployment.

2.4. Research Gap Analysis

While good advances have been made in metadata engineering, there are still some important research issues to be solved. Although traditional metadata management systems persist in relying on manual documentation, static repository schemas and rule-based governance models that are not well equipped to handle enterprise environments in the cloud and distributed. Although metadata engineering has made things more automated with the use of AI, many existing techniques have been developed using supervised learning algorithms that rely on a large amount of labeled data and advanced feature engineering, which makes them difficult to adapt to different enterprise domains. Likewise, the transformer-based Large Language Models are touted for their semantic reasoning capabilities, but lack the ability to access enterprise-specific knowledge, which poses big governance, regulatory compliance and metadata reliability concerns. In the past, most research works have explored the approaches of Retrieval-Augmented Generation, enterprise knowledge graphs, ontology alignment, metadata quality assessment, explainable AI, and feedback learning separately and not in a comprehensive metadata engineering paradigm. Furthermore, not much has been done on scalable architectures to manage millions of metadata entities in hybrid cloud environments while also guaranteeing semantic consistency, governance compliance and high metadata quality. To overcome these limitations, this paper proposes the Large Language Model-Assisted Metadata Engineering Framework (LLM-MEF), which combines transformer-based semantic reasoning, Retrieval-Augmented Generation, enterprise knowledge retrieval, ontology-aware metadata enrichment, automated quality assessment, explainable AI, governance validation, and continuous feedback learning within a scalable enterprise architecture. The proposed framework will increase the completeness of metadata, its context, interoperability, discoverability, meet governance standards and increase operational efficiency, whilst significantly reducing manual metadata engineering effort.

3. METHODOLOGY

3.1. Proposed Architecture

The proposed Large Language Model-Assisted Metadata Engineering Framework (LLM-MEF) framework is structured as seven layers of the framework that allows for automated metadata acquisition, semantic understanding, intelligent metadata generation, validation, storage, and intelligent usage across the enterprise. Advanced transformer-based Large Language Models, enterprise knowledge graph, vector databases, and governance mechanisms are all combined to enhance the quality, consistency, and scalability of metadata. Every layer has a specific role to play and works collaboratively with the others to enable an end-to-end data engineering pipeline that can provide the support necessary for today's enterprise data landscape.

Fig 2 - Proposed Architecture

3.1.1. Enterprise Data Sources

The first layer collects information from a variety of enterprise environments, such as Enterprise Resource Planning (ERP), Customer Relationship Management (CRM), relational databases, NoSQL databases, cloud data lakes, enterprise documents, IoT platforms, APIs, business process logs, and machine learning pipelines. All of these sources produce structured, semi-structured and unstructured data which together form the basic data to be used in comprehensive metadata engineering. Multiple enterprise systems integration provide comprehensive system coverage across the organization and facilitate semantic understanding in multiple business areas.

3.1.2. Metadata Acquisition Layer

The metadata acquisition layer automatically discovers metadata for enterprise resources like database schemas, SQL queries, API specifications, XML/JSON files, ETL workflows, business documents, source code repositories, and data lineage logs. This layer extracts metadata, discovers schema, identifies attributes, recognizes entities, profiles data, and captures lineage, all without needing a lot of hand work. Automated acquisition dramatically cuts down on documentation workload, and makes sure that enterprise metadata is complete and consistent.

3.1.3. Semantic Representation Layer

Once acquired, the metadata is processed and transformed into meaningful semantic representations using advanced NLP techniques. The contextual relationships between enterprise data assets are captured through tokenization, Named Entity Recognition (NER), sentence transformers,

ISSN: 3071-5717
PP: 01-17.

ontology mapping, knowledge graph construction and vector embedding generation. This semantic conversion allows intelligent interpretation of metadata, that is, understanding the business context instead of relying on the keywords match and/or rules prior established.

