Sarcouncil Journal of Multidisciplinary

ISSN(Online): 2945-3445

Volume- 05| Issue- 09| 2025

Research Article **Received:** 27-07-2025 | **Accepted:** 17-08-2025 | **Published:** 14-09-2025

Leveraging Large Language Models for Automated Schema Matching and Data Transformation in Enterprise ETL Pipelines

Annapurneswar Putrevu

Independent Researcher, USA

Abstract: Modern enterprises encounter significant challenges when integrating heterogeneous data sources due to schema variability, inconsistent naming conventions, and noisy real-world datasets that traditionally require extensive manual intervention. This work presents a novel generative AI framework that employs Large Language Models augmented with retrieval-based techniques to automate schema matching, column-level mapping, and transformation rule generation within ETL pipelines. The framework incorporates metadata-aware prompting strategies, domain-specific exemplar retrieval through RAG, and iterative self-refinement mechanisms to produce high-quality mapping suggestions alongside executable transformation code in SQL and pandas formats. Confidence scoring enables effective human-in-the-loop validation while adaptive feedback mechanisms facilitate continuous improvement from user corrections. The system demonstrates the capability of handling evolving schemas and multilingual datasets across diverse enterprise domains. Experimental evaluation on synthetic datasets and established benchmarks reveals substantial improvements in matching accuracy, precision, and recall compared to traditional name-similarity metrics and classical machine learning classifiers. The framework addresses critical deployment considerations, including data privacy compliance, operational scalability, and hallucination mitigation strategies. Results indicate significant potential for reliable large-scale enterprise adoption through transparent, auditable automated ETL workflows that maintain data quality standards while reducing manual overhead.

Keywords: Schema matching, Large Language Models, ETL automation, Retrieval-Augmented Generation, Data integration.

INTRODUCTION

Data Integration Complexities in Contemporary ETL Systems

Modern organizations navigate increasingly complex data environments that include multiple technology platforms within and across, relational database management systems, document NoSQL, cloud-native storage frameworks, and third-party service interfaces. Each platform creates distinct structural paradigms and semantic constructs that introduce considerable barriers to any unified data processing. These technological differences cause complex integration challenges where traditional Extract-Transform-Load architectures cannot create consistent data pathways between heterogeneous systems, particularly when coordinating large sourcing enterprise operations with continuously flowing data that can't be easily staged, transformed, integrated, and loaded into a traditional enterprise data warehouse.

Structural Inconsistencies and Data Quality Challenges

Enterprise data ecosystems often show vast inconsistencies in attribute naming conventions, data type details, and organizational hierarchies that obfuscate meaning from semantically similar information elements. Production datasets frequently contain imperfections, including incomplete records, formatting differences, and structural changes that arise continually as business processes change in response to shifting operational requirements. Therefore, sophisticated identification of correspondences among data characteristics that capture semantic matches in disparate data structures is required, while maintaining accuracy through variations in quality and evolving structure.

Constraints of Established Schema Alignment Methods

Historical schema correspondence techniques have predominantly utilized string similarity calculations, correlation-based statistical measures, and predefined logical rules to establish field relationships (Parciak, M. *et al.*, 2024). While these methodologies function effectively within structured environments containing standardized data formats, they reveal substantial deficiencies when deployed against complex organizational scenarios. Conventional approaches often miss nuanced semantic associations, experience difficulties processing domain-specific abbreviations or specialized terminology, and require comprehensive parameter adjustment to achieve acceptable accuracy across varied application contexts.

***Corresponding Author:** Annapurneswar Putrevu

DOI- https://doi.org/10.5281/zenodo.17117079

Table 1: Traditional vs. Generative AI Schema Matching Approaches (Parciak, M. *et al.*, 2024)

Aspect	Traditional Methods	Generative AI Framework
Matching Basis	String similarity, statistical correlation	Semantic understanding, contextual analysis
Domain Adaptation	Manual rule configuration	Automatic domain knowledge retrieval
Multilingual Support	Limited to predefined dictionaries	Native cross-lingual comprehension
Learning Capability	Static rule-based systems	Adaptive feedback integration
Code Generation	Manual transformation scripting	Automated executable code synthesis
Scalability	Linear performance degradation	Distributed processing capabilities

Human Oversight Dependencies in Data Processing Workflows

Current enterprise data integration operations maintain substantial dependence on manual validation processes, generating operational limitations that affect organizational productivity. Technical staff must allocate considerable time and resources toward reviewing automatically generated correspondence suggestions, correcting misidentified relationships, and developing custom conversion procedures for intricate data transformation scenarios. These human-dependent activities introduce processing delays, limit system expansion capabilities, and may produce variation across concurrent integration projects within organizational boundaries, ultimately restricting the adaptability required for dynamic analytical applications.

