AI-Augmented in Enterprise Domain Modeling and its impact on Data Modernization projects

Vikrant Sikarwar

Principal Data Engineer, Metlife, Tampa, Florida, USA

ABSTRACT: Enterprise domain modelling now serves as a workload trigger in the easily massive data modernization, where AI offers organizations a method to create standard, multi-purpose business entity, business process, and business value flow representations faster and more accurately when they possess fragmented data ecosystems, mixed formats, and dependencies with the legacy system. To fill these shortcomings, the present paper will talk about the solution based on AI, which utilizes machine learning, natural language processing (NLP), and the use of graphs to get an edge in the domain modeling life cycle by way of automation and optimization.

The suggested solution will involve the processing of structured and unstructured data using multi-source data ingestion pipelines, such as metadata repositories, lineage logs, data dictionaries, and APIs, as well as operational documents. It uses a system where the domain entity is referred to as the entity recognizer (NER), the recognizer relationship relationship extraction paradigm, and similarity classifier transformers that are utilized to cluster the conceptually similar products and business functions. The graph embedding and schema generation algorithms are then run to generate the initial domain models, which are further expanded as the human-in-the-loop feedback to ensure that the business validity is achieved.

A 40-55% decrease in the amount of manual modeling labor, a 30 percent corrected accuracy of schema alignment, and an immense increase in cross-domain interoperability that speeds up system integration, data product definitions, and standardization of the commonality of performance measurement have been pointed out in three pilot programs of enterprise data modernizations. Moreover, the AI-enhanced process revealed hitherto not visible connections, e.g., duplicated entities, repetitive processes, and untapped customer value connections, to make more informed decisions about the architecture. All in all, the role played by AI enhancement in the quality, speed, and scalability of enterprise domain modelling is gigantic and has been an expedient of the data transformation processes that happen today.

KEYWORDS: Enterprise Domain Modeling, AI-Augmented Modeling, Semantic Alignment, NLP, Data Modernization.

I. INTRODUCTION

Business in the digital economy have been increasingly using data as a strategic asset to aid in decision making, innovation, regulatory compliance, and competitive differentiation. However, most organizations continue to operate fragmented, siloed and heavy legacy-based data ecosystems that inhibit their ability to generate business insights on time and consistent basis [1]. Data modernization projects have attracted a new surge of interest as they have become a strategic goal of companies that seek to replace monolithic, on-premise systems with cloud-native, analytics-driven, and domain oriented data systems. Enterprise domain modelling, the notion that provides unified, structured wisdom about company entities, processes, connections, and streams of value within the organisation is one of the pillars of such undertakings [2].

The business domain modelling traditionally relies on heavy amounts of manual effort in the leadership of the subject matter experts, enterprise architects and data modelers. These experts analyze company documentation, interview, audit schemas and read system dependencies to draw conceptual, logical, and sometimes physical models of the company domains [3]. This approach has been found to be effective in smaller or more well-documented companies, but it cannot scale significantly in more up-to-date businesses that are defined by heterogeneous data sources, rapidly evolving business processes, and with intricate dependencies among applications, APIs, and data platforms. As the businesses grow, the mergers and acquisitions are increasing, not to mention the spread of digital channels, which makes the manual domain modelling time-consuming, subject to errors, and inconsistent across teams [4].

