MatchIA-GCL: A Hybrid Graph Contrastive Learning Approach for Semantic Schema Matching

Mohamed Raoui, Moulay Hafid El Yazidi, Ahmed Zellou
ENSIAS, Mohammed V University in Rabat, Morocco

***Abstract*—Schema matching remains a fundamental challenge for achieving data interoperability across heterogeneous information systems. Existing deep learning-based approaches often suffer from semantic drift, overlook the structural topology of schemas, and operate as black-box models with limited interpretability. This study introduces MatchIA-GCL, an explainable hybrid framework for semantic schema matching that combines graph contrastive learning (GCL) and Large Language Models (LLMs). The proposed framework leverages Graph Attention Networks (GATs) optimized through a contrastive learning objective to generate robust structure-aware schema embeddings. Unlike conventional approaches that employ LLMs solely as semantic encoders, MatchIA-GCL integrates an LLM as an Autonomous Validation Agent responsible for resolving only complex and ambiguous matching candidates, thereby reducing computational overhead while improving decision reliability. Furthermore, a post-hoc explainability layer provides transparent and human-interpretable justifications for alignment decisions. Experimental evaluations conducted on benchmark schema matching datasets demonstrate that MatchIA-GCL consistently outperforms state-of-the-art methods, achieving up to 9.4% improvement in F1-score while enhancing explainability and robustness. These results highlight the potential of combining graph representation learning, contrastive optimization, and LLM-assisted validation for next-generation semantic schema matching systems.**

***Keywords*—Schema matching; data interoperability; graph contrastive learning; Graph Attention Networks; Large Language Models; explainable artificial intelligence; semantic data integration**

I. Introduction

The integration of large-scale heterogeneous data sources remains a persistent challenge in modern software architectures, enterprise information systems, and cloud-native ecosystems. At the core of this challenge lies schema matching, a fundamental task [1] that aims to identify semantic correspondences between attributes belonging to distinct data structures, including relational databases, NoSQL repositories, and Web APIs. As organizations increasingly rely on distributed and heterogeneous data environments, manual schema alignment becomes difficult to scale, costly, and inherently error-prone. Consequently, accurate and automated schema matching techniques have become essential for ensuring semantic interoperability, improving Extract–Transform–Load (ETL) processes, facilitating data virtualization, and supporting data-driven decision-making.

For several decades, schema matching research has primarily relied on heuristic and lexical approaches. These methods estimate the similarity between schema elements using string comparison metrics, syntactic distance measures, or external linguistic resources. While effective in highly standardized environments, such techniques often struggle when confronted with real-world heterogeneous datasets. In particular, they exhibit limited capability in handling semantic drift, where semantically equivalent concepts are represented using different terminologies, as well as structural ambiguity resulting from abbreviated, automatically generated, or domain-specific attribute names. As a consequence, their ability to capture deeper semantic relationships remains inherently constrained [2].

The emergence of deep learning has significantly transformed the field by replacing manually engineered rules with learned vector representations. Transformer-based language models such as BERT and RoBERTa [3, 4] have demonstrated remarkable effectiveness in capturing contextual semantics by jointly analyzing attribute names, data types, and textual descriptions. At the same time, representing schemas as graphs has encouraged the adoption of Graph Neural Networks (GNNs) [5–7], which can exploit structural dependencies, integrity constraints, foreign-key relationships, and hierarchical organizations. Nevertheless, many existing approaches still process semantic and structural information through separate learning pipelines, resulting in a semantic–structural disconnect that limits the quality of the generated representations. Furthermore, recent studies frequently employ Large Language Models (LLMs) [8–10] merely as semantic encoders, thereby underutilizing their reasoning and decision-making capabilities. The lack of transparency associated with deep neural architectures also remains a critical concern, particularly in domains where explainability, traceability, and auditability are mandatory requirements.

To address these limitations, this study introduces MatchIA-GCL, a hybrid and explainable framework for semantic schema matching that combines graph contrastive learning (GCL), Graph Attention Networks (GATs), and Large Language Models within a unified architecture. The proposed framework models schemas as heterogeneous graphs and learns robust structure-aware representations through a contrastive optimization objective designed to preserve semantic consistency under both structural and contextual perturbations. Unlike conventional approaches that rely extensively on generative models, MatchIA-GCL incorporates a Large Language Model as an Autonomous Validation Agent that is selectively activated only for ambiguous matching candidates and complex many-to-many correspondence scenarios. This strategy allows the framework to benefit from the reasoning capabilities of LLMs while maintaining computational

efficiency. In addition, an explainability layer is integrated to provide interpretable justifications for alignment decisions through attention analysis and gradient-based explanations.

The proposed framework makes several contributions to the schema matching literature. First, it introduces a unified learning paradigm that jointly exploits semantic and structural information through graph contrastive learning, enabling the generation of robust schema representations. Second, it proposes an original validation mechanism in which a Large Language Model acts as an autonomous decision-support agent for resolving uncertain correspondences rather than serving solely as a semantic encoder. Third, it incorporates an explainability layer that enhances the transparency and auditability of matching decisions, thereby increasing trust in automated data integration processes. Extensive experiments conducted on benchmark schema matching datasets demonstrate that MatchIA-GCL consistently outperforms existing approaches while providing improved robustness, interpretability, and generalization capabilities.

The remainder of this study is organized as follows: Section II reviews existing schema matching approaches and identifies the main research gaps. Section III formulates the schema matching problem and introduces its heterogeneous graph representation. Section IV presents the proposed MatchIA-GCL framework, including its semantic encoder, heterogeneous graph attention network, graph contrastive learning mechanism, LLM-based validation agent, explainability layer, and overall algorithm. Section V describes the experimental setup, benchmark datasets, evaluation metrics, comparative analysis, and ablation study. Finally, Section VI concludes the study and outlines future research directions.

II. Related Work

A. Lexical, Heuristic, and Traditional Algorithmic Approaches

Early schema matching approaches relied on lexical similarity measures and handcrafted heuristics. Representative systems such as Cupid, COMA++ [11], and Similarity Flooding estimate correspondences using attribute names, syntactic similarity metrics, and external lexical resources such as WordNet. Although these methods are simple and interpretable, they remain highly sensitive to naming conventions, abbreviations, and domain-specific terminology. To address these limitations, structural techniques such as Similarity Flooding exploit schema topology by propagating similarity scores across neighboring elements. However, their effectiveness largely depends on the quality of the initial lexical matches, and they often struggle with complex schema transformations and heterogeneous data models. These limitations have motivated the development of data-driven approaches capable of learning richer semantic and structural representations [12, 13].