Framework Goals and Technical Contributions

Recent advancements in generative language model architectures have demonstrated substantial proficiency in interpreting natural language semantics and generating contextually appropriate responses across diverse application areas (Sheetrit, E. *et al.*,. 2024). This framework seeks to harness artificial intelligence capabilities for comprehensive automation of schema correspondence identification, field-level mapping generation, and transformation rule construction within enterprise data processing pipelines. Core contributions include developing metadata-enriched prompt engineering strategies that effectively communicate structural information to language models, implementing retrieval-augmented systems for specialized domain knowledge incorporation, and establishing iterative improvement cycles that progressively enhance mapping accuracy through automated verification and adjustment procedures.

TECHNICAL FRAMEWORK AND SYSTEM DESIGN

Artificial Intelligence-Based Integration Architecture

The established architecture is based on generative artificial intelligence technologies to solve schema matching problems in various data spaces. Multiple interconnected modules run together in the whole schema alignment process, starting from source review, up to the creation of rules. The design emphasizes modularity and system extensibility, enabling the system to easily integrate into existing enterprise environments and adapt to future technologies and specific domain needs.

Table 2: Framework Architecture Components and Functions (Al-Onaizan, Y. *et al.*, 2024)

Component	Primary Function	Input Sources	Output Products
Schema Analysis Module	Structural metadata extraction	Source/target databases	Structured schema representations
Retrieval Engine	Domain-specific knowledge access	Historical mappings, examples	Contextual correspondence patterns
Mapping Generator	Correspondence identification	Schema representations, context	Field-level mapping suggestions
Code Synthesizer	Transformation logic creation	Mapping specifications	Executable SQL/pandas scripts
Validation Subsystem	Quality assurance verification	Generated mappings	Confidence scores, error flags
Feedback Processor	Learning from corrections	Human validation inputs	Updated mapping strategies

Enhanced Language Models with Information Retrieval Components

The system utilizes advanced transformer-based architectures supplemented by retrieval components that extend operational knowledge beyond original training datasets (Al-Onaizan, Y. *et al.*, 2024). These enhanced configurations merge inherent natural language comprehension abilities

Publisher: SARC Publisher

with dynamic information access from specialized knowledge repositories. Such integration permits real-time consultation of contextual examples and historical mapping records throughout correspondence identification procedures, consequently improving precision and maintaining consistency across varied integration contexts.

Context-Enriched Prompt Engineering for Structural Comprehension

Effective structural interpretation demands sophisticated prompt construction methodologies capable of transmitting complex organizational information to language processing systems. The framework deploys context-enriched prompting approaches that systematically encode structural characteristics, including field types, validation rules, relational dependencies, and semantic descriptors within organized prompt formats. These communication strategies enable language models to establish a thorough comprehension of source and destination structural properties, supporting accurate semantic correspondence detection and transformation specification across heterogeneous datasets.

Knowledge-Enhanced Generation through Dynamic Information Access

The platform integrates knowledge-enhanced generation capabilities that actively consult repositories containing specialized mapping examples and transformation patterns during correspondence processes (Zhang, J. 2025). This methodology enables the utilization of historical knowledge from previous integration initiatives and established mapping conventions to guide current schema alignment decisions. The information access component maintains indexed collections of validated mapping instances, transformation blueprints, and domain terminology that can be dynamically consulted to improve generated mapping quality and relevance for particular application domains.

Cyclical Enhancement Processes for Correspondence Quality Optimization

The architecture implements cyclical enhancement procedures that facilitate ongoing improvement of mapping quality through automated evaluation and correction mechanisms. These refinement cycles assess initial correspondence suggestions against multiple evaluation criteria, including semantic coherence, structural compatibility, and transformation viability. Upon identifying potential discrepancies, the system automatically initiates corrective procedures that may include alternative correspondence generation, confidence metric adjustment, or supplementary information retrieval operations to identify more appropriate mapping patterns.

Learning-Based Feedback Integration and Expert Validation Workflows

The framework incorporates learning-based adaptation capabilities that capture and utilize expert feedback to progressively enhance mapping precision across operational periods. Expert validation workflows provide structured mechanisms for domain specialists to examine, modify, and approve automatically generated correspondences while contributing valuable correction information to the system. These feedback integration processes enable the platform to learn from user modifications and preferences, gradually adjusting mapping strategies to better conform with organizational standards and domain-specific requirements.