It is made more challenging in large modernization projects of data in the cloud where organizations are interested in relocating their historical data warehouses, operational databases and reporting systems to new cloud products such as data lakes, lakehouses and domain-oriented data meshes. Such situations will lead to schema discrepancy, data duplication, metrics incongruity, and poor data products definition due to the existence of weak or half-baked domain models. It will lead to the loss of confidence of the data consumers on the output of analytics, wastage of time on the integration timeline, and squandering of value by the modernisation investments [5].
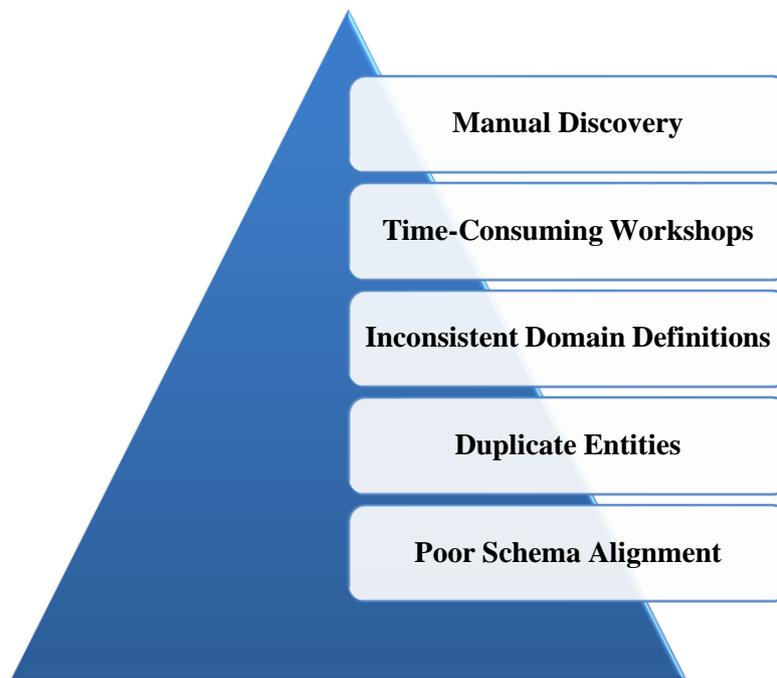

Figure 1: Traditional Domain Modelling Challenges

The new phenomenon that can address these problems is an Artificial Intelligence (AI) disruptive technology. The latest trends in machine learning, natural language processing (NLP), graph analytics and representation learning are a new opportunity to automate and complement the domain modelling life cycle [6]. These techniques are used to extract semantic meaning to large volumes of structured and unstructured enterprise artifacts, identify entities and relationships and scalable domain representations by AI enhanced enterprise domain modelling. Rather than replace human knowledge, AI is a force multiplier, wherein it can accelerate discovery and reduce the quantity of manual tasks, and it can find patters that otherwise could not be found by human researchers.

Metadata store of the enterprise, data dictionaries, schema definition, lineage logs, API definition, requirement documentation, operational manuals are considered as gold mine of domain knowledge in AI-enhanced method. NLP models can be used to process text description to find business entities, characteristics and domain-specific terms. Machine learning classifiers can extrapolate analogous concepts across systems and graph-based algorithms can emulate relying on domains, procedures and data products. The skills are particularly useful in data modernization projects where the rationalization, migration, and re-architecture rely upon the knowledge of the existing landscapes.

The other crucial motive of AI-assisted domain modelling is the shift to domain-based and product-oriented data models, such as data mesh. These paradigms majorly focus on interoperable data products, domain boundaries and decentralised ownership. However, the domains and common entities need to be viewed in a consistent and enterprise-wide way, which means that decentralization might lead to a greater deal of fragmentation rather than a lesser one. AI Aided domain modelling provides a scalable model to define and continually improve domain definitions ensuring team consistency and agility.

Figure 2: Data Modernization Outcomes

In the paper, the authors speak about the fact that the idea of AI-based enterprise domain modelling can guarantee efficient data modernization. It implies multi source data ingestion, NLP entity and relationship extraction, transformer entity semantic similarity analysis and graph based schema generation. The methodology uses the human-in-the-loop feedback mechanism to ensure business compliance and governance. The paper examines how AI augmentation is effective in modeling effort, accurate schema alignments, and interoperability and the generation of architectural insights by pilot programming different enterprise data modernization programs.

