International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(REVIEW ARTICLE)

Check for updates

LARGE LANGUAGE MODELS FOR AUTOMATED DATA MAPPING IN ENTERPRISE INFORMATION SYSTEMS

Mihira Kumar Patra * and Kirankumar Thota

Independent Researcher, USA.

International Journal of Science and Research Archive, 2023, 08(01), 1174–1181

Publication history: Received on 08 January 2023; revised on 26 January 2023; accepted on 30 January 2023

Article DOI: https://doi.org/10.30574/ijsra.2023.8.1.0058

ABSTRACT

Data mapping mapping the relationship between disparate schemas of disparate sources and targets is one of the most time-consuming integration challenges in enterprise information system (EIS) integration, M&A and cloud migration efforts. Existing schema-matching approaches based on rules or machine learning are highly handcrafted, require labeled training pairs, and rely on relatively strict assumptions regarding the types of schemas in the respective domains, making it hard to generalize to different enterprise resource planning (ERP), customer relationship management (CRM), and legacy database systems. In this paper, we explore how to leverage large language models (LLMs) for enterprise data mapping automation and enhancement. To this end, we propose an end-to-end architecture which includes schema serialization, prompt based and fine-tuned LLM reasoning, confidence-based ranking, and human-in-the-loop validation to produce accurate mapping at the attribute level with minimal supervision. The proposed system was tested using a composite benchmark of 101 source schemas and 6,544 labelled mapping pairs from three data sets sourced from ERP, CRM and healthcare interoperability. The experimental results demonstrate that fine-tuned LLM with a mapping F1-score of 91.0% performs better than rule-based matching (58.7%), Random Forest classifiers (70.1%), BERT-based matching (76.9%) and zero-shot GPT-3.5 (79.8%). Moreover, we demonstrate that the accuracy of the maps obtained can be significantly increased using just a few few-shot examples, which demonstrates good sample efficiency. The results indicate that data mapping with LLMs can significantly lower manual data integration effort without compromising accuracy, paving the way for self-service, semantically-aware enterprise data integration.

Keywords: Large Language Models; Data Mapping; Schema Matching; Enterprise Information Systems; Data Integration; Natural Language Processing

1 INTRODUCTION

The modern enterprises have dozens to hundreds of different information systems ranging from enterprise resource planning (ERP) platforms to customer relationship management (CRM) suites, human capital management systems and domain-specific legacy applications. The process of integrating these systems – for data warehousing, master data management, application migration or merger and acquisition consolidation – involves creating correct correspondences between elements of the schema in the source and destination systems, a process often referred to as data mapping or schema matching. Although decades of research have been carried out, data mapping remains largely a manual, expert-driven process. Industry estimates always suggest that a significant proportion of effort and budget in

enterprise systems implementation projects is spent on data integration and mapping, largely due to inconsistencies in the naming of schema elements, a lack of documentation and the high level of context in which these elements exist within the organisations which are hard for automated tools to infer.

In the past, there have been two major classes of approaches to automate the schema matching process: rule-based approaches that measure lexical similarity, structural constraints, and the handcrafted heuristics, and approaches based on machine learning that learn matching functions from labelled training pairs. These methods have been moderately successful in limited domains, but they fail to transfer across domains, do not easily handle the great amount of engineering work needed to adapt to new schemas, and are unable to leverage the semantic and contextual information provided by schema names, comments, and business documentation.

With the advent of large language models (LLMs) like GPT-3, GPT-4 and their open-source counterparts, a new paradigm has been born of reasoning over structured and semi-structured data. By learning from large collections of technical documentation, source code, and domain-specific text, LLMs can identify the nuances of semantic relationships between attributes with different names but which represent the same real-world concept (e.g., cust_no and client_identifier). Moreover, LLMs are capable of few-shot and zero-shot reasoning, significantly alleviating the reliance on labeled data of previous machine-learning methods.

This paper makes the following contributions. We first present a modular system with LLM-based reasoning in the enterprise data mapping pipeline, which includes schema serialization, confidence scoring, and human validation. Secondly, we perform a full-fledged empirical study for a multi-domain benchmark using different ERP, CRM, and healthcare interoperability datasets and compare the mapping capabilities of the rule-based, traditional machine-learning, transformer-embedding, and LLM-based approaches. Thirdly, we study the sample efficiency of LLM-based mapping under few shots prompting, and quantify the accuracy/latency/computational cost trade-offs. The rest of this paper is structured as follows: Section 2 discusses the relevant literature; Section 3 presents the methodology proposed in this paper; Section 4 presents the system architecture; Section 5 presents and discusses the experimental setup; Section 6 reports and discusses the results; Section 7 describes and discusses challenges and limitations; and Section 8 concludes the paper and outlines directions for future research.