B. Deep Learning and Embedding-Based Approaches

To overcome the limitations of lexical and heuristic methods, recent research has increasingly adopted machine learning and deep learning techniques for schema matching. Early learning-based approaches leveraged dense vector representations, commonly referred to as embeddings, to capture semantic relationships between schema elements. Models such as DeepMatcher adapted advances from Natural Language Processing (NLP) by employing distributed word representations, including Word2Vec and FastText, to project attribute names into continuous vector spaces where semantic similarity can be estimated through geometric proximity. These approaches demonstrated improved robustness to lexical variations and implicit synonymy compared with traditional string-based matching techniques. However, static embeddings assign a unique representation to each term regardless of its contextual usage, making them vulnerable to semantic ambiguity and domain-specific terminology.

The introduction of Transformer architectures has significantly advanced semantic schema matching by enabling contextual representation learning. Pre-trained language models such as BERT, RoBERTa, and DistilBERT generate dynamic embeddings that capture semantic relationships based on the surrounding context. In schema matching applications, attributes are commonly transformed into textual descriptions that combine metadata such as attribute names, data types, integrity constraints, documentation, and, in some cases, sampled data values.

C. Graph Neural Networks and Large Language Models

To address the loss of structural information inherent in text-based schema representations, recent research has explored graph-based approaches using Graph Neural Networks (GNNs), including Graph Convolutional Networks (GCNs) and Graph Attention Networks (GATs). By modeling foreign-key relationships, integrity constraints, and hierarchical dependencies, GNNs capture structural patterns that are not accessible to purely semantic models.

Recent hybrid architectures combine Transformer-based semantic embeddings with graph learning mechanisms. While these approaches generally improve matching performance, semantic and structural information are often fused through simple aggregation strategies and optimized separately, limiting their ability to capture complex semantic–structural interactions. Moreover, many graph-based methods remain highly dependent on labeled training data, reducing their generalization across domains [14].

In parallel, Large Language Models (LLMs) [15–17] have demonstrated strong capabilities in semantic understanding and reasoning. However, existing schema matching approaches typically use LLMs as semantic encoders or direct matching predictors, resulting in high computational costs and limited awareness of global schema structure [18].

These limitations reveal a gap between semantic reasoning and structural learning. To address this challenge, MatchIA-GCL integrates graph contrastive learning [19–21], heterogeneous graph attention, and selective LLM-based validation within a unified framework.

D. Research Gap Analysis

The review of existing schema matching approaches reveals several persistent limitations that remain insufficiently addressed in the current literature. First, a significant gap exists between semantic representation learning and structural modeling. While Transformer-based architectures excel at

capturing contextual semantics and Graph Neural Networks effectively exploit topological dependencies, most existing solutions learn these two dimensions through separate optimization processes. Consequently, the resulting representations often fail to fully capture the complex interactions between semantic context and schema topology.

Second, despite the growing adoption of Large Language Models [22], their reasoning capabilities remain largely underutilized in schema matching applications. Most current approaches employ LLMs either as semantic encoders or as direct matching predictors through prompt-based inference. Such usage overlooks their potential to act as intelligent decision-support mechanisms capable of validating uncertain correspondences, resolving ambiguities, and incorporating higher-level reasoning into the matching process. Moreover, applying LLMs indiscriminately to all candidate pairs introduces significant computational overhead and scalability challenges when dealing with large enterprise schemas.

Third, the explainability of schema matching decisions remains an open challenge. Many state-of-the-art deep learning approaches operate as black-box systems [23, 24], providing limited insight into the rationale behind generated correspondences. This lack of transparency can hinder their adoption in industrial environments where alignment decisions must be traceable, interpretable, and auditable by human experts. Table I summarizes the main characteristics of representative categories of schema matching approaches and highlights the positioning of the proposed MatchIA-GCL framework.

TABLE I. COMPARATIVE ANALYSIS OF EXISTING SCHEMA MATCHING APPROACHES.

Approach Category	Contextual Semantics	Structural Dependencies	Agent-Based Reasoning	Explainability
Heuristic Methods (COMA++, Cupid)	No	Partial	No	Yes
Deep Learning (DeepMatcher)	Yes	No	No	No
Graph-Based Approaches (GNNMatcher)	Yes	Yes	No	Limited
LLM-Based Approaches	Yes	No	No	No
MatchIA-GCL (Proposed)	Yes	Yes	Yes	Yes

Motivated by these observations, the next section formally defines the schema matching problem and introduces the mathematical foundations underlying the proposed MatchIA-GCL framework.

III. PROBLEM FORMULATION

A. Heterogeneous Graph Representation of Data Schemas

Traditional schema matching approaches often represent schemas as collections of independent attributes, thereby overlooking the structural dependencies that naturally exist within modern information systems. However, semantic relationships between schema elements are frequently influenced by their surrounding structural context, including table hierarchies, integrity constraints, and inter-entity relationships. To preserve both semantic and topological information, we model each data schema as a heterogeneous attributed graph.

Fig. 1 illustrates a representative schema matching scenario between two heterogeneous schemas describing the same entity. This example motivates the heterogeneous graph representation introduced in the following formulation.

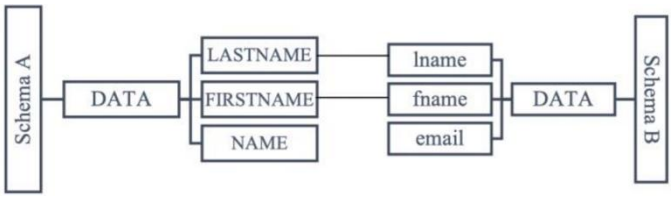

Fig. 1. Illustrative example of schema matching between two heterogeneous schemas describing the same entity.

Let S denote a data schema represented by a heterogeneous directed graph defined in:

$$G = (V, E, \tau_v, \tau_e, \mathcal{F}) \quad (1)$$

where, V denotes the set of nodes, $E \subseteq V \times V$ represents the set of directed edges, τ_v and τ_e are the node-type and edge-type mapping functions, respectively, and $\mathcal{X}$ defines the feature space associated with graph elements.

