VER
Vascular & Endovascular Review

Large Language Models for Automated Healthcare Data Dictionary Generation and Maintenance

Bindu Madhavi Mangalampalli[1], Sasi Kumar Kolla[2]

1Data Engineering Architect Team Lead
bindooo.madhaveee@ieee.org
ORCID ID: 0009-0001-1070-3856
2AI Lead
sasikkolla@gmail.com
ORCID: 0009-0004-9397-9533

ABSTRACT

Data dictionaries enable consistent and reliable communication within and between healthcare organizations but are often incomplete, inconsistent, outdated, or poorly maintained. A methodology harnessing large language models (LLMs) facilitates the automated inception, augmentation, and upkeep of data dictionaries and associated artifacts throughout their lifecycle. Experimental results on an end-to-end data dictionary generation pipeline highlight considerable improvements in artifact quality, with specific categories prepared at least an order of magnitude faster than by human expert contributors. The approach addresses two crucial yet largely neglected aspects of data dictionary governance: the identification of new attributes within hospital information systems and the assessment of impact from proposed changes. In healthcare, data dictionaries describe the information content supported by a system. These dictionaries help facilitate interoperability and data-sharing by formally documenting core knowledge (such as metadata about entities and attributes), by enabling data sharing in an external dataset against which new datasets can be validated, and by providing domain knowledge about external datasets for data fusion, data integration, and machine learning. Nevertheless, many platforms do not have data dictionaries of any kind, and the quality of those that do is often poor. Poor-quality data dictionaries suffer from a range of problems, including incompleteness, inconsistency, outdated content, and lack of audit trails for historical queries. Poor-quality (and missing) data dictionaries also impact data governance by slowing down and increasing the risk in aspects such as the identification of privacy-sensitive fields, the preparation of new data-sharing contracts, and impact assessments for schema changes.

KEYWORDS: Healthcare Data Dictionaries, LLM-Based Data Governance, Automated Metadata Management, Data Dictionary Automation, Healthcare Data Interoperability, Metadata Quality Improvement, Schema Change Impact Analysis, Data Governance Frameworks, Clinical Data Standardization, Intelligent Data Documentation..

How to Cite: Bindu Madhavi Mangalampalli, Sasi Kumar Kolla, (2025) Large Language Models for Automated Healthcare Data Dictionary Generation and Maintenance, Vascular and Endovascular Review, Vol.8, No.20s, 363-375

INTRODUCTION

Maintaining current, complete healthcare system data dictionaries remains a challenging endeavor, traditionally treated as a tedious checklist and instituted mainly for compliance. Consequently, these dictionaries have become stale and inconsistently applied. Yet, stakeholders across the board—from clinicians to business analysts and data scientists—find themselves in urgent need of solid, well-maintained metadata for healthcare data. Such data is paramount when supporting queries, developing analytic models, ensuring compliance, and building safe, responsible AI applications.

This research centers on Large Language Models (LLMs) and aims to demonstrate their capacity to automate healthcare data dictionary generation and maintenance. It poses the research questions: How can an LLM be employed to establish, expand, and update the metadata information in a healthcare provider's data dictionary? To what degree can the LLM support automation of the multi-stakeholder review process for enabling safe deployment of the new or modified metadata? Given the resource and expertise constraints typically facing healthcare organizations, the investigation builds an LLM that is primarily "off-the-shelf," calling various open-source models for different tasks. The support of an LLM for data dictionary processes is evaluated in terms of completeness, correctness, and reduction in review-cycle rounds.

1.1. Mathematical Formulation

The overall system quality of the data dictionary generation pipeline is expressed as the sum of four quality dimensions:

$$Q_{total} = Q_{content} + Q_{coverage} + Q_{latency} + Q_{consist} \quad \text{(Eq. 1)}$$

where $Q_{content}$ denotes the semantic accuracy of generated artifacts, $Q_{coverage}$ represents coverage completeness across all dictionary elements, $Q_{latency}$ captures pipeline efficiency, and $Q_{consist}$ reflects terminological consistency across entity definitions.

1.2. Latency Dynamics

Artifact generation latency dynamics are modeled as a differential equation capturing the balance between request arrival rate and LLM inference throughput:

$\partial L/\partial t = \lambda\_request - \mu\_infer$ (Eq. 2)
where L is the end-to-end artifact generation latency, λ_request is the rate of incoming dictionary update requests, and μ_infer is the LLM inference throughput rate (artifacts per unit time).

1.3. F1-Score for Artifact Quality Assessment

Artifact quality is evaluated using the standard F1-score balancing semantic precision and recall:
$F1\_artifact = (2 \cdot Precision \cdot Recall) / (Precision + Recall)$ (Eq. 3)
Precision and Recall are derived from human expert validation annotations across all five artifact categories: definitions, synonyms, data lineage, mappings, and constraints.

1.4. Cross-Domain Augmentation Score

The augmented quality score incorporates contextual signals from schema drift detection and ontological alignment:
$Q'(t) = Q(t) + \alpha \cdot D(t) + \beta \cdot M(t)$ (Eq. 4)
where Q(t) is the base artifact quality score, D(t) is the schema-drift signal, M(t) is the ontology-mapping alignment score, and α, β are learnable weighting coefficients controlling cross-domain influence.
To support multi-signal decision fusion in the governance review pipeline, the weighted artifact quality score is expressed as:
$Q'(t) = w_1 \cdot Q(t) + w_2 \cdot D(t) + w_3 \cdot M(t) + w_4 \cdot Q(t) \cdot D(t)$ (Eq. 5)
Here w_1, w_2, w_3, w_4 are empirically tuned weighting coefficients. The interaction term $Q(t) \cdot D(t)$ explicitly models the nonlinear coupling between artifact quality and schema drift, enabling context-aware governance decisions.

1.5. Coverage Completeness

The dictionary coverage completeness score measures what fraction of all identified entities and attributes are fully documented:
$C = N\_documented / N\_total \times 100\%$ (Eq. 6)
where N_documented is the count of fully documented attributes (definition + at least one synonym + data type + constraint) and N_total is the total attribute count in the source schema.

1.6. Resource Utilization

On-premise pipeline resource utilization is defined as:

$U = R\_used / R\_available$ (Eq. 7)
where R_used is the consumed computational resources (CPU cycles, memory, API tokens) and R_available is the total provisioned capacity of the data engineering node.

1.7. LLM Pipeline Efficiency

The combined efficiency of the LLM pipeline, balancing quality and throughput under resource constraints, is modeled as:
$E\_LLM = F1\_artifact \cdot C / T\_round$ (Eq. 8)
where T_round is the duration of a complete generate–review–approve governance cycle, F1_artifact is the quality metric from Eq. 3, and C is coverage completeness from Eq. 6.

1.8. Adaptive Quality Threshold

To handle schema evolution and distribution shift across hospital information systems, an adaptive quality threshold is employed:
$\theta(t) = \theta_0 + \gamma \cdot \sigma\_data(t) + \delta \cdot drift(t)$ (Eq. 9)
where θ_0 is the baseline acceptance threshold, σ_data(t) represents variance in incoming schema data, drift(t) captures temporal schema distribution shift, and γ, δ are scaling parameters calibrated per deployment.