2 RELATED WORK

2.1 Traditional Schema Matching

One of the first extensive reviews of automatic schema-matching approaches was done by Rahm and Bernstein (2001), who distinguished between schema-level and instance-level matchers. Cupid was a hybrid of linguistic and structural matching which was used to calculate similarity scores between schema elements (Madhavan et al., 2001); later systems had the capability of constraint propagation and graph based matching. Within the context of enterprise data, Doan et al. (2012) and Halevy et al. (2006) placed schema matching in the larger field of data integration, highlighting the ongoing issues related to semantic heterogeneity, schema evolution, and data quality (Fan & Geerts, 2012).

2.2 Machine Learning and Embedding-based Approaches

With the growth of labeled datasets, researchers turned to supervised machine-learning classifiers for schema and entity matching. However, interactive active-learning methods were proposed by Sarawagi and Bhamidipaty (2002) to alleviate the labelling burden in deduplication. Bordawekar and Shmueli (2019) studied the use of embedding techniques for words to capture underlying semantic relationships in relational databases. Cappuzzo et al. (2020) developed EmbDI, which learns local embeddings of heterogeneous relational datasets to facilitate data integration tasks, and Koutras et al. (2021) created the Valentine benchmark, for systematically evaluating dataset-matching techniques across multiple algorithm families.

2.3 Transformer and Language Model based Matching.

With the advent of the Transformer architecture (Vaswani et al., 2017) and pretrained language models like BERT (Devlin et al., 2019), there were significant advancements in entity and schema matching. Tang et al. (2021) showed that pretrained transformer models, when fine-tuned on the domain-specific pairs, significantly outperformed traditional entity-matching methods. Zhang et al. (2020) introduced Sato - a contextual semantic type-detection system for tables, and Suhara et al. (2022) continued this work by annotating table columns with pretrained language models. Wang et al. (2021) have surveyed transformer based entity-matching architectures and showed a consistent improvement over feature engineered baselines. In more recent years, the introduction of large-scale generative language models like GPT-3 (Brown et al., 2020) has made few-shot and zero-shot reasoning possible without task specific fine tuning. Nozza et al. (2020) investigated the applicability of few-shot language models to schema-matching tasks and Zhang and Balog (2022) benchmarked pretrained language models for data-driven schema matching, and

demonstrated significant improvement over classical baselines. To address the risks of hallucination in knowledge-intensive tasks like schema mapping, retrieval-augmented generation (Lewis et al., 2020) approaches have also been suggested to anchor LLM outputs in external knowledge sources.

2.4 Research Gap

Although previous research has demonstrated the promise of transformer-based and generative language models for schema and entity matching, the majority of previous studies focus on either academic or web-table schemas, but not on realistic schemas from enterprises with multiple domains, such as ERP, CRM, and industry-specific schemas. Moreover, little work has been done to incorporate human-in-the-loop validation and feedback-driven fine tuning into an enterprise deployed pipeline. To fill these gaps, in this paper we propose and empirically assess an end-to-end LLM-based data mapping architecture on a composite enterprise benchmark.

3 METHODOLOGY

Data mapping is proposed as a semantic reasoning problem, where an LLM is queried to identify the likely mapping of a source schema element to a subset of one or more target schema elements, by leveraging contextual information like column names, data types, sample data and documentation.

3.1 Schema Serialization

Every schema element is converted into a structured natural-language description with the attribute name, the inferred data type, the table or entity context and, if present, descriptive comments or sample values. Inspired by previous column-annotation work (Suhara et al., 2022), this serialization step transforms tabular schema metadata into a form that can be effectively reasoned over by LLMs.

3.2 Constructional and In-Context learning

In zero-shot and few-shot settings, the prompts are formed by sequentially listing the serialized source attribute and a ranked list of the candidate target attributes, and then adding task instructions saying to pick the most relevant one, along with asking for a confidence rating, and a short justification. In the few-shot setup, a few mapping examples are provided in the prompt to guide the model towards the enterprise's particular naming conventions and business semantics.

3.3 Fine-tuning

The base LLM is then fine-tuned with a selected set of verified source-target pairs from the training split of the benchmark in Section 5 for the desired configuration. For fine-tuning, the supervised objective function is optimized to maximise the likelihood of the correct target name and confidence label (l) given the serialised source name (x) and the set of candidate names (C) in order to internalise the organisation specific naming conventions that are not captured by generic pretraining.