The node set is defined by:

$$V = V_T \cup V_A \quad (2)$$

where, V_T represents schema entities such as relational tables or document collections, while V_A denotes schema attributes such as columns or fields.

The table node set is formally defined in:

$$V_T = \{t_1, t_2, \dots, t_n\} \quad (3)$$

Similarly, the attribute node set is defined in:

$$V_A = \{a_1, a_2, \dots, a_m\} \quad (4)$$

The node-type mapping function is expressed by:

$$\tau_v : V \to T_v \quad (5)$$

where, each node belongs to one of the predefined categories given in:

$$T_v = \{Table, Attribute\} \quad (6)$$

The edge set captures the structural relationships among schema elements. The edge-type mapping function is defined by:

$$\tau_e : E \to T_e \quad (7)$$

The set of structural relation types is defined in:

$$T_e = \{e_{has}, e_{pk}, e_{fk}\} \quad (8)$$

where, e_{has} represents containment relationships between tables and attributes, e_{pk} denotes primary-key associations, and efk models foreign-key dependencies between related entities. These relations collectively preserve the structural organization of the schema and provide valuable contextual information for downstream matching tasks.

Each node is initialized with a feature vector:

$$x_i \in X \tag{9}$$

This feature vector aggregates the semantic metadata describing the corresponding schema element. Depending on data availability, these features may include attribute names, physical data types, textual descriptions, integrity constraints, and representative data values. Consequently, the graph representation offers a unified view that simultaneously captures both the semantic content and the structural organization of a schema.

Given two heterogeneous schema graphs representing the source and target schemas, respectively, the source graph is defined by:

$$G_s = (V_s, E_s) \tag{10}$$

The target graph is defined by:

$$G_t = (V_t, E_t) \tag{11}$$

The schema matching problem consists of identifying semantic correspondences between elements of the source graph and the target graph while preserving the structural consistency of both graph representations.

B. Extended Alignment Matrix and Correspondence Space

The primary objective of schema matching is to discover semantic correspondences between two heterogeneous schema graphs, namely the source schema (G_s) and the target schema (G_t). Formally, the matching process aims to learn a global correspondence matrix (M) that quantifies the semantic similarity between schema elements belonging to the two graphs.

The correspondence matrix is defined as a real-valued matrix of $size$ $(|V_s| \times |V_t|)$, as shown in:

$$M \in R^{|V_s| \times |V_t|} \tag{12}$$

Each element (M_{ij}) represents the confidence score associated with the semantic correspondence between a source node $(v_i \in V_s)$ and a target node $(v_j \in V_t)$. The confidence score is bounded within the interval defined in:

$$0 \leq M_{ij} \leq 1 \tag{13}$$

A value close to 1 indicates a strong semantic correspondence, whereas a value close to 0 suggests little or no semantic relationship between the two schema elements. The matching score can therefore be interpreted as the conditional probability defined in:

$$M_{ij} = P\big(match(v_i, v_j) \big| G_s, G_t\big) \tag{14}$$

While many schema matching methods assume one-to-one correspondences, real-world schemas frequently contain composite relationships such as one-to-many, many-to-one, and many-to-many mappings. Therefore, MatchIA-GCL considers both simple and composite correspondences when identifying semantic alignments between heterogeneous schemas. The resulting optimization problem is formally defined in:

$$M^* = argmax_M \sum_{v_i, v_j} M_{ij} \tag{15}$$

subject to semantic consistency and structural compatibility constraints across both schema graphs.

The objective is to maximize the confidence scores associated with valid correspondences while minimizing incorrect alignments and preserving the structural coherence of the source and target schemas. This formulation provides the mathematical foundation for the graph contrastive learning framework introduced in the next section.

C. Global Learning Objective and Optimization Criteria

MatchIA-GCL adopts a graph contrastive learning strategy to jointly learn semantic and structural representations, enabling robust schema matching across heterogeneous datasets.

Let (z_i) and (z_j) denote the latent representations of two positive nodes, corresponding either to aligned schema elements or to two augmented views of the same node. Following the InfoNCE principle, the contrastive loss associated with node (v_i) is defined as:

$$L\_(i,j) = -log\left(\frac{exp\left(\frac{sim(z_i, z_j)}{\tau}\right)}{exp\left(\frac{sim(z_i, z_j)}{\tau}\right) + \Sigma_k exp\left(\frac{sim(z_i, z_k)}{\tau}\right)}\right) \tag{16}$$

where, $sim(\cdot)$ denotes the cosine similarity function, (τ) is the temperature hyperparameter, and (z_k) represents the latent representations of negative samples that do not correspond semantically to node (v_i).

The objective of the contrastive learning process is to maximize the similarity between positive pairs while simultaneously separating negative pairs in the latent embedding space. This mechanism encourages the generation of discriminative representations that remain stable under both semantic and structural perturbations [25, 26].

The overall optimization objective of MatchIA-GCL combines the contrastive loss with an $L2$ regularization term to improve model generalization and reduce overfitting. The global loss function is defined as:

$$L_{global} = \left(\frac{1}{N}\right) \Sigma_i L_{i,match(i)} + \lambda ||W||^2 \tag{17}$$

where, (N) denotes the number of training pairs within a mini-batch, $(match(i))$ identifies the positive correspondence associated with node (v_i) , (W) represents the trainable parameters of the Graph Attention Network, and (λ) is the regularization coefficient controlling the contribution of the weight decay term. The optimization process aims to minimize $(L_{\{global\}})$, thereby learning schema representations that simultaneously preserve semantic consistency, structural coherence, and robustness to heterogeneous schema transformations. These learned embeddings constitute the

foundation of the MatchIA-GCL architecture presented in the following section.

IV. Proposed MatchIA-GCL Approach

A. Overall Architecture

The proposed MatchIA-GCL framework is organized into three sequential processing stages, as illustrated in Fig. 2. In Stage 1 (Semantic Representation Learning), heterogeneous source and target schemas are transformed into contextual semantic embeddings using a Transformer-based encoder that captures lexical and contextual information from schema metadata. In Stage 2 (Structural Representation Learning), a Heterogeneous Graph Attention Network (HGAT) combined with Graph Contrastive Learning (GCL) models structural dependencies and learns robust hybrid representations by jointly optimizing semantic and structural information. These hybrid embeddings are then used to compute the correspondence matrix between source and target schema elements. Finally, in Stage 3 (Validation and Explainable Alignment), high-confidence correspondences are directly accepted, whereas ambiguous candidate pairs are selectively validated by the LLM-based Autonomous Validation Agent. An explainability layer subsequently provides interpretable justifications for the validated alignments, improving transparency and supporting human interpretation of the matching decisions.