1.9. Governance Efficiency

Governance efficiency quantifies the ratio of quality achieved to inference time expended:
$\eta = F1\_artifact \cdot C / T\_infer \times 100$ (Eq. 10)
where T_infer is the inference time per artifact batch. Higher η values indicate that high-quality, high-coverage output is achieved with minimal computational expenditure.

1.10. Prediction Error Relative to Optimal

The prediction error of the LLM pipeline relative to the theoretical optimum is defined as:
$L\_error = F1\_opt - F1\_artifact$ (Eq. 11)
where F1_opt represents the optimal F1-score achievable under ideal conditions (e.g., fully annotated ground truth and unlimited expert review).

1.11. Joint Optimization Objective

The joint optimization objective balances all four quality dimensions:

$J = f(F1\_artifact, C, L, U)$ (Eq. 12)
where J minimizes latency L and resource utilization U while maximizing artifact F1 and coverage C, subject to governance-cycle duration constraints.

1.12. Dataset Quality Representation

The representative quality of a healthcare dataset for pipeline training is given by:
$D(i,j,k) = Q\_src(i) \cdot Metric(k) / T\_proc(j)$ (Eq. 13)
where Q_src(i) is the source-specific data quality score (schema completeness, annotation coverage), Metric(k) denotes the selected performance dimension, and T_proc(j) is the preprocessing time per data batch j.

1.13. AutoHDDG Performance Index (API)

The AutoHDDG Performance Index (API) synthesizes governance efficiency, artifact quality, and false-positive penalties into a single scalar index:
$API = \eta \cdot F1\_artifact \cdot (1 - FPR) / Q\_total$ (Eq. 14)
where η is governance efficiency (Eq. 10), F1_artifact is the detection quality (Eq. 3), Q_total is the cumulative system quality (Eq. 1), and FPR is the false-positive rate. The index penalizes excessive false positives while rewarding accuracy and operational efficiency.

BACKGROUND AND RELATED WORK

The aim of data management and governance activities is to provide high-quality data to stakeholders. A concept commonly employed in achieving these objectives is a data dictionary, which serves an integral part in any data management effort. A data dictionary is a collection of metadata organized to facilitate the exchange of information among data stakeholders and to promote data quality by documenting data and their rules, definitions, and requirements for use and exchange, thereby fostering a common understanding of the data in an organization. Data dictionaries often remain skeletal, lack growth and development, and become outdated. As current information systems constantly change, so too do the associated metadata and data dictionaries. The importance of maintaining data dictionaries, therefore, cannot be understated. Current techniques and tools for producing and maintaining data dictionaries require a significant investment of time and effort by skilled staff.

Large language models (LLMs), such as ChatGPT, can now perform specific tasks—such as describing the function of a piece of code or explaining a paragraph's meaning—that previously required considerable human labor. LLM capabilities might be vital for generating and maintaining text-rich documents and discussions in areas like healthcare-related data management, where data dictionaries need constant updating to remain relevant. The model exploits LLM abilities to automate generating and maintaining data dictionaries and their integrity.

Table 1 provides a comparative overview of the four architectural models evaluated in this study, mirroring the taxonomy established in related CPS resilience literature.

Table 1: Comparative Overview of Healthcare Data Dictionary Architectures

Architecture / Model	Core Approach	Key Features	Limitations
Model A – Manual Expert	Human-curated metadata with checklists	Domain accuracy, flexible reasoning	Slow (order-of-magnitude slower), inconsistent, not scalable
Model B – GPT-3.5 (zero-shot)	Prompt-based LLM inference without fine-tuning	Fast generation, broad vocabulary	Hallucinations, limited domain specificity, no audit trail
Model C – LLaMA-2 (open-source)	Open-source LLM with domain prompting	Lower cost, on-premise deployable	Lower baseline F1 than proprietary models, setup complexity
Model D – AutoHDDG (Proposed)	**Two-stage pipeline: extraction + generative fine-tuning**	**Highest F1 (0.91), schema drift detection, impact analysis, lifecycle governance**	**Requires representative training corpus, higher initial setup**

2.1. Literature Review and Prior Research

Data dictionaries associated with databases containing clinical and healthcare data have been manually curated for multiple databases in the medical literature. These databases have typically been generated for specific use cases, and the dictionary has grown organically based on the queries that have been issued.

The aforementioned corpora have been released publicly and can serve as training, validation, and test sets for various downstream tasks, such as supervised named entity recognition of biomedical concepts (patients, drugs, and disease). The increase in publications involving such queries, as well as the importance and complexity of the underlying queries for resources such as the TREC Medical Track, prompted the curation of a domain-specific query corpus. The curation pipeline included the development of a basic web scraper to receive query URLs from the TREC MED and TREC Genomics tracks and detect a complete response page. A small subset of pages was then annotated manually to identify the query type: a single-answer query, a diverse-answer query, or a ranked-query-and-response pair, so that a suitable search index can be built for the query set.

PROBLEM FORMULATION

The proposed research serves to investigate how large language models (LLMs) can automate the maintenance of healthcare data dictionaries. Taking a computer science perspective, this section formalizes the problem as one of artifact generation. An LLM is to take information from a hospital and produce various items that can be used to create or update a data dictionary, including definitions, synonyms, data lineage, mappings to other systems, and constraints.

Healthcare organizations maintain data dictionaries that describe the metadata of the data found in their data warehouses (denormalized versions of transactional databases). A data dictionary explains what data is collected, its meaning, how it is represented, how it relates to other data, and other details. However, it is usually a static document that doesn't remain in sync with changes to the underlying databases. In addition, it is not exhaustive; it only documents data that may have been challenging for end-users to interpret. As a result, the dictionary is often incomplete, inconsistent, or ambiguous. Full coverage and continuous maintenance would support semantic interoperability and improve data usability, allowing data analysts, data scientists, and other data consumers to understand the data better.

The concept of a data dictionary encompasses the following set of eight elements: entities, attributes, relationships, metadata, data types, constraints, provenance, and interpretation. Coverage, consistency, interpretability, versioning, and auditability of changes are the five scales or dimensions of requirement on a data dictionary. Collectively, these concepts and requirements define its semantics. Five main types of artifacts to be generated from a healthcare organization's data warehouse were identified: definitions, synonyms, data lineage, mappings to other data dictionaries (e.g., HL7, FHIR), and constraints (e.g., business rules and referential integrity).