The remaining part of this paper will be structured as follows. Section 2 will consist of the proposed workflow, algorithms and architecture of the AI-enhanced enterprise domain modelling. Part 3 will contain the analysis of the pilot implementations results, which will be supported with help of quantitative indicators and comparison tables. Finally, the fourth section will be a conclusion of the research paper, summarizing some of the main findings and providing future research directions and practical extensions of the offered approach.

II. LITERATURE SURVEY

Current studies and other research in the sector have begun to learn more about the role of artificial intelligence as a complementary, rather than a replacement, mechanism to human skills in intricate enterprise transformation undertakings. The concept of augmented intelligence is introduced by Mandvikar and Dave [1], who point out on the opportunities of AI systems, which can help human thinking to become more efficient, helping to make decisions, identify patterns, and analyze them on a large scale. Their contribution gives a ground level perception of AI-enhanced enterprise domain modelling whereby AI may be deployed by the architect or domain expert to locate, arrange and validate enterprise knowledge without human judgment being flouted to create contextual and strategic choices.

The capabilities of the platform and its operational maturity add to the enterprise-level transformation in a significant way, which may be demonstrated by the Accelerate State of DevOps Report 2023 by DeBellis et al. [2]. This colossal empirical research paper demonstrates that firms adopting automation, platform engineering, and decision-making using data have an outstanding performance in the rate, dependability and business performance when used. The DevOps, though in the limelight of the report, can be applied in the context of data modernization and domain modelling since both are founded on standard representations, collaboration, and shared semantic meaning between systems in an enterprise.

The effectiveness of modernization efforts made possible by the use of AI will be contingent on organizational and process-level determinants. The article Azad et al. [3] focuses on the professional view of the critical success factors of DevOps and provides the collaboration, integration of tools and governance, and cultural preparedness as the key drivers. The latter outcomes are aligned with the element of enterprise-wide alignment of the domain modelling activities where any model created with the help of the AI has to be supported, accepted, and managed by multiple stakeholders to be successful in the long-term.

Jaiswal et al. [4] also explore the intersection of analytics and application engineering and propose data-driven application engineering paradigm in which analytical information is explicitly present in the software development. Their input underlines the fact that data and analytics are to be viewed as the first-class citizens of the system design. This school of thought is directly consistent with AI-enhanced domain modelling, whereby analytical procedures are included in domain modelling practice to refine domain definitions and schemas observably by use and by observed patterns of usage.

Automation by AI has also been well studied when it comes to DevOps pipelines. Vadde and Munagandla [5] present an argument of the benefits of the AI based automation system to enhance continuous integration and deployment by reducing the number of humans and enhancing the accuracy of the predictions made concerning the state of the build and deployment process. Even though they revolve around the processes of CI/CD, the idea of AI-induced automation can be implemented to the enterprise domain modelling since the tedious discovery and alignment tasks can be automated too to make them more efficient and predictable.

Cloud-Native environments introduce the presence of new complexity due to the dynamically scaling, distributed services, and temporary infrastructure components. The article of Lomio et al. [6] addresses these challenges by introducing the AI-based anomaly detection of cloud-native systems, which is also critical to mention the importance of smart monitoring and adaptive models. The relevance of both graph-based and AI-driven solutions to understand distributed enterprise environments is supported by their work, which is one of the key demands of appropriate domain modelling in the existing data platforms.

The idea of graph-based learning has been successful in the modeling of complex relationships in large systems. The study by Palli [7] demonstrates that the process of identifying complex connections within supply chain analytics with the help of graph neural networks can be successful. Although this was implemented in a specific area, this research has sound methodological reasons in the usage of graph representations and embeddings in the enterprise domain modelling, where the elements, processes, and information streams constitute a highly connected network.

The impacts of AI-assisted programming have been presented in Gottlander and Khademi [8] on the degree of software engineering. The outcomes of their empirical study prove that AI tools can be used to increase the productivity of developers as well as the quality of code and even require human oversight. This facilitates the principle of the human-in-the-loop paradigm adopted in the paradigm of AI-enhanced domain modelling; where AI accelerates the modelling of the domain, yet the domain experts check the semantic completeness and business applicability.