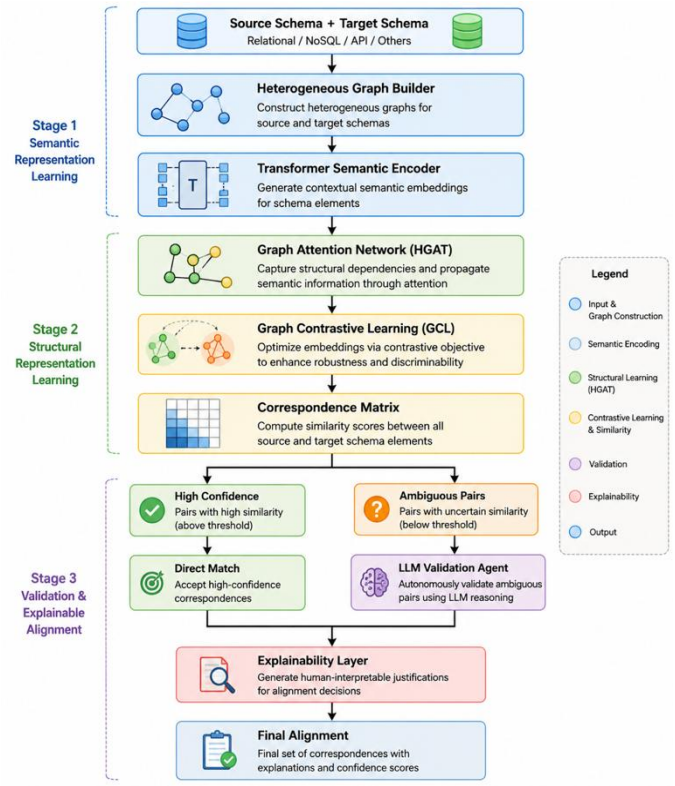

Fig. 2. Overall architecture and execution workflow of the proposed MatchIA-GCL framework.

B. Contextual Semantic Encoder

The first stage of MatchIA-GCL generates contextual semantic embeddings from schema metadata using a Transformer-based encoder. By leveraging attribute names, data types, and additional metadata, the model captures contextual semantics and reduces ambiguity between semantically related schema elements.

Let (C_i) denote the contextual description associated with node (v_i). The semantic embedding of the node is obtained by applying the Transformer encoder as defined in:

$$h_{sem_i} = TransformerEncoder(C_i) \quad (18)$$

where, $(h_{\{sem_i\}})$ represents the contextual semantic embedding generated for node (v_i).

The resulting embedding has a fixed dimensionality (d) and captures semantic relationships among schema elements independently of their structural position within the graph. These embeddings constitute the initial node features that are subsequently propagated through the graph attention layers described in the next subsection.

C. Heterogeneous Graph Attention Network

After obtaining the initial semantic embeddings, the framework activates its structural learning pipeline to incorporate topological information into the representation space. The semantic embeddings generated by the Transformer encoder serve as the initial node features of a Heterogeneous Graph Attention Network (HGAT).

To handle heterogeneous node types and relation categories, the attention mechanism is parameterized according to the relation type. For a node (v_i) connected to a neighboring node (v_j) through relation type (r), an attention coefficient is computed to quantify the structural importance of node (v_j) with respect to node (v_i). The unnormalized attention score is defined as:

$$e_{ij}^{r} = \text{LeakyReLU}(W_r[h_i||h_j]) \quad (19)$$

where, (W_v) denotes a shared linear projection matrix, (W_r) represents the attention vector associated with relation type (r), and (||) denotes vector concatenation.

The attention coefficients are subsequently normalized through a Softmax operation over the neighborhood of node (v_i), as defined in:

$$\alpha_{ij}^{r} = \frac{exp(e_{ij}^{r})}{\sum_{k \in N_i^r} exp(e_{ik}^{r})} \quad (20)$$

where, the summation is performed over all neighboring nodes connected to (v_i) through relation type (r).

The structural representation of node (v_i) is then obtained by aggregating the information received from its neighbors according to the normalized attention weights. This aggregation process is defined in:

$$h_i^{struct} = ELU\left(\sum_{j \in N_i^r} \alpha_{ij}^{r} W_v h_j\right) \quad (21)$$

To improve learning stability and capture multi-scale structural patterns, MatchIA-GCL employs a multi-head attention mechanism.

To better illustrate the structural learning process implemented in MatchIA-GCL, Fig. 3 presents the architecture

of the proposed multi-head Heterogeneous Graph Attention Network (HGAT) used for structural representation learning.

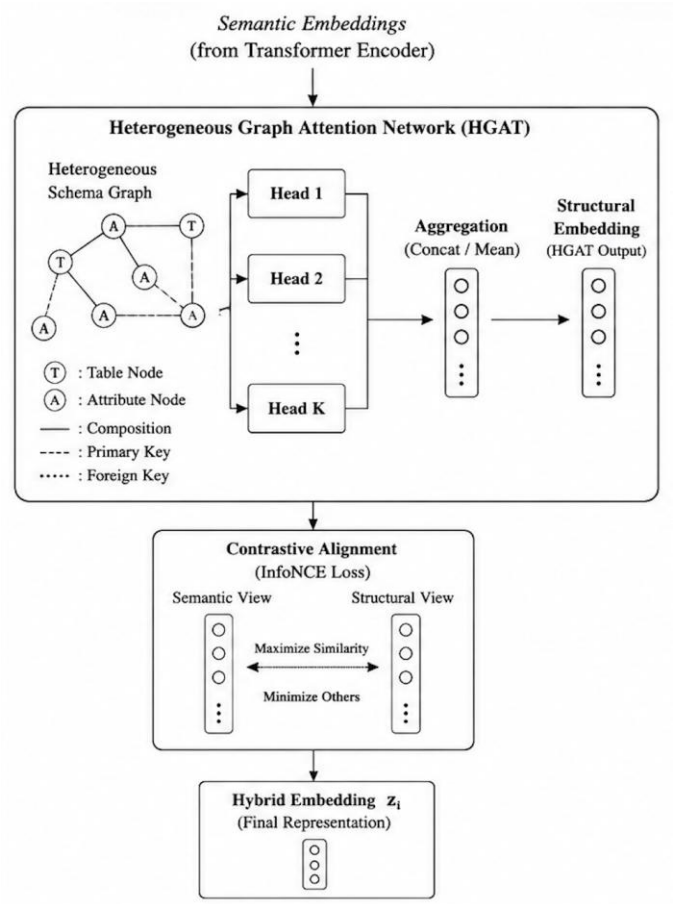

Fig. 3. Architecture of the proposed multi-head Heterogeneous Graph Attention Network (HGAT) used in MatchIA-GCL.