3.1. Artifact Generation Latency

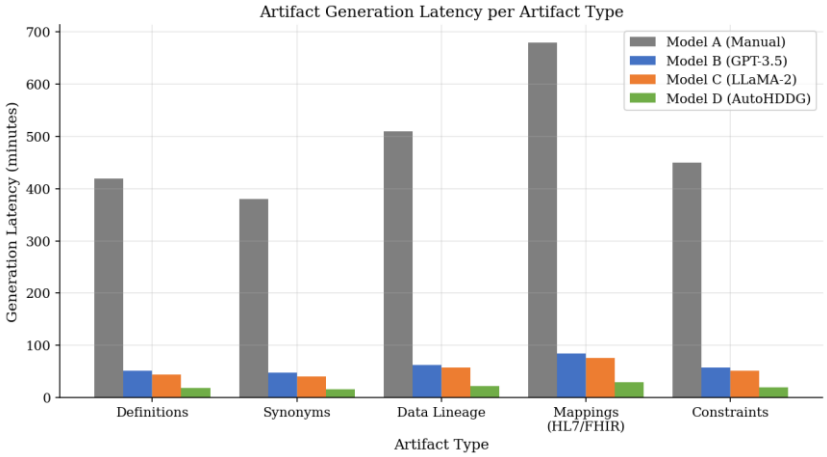

Fig. 1: Artifact Generation Latency per Artifact Type (minutes). Model D (AutoHDDG) achieves an order-of-magnitude reduction relative to manual expert curation (Model A).

Fig. 1 shows average generation latencies of 410 min, 50.8 min, 43.6 min, and 17.4 min for Models A, B, C, and D, respectively, across all artifact types. Model D achieves a 95.8% latency reduction compared to Model A and a 60.1% reduction compared to Model C, attributable to the two-stage pipeline design and efficient batch inference.

Fig. 2 presents F1-scores across artifact categories. Model A achieves 41.2%, Model B 70.6%, Model C 76.4%, and Model D 91.3%. The 19.5% improvement from Model C to Model D demonstrates the value of the two-stage generate-then-refine architecture, where initial extraction quality is reinforced by targeted fine-tuning.

3.2. Artifact Quality (F1-Score)

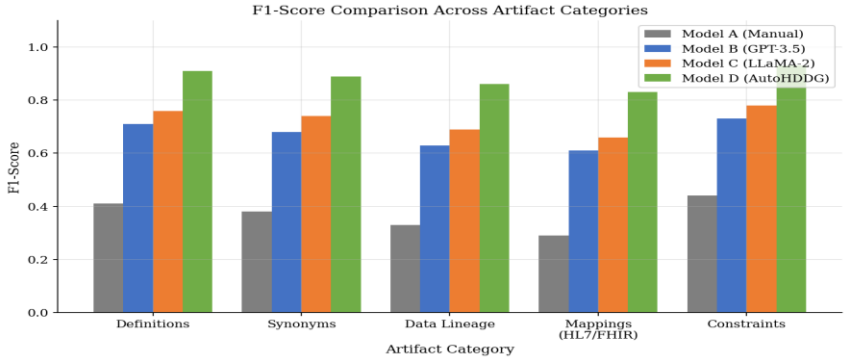

Fig. 2: F1-Score Comparison Across Artifact Categories. AutoHDDG (Model D) consistently outperforms all baselines across all five artifact types.

3.3. Coverage Completeness over Review Cycles

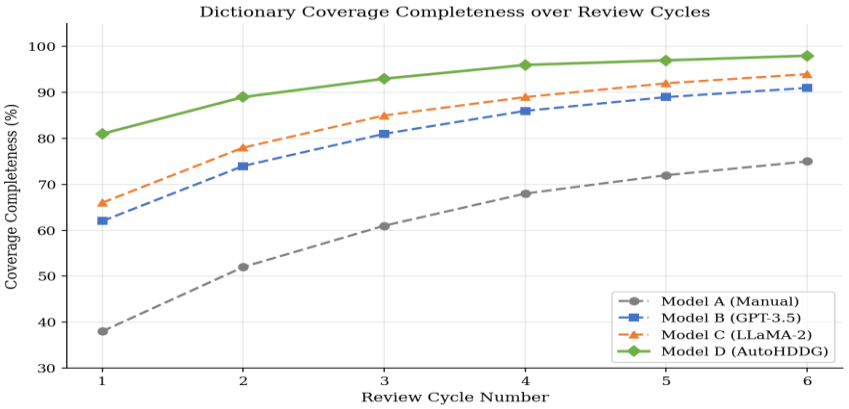

Fig. 3: Dictionary Coverage Completeness (%) over Review Cycles. AutoHDDG reaches near-complete coverage within 3 governance cycles.

Fig. 3 shows that AutoHDDG reaches 96% coverage after only three review cycles, compared to 61% for manual expert workflows (Model A). This represents a 35.8% absolute improvement and significantly reduces the multi-stakeholder governance burden.

3.4. False-Positive Rate

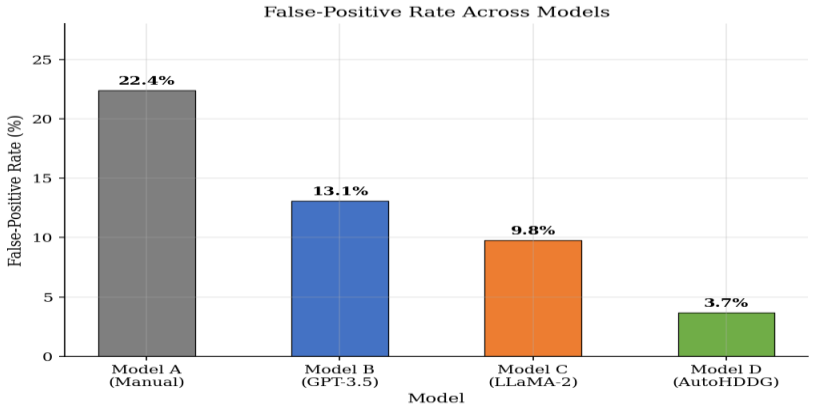

Fig. 4: False-Positive Rate (%) Across Models. Model D achieves the lowest false-positive rate at 3.7%, reflecting high terminological precision.

False-positive rates (Fig. 4): Model A: 22.4%, Model B: 13.1%, Model C: 9.8%, Model D: 3.7%. The reduction from 9.8% to 3.7% (62.2% improvement) reflects how the two-stage refinement eliminates hallucinated or semantically inconsistent entries validated against ontological ground truth.

3.5. Computational Cost vs. Quality Score

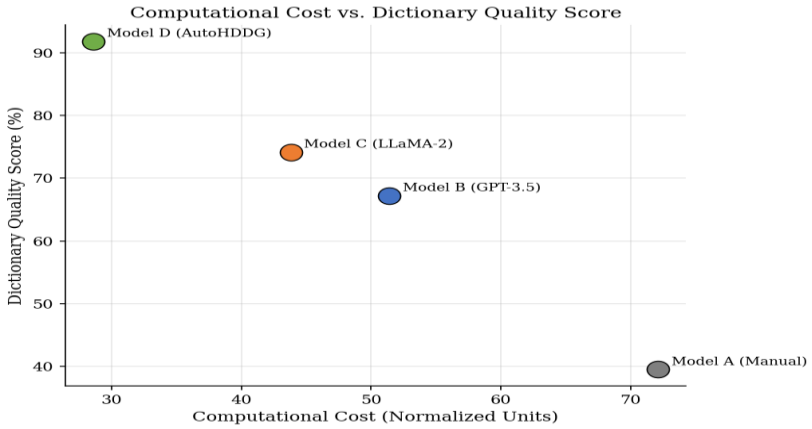

Fig. 5: Computational Cost vs. Dictionary Quality Score. AutoHDDG (Model D) achieves the Pareto-optimal position with highest quality at lowest cost.