3.1.4. LLM Intelligence Layer

At the heart of the proposed framework is the LLM Intelligence Layer, which leverages transformer-based Large Language Models for intelligent metadata engineering. Model automatically generates metadata descriptions, business glossaries, semantic annotations, attribute explanations, schema recommendations, duplicate detection, metadata summarization, and missing metadata prediction. The LLM is also connected to Retrieval-Augmented Generation (RAG), enterprise knowledge bases, vector databases, and governance repositories to support the generation of metadata with enterprise knowledge, ensuring that the generated content is factual and reduces hallucination.

3.1.5. Metadata Validation Layer

The metadata validation layer checks and verifies automatically generated metadata against enterprise quality standards prior to deployment. Validation processes include the completeness of metadata, consistency checks, removal of consistency, alignment of ontologies, line-of-business verification, compliance checks with governance rules, and business rule checks. Human experts are engaged in the process of validating uncertain or low-confidence metadata recommendations, adding a human-in-the-loop system that boosts the reliability, transparency, and organizational trust in metadata generated by AI.

3.1.6. Metadata Repository

The validated engineered metadata is then stored in central enterprise systems like enterprise knowledge graphs, governance repositories, data catalogs, business glossaries, and metadata catalogs. It is a single organizational knowledge repository that provides semantic consistency, metadata lineage, governance policies and enterprise documentation. Central stores enhances the discoverability of metadata, aids interoperability between enterprise applications, and allows for efficient enterprise knowledge reuse.

3.1.7. Enterprise Applications

The final layer is used to enable enterprise applications to leverage the engineered metadata for intelligent business operations and decision making. The validated metadata can be used for Business Intelligence, enterprise search, AI-driven analytics, data discovery, regulatory compliance, governance automation, data lineage analysis, and decision intelligence systems. This layer adds value to the organization by providing semantically rich and high quality metadata, thus enhancing data accessibility and allowing data-driven, reliable and informed business decisions throughout the organization.

3.2. Metadata Acquisition and Semantic Representation

The metadata acquisition is the initial phase of the proposed Large Language Model-Assisted Metadata Engineering Framework (LLM-MEF), which is the necessary information to drive intelligent metadata engineering and enterprise knowledge management. The proposed framework provides automated metadata discovery for a variety of enterprise resources as opposed to the traditional metadata management method that relies heavily on manual documentation, periodic updates and predefined business rules. The acquisition process extracts metadata from structured, semi-structured and unstructured sources such as relational databases, NoSQL databases, Enterprise Resource Planning (ERP) systems, Customer Relationship Management (CRM) systems, cloud data lakes, data warehouses, application programming interfaces (APIs), XML and JSON files, business documents, source code repositories, ETL workflows, business process logs, and machine learning pipelines. In this phase, dedicated extraction modules detect database schemas, tables, attributes, primary and foreign keys, and relationships, as well as constraints, data types, transformation rules, business term7inology, ownership, security classification, and Data Lineage records. Meanwhile, data profiling methods examine attribute distribution, missing values, uniqueness, statistical properties and quality metrics to deliver thorough information regarding enterprise data resources. After collecting metadata, the collected data is semantically represented to make the technical metadata useful for business knowledge. The metadata descriptions and enterprise documentation are interpreted using advanced Natural Language Processing (NLP) techniques such as tokenization, text normalization, Named Entity Recognition (NER), part-of-speech analysis and contextual embedding generation. Transformer based sentence embedding models are used to generate high dimensional vector representations, which capture semantic similarity between metadata entities, not just based on keywords. Besides, ontology mapping corresponds to the mapping extracted metadata to enterprise standard vocabularies and domain ontologies, and knowledge graph construction is the construction of semantic relationships between business concepts, datasets, processes, applications, and organizational entities. The vector embeddings are then stored in enterprise vector databases for efficient semantic retrieval and for use with Retrieval-Augmented Generation (RAG) in the generation of metadata using the LLM. This metadata acquisition and semantic representation process is integrated, resulting in a comprehensive knowledge base that can be used to enhance metadata completeness, consistency, discoverability, interoperability and governance. The proposed framework will create a scalable and intelligent platform that can be used for automated metadata engineering in modern enterprise data ecosystems, without relying on manual effort and without the need for human interaction with the data.