The HGAT module enables the propagation of semantic information across heterogeneous structural dependencies, including table–attribute relationships and referential constraints. The resulting structural embeddings are subsequently aligned with the semantic representations through the graph contrastive learning mechanism, producing robust hybrid embeddings suitable for schema matching.

D. Semantic–Structural Contrastive Alignment

The semantic embeddings generated by the Transformer encoder and the structural representations learned by the HGAT are subsequently aligned through a graph contrastive learning strategy. Rather than combining both representations through a simple aggregation operation, the proposed framework explicitly maximizes the agreement between the semantic and structural views of the same schema element.

The contrastive objective encourages semantically equivalent and structurally related nodes to occupy nearby regions in the latent space while simultaneously separating unrelated elements. Consequently, schema attributes sharing similar semantic roles or structural functions become geometrically closer even when their naming conventions differ significantly.

Positive pairs are constructed either from aligned schema elements available in the reference mappings or from two augmented views of the same schema node generated through feature masking and edge dropout. Negative pairs are randomly sampled from semantically unrelated nodes belonging to different schemas while ensuring that no reference correspondence exists between them. This sampling strategy enables the contrastive objective to maximize the similarity of semantically equivalent representations while effectively separating non-matching schema elements in the latent embedding space.

The final hybrid representation associated with node $(v_i)$is obtained by combining the semantic embedding and the structural embedding as defined in:

$$z_i = h_i^{\{sem\}} + h_i^{\{struct\}} \quad (22)$$

where, (z_i) denotes the final hybrid embedding used for correspondence estimation and downstream alignment tasks. These representations constitute the input of the matching and validation stage described in the next subsection.

E. LLM-Based Autonomous Validation Agent

Although the hybrid embeddings generated by MatchIA-GCL successfully identify most schema correspondences, ambiguous mappings often require additional semantic reasoning. To address this challenge, the framework incorporates a Large Language Model (LLM) as an Autonomous Validation Agent. Unlike conventional approaches that apply LLMs to all candidate pairs, the proposed agent is activated only for uncertain correspondences, thereby reducing computational overhead while maintaining scalability. For each ambiguous candidate, the agent analyzes the source and target schema elements together with their metadata and structural context. It then determines the validity and type of correspondence $(1{:}\,1,1{:}\,n,n{:}\,1,or\ n{:}\,m)$ and produces a concise justification for the decision. The validated correspondences are subsequently integrated into the global correspondence matrix, enabling the framework to combine graph-based representation learning.

F. Post-Hoc Explainability Layer

To improve transparency and facilitate adoption in industrial environments, MatchIA-GCL incorporates a post-hoc explainability layer that provides human-interpretable justifications for alignment decisions. The module combines attention-based structural analysis and semantic feature attribution to identify the factors that contribute to each correspondence. At the structural level, attention coefficients extracted from the HGAT layers highlight the influence of neighboring nodes and schema relationships. At the semantic level, feature attribution methods estimate the contribution of attribute names, data types, descriptions, and representative values to the matching decision.

The resulting explanations are aggregated into interpretable audit reports that provide both structural and semantic evidence supporting the generated correspondences. Fig. 4 illustrates the validation and explainability workflow integrated into MatchIA-GCL.

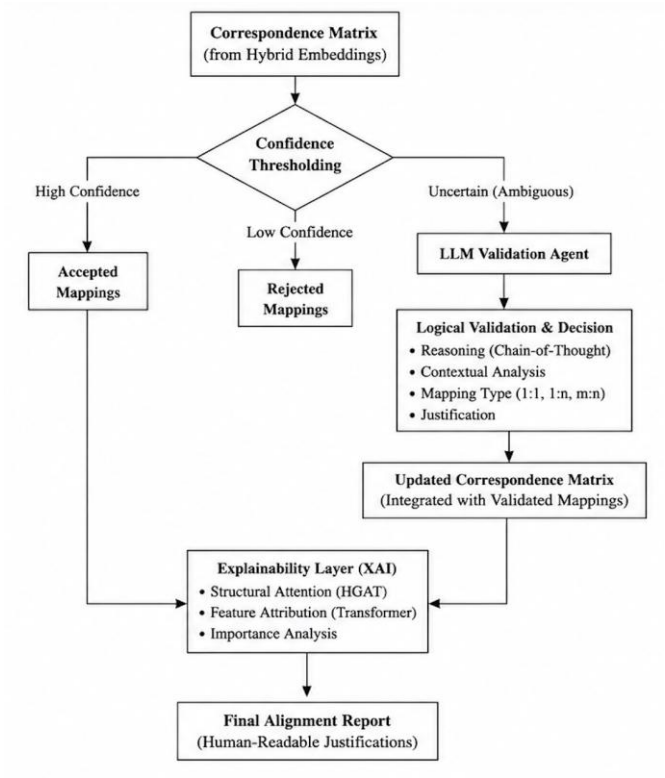

Fig. 4. Autonomous validation and explainability framework for schema matching in MatchIA-GCL.

G. Overall MatchIA-GCL Algorithm

The overall workflow of MatchIA-GCL is summarized in Algorithm 1. The algorithm integrates semantic representation learning, structural graph modeling, contrastive optimization, and selective LLM-based validation within a unified framework. Contextual semantic embeddings are first generated using a Transformer-based encoder and refined through a Heterogeneous Graph Attention Network (HGAT) to incorporate structural dependencies between schema elements. A graph contrastive learning objective is then optimized to improve the separability of semantically equivalent and non-equivalent representations in the latent space. Candidate correspondences are subsequently estimated using cosine similarity between the learned hybrid embeddings, whereas only uncertain candidates are subjected to LLM-based semantic validation. The validated correspondences constitute the final schema alignment produced by MatchIA-GCL.