IMPLEMENTATION ARCHITECTURE AND CORE COMPONENTS

Correspondence Detection and Field-Level Association Procedures

The implementation establishes advanced procedural mechanisms that autonomously detect relationships between source and destination structural elements through contextual interpretation and pattern analysis techniques. These procedures examine structural metadata, attribute nomenclature, data classifications, and validation specifications to create precise field-level associations without requiring human guidance. The platform employs diverse procedural methodologies, including proximity calculations, contextual vector comparisons, and rule-based verification to guarantee thorough identification of potential association scenarios across varied structural configurations (Kanagarla, K. 2025).

Executable Script Creation and Transformation Logic Development

The platform constructs syntactically accurate and operationally viable transformation scripts across multiple programming environments and database query languages, encompassing SQL command generation and pandas framework operations for Python development contexts. Script creation modules examine detected associations and autonomously develop transformation procedures that manage data classification conversions, format standardization, and structural reorganization necessary for effective data consolidation. These constructed scripts maintain operational

Publisher: SARC Publisher

compatibility with established data processing frameworks while incorporating exception management and performance optimization techniques to guarantee dependable execution within production systems.

Reliability Assessment Systems for Association Recommendations

The platform establishes thorough reliability evaluation systems that assess the dependability and precision of created association recommendations through diverse assessment parameters. These evaluation procedures consider elements including contextual similarity strength, structural compatibility measurements, historical association performance records, and compliance verification to generate numerical reliability indicators for each suggested correspondence. The reliability measurements enable recommendation prioritization and facilitate informed decision-making throughout expert review procedures (Saripalle, R. *et al.*, 2025).

Dynamic Structure Management and International Dataset Processing

The implementation manages dynamic structural evolution through version monitoring capabilities and flexible association strategies that accommodate structural modifications across operational periods. The platform maintains structural version records and automatically identifies changes in source or destination frameworks, initiating re-association procedures when required. Furthermore, international dataset compatibility enables processing of global data sources through language-sensitive contextual interpretation and cross-linguistic correspondence detection, expanding the framework's utility across multinational enterprise environments.

Multi-Domain Enterprise Connectivity Features

The framework exhibits robust multi-domain connectivity features that facilitate structural matching across different business sectors and industry categories. Domain-specialized knowledge databases and terminology collections support accurate correspondence detection between specialized vocabularies and technical naming systems. The platform accommodates varying data modeling practices and industry specifications while maintaining uniformity in association quality across different organizational environments and application sectors.

Development Prototype Specifications and Architectural Framework

The development prototype creates a component-based architecture containing separate processing elements including structural analysis modules, association creation engines, script development units, and verification subsystems. Each element functions independently while maintaining standardized communication protocols for seamless integration and data exchange. The architecture enables horizontal expansion through distributed processing capabilities and provides enhancement mechanisms for incorporating additional language processing models, retrieval frameworks, and domain-specialized knowledge resources as organizational needs develop.

Table 3: System Implementation Technical Specifications (Kanagarla, K. 2025)

Technical Component	Implementation Details	Supported Formats	Performance Characteristics
Language Models	Transformer-based architectures	Natural language, structured data	Context window optimization
Retrieval Systems	Vector similarity search	Embeddings, metadata	Sub-second query response
Code Generation	Template-based synthesis	SQL, Python pandas	Syntax validation included
Confidence Scoring	Multi-criteria assessment	Numerical confidence ranges	Threshold-based filtering
Version Management	Schema change detection	Temporal schema snapshots	Incremental update support
API Integration	REST/GraphQL endpoints	JSON, XML data formats	Horizontal scaling enabled

PERFORMANCE TESTING AND COMPARATIVE EVALUATION

Testing Protocols on Constructed and Established Benchmark Collections

The validation approach establishes thorough examination protocols utilizing both artificially generated datasets and recognized public benchmark repositories to guarantee

Publisher: SARC Publisher

comprehensive performance assessment across varied operational scenarios. Artificially constructed collections provide controlled testing environments for systematic evaluation of specific platform capabilities, while established benchmark repositories deliver standardized comparison foundations against current methodologies. These testing environments incorporate multiple data complexity dimensions, structural variation patterns, and specialized domain characteristics to comprehensively assess platform effectiveness under realistic operational circumstances (Zhang, Z. *et al.*, 2025).