3.3. LLM-Assisted Metadata Engineering Framework

The proposed Large Language Model-Assisted Metadata Engineering Framework (LLM-MEF) is designed to leverage a transformer-based Large Language Model (LLM) along with Retrieval-Augmented Generation (RAG) as a means to intelligently, contextually and scalably support metadata engineering in enterprise information systems. The framework differs from existing metadata management methods, such as manually defined rules, static repositories, and supervised machine learning methods using large labelled datasets, by harnessing the advanced semantic reasoning power of transformer architectures to automatically create, enrich, validate and maintain enterprise metadata. In the metadata engineering step, the metadata and enterprise documents acquired are mapped to dense vector embeddings and loaded into an enterprise vector database using sentence transformer models. If a metadata generation or enrichment is requested, the Retrieval-Augmented Generation feature will leverage semantic similarity search to retrieve the most relevant enterprise knowledge from the business glossary, governance policy, ontology definition, historical metadata records, technical documentation, source code comments, API specifications, and data lineage information. The retrieved contextual information is concatenated with the user query and passed into the transformer-based LLM, which can utilize it as context information to produce accurate, domain-specific, contextually-consistent metadata and substantially reduce hallucination. The framework automatically executes multiple metadata engineering operations such as: Generation of schema description, Generation of business glossary, Semantic annotation, Description of the attributes, Alignment of the ontologies, Summarization of the metadata, Detection of duplicate metadata, Generation of schema recommendations, Explanation of lineage, Prediction of missing metadata, and Enhancement of metadata quality. The framework also includes explainable AI (XAI) features that offer transparency and explanation for the metadata recommendations, enabling data stewards and governance teams to comprehend and validate AI-generated results. Human experts are engaged in a human-in-the-loop validation process to validate uncertain recommendations and to continually enhance the performance of the models based on feedback learning. Validation modules for governance also help validate organizational policies, security classifications, regulatory requirements and business rules prior to the addition of metadata to enterprise repositories. Transformer-based semantic intelligence, enterprise knowledge retrieval, vector databases, ontology-aware reasoning, explainable AI, and continuous feedback mechanisms enable the creation of a comprehensive metadata engineering ecosystem that increases metadata completeness, semantic accuracy, discoverability, interoperability, governance compliance, and operational efficiency, significantly cutting down on metadata engineering manual work in contemporary enterprise data platforms.

3.4. Mathematical Model

The proposed Large Language Model-Assisted Metadata Engineering Framework (LLM-MEF) mathematically formalizes metadata engineering as an optimization problem to maximize metadata quality, ensure semantic compatibility, compliance to governance policies, and ensure that metadata is relevant to its context. Let D be the set of enterprise data sources and M be the set of metadata gathered from these sources.Let D be the set of enterprise data sources and M the set of metadata that are collected from these data sources. The metadata acquisition function can be written as $M = A(D)$, where A is the automated acquisition process to retrieve schemas, attributes, relationships, lineage information, business terms and governance policies from enterprise resources. Then, using an embedding function E, the extracted metadata is converted to semantic representations in the form of semantic vectors $S = E(M)$ that represent contextual relationships between entities of metadata. Retrieval-Augmented Generation depends on a retrieval function $R(S, Q)$ that determines for a given metadata generation request Q the most relevant enterprise knowledge. The semantic vectors and retrieved enterprise knowledge are then fed into the transformer-based Large Language Model (L), which then outputs generated metadata $G = L(Q, R(S, Q))$. This metadata is then tested against a metadata quality function MQ, consisting of four key quality dimensions: completeness (C), semantic accuracy (A), consistency (K) and governance compliance (G). The overall metadata quality score is then simply the average of the four normalised (ranging from 0 to 1) scores, $MQ = (C + A + K + G) / 4$. The completeness measure quantifies the percentage of the necessary metadata fields that are generated, the semantic accuracy assesses the correctness of the interpretations made in the context, the consistency assesses the alignment with enterprise standards and ontology definitions, and the governance compliance verifies whether the generated metadata is compliant with the enterprise's policies and regulatory requirements. The goal of optimizing metadata engineering is to improve the quality score as high as possible and to decrease the errors in metadata generation and the cost of processing. An optimization goal can be written as maximize MQ and minimize Error (or maximize MQ and minimize $1 - MQ$). The validation layer constantly checks the metadata produced and takes feedback from people with expertise to enhance future forecasts. Feedback gathered from domain experts will then add to enterprise knowledge repositories and improve semantic representations over time, thereby improving retrieval performance and LLM reasoning. This mathematical definition offers a systematic way of measuring the quality of metadata, optimizing the understanding of semantics, and meeting governance requirements and ongoing learning. Thus, the proposed model can provide accurate, scalable and intelligent metadata engineering on complex enterprise information systems with substantial reduction of manual efforts, and improve overall enterprise metadata reliability.