Algorithm 1: MatchIA-GCL: Schema Matching Framework

Input : Source schema S_s, Target schema S_t
Parameters θ (encoder, HGAT, projection, LLM),confidence threshold τ
Output : Set of correspondences C

Initialize $C \leftarrow \emptyset$
Construct heterogeneous schema graph *G = (V, E, R)* from S_s *and* S_t
/ Stage 1: Contextual Semantic Encoding */*
for each node $v \in V$ **do**
  $x_v \leftarrow build\ v(name, type, parent, description, samplevalues)$
  $h_v^{(0)} \leftarrow TransformerEncoder(x_v; \theta_e nc)$
end for
/ Stage 2: Heterogeneous Graph Attention Network */*
for $l = 1\ to\ L$ **do**
  Update node representations using HGAT layer l
end for
/ Stage 3: Graph Contrastive Learning */*
Construct positive and negative pairs from S_s and S_t
Compute contrastive loss L_{cl} and update parameters θ
/ Stage 4: Similarity Computation */*
for each pair $(u \in Ss, v \in St)$ **do**
  $sim(u,v) \leftarrow cosinSimilarity\left(h_u^{(L)}, h_v^{(L)}\right)$
end for
/ Stage 5: LLM-based Validation */*
for each candidate pair (u, v) with $sim(u,v) \geq \tau$ **do**
  $decision, score \leftarrow LLMValidation(u, v, context_{uv}, sim(u,v))$
  if decision == match **then**
    $C \leftarrow C \cup (u, v, score)$
  end if
end for
return *C*

The following section presents the experimental evaluation of the proposed framework.

V. Experimental Evaluation

A. Experimental Setup and Evaluation Protocol

To assess the effectiveness of the proposed MatchIA-GCL framework, a comprehensive experimental protocol was designed to evaluate its matching accuracy, robustness, scalability, and generalization capabilities across heterogeneous schema matching scenarios. All experiments were conducted using a Python-based implementation relying on the PyTorch and PyTorch Geometric frameworks for deep learning and heterogeneous graph processing.

The semantic encoding module employs a pre-trained RoBERTa-base model to generate contextual embeddings from schema metadata. The structural learning component is implemented using a Heterogeneous Graph Attention Network composed of three attention layers with eight attention heads per layer. The graph contrastive learning module is optimized using the InfoNCE loss function, with the temperature parameter empirically fixed to 0.07 following a hyperparameter tuning phase.

For ambiguous correspondences requiring higher-level semantic reasoning, the autonomous validation module relies on OpenAI GPT-4o accessed through the OpenAI API. The model is integrated through a standardized inference interface and is selectively activated only for candidate correspondences whose similarity score falls within a predefined uncertainty interval.

B. Benchmark Datasets

The experimental evaluation was conducted on three benchmark datasets commonly used in schema matching research. These datasets cover diverse semantic and structural challenges, including noisy schemas, semantic heterogeneity, multilingual metadata, and complex correspondences.

The first benchmark considered in this study is the Valentine Benchmark, a well-established evaluation framework frequently used in schema matching research [1]. This benchmark includes both academic and industrial datasets originating from domains such as finance, healthcare, and business information systems. A key advantage of the Valentine benchmark lies in its ability to generate degraded schema variants through controlled transformations, including attribute renaming, abbreviation insertion, typographical noise, and structural perturbations. Such characteristics make this benchmark particularly suitable for evaluating the robustness of representation learning methods under realistic and adverse conditions.

The second benchmark is the Open Public Data (OPD) dataset, which consists of real-world schemas extracted from governmental open-data portals and public information repositories. Compared with curated academic datasets, OPD exhibits significantly higher semantic heterogeneity due to inconsistent naming conventions, multilingual terminology, undocumented abbreviations, and varying normalization levels. Consequently, this benchmark provides a realistic environment for assessing the effectiveness of semantic representation learning approaches in large-scale data integration scenarios [27].

The third benchmark is the Wikidata-to-Schema dataset, which focuses on aligning relational schemas with semantic web and knowledge graph structures derived from Wikidata. It includes hierarchical relationships and many-to-many correspondences [28], making it particularly suitable for evaluating the reasoning capabilities of the proposed LLM-based validation component. The three benchmarks provide complementary evaluation scenarios covering robustness, semantic generalization, and complex semantic reasoning. Table II summarizes the main characteristics of the benchmark datasets used throughout the experimental evaluation.

TABLE II. CHARACTERISTICS OF THE BENCHMARK DATASETS

Dataset	Application Domain	Number of Schemas	Schema Size (Attributes / Tables)	Correspondence Type
Valentine Benchmark	Academic & Industrial	14 pairs	43 to 70 attributes / 2 to 15 tables per schema	1:1, Textual & Complex
Open Data (OPD)	Public Governance	22 pairs	30 to 120 attributes / 1 table per flat file	1:1, Abbreviations & Synonyms
Wikidata-to-Schema	Semantic Web	8 Graphs	> 500 relational properties / Graph-Ontologies	m:n, Composite & Hierarchical

As shown in Table II, the selected benchmark datasets encompass a wide range of schema matching challenges, from noisy relational schemas to highly complex ontology-driven structures. This diversity enables a rigorous assessment of the semantic, structural, and reasoning capabilities of the proposed MatchIA-GCL framework.

C. Evaluation Metrics

The performance of MatchIA-GCL was evaluated using the standard metrics commonly adopted in schema matching research, namely Precision, Recall, and F1-Score. The generated correspondences were compared against manually validated reference alignments (ground truth) to assess matching quality. Precision measures the proportion of correctly identified correspondences among all predicted matches and is defined by:

$$Precision = TP/(TP + FP) \tag{23}$$

where, TP denotes the number of true positive correspondences and FP denotes the number of false positive correspondences.

Recall evaluates the ability of the framework to identify all valid correspondences present in the reference alignment. This metric measures the sensitivity of the model with respect to missed matches and is defined by:

$$Recall = TP/(TP + FN) \tag{24}$$

where, FN represents the number of false negative correspondences.