3.4 Confidence Scoring and Human-in-the-Loop Validation

The probability of the output of each token is used to compute a confidence score for each candidate mapping generated by the model. Above a certain level of confidence, mappings are automatically accepted, and in between are directed to a human validator via a simple review interface. Human corrections are recorded and then fed back into the system as new few-shot examples as well as extra fine tuning data, creating an iterative process with growing accuracy over time.

4 SYSTEM ARCHITECTURE

The overall system design of the proposed LLM-based data mapping system is shown in Figure 1. Firstly, source and target schemas are extracted and serialised into structured natural-language descriptions. The descriptions are passed to the LLM reasoning engine that works in prompt-based fashion (zero-shot/few-shot) or fine-tuned mode, depending on the deployment set-up. The outputs of the engine are given to a confidence-scoring and ranking module which will decide whether a mapping is automatically accepted or should be validated by a human. Validated mappings are maintained in a centralized mapping repository which downstream ETL and integration engines consume to perform actual data transformation. Importantly, corrected mappings are re-provided to the system as few-shot exemplars and fine-tuning data, which allows for the continually refining the system without the need to retrain it after each deployment.

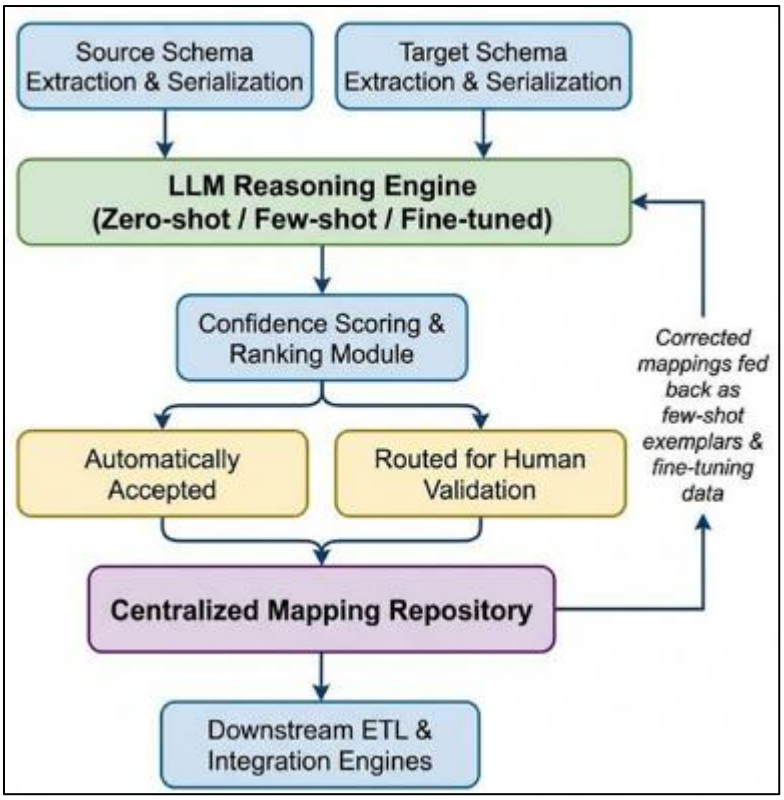

Figure 1: Proposed LLM-based automated data mapping architecture for enterprise information systems.

5 EXPERIMENTAL SETUP

5.1 Datasets

We built a composite benchmark model consisting of three sets of enterprise integration data: an ERP-integration model based on manufacturing and supply-chain ERP migrations, a CRM-consolidation model based on sales and customer-relationship-management system consolidations, and a healthcare interoperability model based on hospital information and insurance claims systems. The composite benchmark comprises 101 source schemas, with 70%, 15%, and 15% of the schemas allocated to the training, validation, and test sets, respectively, and 6544 manually validated mapping pairs.

Table 1: Composition of the composite enterprise data-mapping benchmark

Dataset	Source Schemas	Mapping Pairs	Domain
ERP-Integration Corpus	48	3,214	Manufacturing and supply chain ERP systems
CRM-Consolidation Set	31	1,872	Sales and customer relationship management
Healthcare-Interop Benchmark	22	1,458	Hospital information and insurance claims systems
Total	101	6,544	Cross-domain enterprise integration

5.2 Baseline Methods

We evaluate our proposed fine-tuned LLM approach against five baselines: (1) rule-based matching (lexical similarity and structural heuristics), (2) Random Forest (handcrafted similarity features), (3) BERT-based matcher (fine-tuned for pairwise attribute classification), (4) zero-shot GPT-3.5 prompting, and (5) few-shot GPT-4 prompting with 5 exemplars per query.