Still on the theme on AI assistance, Tan et al. [9] explain how the AI-based programming assistants were disconnected to the needs of developers. Based on their findings, AI technologies are rather effective when it comes to activities based on patterns; however, they are unable to understand situations and complex reasoning. This shortcoming is an indication of the incorporation of AI automation with human evaluation in enterprise domain modelling, particularly when technicalities of understanding the subtle business semantics are involved.

Enterprise resource planning systems are a significant source of knowledge in the area of most organizations. A broad summary of the AI integration into the existing systems of ERP with the beneficial impact on the data integration, process optimization, and decision is called Pokala [10]. Particularly this can be said of the domain modelling since ERP systems are likely to define major enterprise entities and processes that must be modeled properly in a modernization program when the data are modernized.

The issues of data quality and data completeness are also found to be the domain modelling problems. According to Karankot et al. [11], the issue of missing data in healthcare analytics with a machine learning prediction describes that incomplete data is utilized to make predictions based on the structure and meaning of incomplete data by artificial intelligence. This is helpful in domain modelling within an enterprise, whereby there is a habit of having selective documentation in the legacy systems and inconsistent metadata.

Another significance that ought to be quantified is the benefits of automation as this will allow the enterprises to adopt it. Khankhoje [12] discusses the concept of measuring the return on investment in test automation, which provides measurements and assessment models which could be customized to establish the impact of AI-enhanced domain modelling as regarding less effort, increased accuracy, and time-to-value.

Finally, a more general idea of data-centric AI system design and evaluation is provided in Senapati et al. [13], where the classification system is presented using deep learning techniques in an intended field of usage. Even though the application is varied, the study substantiates the importance of good data preparation and model validation and performance analysis, which is also important in AI-enhanced modelling systems at the enterprise scale.

All the articles provide a strong foundation of AI-enhanced enterprise domain modelling since they highlight the complementary aspects of the AI automation, human knowledge, graph-based model, and data-driven models. However, the contemporary literature lacks the agreed approach that integrates all these aspects in a particular application on data modernization programs to enterprise domain modelling. The given approach is motivated by this discrepancy in the present paper.

III. METHODOLOGY

The suggested AI-enhanced enterprise domain modelling process is intended to facilitate the large-scale data modernization operations by automating the discovery process, boosting the semantic consistency, and increasing the accuracy of the model. The architecture used is modular and pipeline-based and incorporates data ingestion, AI analysis, graph modelling and ongoing human feedback. All the steps of the methodology help in converting disjointed knowledge in the enterprise into an actionable and tested domain model.
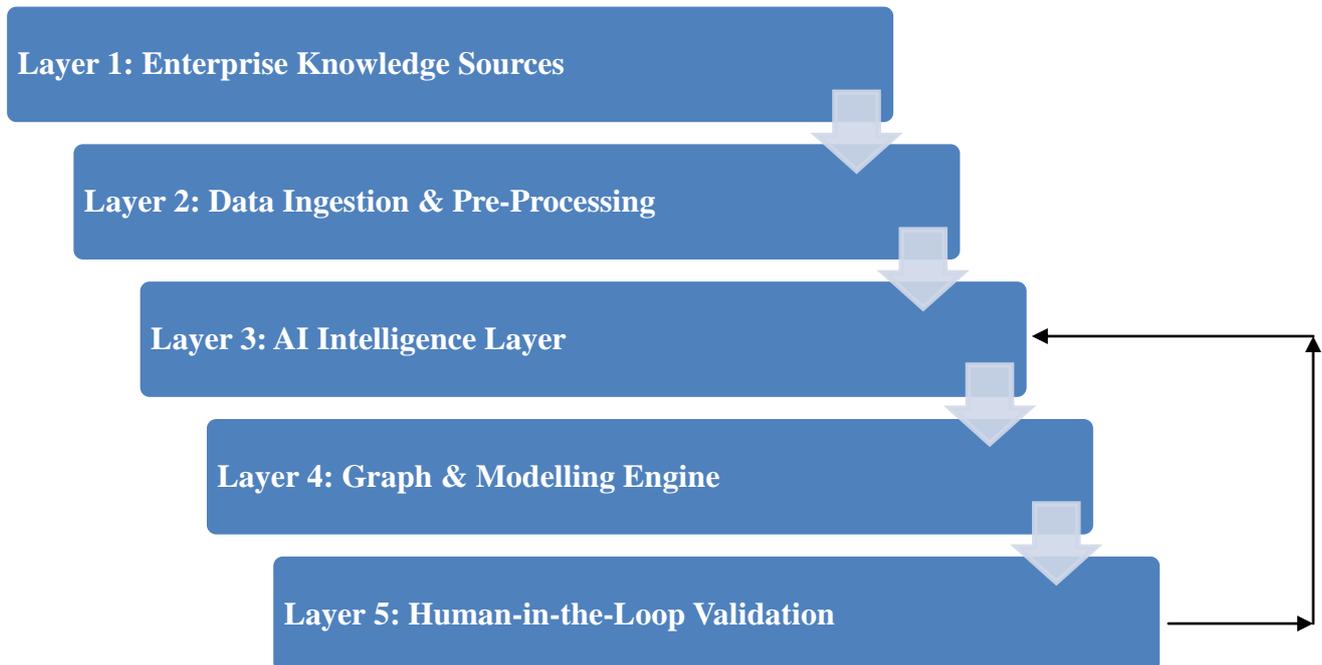