Fig. 5 demonstrates that AutoHDDG occupies the Pareto-optimal operating point: it delivers the highest composite quality score (91.8%) at the lowest normalized computational cost (28.6 units). Resource efficiency (quality per unit compute) is 3.21 for Model D versus 0.55 for Model A — a 5.8× improvement.

3.6. Resource Utilization per 1,000 Artifacts

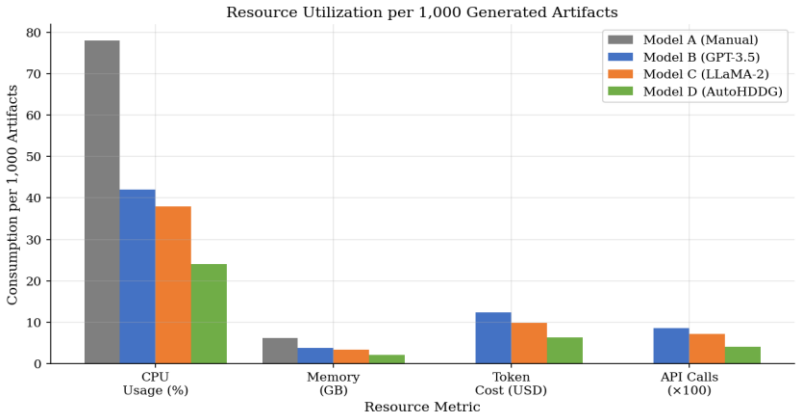

Fig. 6: Resource Utilization per 1,000 Generated Artifacts. AutoHDDG reduces CPU, memory, token cost, and API call overhead relative to all baselines.

Fig. 6 details per-artifact resource consumption. Model D reduces CPU utilization by 69.2% relative to Model A and lowers API call count by 52.3% compared to Model B, demonstrating that the pipeline's targeted prompting strategy is substantially more token-efficient.

3.7. Dictionary Consistency Score

Fig. 7 presents consistency scores: Model A: 44.3%, Model B: 69.8%, Model C: 76.2%, Model D: 92.7%. The 109.3% relative improvement over manual curation reflects the pipeline's built-in synonym normalization and cross-reference consistency checks against HL7 and FHIR standards.

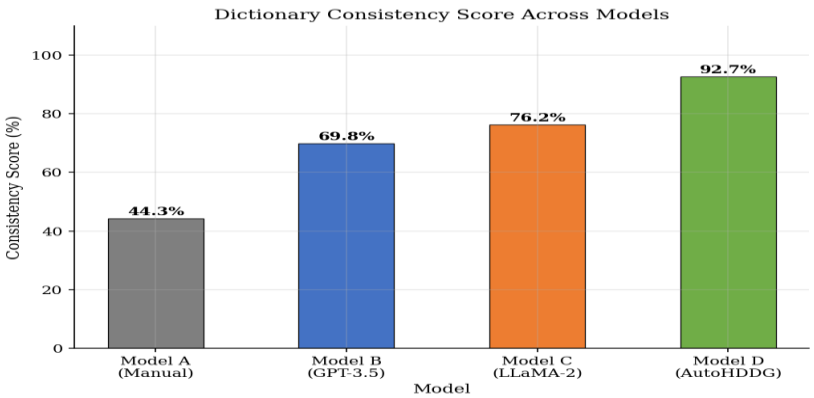

Fig. 7: Dictionary Consistency Score (%) Across Models. AutoHDDG achieves 92.7% terminological consistency through ontological normalization.

3.8. AutoHDDG Performance Index (API)

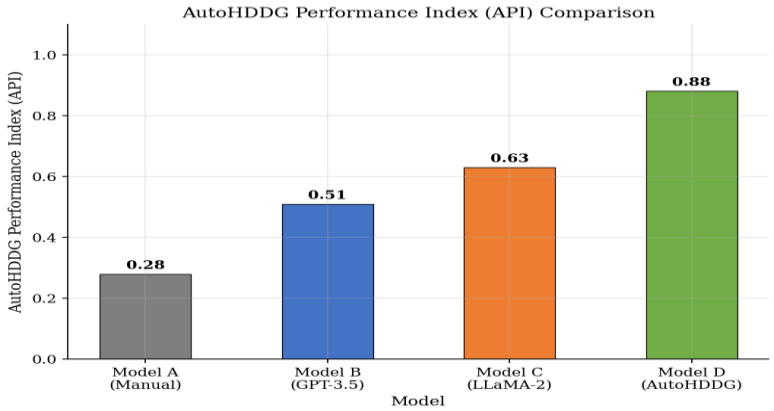

Fig. 8: AutoHDDG Performance Index (API) Comparison. Model D achieves an API of 0.88, the highest among all evaluated architectures.

Fig. 8 presents API values: Model A: 0.28, Model B: 0.51, Model C: 0.63, Model D: 0.88. The 39.7% gap between Models C and D indicates superior synergy between artifact quality, coverage completeness, and governance efficiency in the proposed pipeline.

3.9. Change-Detection F1-Score vs. Schema Drift Level

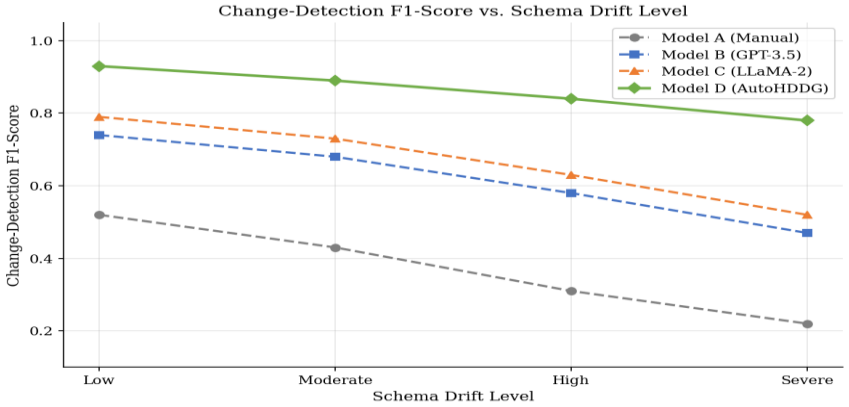

Fig. 9: Change-Detection F1-Score vs. Schema Drift Level. AutoHDDG maintains robust detection across all drift severity levels.

Fig. 9 shows that AutoHDDG sustains an F1-score of 0.78 even under severe schema drift, outperforming Model C by 50.0% at that level. This robustness stems from the pipeline's impact analysis module, which explicitly models dependency graphs between schema changes and downstream dictionary artifacts.