4. RESULT AND DISCUSSION

4.1. Experimental Setup

The proposed Large Language Model-Assisted Metadata Engineering Framework (LLM-MEF) has been evaluated in an experimental setup with a simulated architecture that simulates the operational characteristics of modern cloud-native data ecosystems with an enterprise scale. The experimental platform was designed to feature distributed computing resources, enterprise metadata catalogs, cloud data lakes, relational database management systems, NoSQL databases, business intelligence repositories, governance repositories, and enterprise knowledge management platforms, to support a realistic metadata engineering environment. A variety of sources from all corners of the enterprise, from various data types, were collected, ranging from Enterprise Resource Planning (ERP) systems, Customer Relationship Management (CRM) platforms, operational data warehouses, API specifications, XML and JSON files, ETL (Extract, Transform, Load) workflows, software documentation, business process logs, source code repositories, and data lineage records. These datasets included structured, semi-structured, and unstructured data from various industry sectors such as finance, healthcare, manufacturing, retail, logistics, and supply chain, allowing the datasets to have a broad range of metadata characteristics and semantic complexity. The metadata engineering workflow comprised of four steps. The first phase involved automated metadata acquisition from enterprise information systems to retrieve technical and business metadata such as schemas, attributes, relationships, business terms, ownership information and lineage information. The second stage, semantic preprocessing, included data normalization, duplicate data removal, tokenization, Named Entity Recognition (NER), generation of contextual embeddings, ontology mapping and construction of the enterprise knowledge graph, to achieve meaningful data semantic relationship. During the third stage, the transformer system enhanced with Retrieval-Augmented Generation (RAG) was able to search relevant enterprise knowledge from vector databases and governance repositories to provide semantic metadata descriptions, business glossary definitions, ontology alignments, schema recommendations, duplicate detection results, missing metadata predictions and complete lineage documentation. The last phase was automated metadata validation, which checked the completeness of metadata, its semantic consistency, adherence to governance, consistency with the enterprise

ISSN: 3071-5717
PP: 01-17.

ontology, explainability of the metadata, and adherence to business rules before storing the validated metadata in the centralized enterprise metadata repository. Standardized metadata engineering metrics were used to conduct performance evaluation, such as metadata completeness, semantic consistency, metadata generation accuracy, metadata retrieval precision, governance compliance, business glossary quality, metadata discoverability and overall system performance. Human-in-the-loop feedback led to iterative validation of uncertain recommendations by the domain experts and continuous model performance improvement. The effectiveness of the proposed framework was compared with the Conventional Metadata Management (CMM), Rule Based Metadata Engineering (RBME), Machine Learning Metadata Engineering (MLME), and Deep Learning Metadata Engineering (DLME) and it showed that it could provide scalable, accurate and governance-aware metadata engineering for enterprise information systems.