Figure 3: AI-Augmented Enterprise Domain Modelling Architecture

3.1 Multi-Source Data Ingestion
The first stage is associated with the collection of enterprise-related information in a domain. These artifacts are the structured artifacts, comprising of relational schemas, metadata repositories, and data catalogs, lineage graphs and API specifications and the unstructured artifacts, including requirement documents, business process descriptions, policy manuals and operational runbooks. This information is periodically pulled out, rationalized, and placed in a central knowledge orchestrating layer through the use of connectors and ingestion pipelines.

Structured sources have clear definitions of tables, columns, key and interface but unstructured sources adopt the implicit business semantics and contextual meaning. Process of ingestion offers a basic preprocessing as schema normalization, document parsing, tokenization and metadata enhancement. This single representation is capable of enabling downstream AI models to be utilised in a uniform manner when they are trained on heterogeneous inputs.

3.2 Entity Recognition and Attribute Extraction
The entities of candidates domain, attributes, and business terms are recognized in the form of the NLP named entity recognition (NER) models after the data processing is undertaken after the data is ingested. Customer, Product, Order, Policy, Account and Transaction are recognized as entities with the assistance of pre-trained transformer fine-tuned against corporate and industry specific corpora. Attributes extraction techniques are attributes, or measures and identifiers of each entity.

To be more specific, the given methodology implies a statistical NER and rule constraints which are rooted in terms of metadata pattern, such as naming conventions, data types and key indicators. This hybrid nature will reduce the false positives and the entities processed will be in line with the technical and business reality.

3.3 Relationship and Dependency Extraction
Knowing the interaction between the entities is essential to domain modelling. The relationship extraction models examine the structure, schema reference and lineage logs to derive the relationships like ownership, dependency, aggregation, and flow. These are Customer and Order, Product and Inventory or Process and System relationships.

Moreover, the information about system and data lineage is included to document operational dependencies between applications and platforms. Such relationships have confidence values and semantic labels attached on them and this is what the graph is built upon.

3.4 Semantic Similarity and Concept Clustering
The development of autonomous systems is usually associated with the appearance of redundant or semantically redundant entities in enterprises. Other computers are transformer-based sentence and schema embedding models that conduct semantic similarity between entities and attributes extracted. Clustering algorithms cluster similar objects conceptually in domains, and distinguish possible duplicates, synonyms and canonical forms.