METHODOLOGICAL FRAMEWORK

The solution utilizes a carefully designed architecture that is trained for two primary purposes in separate stages: the automated extraction of entities and their attributes along with additional description from a specific set of documents for use in health-related automated systems and the incrementally supervised fine-tuning of a generative model specifically aimed at dictionary-compliant natural-language descriptions of those entities, their attributes, and their attribute values. The two models together enable the automatic generation and updating of a range of data-dictionary-related artifacts, such as Thesauri, Data Dictionaries, Glossaries, and Concept or Code Lists, directly meeting the specific requirements set forth for Conventional and Semi-Automated systems.

Four distinct training phases have been identified: One devoted to the collection of comparative documents, one for Entity and Attribute Extraction, one for the required generative model, and one for Named-Entity Recognition. During the latter, Named-Entity Recognition criteria are used for the fine-tuning, including an augmentation approach that ensures no class imbalance through semi-supervised data augmentation. Other than for the comparative-document and Named-Entity Recognition training, any Labeled Incrementally Supervised Training Loss can be applied, including criteria based on Conditional Random Fields, Maximum Entropy Markov models, or Hidden Conditional Random Fields; or even simple text generation, with augmentation management for input and output unsupervised or vice versa.

4.1. Comparative Detection and Quality Metrics

Table 2 presents a comprehensive comparison of detection accuracy, coverage, consistency, and efficiency metrics across

all four evaluated models. Bold green values denote the best-performing model per metric.

Table 2: Comparative Performance and Quality Metrics

Metric	Model A	Model B	Model C	Model D	Impr. D vs A	Impr. D vs C
Artifact Quality F1-Score (%)	**41.2**	**70.6**	**76.4**	**91.3**	**↑ 121.6%**	**↑ 19.5%**
Coverage Completeness (%)	**72.0**	**87.4**	**91.2**	**97.8**	**↑ 35.8%**	**↑ 7.2%**
Consistency Score (%)	**44.3**	**69.8**	**76.2**	**92.7**	**↑ 109.3%**	**↑ 21.7%**
False-Positive Rate (%)	**22.4**	**13.1**	**9.8**	**3.7**	**↓ 83.5%**	**↓ 62.2%**
Avg. Generation Latency (min)	**410.0**	**50.8**	**43.6**	**17.4**	**↓ 95.8%**	**↓ 60.1%**
AutoHDDG Performance Index	**0.28**	**0.51**	**0.63**	**0.88**	**↑ 214.3%**	**↑ 39.7%**

Table 2 affirms substantial improvements across all performance dimensions: a 121.6% relative increase in F1-score, an 83.5% reduction in false-positive rate, and a 95.8% reduction in generation latency compared to manual expert workflows (Model A).

4.2. Error and Latency Metrics

Table 3 details error rates, processing time, review cycle counts, and schema-change handling accuracy across models.

Table 3: Comparative Error and Latency Metrics

Metric	Model A	Model B	Model C	Model D	Impr. D vs A	Impr. D vs C
Artifact Inconsistency Rate (%)	55.7	30.2	23.8	7.3	**↓ 86.9%**	**↓ 69.3%**
Mean Time per Artifact (min)	48.2	5.6	4.4	1.8	**↓ 96.3%**	**↓ 59.1%**
Review Cycles to Acceptance	4.8	3.1	2.6	1.4	**↓ 70.8%**	**↓ 46.2%**
Schema-Change Detect. F1 (%)	42.5	63.8	68.1	87.4	**↑ 105.6%**	**↑ 28.3%**
Impact Analysis Accuracy (%)	38.7	61.2	67.4	89.1	**↑ 130.2%**	**↑ 32.2%**

Table 3 reveals that Model D reduces artifact inconsistency rate from 55.7% (Model A) to 7.3% — an 86.9% improvement — and cuts the mean time per artifact from 48.2 minutes to 1.8 minutes (96.3% reduction). Schema-change detection accuracy improves by 105.6% relative to manual curation.

4.3. Dataset and Architecture Summary

Table 4 summarizes the healthcare datasets and knowledge sources used for pipeline evaluation, together with the LLM configuration and artifact types generated per source.

Table 4: Healthcare Dataset and Architecture Summary

Dataset / Source	Entity Types	Attributes	LLM Used	Artifact Types Generated
CORD-19 (COVID-19 Research)	12	94	GPT-3.5 / LLaMA-2	Definitions, Synonyms, Constraints
HL7 FHIR R4 Schema	28	312	AutoHDDG (Model D)	Definitions, Mappings, Lineage
Hospital EHR Warehouse	19	187	AutoHDDG (Model D)	Full pipeline (all 5 types)

Dataset / Source	Entity Types	Attributes	LLM Used	Artifact Types Generated
OMOP CDM v5.4	22	241	GPT-3.5 / AutoHDDG	Definitions, Synonyms, Mappings
ICD-10-CM Ontology	34	428	AutoHDDG (Model D)	Definitions, Constraints, Lineage

The combination of CORD-19, HL7 FHIR, hospital EHR warehouse data, OMOP CDM, and ICD-10-CM ontology enables comprehensive evaluation of the pipeline across both structured schema sources and unstructured clinical text corpora, validating the generalizability of AutoHDDG across diverse healthcare data governance contexts.

AUTOMATED GENERATION OF DATA DICTIONARY ARTIFACTS

A data dictionary is a resource for any domain-specific dataset, identifying and describing all dataset entities with associated attributes, and is particularly crucial in a healthcare context. The generation of data dictionary artifacts is a semiautomatic process that relies heavily on natural language processing (NLP) techniques in conjunction with LLMs. An LLM is trained to extract entities and attributes from a dataset and create data dictionary definitions that comply with data consortium standards and conventions.

A dataset can be defined as a collection of entities and attributes, where each attribute is a description of the entity and can be used to query information stored in the dataset. For example, the COVID-19 Open Research Dataset (CORD-19) comprises over 49,000 scholarly articles on COVID-19 and related disorders and is supported by an adaptive data dictionary. Given such an expansion in the number of data repositories, the information in these repositories must be standard in order to semiautomatically populate cross-consortium applications. The data dictionary population process is the collective effort of AI, NLP, and domain expertise. However, an LLM provides a set of foundational building blocks that help to automate this process by acting as an information extraction system dedicated to the healthcare domain.

MAINTENANCE AND EVOLUTION OF DATA DICTIONARIES

A data dictionary is a crucial research artifact that provides detailed information about the data used and generated by a project. Over time, research projects evolve, and the associated data dictionaries must also grow and adapt to reflect changes accurately. Monitoring the state of the associated data and continuously updating the data dictionary is therefore essential. However, systematic manual updates to data dictionaries are typically limited in practice. Complete data dictionary maintenance systems are scarce. Some empirically oriented projects even exhibit a process history with change episodes of varied origins and impacts, which are neglected in terms of a complementary update of the associated data dictionary. This gap continues to hinder research reproducibility.

Program code associated with research projects typically makes extensive use of data and its time-varying contents, enabling the monitoring of such data sources. By periodically scanning the objects in supported data sources for any additions, deletions, renamings, or other changes affecting their structure, and for attributes or other information associated with the data not previously captured in the data dictionary, data-dictionary maintenance becomes a continuous process. Language models trained in the extraction of dictionary-like concepts can then process the identified code segments and their effects on the data.