Methodological Comparison: Conventional Similarity Calculations versus Classification Algorithms versus Prompt-Based Models versus Augmented Information Access Systems

The examination framework executes systematic evaluations across four separate methodological classifications including conventional string proximity calculations, established machine learning classification methods, prompt-engineered model configurations, and information-augmented retrieval systems. Each methodological classification receives identical testing protocols using uniform datasets and assessment standards to guarantee equitable performance evaluations. This thorough comparison structure facilitates objective evaluation of relative capabilities and constraints across different technological methodologies for structural correspondence detection challenges.

Table 4: Performance Comparison Results (Zhang, Z. *et al.*, 2025).

Methodology	Matching Accuracy	Precision Score	Recall Score	Processing Time
String Similarity	Moderate	Low-Moderate	Moderate	Fast
Classical ML	Moderate-High	Moderate	Moderate-High	Moderate
Prompt-based LLMs	High	High	High	Moderate-Slow
Retrieval-augmented	Very High	Very High	Very High	Moderate
Self-improving Systems	Highest	Highest	Highest	Variable

Effectiveness Measurements: Correctness Ratios, Specificity, and Coverage Calculations

The evaluation structure utilizes recognized effectiveness measurements, including correctness ratios, specificity assessments, and coverage calculations, to measure system performance across different operational contexts. Correctness measurements evaluate the comprehensive accuracy of produced correspondences, while specificity indicators examine the fraction of accurate correspondences among all produced recommendations. Coverage calculations establish the platform's capacity to detect all legitimate correspondences within specified datasets, delivering a comprehensive scope evaluation of platform capabilities (Kadakia, J. J. 2025).

Structured Assessment of Information-Enhanced and Adaptive Language Processing Systems

The validation framework incorporates specialized examination protocols for information-enhanced configurations and adaptive language processing implementations to assess their performance benefits over standard methodologies. Information-enhanced systems receive testing to establish the effectiveness of dynamic knowledge consultation during correspondence detection operations. Adaptive implementations undergo evaluation through cyclical testing situations that examine improvement capabilities across multiple operational periods and learning cycles.

Numerical Results and Statistical Reliability Verification

The experimental framework incorporates thorough statistical evaluation protocols to confirm the significance and dependability of observed performance distinctions between methodological strategies. Statistical reliability verification encompasses confidence range calculations, hypothesis examination protocols, and significance threshold establishments to guarantee that observed performance enhancements represent authentic methodological benefits rather than random fluctuation. Numerical results undergo a comprehensive statistical assessment to support dependable conclusions concerning platform effectiveness and comparative performance across different operational environments.

PRODUCTION DEPLOYMENT AND ORGANIZATIONAL COMPLEXITIES

Data Governance Adherence and Legal Framework Compliance

Enterprise deployment requires complete compliance with data governance policies and legal regulations that govern information usage in multiple jurisdictions. Organizations must build compliance structures to handle territorial data requirements, permissions control policies, and use documentation requirements during

Publisher: SARC Publisher

correspondence detection activities. The platform will need to use privacy-preserving methods to limit exposure to sensitive information during association identification, while preserving working effectiveness within a variety of legal jurisdictions (Benedetti, A. U. 2025).

Economic Investment and Infrastructure Scaling Evaluation

Implementation contexts require a comprehensive evaluation of economic investment patterns and infrastructure scaling potential to guarantee sustainable deployment across organizational environments. Economic considerations include computational resource consumption, data storage needs, network capacity utilization, and staff development investments required for effective platform operations. Infrastructure scaling evaluation must examine capacity limitations, performance deterioration trends, and technology requirements that develop as organizational data quantities and integration complexity expand across operational periods.

Inaccuracy Prevention Methods and System Reliability Measures

The platform requires comprehensive inaccuracy prevention methods to address potential errors and false outputs that may occur during automated association detection procedures. System reliability measures must incorporate verification stages, cross-validation processes, and confidence parameter modifications to reduce incorrect mapping recommendation risks. These methods include multiple confirmation layers encompassing semantic coherence verification, structural compatibility testing, and historical trend analysis to guarantee dependable operational effectiveness (Ajibode, A. *et al.*, 2025).

Uniform Evaluation Procedure Development

Effective organizational adoption requires the development of uniform evaluation procedures that facilitate consistent performance assessment across different organizational environments and data contexts. Procedure development includes creating standardized testing methodologies, reference dataset identification, effectiveness measurement specifications, and assessment standards that enable objective evaluation across different implementation situations. These procedures must accommodate diverse organizational needs while maintaining evaluation uniformity and reliability.