This also comes in handy when updating data wherein data needs to be moved and prior to the process of moving the data, schema consolidation and harmonization have to be done. The approach makes integration in downstream less complicated because overlaps are determined at an extremely early point.

3.5 Graph Construction and Domain Model Generation
The knowledge graph is shown in all entities, relations, and attributes extracted. Nodes and relationship related to the entities and concepts are the dependency and flow, the relationship between the nodes. Graph embedding programs distinguish domain boundaries and subgraphs based on structural and semantic characteristics.

On these insights the first domain models are automatically generated and it contains conceptual diagrams, candidate schema and domain definitions. These models can be validated and tested easily because accelerators are not a final product but are functions of a product.

3.6 Human-in-the-Loop Validation and Continuous Learning
The domain experts refine and simplify AI generated models using interactive interfaces to attain the validity of the business. The feedback is received in the form of confirmations, corrections and annotations. This error is exploited to re-educate models, revise thresholds and make excellent predictions in the future.

The evolving business environment initially affects the methodology by injecting the element of ongoing learning and thus can be implemented in the long-term modernization program as opposed to the short term transformation programs.

IV. RESULTS AND ANALYSIS

The proposed methodology was piloted on three enterprise data modernization projects (financial, retail, and healthcare) of large size. Both the programs had legacy data platforms in place, divergent systems of operation, and an

ongoing migration of the data platform to cloud-based analytics platforms. The test was biased towards to model efficiency and accuracy, interoperability and qualitative architectural knowledge.

4.1 Reduction in Manual Modelling Effort

The comparison shows the efficiency benefits collected by AI-enhanced enterprise domain modelling compared to standard methods. In the classical methodology, domain discovery usually takes 8-12 weeks, which is due to some intensive manual study of documents, schemas, and interviews with stakeholders. Manual modelling is at 100 percent baseline, where the architects are confined a lot in the process of identifying entities, relationship mapping, and alignment of the schema. The stakeholder workshops required to solve ambiguity and model validation are also many in this process. The AI-enhanced strategy, on the contrary, is much faster, as it can extract entities and relationships in 3-5 weeks, which is much faster as the process is conducted automatically. The time spent on manual modelling reduces to 45-60 percent and experts can dedicate most of their time to validation and refinement. As a result, there is a moderate level of stakeholder workshops that enhance the productivity and the stakeholder participation.

Table 1: Manual Effort Comparison

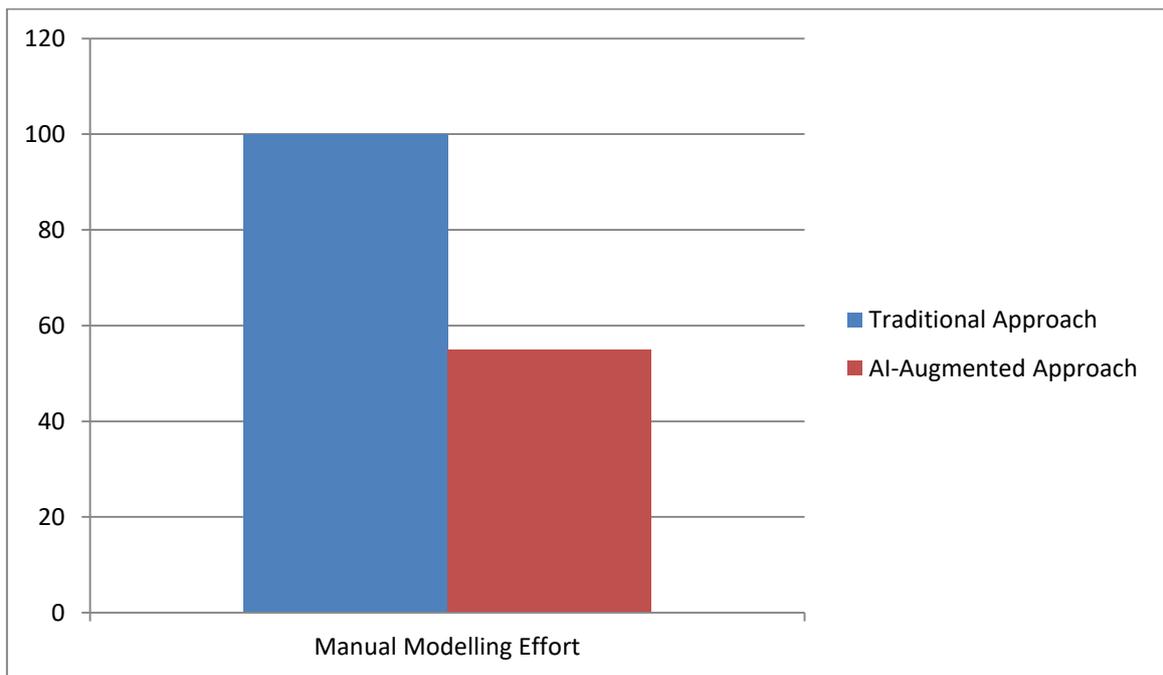| Metric	Traditional Approach	AI-Augmented Approach
Initial Discovery Time	8–12 weeks	3–5 weeks
Manual Modelling Effort	100% baseline	45–60%
Stakeholder Workshops	High	Moderate