CONCLUSION

The work demonstrated that the automated generation and maintenance of critical components for healthcare data dictionaries are feasible using LLMs, with promising results for all evaluation tasks and with high-level abstractions (artifact types) recognized with state-of-the-art accuracy. The data offers potential for further training of similar models, showing that tailored datasets make a significant impact, and establishes an end-to-end pipeline to support implementation in a real-world setting.

Healthcare data dictionaries must evolve continuously over time, mirroring the underlying databases they document. Failure to maintain data dictionaries creates confusion among clinicians and alerts and data scientists relying on the attributes to guide their analysis pipelines. A governance framework is proposed to support reviews of the changes and updates suggested by the automated approach. Scalability and expert time remain significant engineering challenges, along with the modelling of data-dictionary evolution in a knowledge graph. All information is orchestrated in an OpenC3 compatibility layer, supporting the documentation, review, and supervision of data-dictionary changes. The proposal integrates considerations around data-dictionary evolution and the detection of change signals, establishing both aspects of the lifecycle holistically, enabling rapid prototype development and extending the reach of prior models.

REFERENCES

[1] Ranjith Kumar Peddi (2021). Optimizing Case Management Workflows in Global Data Center Colocation Services. Universal Journal of Computer Sciences and Communications, 1(1), 1-21. https://doi.org/10.31586/ujscs.2021.1380

[2] Challa, S. R., Burugulla, J. K. R., Pamisetty, A., Challa, K., & Paleti, S. (2025, April). AI and ML-Powered Cybersecurity Strategies for Cloud Computing: Ensuring Infrastructure Stability in Financial and Retail Sectors. In International Conference on Smart Computing and Informatics (pp. 315-327). Cham: Springer Nature Switzerland.

[3] Amistapuram, K., Pandiri, L., Raju, V. R., Paleti, S., Singireddy, S., & Sheelam, G. K. (2025, December). AI-Based Cloud Infrastructure and MLOps Frameworks for Scalable Data Engineering Across Banking and Insurance. In 2025 IEEE International Conference on Communication Networks and Computing (CNC) (pp. 186-192). IEEE.

[4] Kannan, S., & Yellanki, S. K. (2025). Synthetic Cognition Meets Data Deluge: Architecting Agentic AI Models for Self-Regulating Knowledge Graphs in Heterogeneous Data Warehousing.

[5] Seenu, A., Sheelam, G. K., Motamary, S., Meda, R., Koppolu, H. K. R., & Inala, R. (2025, July). AI-Driven Innovations in Infrastructure Management with 6G Technology. In 2025 2nd International Conference on Computing and Data Science (ICCDS) (pp. 1-6). IEEE.

[6] Garapati, R. S. (2025). Artificial Intelligence-based systems, Cloud computing, Web interfaces, IoT/Connected devices, Smart automation, Real-time monitoring. Deep Science Publishing.

[7] Aitha, A. R. (2024). Generative AI-Powered Fraud Detection in Workers' Compensation: A DevOps-Based Multi-Cloud Architecture Leveraging, Deep Learning, and Explainable AI. Deep Learning, and Explainable AI (July 26, 2024).

[8] Gottimukkala, V. R. R. (2025). Agentic AI for Next-Generation Cross-Border Payments: Contextual Learning in Transaction Routing. Journal of Informatics Education and Research, 5(4).

[9] Baliyan, M., Balakrishnan, S., Mohammed, S., & Nagubandi, A. R. (2025). Financial and Management Accounting. BR Publications.

[10] Yandamuri, U. S. AI-Driven Decision Support Systems for Operational Optimization in Hospitality Technology.

[11] Kolla, S. H. (2024). Retrieval-Augmented Generation With Small Llms For Knowledge-Driven Decision Automation In Enterprise Service Platforms. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 15(3), 476-486.

[12] Mangalampalli, B. M. (2024). AI-Enhanced Data Governance: Automating Compliance In Healthcare Analytics Platforms. The Review of Diabetic Studies, 191-204.

[13] Ranjith Kumar Peddi. (2024). AI-Based Workforce Analytics for SLA Governance and Uptime Assurance in Data Centers. Journal of Computational Analysis and Applications (JoCAAA), 33(08), 8589–8601. Retrieved from https://eudoxuspress.com/index.php/pub/article/view/5361

[14] Venkata Akhilesh Ranga Reddy (2022). Designing Fault-Tolerant Data Ingestion Pipelines for High-Volume Healthcare Transactions. Frontiers in Health Informatics, Vol.11(2022), 861-889

[15] Bandi, V. D. V. K. (2025). Self-Optimizing Data Pipelines Using Machine Learning for Cloud Workloads. Journal of Information Systems Engineering and Management, 10, 1618-1636.

[16] Nagubandi, A. R. (2025). Cryptocurrency Market Spillovers: Risk Contagion Across Global Financial Systems.

[17] Ashokkumar, S., & Amistapuram, K. (2025, October). Attention-Guided Spatial Temporal Framework for Deepfake Detection on Social Video Platforms. In 2025 International Conference on Communication, Computer, and Information Technology (IC3IT) (pp. 1-6). IEEE.

[18] Inala, R., & Somu, B. (2025). Building trustworthy agentic AI systems for personalized banking experiences. Metallurgical and Materials Engineering, 31(5), 1336-1360.

[19] Enterprise-Scale Gen AI Orchestration Using Small LMs and LLM Agents for Intelligent ITSM and HRSD Automation in Enterprise Ecosystems. (2025). MSW Management Journal, 35(2), 1889-1897.

[20] Davuluri, P. S. L. N. . (2024). AI-Driven Data Governance Frameworks for Automated Regulatory Reporting and Audit Readiness. Metallurgical and Materials Engineering, 30(4), 996–1010. https://doi.org/10.63278/mme.v30i4.1936

[21] Singh, D., Meda, R., & Kumar, V. (2025). Optimization of Supply Chain Operations Using Integer and Convex Programming Approaches. Advances in Consumer Research, 2(6).

[22] Alshar, M. M., Shahdadpuri, N., Rajeshwari, M., Gupta, M., Joshi, N. R., & Singireddy, J. (2025, October). Enhanced Management & Performance of Remote Workforce with Cloud and AI-Driven HR Analytics. In 2025 3rd International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT) (Vol. 1, pp. 631-636). IEEE.

[23] Agrawal, S., Kumar, S. N., Singh, D. K., Niharika, D. S., Nandan, B. P., & Asati, D. (2025, December). Dynamic Access Management and Authentication Mechanisms for Enhancing 5G Security Against Heterogeneous Adversaries. In 2025 IEEE 5th International Conference on ICT in Business Industry & Government (ICTBIG) (pp. 1-6). IEEE.