The F1-Score, defined as the harmonic mean of Precision and Recall, is adopted as the primary evaluation metric because it provides a balanced assessment of matching performance by jointly considering false positive and false negative correspondences. It is particularly suitable for schema matching, where both accurate correspondence identification and error minimization are equally important. It is computed according to:

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{25}$$

In addition to matching accuracy, execution time and inference efficiency were evaluated to assess the computational scalability of MatchIA-GCL. These metrics quantify the practical cost of semantic representation learning and selective LLM-based validation, providing insight into the framework's suitability for large-scale heterogeneous schema matching scenarios.

D. Comparative Results and Quantitative Analysis

The Precision results obtained by COMA++, DeepMatcher, GNNMatcher, LLM-Centric, and the proposed MatchIA-GCL are presented in Fig. 5. MatchIA-GCL achieves the highest Precision across all benchmark datasets, reaching 90.3%, 92.7%, and 88.6% on the Valentine, OPD, and Wikidata-to-Schema datasets, respectively. The consistent improvement over both traditional and deep learning-based baselines indicates that the joint integration of contextual semantic representations, heterogeneous graph attention, and graph contrastive learning effectively reduces false positive correspondences. Furthermore, the selective activation of the LLM-based validation agent refines ambiguous candidate mappings without introducing

unnecessary validation overhead, leading to more reliable semantic alignments across heterogeneous schemas.

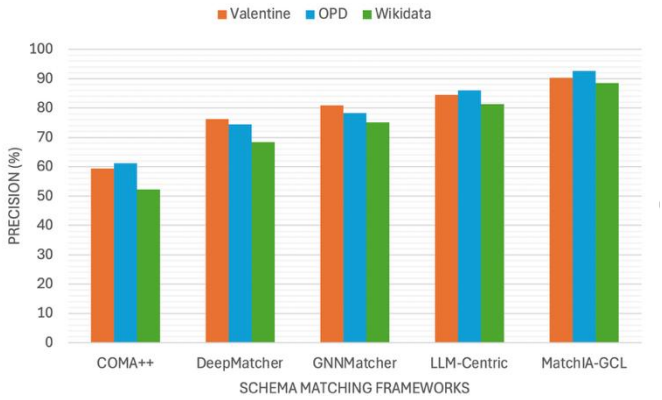

Fig. 5. Precision comparison across benchmarks.

The Recall results are reported in Fig. 6. MatchIA-GCL consistently outperforms competing approaches, achieving Recall values of 88.1%, 90.1%, and 86.4% on the three benchmark datasets. This demonstrates its effectiveness in identifying a larger proportion of valid correspondences.

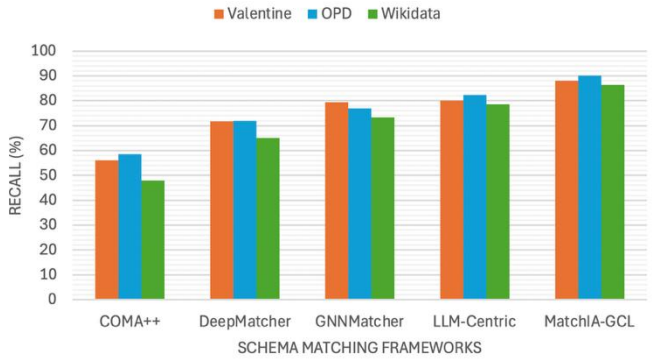

Fig. 6. Recall comparison across benchmarks.

The overall matching performance measured by the F1-Score is presented in Fig. 7. MatchIA-GCL consistently achieves the highest results across all benchmark datasets, reaching 89.2%, 91.4%, and 87.5% on the Valentine, OPD, and Wikidata-to-Schema benchmarks, respectively. The observed improvements demonstrate the effectiveness of jointly learning semantic and structural representations through the proposed hybrid architecture. The largest gain is obtained on the Wikidata-to-Schema benchmark, where MatchIA-GCL outperforms the strongest baseline by 7.6 percentage points in terms of F1-Score. This benchmark contains ontology-driven structures, hierarchical dependencies, and many-to-many semantic correspondences, making schema alignment particularly challenging. The results indicate that the integration of heterogeneous graph attention learning, graph contrastive optimization, and LLM-assisted validation enables the framework to better capture complex semantic and structural relationships than existing approaches.

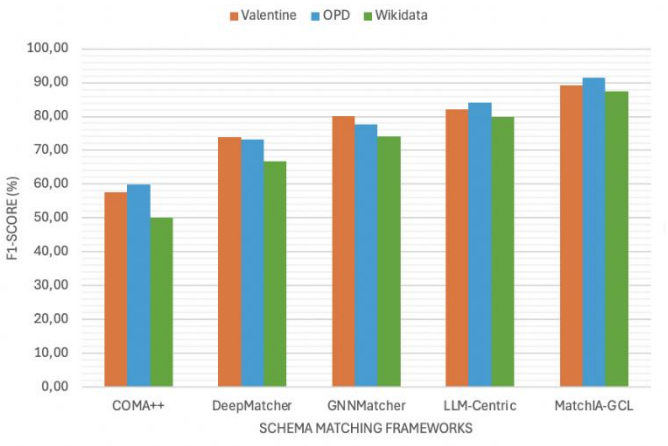

Fig. 7. F1-Score comparison across benchmarks.

Overall, the results confirm the effectiveness of combining contextual semantic encoding, heterogeneous graph attention learning, graph contrastive optimization, and LLM-assisted validation within a unified schema matching framework.

E. Ablation Study and Component Contribution Analysis

To assess the contribution of each component of MatchIA-GCL, an ablation study was conducted by removing one module at a time while keeping the remaining architecture unchanged. Four degraded variants were evaluated.

- MatchIA w/o HGAT removes the Heterogeneous Graph Attention Network and relies solely on Transformer-based semantic embeddings, thereby evaluating the contribution of structural information.
- MatchIA w/o GCL removes the Graph Contrastive Learning objective while preserving the Transformer encoder and the HGAT module, thereby evaluating the contribution of contrastive optimization to robust semantic-structural representation learning.
- MatchIA w/o LLM disables the LLM-based Autonomous Validation Agent and relies exclusively on embedding similarity scores, thereby evaluating the contribution of semantic reasoning to resolving ambiguous correspondences.
- MatchIA w/o Context replaces the Transformer-based contextual semantic encoder with basic attribute representations, thereby evaluating the importance of contextual semantic information in generating discriminative schema embeddings.