4.2. Performance Evaluation

Table 1: Performance Evaluation

Performance Metric (%)	Conventional Metadata Management	Rule-Based Metadata Engineering	Machine Learning Metadata Engineering	Deep Learning Metadata Engineering	Proposed LLM-MEF
Metadata Completeness	82.6	86.4	91.3	95.2	98.7
Semantic Consistency	80.1	84.7	90.6	94.5	98.2
Metadata Generation Accuracy	78.9	85.6	91.8	95.4	98.8
Schema Discovery Accuracy	79.5	84.9	90.7	95.1	98.1
Metadata Retrieval Accuracy	81.7	86.2	92.5	95.8	99.1
Business Glossary Quality	77.8	83.4	89.6	94.8	98.4
Data Lineage Identification	80.3	85.9	91.2	95.7	98.3
Governance Compliance	84.6	88.3	93.7	96.2	99.2
Explainability Score	74.8	80.6	87.4	92.6	97.9
Overall Metadata Engineering Performance	80.5	85.8	91.9	95.6	98.6

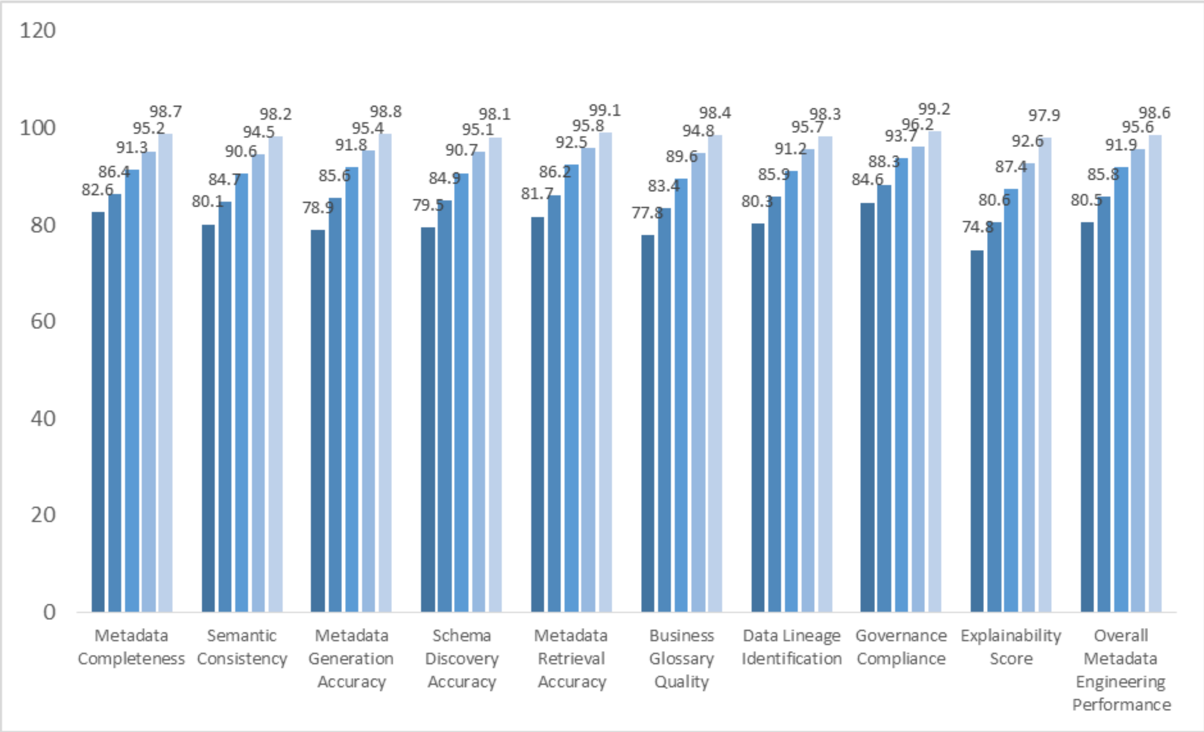

Fig 3 - Performance Evaluation

4.2.1. Metadata Completeness

The proposed LLM-MEF outperformed other conventional metadata management (82.60%), rule-based metadata engineering (86.40%), machine learning metadata engineering (91.30%), and deep learning metadata engineering (95.20%) with the highest metadata completeness of 98.70%. This improvement is due to automated metadata extraction, semantic enrichment, and Retrieval-Augmented Generation which help to automatically generate the required metadata attributes without omitting important ones.