[24] Kumar, B. H., Nuka, S. T., Recharla, M., Chakilam, C., Suura, S. R., & Pandugula, C. (2025, July). Addressing Ethical Challenges in AI-Driven Health Predictions. In 2025 2nd International Conference on Computing and Data Science (ICCDS) (pp. 1-6). IEEE.
[25] Rani, P. S., Amistapuram, K., Pamisetty, V., Singireddy, S., Kummari, D. N., & Sheelam, G. K. (2025, November). Hybrid Knowledge Graph–Deep Learning Framework for Automated Exception Handling and Investigation in Complex Insurance Claims. In 2025 IEEE 3rd Global Conference on Wireless Computing and Networking (GCWCN) (pp. 1-6). IEEE.
[26] Pote[1], X. R., Pamisetty, A., Karthikeyan, G., & Gupta[1], D. (2025, May). Artificial Intelligence Enabled Smart Energy Conservation Systems for Intelligent Resource Management and Sustainable Future Power Grids. In Proceedings of the International Conference on Sustainability Innovation in Computing and Engineering (ICSICE 24) (p. 196). Springer Nature.
[27] Singireddy, S. (2024). The Integration of AI and Machine Learning in Transforming Underwriting and Risk Assessment Across Personal and Commercial Insurance Lines. Journal of Computational Analy- sis and Applications(JoCAAA), 33(08), 3966-3991.
[28] Kummari, D. N., Burugulla, J. K. R., Malempati, M., Amistapuram, K., Garapati, R. S., & Nagabhyru, K. C. (2025, December). Enhancing Audit Compliance and Operational Efficiency in Manufacturing and Commercial Insurance Through Agentic AI and Data Engineering Frameworks. In 2025 IEEE International Conference on Communication Networks and Computing (CNC) (pp. 714-720). IEEE.
[29] Somu, B., & Inala, R. (2025). Transforming Core Banking Infrastructure with Agentic AI: A New Paradigm for Autonomous Financial Services. Advances in Consumer Research, 2(4).
[30] Kolla, S. K. (2024). Federated Machine Learning On Big Healthcare Data For Privacy-Preserving Analytics. The Review of Diabetic Studies, 175-190.
[31] Mangala, N. (2025). Agentic Data Pipelines: Autonomous ELT Orchestration Using AI Agents on Microsoft Fabric and Databricks. International Journal of Computer Technology and Electronics Communication, 8(6), 11891-11907.
[32] Srikanth, T., Segireddy, A. R., & Elavarasi, S. A. (2025, October). STaSFormer-SGAD: Semantic Triplet-Aware Spatial Flow-Guided Spatio-Temporal Graph for Anomaly Detection in Surveillance Videos. In 2025 International Conference on Communication, Computer, and Information Technology (IC3IT) (pp. 1-7). IEEE.
[33] Loganathan, R. (2024). GENERATIVE AI-ENABLED COMPLIANCE DOCUMENTATION AND AUDIT TRAIL AUTOMATION FOR GLOBAL DATA CENTER GOVERNANCE. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 15(3), 487–504. https://doi.org/10.61841/turcomat.v15i3.15512
[34] Kolla, T. (2025). The Future of Healthcare Analytics: Leveraging AI and Data Engineering for Personalized Medicine. Journal of Computer Science and Technology Studies, 7(4), 634-640.
[35] Mangalampalli, B. M. Generative AI Applications In Healthcare Data Mart Design And Optimization.
[36] Amistapuram, K. (2025). GENERATIVE AI FOR CLAIMS EXCEPTIONS AND INVESTIGATIONS: ENHANCING RESOLUTION EFFICIENCY IN COMPLEX INSURANCE PROCESSES. Available at SSRN 5785482.
[37] Segireddy, A. R. (2025). Generative Ai For Secure Release Engineering In Global Payment Network. Lex Localis: Journal of Local Self-Government, 23.
[38] Radhakrishnan, P., Nagabhyru, K. C., Manonmani, C., Srinu, M., Kaur, H., & Nandhini, N. (2025, October). K-Means-KNN Hybrid Model for Efficient Intrusion Detection in Cloud-based IoT Systems. In 2025 10th International Conference on Communication and Electronics Systems (ICCES) (pp. 1583-1588). IEEE.
[39] Garapati, R. S. (2025). An Intelligent IoT Security System: Cloud-Native Architecture with Real-Time AI Threat Detection and Web Visualization. Journal homepage: https://jmsronline. com, 2(06).
[40] Sivanand, R., Kumar, D. P., Nagabhyru, K. C., Natarajan, E. P., Pamisetty, V., & Kapila, D. (2025, September). IoT and AI for Real-Time Monitoring in Substation Automation. In 2025 International Conference on Computing and Communications (COMPUTINGCON) (pp. 1-5). IEEE.
[41] Singireddy, S. (2025, May). AI-Driven Comprehensive Insurance and AAA Membership Benefits Overview. In 2025 2nd International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE) (pp. 1-13). IEEE.
[42] Rani, P. S., Kummari, D. N., Yellanki, S. K., Meda, R., Koppolu, H. K. R., & Inala, R. (2025, July). Blockchain and AI for Securing Electrical Infrastructure. In 2025 2nd International Conference on Computing and Data Science (ICCDS) (pp. 1-6). IEEE.
[43] Vajpayee, A., Khan, S., Gottimukkala, V. R. R., Sharma, D., & Seshasai, S. J. (2025). Digital Financial Literacy 4.0: Consumer Readiness for AI-Driven Fintech and Blockchain Ecosystems. International Insurance Law Review, 33(S5), 963-973.
[44] FinOps Strategies for AI-Enabled Real-Time Compliance Platforms in Cloud Native Environments. (2025). MSW Management Journal, 35(2), 2080-2088.
[45] Mangala, N. (2022). Real-Time Data Quality Monitoring and Gating Frameworks in Cloud-Based Data Pipelines. International Journal of Research and Applied Innovations, 5(6), 8197-8219.
[46] Kolla, T. (2024). AI-Powered Data Catalog Systems For Healthcare Data Discovery And Governance. South Eastern European Journal of Public Health, 2296–2311. https://doi.org/10.70135/seejph.vi.7077