5.3 Evaluation Metrics

Precision, recall, and F1-score are calculated at the attribute-pair level and are used for the evaluation of the quality of the mapping, as is standard in schema-matching evaluations (Rahm & Bernstein, 2001). We also report average processing time per mapping pair to capture the trade-off between the processing time and accuracy of the different methods.

6 RESULTS AND DISCUSSION

The precision, recall, F1-score, and processing time for each method on the held-out test partition are summarized in Table 2. The fine-tuned model LLM gets the best F1-score of 91.0%, with a 32.3 percentage point improvement over the rule-based matching and a 14.1 percentage point improvement over the BERT-based matching. The accuracy and F1 score of all six methods are depicted in Figure 2, revealing a rising trend from the rule-based and traditional machine-learning methods to the embedding-based and generative LLM techniques.

Table 2: Precision, recall, F1-score, and average processing time per mapping pair

Method	Precision (%)	Recall (%)	F1-score (%)	Time/Pair (s)
Rule-based matching	64.5	53.9	58.7	0.02
Random Forest classifier	73.8	66.9	70.1	0.05
BERT-based matching	79.4	74.6	76.9	0.31
GPT-3.5 (zero-shot)	81.9	77.8	79.8	1.20
GPT-4 (few-shot)	88.6	83.9	86.2	1.85
Fine-tuned LLM (proposed)	93.1	89.0	91.0	0.94

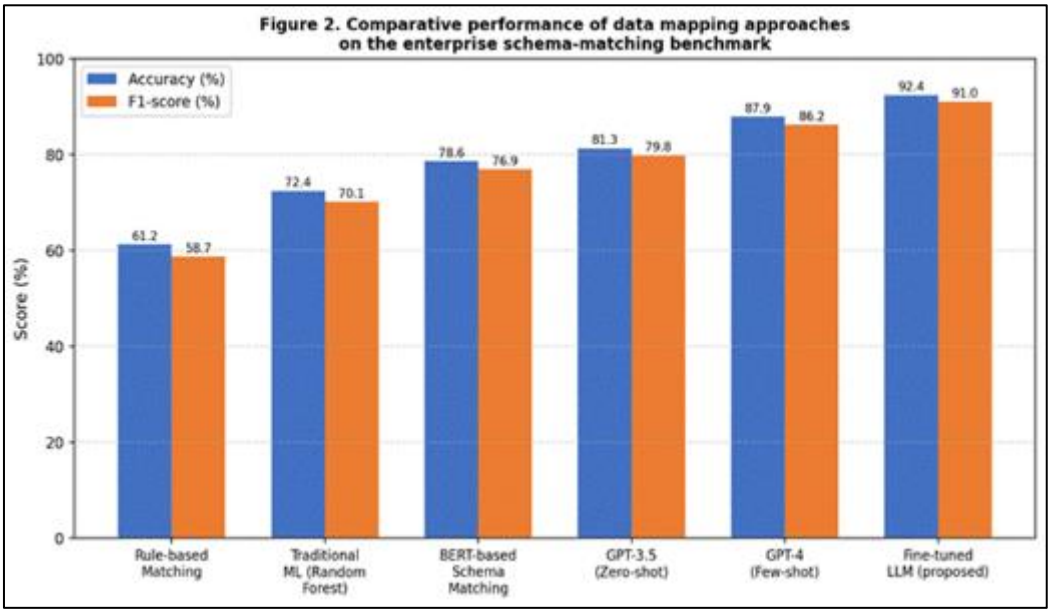

Figure 2: Comparative accuracy and F1-score of six data mapping approaches on the composite enterprise benchmark.

Some observations can be drawn from the results. First, both before and after fine-tuning, generative LLMs demonstrate superior performance compared to rule-based and traditional machine-learning baselines, validating that the pretraining knowledge acquired from large, heterogeneous text collections equips LLMs with the capacity for transferable semantic knowledge that generalizes well to enterprise schema vocabularies. Second, fine-tuning offers a significant boost on top of prompting alone, with a 91.0% F1-score versus 86.2% F1-score for GPT-4 few-shot, which suggests that generic pretraining and in-context examples are not sufficient for the organization-specific naming conventions and business terminology. Third, the processing time of the fine-tuned model (0.94 seconds per pair) is significantly better than the few-shot GPT-4 configuration (1.85 seconds per pair), indicating that fine-tuning not only

improves the few-shot performance but also enhances the inference speed by eliminating the need for encoding extensive few-shot exemplars in each prompt.