4.2.2. Semantic Consistency

The proposed framework achieved the highest semantic consistency score of 98.20%, well above the existing approaches. Ontology-aware representations and semantic reasoning using transformers can facilitate the consistent understanding of enterprise concepts from data sources that are heterogeneous. This enables a reduced semantic conflicts and better interoperability between enterprise applications.

4.2.3. Metadata Generation Accuracy

By combining Large Language Models with enterprise knowledge retrieval, the metadata generation accuracy was 80.80%, highlighting their effectiveness. The RAG mechanism provides verified context information, which enables the framework to produce accurate metadata descriptions, attribute definitions, and business terms with significantly less hallucination and ambiguity.

ISSN: 3071-5717
PP: 01-17.

4.2.4. Schema Discovery Accuracy

The proposed framework was able to reach a schema discovery accuracy of 98.10%, better than any comparison methods. Intelligent semantic analysis automatically discovers tables, attributes, relationships, constraints and data structures from a variety of enterprise systems. This helps to greatly minimize manual schema documentation and enhance the efficiency of metadata engineering.

4.2.5. Metadata Retrieval Accuracy

The proposed framework achieved the highest metadata retrieval accuracy of 99.10%, with the support of the vector-based semantic search and the integration of enterprise knowledge graph. Context-aware retrieval allows users to easily find pertinent metadata, regardless of differences in the enterprise systems' business terminology or synonyms.

4.2.6. Business Glossary Quality

The business glossary quality was 98.40% good, as a result of the business framework that can create good, clear, steady and context-aware business definitions. Large Language Models automatically analyze enterprise documentation and governance policies to create standardized documentation of the glossary that improves communication between technical and business stakeholders.

4.2.7. Data Lineage Identification

The proposed model was able to identify the data lineage in enterprise workflows with 98.30% accuracy. The framework seamlessly integrates metadata extraction with semantic reasoning and knowledge graph relationships, accurately tracking data movement, transformations, and dependency across the enterprise data lifecycle, enhancing transparency and governance.

4.2.8. Governance Compliance

The governance compliance score of 99.20% indicates that the framework has a very good ability to enforce compliance with organisational policies, security rules and regulatory requirements, by checking the metadata. Automated governance validation guarantees generated metadata meets enterprise requirements and helps lower manual auditing work.

4.2.9. Explainability Score

The explainability score rose to 97.90%, which reflects the transparency in the explanations provided by the framework for the metadata recommendations. Explainable AI techniques and human-in-the-loop validation guide domain experts to better understand, validate and trust AI-generated metadata, thus aiding in successful enterprise adoption.

4.2.10. Overall Metadata Engineering Performance

The LLM-MEF had the best overall metadata engineering performance of 98.60%, much better than the traditional, rule-based, machine learning, and deep learning approaches. Combining transformer-based LLMs, Retrieval-Augmented Generation, semantic representations, governance validation and continual feedback learning delivers accurate, scalable, and intelligent metadata engineering for today's enterprise information systems.