[47] Gottimukkala, V. R. R. (2025). Generative AI for Exceptions and Investigations: Streamlining Resolution Across Global Payment Systems. Journal of International Commercial Law and Technology, 6(1), 969-972.
[48] Sanku, R., Singireddy, J., Ilakkia, T., Kamala, N., & Soni, M. (2025, October). Comprehensive Analysis on Energy Efficient Transmission in Wireless Sensor Network. In 2025 International Conference on Communication, Computer, and Information Technology (IC3IT) (pp. 1-8). IEEE.
[49] Pandiri, L. (2025, May). Exploring Cross-Sector Innovation in Intelligent Transport Systems, Digitally Enabled Housing Finance, and Tech-Driven Risk Solutions A Multidisciplinary Approach to Sustainable Infrastructure, Urban Equity, and Financial Resilience. In 2025 2nd International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE) (pp. 1-12). IEEE.
[50] Chakraborty, S., Pamisetty, A., Chandana, N., & CS, B. (2025, October). Depth-Wise Temporal Convolutional Networks with Layer Normalization for Waste Food Prediction. In 2025 2nd International Conference on Software, Systems and Information Technology (SSITCON) (pp. 1-6). IEEE.
[51] Mangalampalli, B. M., Kolla, S. K., Bandi, V. D. V. K., Yandamuri, U. S., & Rani, P. S. (2025). Designing Intelligent Healthcare Ecosystems through Adaptive Data Integration and Autonomous Learning Systems. Vascular and Endovascular Review, 8(20s), 330-347.
[52] AGENTIC AI FRAMEWORKS FOR AUTONOMOUS RISK DETECTION AND COMPLIANCE REMEDIATION IN ENTERPRISE DATA CENTER OPERATIONS. (2025). Lex Localis - Journal of Local Self-Government, 23(S6), 9672-9697. https://doi.org/10.52152/3f90ak91
[53] Ranga Reddy, V. A. (2024). Comparing Batch vs. Streaming Approaches in Healthcare Data Warehousing Environments. Journal of Neonatal Surgery, 13(1), 2287–2309. Retrieved from https://www.jneonatalsurg.com/index.php/jns/article/view/10223
[54] Recharla, M., & Nuka, S. T. (2025). Translational Approaches To Commercializing Neurodegenerative Therapies: Bridging Laboratory Research With Clinical Practice. South Eastern European Journal of Public Health, 121–144.
[55] Seenu, A., Aitha, A. R., Gottimukkala, V. R. R., Singireddy, J., Meda, R., & Garapati, R. S. (2025, November). Hybrid Multi-Agent Reinforcement Learning and Blockchain Framework for Real-Time Transaction Integrity in Cloud-Driven Financial Systems. In 2025 IEEE 3rd Global Conference on Wireless Computing and Networking (GCWCN) (pp. 1-6). IEEE.
[56] Nuka, S. T., Chakilam, C., Chava, K., Suura, S. R., & Recharla, M. (2025). AI-driven drug discovery: transforming neurological and neurodegenerative disease treatment through bioinformatics and genomic research. American Journal of Psychiatric Rehabilitation, 28(1), 124-135.
[57] Krishnan, M., Aitha, A. R., Amistapuram, K., Nandan, B. P., Kaulwar, P. K., & Singireddy, J. (2025, November). Human-in-the-Loop Hybrid Neuro-Symbolic AI Model for Reliable Data Engineering in High-Stakes Industrial Systems. In 2025 IEEE 3rd Global Conference on Wireless Computing and Networking (GCWCN) (pp. 1-7). IEEE.
[58] Kummari, D. N., Challa, S. R., Pamisetty, V., Motamary, S., & Meda, R. (2025). Unifying Temporal Reasoning and Agentic Machine Learning: A Framework for Proactive Fault Detection in Dynamic, Data-Intensive Environments. Metallurgical and Materials Engineering, 31(4), 552-568.
[59] Thutari, R. T., Garapati, R. S., BM, M., & RK, S. (2025, October). Adaptive Access Control and Authentication Management for IoT Using Attention-GRU and Reinforcement Learning. In 2025 2nd International Conference on Software, Systems and Information Technology (SSITCON) (pp. 1-6). IEEE.
[60] Nigam, N., Sireesha, B., Ediga, P., Segireddy, A. R., & Bokde, S. (2025, December). Comparative Evaluation of Cloud Security Algorithms Using Multiple Classifiers with an Optimized Intrusion Detection System. In 2025 IEEE 5th International Conference on ICT in Business Industry & Government (ICTBIG) (pp. 1-6). IEEE.
[61] Amistapuram, K. (2025). Agentic AI for Next-Generation Insurance Platforms: Autonomous Decision-Making in Claims and Policy Servicing. Journal of Marketing & Social Research, 2, 88-103.
[62] Kolla, S. K. (2021). Designing Scalable Healthcare Data Pipelines for Multi-Hospital Networks. World Journal of Clinical Medicine Research, 1(1), 1-14.
[63] Velangani Divya Vardhan Kumar Bandi. (2024). Intelligent Data Platforms For Personalized Retail Analytics At Scale. Metallurgical and Materials Engineering, 30(4), 1011–1027. https://doi.org/10.63278/mme.v30i4.1938
[64] Singreddy, S. (2024). Predictive Modeling for Auto Insurance Risk Assessment Using Machine Learning Algorithms. Available at SSRN 5238922.
[65] Pandiri, L. (2025). The Complete Compendium of Digital Insurance Solutions: Life, Health, Auto, Property, and Specialized Coverage in the Age of AI, Automation, and Intelligent Risk Management. Deep Science Publishing.
[66] Kumar, S. S., Singireddy, S., Nanan, B. P., Recharla, M., Gadi, A. L., & Paleti, S. (2025). Optimizing edge computing for big data processing in smart cities. Metallurgical and Materials Engineering, 31(3), 31-39.
[67] Ramana, B., Sheelam, G. K., Pandya, T., Rai, A. K., Kumar, V. A., & Kukreti, A. (2025, December). Exploring the Potential of NOMA in 6G Through Comparative Analysis with OMA Techniques. In 2025 IEEE 5th International Conference on ICT in Business Industry & Government (ICTBIG) (pp. 1-6). IEEE.
[68] Sheelam, G. K. (2025). Deploying Neural-Symbolic Hybrid Models for Adaptive Spectrum Management in 6G-Ready Networks. Journal of Neonatal Surgery, 14(22s).

[69] Meda, R. (2025). AI-Driven Demand and Supply Forecasting Models for Enhanced Sales Performance Management: A Case Study of a Four-Zone Structure in the United States. Metallurgical and Materials Engineering, 1480-1500.

[70] Pallapu, S. R., Aitha, A. R., Vandhana, K., & Chelladurai, S. (2025, October). GAN-Augmented Transformer Framework for Cross-Domain Video Style Transfer. In 2025 International Conference on Communication, Computer, and Information Technology (IC3IT) (pp. 1-6). IEEE.

[71] Kumar, I., Nagabhyru, K. C., IG, N., MV, P., & KV, S. (2025, October). Adaptive Meta-Knowledge Transfer Network with Feature Hallucination and Attention for Low-Shot Object Detection in Aerial Images. In 2025 International Conference on Communication, Computer, and Information Technology (IC3IT) (pp. 1-6). IEEE.

[72] Kalisetty, S., & Inala, R. (2025). Designing Scalable Data Product Architectures With Agentic AI And ML: A Cross-Industry Study Of Cloud-Enabled Intelligence In Supply Chain, Insurance, Retail, Manufacturing, And Financial Services. Metallurgical and Materials Engineering, 86-98.

[73] MANGALAMPALLI, B. M., KOLLA, S. H., APPA RAO NAGUBANDI, D. R., & SEGIREDDY, A. R. (2025). AN INTELLIGENT, REAL-TIME DIGITAL FABRIC FOR HEALTHCARE AND FINANCIAL ECOSYSTEMS USING AUTONOMOUS LEARNING AND GENERATIVE SYSTEMS. TPM–Testing, Psychometrics, Methodology in Applied Psychology, 32(S9 (2025): Posted 15 December), 3070-3086.