Protected, Observable, and Accountable Automated Processing Systems

Implementing at the organizational level requires the capability to establish 'protected' systems of processing that maintain formal observability and accountability throughout the automated association identification process. System design should allow the organization to maintain sufficient record-keeping, decision-making records, and validation steps, so these practices will enable organizational oversight and potential legal compliance. Observability means thorough documentation of mapping justification, confidence metric calculations, and justification and processes for human examination and verification.

Large-Scale Organizational Implementation Readiness Assessment

Full organizational deployment necessitates an extensive readiness assessment that encompasses the following elements: technical infrastructure readiness, human resource development needs, and capability with regard to operational integration. Preparation for implementation entails assessing existing system interoperability, complementary data management framework coordination, and the levels of change necessary to enable viable platform deployment for potential users. Assessment methods must scrutinize organizational capacity factors, technical knowledge availability, and available support infrastructure to ensure sustained implementation success.

CONCLUSION

In this work, developed a generative AI framework that solves significant problems for schema matching and data transformation for ETL pipelines in enterprises. The framework successfully brought together large language model capabilities with retrieval-augmented generation, metadata-aware prompting, and iterative self-improvement techniques to automate complex data integration processes that largely require human resources, for example, subject matter expertise to map the original schema with the target schema. The implementation outcomes provided improvements to matching accuracy, precision, and recall over similarity-based methods and classical machine learning classifiers. The system successfully managed dynamic schemas, cross-lingual datasets, and cross-domain integrations and produced executable transformation code inside many coding environments. Deployment topics related to data privacy, operational costs, and hallucination mitigation strategies were documented, with

Publisher: SARC Publisher

enterprise readiness considerations completed. As a component of the framework, adaptive feedback approaches will allow the framework to be a continuous learning system through humans correcting their mistakes, with the mapping to rise in quality after operational time. The assessment of the system's performance with synthetic and publicly available benchmark datasets confirmed the retrieval-enhanced configurations and self-improving language model systems. This new generative AI-enabled integration framework has achieved a significant milestone in automation of data integration, which gives a pathway to future organizations with more efficient, accurate, and scalable ETL operations.

REFERENCES

1. Parciak, M., Vandevoort, B., Neven, F., Peeters, L. M., & Vansummeren, S. "Schema matching with large language models: an experimental study." *arXiv preprint arXiv:2407.11852* (2024).
2. Sheetrit, E., Brief, M., Mishaeli, M., & Elisha, O. "Rematch: Retrieval enhanced schema matching with llms." *arXiv preprint arXiv:2403.01567* (2024).
3. Al-Onaizan, Y., Bansal, M., & Chen, Y. N. "Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing." *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. (2024).
4. Zhang, J., Zhang, H., Chakravarti, R., Hu, Y., Ng, P., Katsifodimos, A., ... & Halevy, A. "CoddLLM: Empowering Large Language Models for Data Analytics." *arXiv preprint arXiv:2502.00329* (2025).
5. Kanagarla, K. "Data Engineering with Generative AI Automating Pipelines and Transformations Using LLMS like GPT." *Available at SSRN 5107348* (2025).
6. Saripalle, R., Foulger, R., & Dooda, S. "Leveraging LLMs and RAG for Schema Alignment: A Case Study in Healthcare." *SCITEPRESS*, (2025).
7. Zhang, Z., Groth, P., Calixto, I., & Schelter, S. "A Deep Dive Into Cross-Dataset Entity Matching with Large and Small Language Models." *EDBT*. (2025).
8. Kadakia, J. J. "Demystifying Modern Data Engineering: From ETL to AI-Driven Pipelines." *Journal Of Engineering And Computer Sciences* 4.7 (2025): 948-856.
9. Benedetti, A. U. "Automation of ETL Pipelines in DataStage." *Diss. Politecnico di Torino*, (2025).
10. Ajibode, A., Bangash, A. A., Cogo, F. R., Adams, B., & Hassan, A. E. "Towards semantic versioning of open pre-trained language model releases on hugging face." *Empirical Software Engineering* 30.3 (2025): 1-63.

Source of support: Nil; **Conflict of interest:** Nil.

Cite this article as:

Putrevu, A. "Leveraging Large Language Models for Automated Schema Matching and Data Transformation in Enterprise ETL Pipelines." *Sarcouncil Journal of Multidisciplinary 5.9* (2025): pp 98-104.

Publisher: SARC Publisher