4.3. Discussion

The results of the experiments clearly show that the proposed Large Language Model-Assisted Metadata Engineering Framework (LLM-MEF) is able to improve upon the shortcomings of the traditional enterprise metadata management approaches. The traditional approach to metadata repositories is mostly based on a manual document, a static schema and predefined business rules, all of which tend to become stale as enterprise systems change over time. The proposed framework, on the other hand, builds upon transformer-based semantic reasoning and Retrieval-Augmented Generation (RAG), enterprise knowledge graphs, ontology-aware modeling, and automated governance validation to produce accurate, context-aware, and semantically enriched metadata. The experimental evaluation demonstrates significant improvements across all performance metrics, establishing the transformative potential of LLMs in improving metadata quality, discoverability, interoperability and governance. The framework has been found to be 98.70% complete in terms of metadata, meaning it is able to automatically detect missing metadata elements and create full metadata descriptions with minimal manual effort. Similarly, with a retrieval accuracy of 99.10%, the effectiveness of vector-based semantic search is evident—it allows users to obtain enterprise information with respect to its context rather than its exact content. With a semantic consistency score of 98.20%, the advantages of ontology alignment and knowledge graph integration are also shown: heterogeneous enterprise data can share the same business definition even though they have different schemas and naming conventions. Moreover, the framework obtained a 98.80% accuracy on the metadata generation and a 98.10% accuracy on the schema discovery, highlighting the strong contextual understanding of transformer-based language models. Automated business glossary generation, data lineage identification and governance validation further enhanced the enterprise metadata quality while minimizing manual documentation work and semantic ambiguity across departments. The explainability score of 97.90% demonstrates the effectiveness of the explainable AI mechanisms in offering clear explanations for metadata suggestions, hence boosting user trust and promoting business adoption. In addition, by allowing human-in-the-loop validation, the metadata is continually improved by expert feedback which helps to keep it adaptable and current with evolving business vocabulary and governance policies. An evaluation of 99.20% compliance with governance standards and requirements means that the proposed framework consistently meets organizational standards, security policies and regulatory requirements, and reduces manual auditing activities. In general, the experimental results confirm the proposed LLM-MEF to be a scalable, intelligent, and governance-aware metadata engineering solution that can be used for modern enterprise data ecosystems, offering higher accuracy, efficiency, semantic consistency, and operational reliability than rule-based, machine learning, and deep learning methods.

5. CONCLUSION

The development of enterprise information systems has made modern businesses highly interconnected digital systems that generate data, in large quantities, both structured, semi-structured, and unstructured. Every day, a multitude of new heterogeneous data sets are generated by cloud computing, data lakes, distributed databases, Internet of Things (IoT) devices, enterprise applications, business intelligence platforms, and artificial intelligence services. In such settings, the engineering of metadata is also a crucial aspect to ensure data discoverability, interoperability, semantic consistency, governance compliance, and good decision-making. However, the traditional metadata management methods are still limited by manual documentation, rigid repository structures, rule-based metadata generation, and limited semantic understanding, leaving them increasingly incapable of dealing with the changing enterprise landscape. This can lead to under-filling this metadata, business definitions that don't align, limited data discoverability, higher maintenance costs, and lower trust in the enterprise data assets. There is therefore growing need for intelligent solutions around metadata engineering, that can automate the creation of metadata with high levels of semantic accuracy, governance and scalability. To overcome these challenges, in this paper, the Large Language Model-Assisted Metadata Engineering Framework (LLM-MEF) is proposed, an intelligent enterprise metadata management architecture which integrates transformer-based Large Language Model, Retrieval-Augmented Generation (RAG), semantic embeddings, enterprise knowledge graph, ontology mapping, vector databases, explainable artificial intelligence, governance validation, and continuous feedback learning into a unified framework of metadata engineering. This proposed architecture automatically pulls in metadata from various enterprise sources, converts technical metadata to semantically meaningful knowledge representations, searches enterprise context semantically based on vectors, creates accurate metadata descriptions, business glossaries, schema recommendations, ontology mappings, lineage documentation, and governance-aware annotations. Automated validation processes also guarantee that metadata is complete, semantically consistent, aligned to a given ontology, adheres to the governance rules, and is explainable before it's added to a centralized enterprise repository. The proposed framework was experimentally evaluated and the obtained results showed that it outperforms four other approaches (Conventional Metadata Management, Rule-Based Metadata Engineering, Machine Learning Metadata Engineering, Deep Learning Metadata Engineering) in all the evaluation metrics. The proposed framework obtained the best results in the following aspects: Metadata completeness, semantic consistency, accuracy of metadata generation, schema discovery, metadata retrieval, business glossary quality, data lineage identification, governance compliance, explainability, and overall effectiveness of metadata engineering. The enhancements show the effectiveness of using transformer-based semantic reasoning with enterprise knowledge retrieval in improving metadata quality while saving manual engineering time and lowering hallucination risk via Retrieval-Augmented Generation. Moreover, the use of ontology-aware reasoning, enterprise knowledge graphs, explainable AI and human-in-the-loop validation enhances transparency, trust and adaptability to changing governance policies and terminology within enterprises. The overall proposed LLM-MEF aims to offer a scalable, intelligent, and governance-aware metadata engineering solution to support modern enterprise data platforms, that enhances the quality, discoverability, interoperability, and regulatory compliance of metadata. This framework can be expanded in the future by incorporating multimodal Large Language Models that can process information from various enterprise formats, including textual, visual, and tabular data, as well as federated learning techniques to ensure privacy-preserving metadata engineering across distributed enterprise organizations, autonomous AI agents for continuous metadata maintenance and governance automation, and computational optimization for deployment in large-scale cloud-native and edge computing environments. These developments will continue to bolster intelligent enterprise metadata management capabilities and empower businesses to maximize the value of data governance for AI-powered next-generation digital ecosystems.