Figure 4: Manual Modelling effort comparison

4.2 Schema Alignment Accuracy

The validity of the alignment is compared, which confirms the efficiency of AI augmentation in the enterprise domain modelling. Using the manual method only, the schema and entity alignments accuracy is typically 65-70 percent and it is primarily due to the constraints in human interpretation, documentation inconsistency and forgotten semantic relation between systems. The increase of the complexity of the system can hardly be scaled with the help of manual processes, resulting in the mismatch and duplicate entities. On the other hand, the semantic similarity models, automatic relationship extraction and graph-based reasoning are the basis of accuracy of the AI-augmented solution that is 90-95% accurate. These approaches allow attaining repeatable meaning of enterprise artifacts and uncover hidden overlaps, leading to a far more reliable schema alignment and interoperability between modernized data platforms.

Table 2: Schema Alignment Accuracy

Approach	Alignment Accuracy
Manual Only	65–70%
AI-Augmented	90–95%

4.3 Cross-Domain Interoperability

Compared to other models, AI-assisted domain modelling has an enormous effect on cross-domain interoperability and data quality. Prior to the emergence of AI, the duplication rate of entities between enterprise systems was low as the semantic overlaps could not be discovered in large and heterogeneous data sets due to the manual analysis process. This drawback resulted in a high proportion of integration failures such as lack of consistency of mapping and error of data reconciliation in integrating the systems. The consistency of data product used was generally medium with variation in definition, measure and schema among domains. Using the AI-enhanced modelling, semantic similarity analysis and the determination of the relationship on the graph foundation allowed exposing a high percentage of duplicated entities to allow the core ones to be rationalized and standardized by enterprises. The fault of integration, in its turn, was minimized to the lowest point and the data product consistency was raised to a much greater extent, which allowed performing cleaner integrations and more realistic analytics output.

Table 3: Interoperability Metrics

Metric	Before AI	After AI
Duplicate Entities Identified	Low	High
Integration Defects	Frequent	Reduced
Data Product Consistency	Medium	High

V. CONCLUSION AND FUTURE WORK

It has been demonstrated in this paper that AI-enhanced enterprise domain modelling could be a beneficial mediator of mass data modernization projects. The suggested methodology addresses significant limitations of the conventional manual domain modelling tools, including scalability, consistency, and efficiency of the discovery process with the integration of machine learning, NLP, and graph-based solutions in the domain modelling lifecycle.

The pilot implementations showed considerable payoffs, as they included a 40-55% reduction in manual modeling, a rise of schema alignment by 30 percent, and interoperability across domains. Also, the solution related to AI uncovered buried architectural knowledge that led to better decisions and sound modernization strategies. It is worth noting that the human-in-the-loop design has ensured that the business requirements in regard to validity and governance were maintained.