The performance differences between these variants and the complete MatchIA-GCL framework provide insights into the contribution of semantic context, structural learning, contrastive optimization, and LLM-based reasoning to the overall matching performance. Table III summarizes the performance degradation observed after removing each component of the proposed architecture.

TABLE III. ABLATION STUDY: F1-SCORE DEGRADATION (%)

Framework Variant	Valentine	OPD	Wikidata
MatchIA w/o HGAT	-9.0	-6.8	-11.3
MatchIA w/o GCL	-5.4	-4.1	-5.9
MatchIA w/o LLM	-4.2	-7.9	-12.1
MatchIA w/o Context	-7.1	-5.3	-3.6

Overall, the ablation study validates the complementarity of the proposed components.

VI. CONCLUSION AND FUTURE WORK

MatchIA-GCL introduces a unified framework for heterogeneous schema matching that combines contextual semantic encoding, heterogeneous graph attention learning, contrastive optimization, and LLM-assisted validation. By jointly modeling semantic and structural information, the proposed approach effectively addresses the limitations of existing schema matching methods and improves the identification of complex correspondences across heterogeneous data sources. Experimental evaluations on the Valentine, OPD, and Wikidata-to-Schema benchmarks demonstrate consistent improvements over representative state-of-the-art approaches, while the ablation study confirms the contribution of each architectural component. Furthermore, the integration of an explainability layer enhances the transparency and auditability of alignment decisions, facilitating the adoption of the framework in real-world data integration environments. Future work will focus on scaling the framework to large enterprise data ecosystems and incorporating adaptive learning mechanisms that continuously refine matching decisions through expert feedback and autonomous reasoning.

REFERENCES

[1] Koutras, C., Siachamis, G., Ionescu, A., Psarakis, K., Brons, J., Fragkoulis, M., ... & Katsifodimos, A. (2021, April). Valentine: Evaluating matching techniques for dataset discovery. In 2021 IEEE 37th International Conference on Data Engineering (ICDE) (pp. 468-479). IEEE.

[2] Li, H., Govind, Y., Mudgal, S., Rekatsinas, T., & Doan, A. (2021). Deep learning for semantic matching: A survey. Journal of Computer Science and Cybernetics, 37(4), 365-402.

[3] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).

[4] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[5] Hamilton, W. L. (2020). Graph representation learning. Morgan & Claypool Publishers.

[6] Vrahatis, A. G., Lazaros, K., & Kotsiantis, S. (2024). Graph attention networks: a comprehensive review of methods and applications. Future Internet, 16(9), 318.

[7] Shi, C. (2022). Heterogeneous graph neural networks. In Graph neural networks: Foundations, frontiers, and applications (pp. 351-369). Singapore: Springer Nature Singapore.

[8] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

[9] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

[10] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Rozière, B., ... & Szafraniec, M. (2024). Llama 3: Open foundation and instruction models. Meta AI.

[11] Aumueller, D., Do, H. H., Massmann, S., & Rahm, E. (2005, June). Schema and ontology matching with COMA++. In Proceedings of the 2005 ACM SIGMOD international conference on Management of data (pp. 906-908).

[12] Yousfi, A., El Yazidi, M. H., & Zellou, A. (2020, November). An Efficient Holistic Schema Matching Approach. In International Conference on Information and Communication Technology and Applications (pp. 588-601). Cham: Springer International Publishing.

[13] Yousfi, A., El Yazidi, M. H., & Zellou, A. (2020, December). CSSM: A Context-Based Semantic Similarity Measure. In 2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS) (pp. 1-6). IEEE.

[14] Raoui, M., Ennaouri, M., El Yazidi, M. H., & Zellou, A. (2024). SchemaLogix: Advancing Interoperability with Machine Learning in Schema Matching. parameters, 6, 7.

[15] Parciak, M., Vandevoort, B., Neven, F., Peeters, L. M., & Vansummeren, S. (2024). Schema matching with large language models: an experimental study. arXiv preprint arXiv:2407.11852.

[16] Raoui, M., Ennaouri, M., El Yazidi, M. H., & Zellou, A. (2025, July). Next-Gen Schema Matching: How Machine Learning Enhances Data Interoperability. In 2025 9th International Conference on Computer, Software and Modeling (ICCSM) (pp. 175-180). IEEE.

[17] Seedat, N., & van der Schaar, M. (2025). Matchmaker: Schema matching with self-improving compositional LLM programs.

[18] Liu, Y., Pena, E., Santos, A., Wu, E., & Freire, J. (2024). Magneto: Combining small and large language models for schema matching. arXiv preprint arXiv:2412.08194.

[19] You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., & Shen, Y. (2020). Graph contrastive learning with augmentations. Advances in neural information processing systems, 33, 5812-5823.

[20] Zhu, Y., Xu, Y., Yu, F., Liu, Q., Wu, S., & Wang, L. (2020). Deep graph contrastive representation learning. arXiv preprint arXiv:2006.04131.

[21] Ju, W., Wang, Y., Qin, Y., Mao, Z., Xiao, Z., Luo, J., ... & Zhang, M. (2024). Towards graph contrastive learning: A survey and beyond. arXiv preprint arXiv:2405.11868.

[22] Freire, J., Fan, G., Feuer, B., Koutras, C., Liu, Y., Peña, E., ... & Wu, E. (2025). Large language models for data discovery and integration: Challenges and opportunities. IEEE Data Engineering Bulletin.

[23] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM computing surveys (CSUR), 51(5), 1-42.

[24] Lundberg, S., & Lee, S. I. (2017). Shap: A unified approach to interpreting model predictions. Advances in neural information processing systems, 1-10.

[25] Ju, W., Gu, Y., Mao, Z., Qiao, Z., Qin, Y., Luo, X., ... & Zhang, M. (2025). GPS: Graph contrastive learning via multi-scale augmented views from adversarial pooling. Science China Information Sciences, 68(1), 112101.

[26] Liu, Y., Zhao, Y., Wang, X., Geng, L., & Xiao, Z. (2024). Multi-scale subgraph contrastive learning. arXiv preprint arXiv:2403.02719.

[27] Miller, R. J. (2018). Open data integration. Proceedings of the VLDB Endowment, 11(12), 2130-2139.

[28] Vrandečić, D., & Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. Communications of the ACM, 57(10), 78-85.