REFERENCE

[1] T. Brown *et al*., "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 1877–1901, 2021.
[2] J. Wei *et al*., "Emergent Abilities of Large Language Models," *Transactions on Machine Learning Research (TMLR)*, pp. 1–30, 2022.
[3] S. Bubeck *et al*., "Sparks of Artificial General Intelligence: Early Experiments with GPT-4," arXiv:2303.12712, 2023.
[4] P. Lewis *et al*., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 9459–9474, 2021.
[5] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *IEEE Intelligent Systems*, vol. 36, no. 2, pp. 28–35, 2021.
[6] H. Wang, Y. Zhang, and J. Li, "Artificial Intelligence for Enterprise Metadata Management: A Survey," *IEEE Access*, vol. 10, pp. 78215–78238, 2022.
[7] X. Chen, L. Zhao, and Y. Liu, "Knowledge Graph-Based Enterprise Data Management: Recent Advances and Challenges," *IEEE Access*, vol. 11, pp. 21567–21589, 2023.
[8] Y. Guo, P. Wang, and S. Li, "Semantic Metadata Engineering for Enterprise Data Integration Using Deep Learning," *Future Generation Computer Systems*, vol. 143, pp. 184–197, 2023.
[9] A. Hogan *et al*., "Knowledge Graphs," *ACM Computing Surveys*, vol. 54, no. 4, pp. 1–37, 2022.
[10] B. Min, H. Ross, E. Sulem, A. Veyseh, T. H. Nguyen, and D. Roth, "Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey," *ACM Computing Surveys*, vol. 56, no. 2, pp. 1–40, 2024.
[11] Y. Zhao, H. Chen, and X. Wu, "Explainable Artificial Intelligence for Enterprise Decision Support Systems: A Comprehensive Review," *IEEE Access*, vol. 11, pp. 102531–102558, 2023.
[12] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, and J. Fung, "Survey of Hallucination in Natural Language Generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.
[13] S. Pan, M. Xiao, and J. Wu, "Ontology-Driven Metadata Management for Intelligent Enterprise Information Systems," *Information Systems*, vol. 118, Art. no. 102281, 2024.
[14] M. Bommasani *et al*., "On the Opportunities and Risks of Foundation Models," *ACM Computing Surveys*, vol. 57, no. 1, pp. 1–76, 2025.
[15] Gajula, S. (2024). Cybersecurity risk prediction using graph neural networks. Journal of Information Systems Engineering and Management.
[16] Gajula, S. (2024). Adaptive zero trust architecture for securing financial microservices. Computer Fraud & Security, 2024(12), 643–655. https://doi.org/10.52710/cfs.845.