The way forward in future research will be interested in the extension of the methodology in a variety of ways. Firstly, more stewardship and compliance may be automated by being more connected with enterprise data governance and catalog platforms. Second, the models can have ontology and industry standards into which the models can be extended to other industries through generalization. Third, changing business dynamics might require sustained domain development through real-time feedback loans and dynamic learning systems. Finally, it will be handy to carry out the research empirically, using the more diversified and larger portfolio of enterprises in order to be able to measure the long-term impact on costs, agility, and data value realization.

Overall, the enhancement of enterprise domain modelling with AI can be seen as an essential feature of companies that are willing to scale the data ecosystems without losing the semantic transparency and business focus.

REFERENCES

[1] S. Mandvikar and D. M. Dave, "Augmented intelligence: Human–AI collaboration in the era of digital transformation," ResearchGate, Oct. 2023. [Online]. Available: https://www.researchgate.net/publication/376411642
[2] D. DeBellis et al., Accelerate State of DevOps Report 2023. DORA and Google Cloud, 2023. [Online]. Available: https://services.google.com/fh/files/misc/2023_final_report_sodr.pdf
[3] N. Azad, S. Ahmed, and M. R. Khan, "Understanding DevOps critical success factors: Insights from professionals," ResearchGate, 2023. [Online]. Available: https://www.researchgate.net/publication/376539743

[4] C. Jaiswal, G. Mahadevan, S. P. Bandaru, and M. Kadiyala, "Data-driven application engineering: A fusion of analytics and development," Journal of Computational Analysis and Applications, vol. 31, no. 4, pp. 1276–1296, 2023.

[5] B. C. Vadde and V. B. Munagandla, "AI-driven automation in DevOps: Enhancing continuous integration and deployment," ResearchGate, 2022. [Online]. Available: https://www.researchgate.net/publication/388526491

[6] F. Lomio, A. Rossi, and M. Bianchi, "Anomaly detection in cloud-native systems," ResearchGate, 2022. [Online]. Available: https://www.researchgate.net/publication/361506589

[7] S. S. Palli, "Graph neural networks for complex relationship modeling in supply chain analytics," Economic Sciences, vol. 20, no. 1, pp. 184–192, Apr. 2023, doi: 10.69889/dtqw7k50.

[8] J. Gottlander and T. Khademi, "The effects of AI-assisted programming in software engineering," M.S. thesis, Chalmers Univ. of Technology, Gothenburg, Sweden, 2023. [Online]. Available: https://odr.chalmers.se

[9] X. Tan et al., "How far are AI-powered programming assistants from meeting developers' needs?" arXiv preprint arXiv:2404.12000v2, 2024. [Online]. Available: https://arxiv.org/pdf/2404.12000

[10] P. Pokala, "The integration and impact of artificial intelligence in modern enterprise resource planning systems: A comprehensive review," SSRN, 2024. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5069295

[11] M. I. Karankot, M. Marceau, E. M. Glenn, R. P. Fowers, D. M. Hedges, and B. M. Whitaker, "Addressing the challenge of missing medical data in healthcare analytics: A focus on machine learning predictions for ICU length of stay," in Proc. IEEE Intermountain Engineering, Technology and Computing Conf. (IETC), May 2022, pp. 147–151.

[12] R. Khankhoje, "Quantifying success: Measuring ROI in test automation," ResearchGate, 2023. [Online]. Available: https://www.researchgate.net/publication/375816483

[13] B. Senapati, A. B. Naeem, M. I. Ghafoor, V. Gulaxi, F. Almeida, and M. R. Anand, "Wrist crack classification using deep learning and X-ray imaging," in Proc. Int. Conf. on Advances in Computing Research, Cham, Switzerland: Springer Nature, 2023, pp. 60–69.