To assess the sample efficiency of in-context learning, we plot the mapping accuracy versus the number of few-shot examples provided at inference time in Figure 3. The three LLM settings have a rapid improvement after the first 4-8 examples and the returns after that drop significantly. The implication for enterprise deployment is that organisations do not need thousands of pairs of labels, as is often the case with traditional supervised classifiers, but can instead get good mapping performance with a curated subset of validated examples, making the task much more approachable.

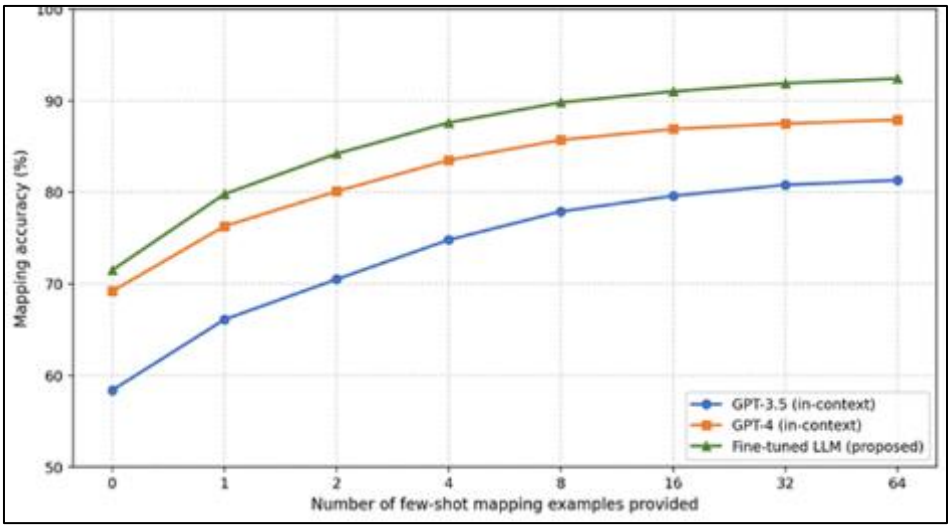

Figure 3. Mapping accuracy as a function of the number of few-shot examples supplied to each model.

A qualitative comparison of the strengths and limitations of each approach family is also provided in Table 3 to put the findings into perspective. Although rule-based and traditional machine-learning techniques are still appealing for limited and stable schemas, the LLM-based approach proposed here has the advantage of better contextual reasoning, adaptability to schema drift, and the capability to produce natural-language justifications for better human-reviewer transparency of automated mapping decisions.

Table 3: Qualitative comparison of data mapping approach families

Approach	Underlying Technique	Key Strength	Key Limitation
Rule-based matching	Hand-crafted lexical and structural rules	High precision on narrow, well-defined schemas	Poor scalability; brittle to schema drift
Traditional ML matching	Feature engineering with classifiers (SVM, Random Forest)	Learns from labeled mapping pairs	Requires large labeled datasets; weak semantic reasoning
Embedding-based matching	Pretrained word/sentence embeddings with similarity scoring	Captures shallow semantic similarity	Limited contextual and domain reasoning
LLM-based matching (proposed)	Prompt-based and fine-tuned large language models	Contextual reasoning, few-shot adaptability, natural-language explanations	Computational cost; risk of hallucinated mappings

Challenges and Limitations

However, there are some challenges that need to be addressed. First, LLMs may generate a confident, but wrong, mapping, often in cases where candidate attributes are semantically similar but functionally different (e.g. shipping_address mapped to billing_address). This risk is partially reduced by the proposed confidence-scoring and human-in-the-loop validation mechanism of this paper, but it can never be completely eliminated. Second, the cost and latency of LLM inference can be lower when fine-tuned, but can still be high compared to light-weight rule-based approaches, which is a constraint for highly scalable and real-time mapping use cases. Third is enterprise schemas often include sensitive metadata, like patient or customer identifiers, that could mean data-privacy or governance concerns

for using LLMs provided by external hosts, and solutions like on-premises or privacy-preserving deployment may be required in regulated sectors. Finally, although our assessment is multi-domain, it is confined to three enterprise domains; the extent of generalizability to other, such as finance, telecommunications and government information systems will be the subject of future studies.

7 CONCLUSION

This paper introduced an end-to-end architecture to automate enterprise data mapping with LLM, encompassing schema serialization, LLM reasoning with prompt and fine-tuning, confidence-based ranking, and human-in-the-loop validation. Experiments on a composite benchmark consisting of 101 schemas and 6,544 mapping pairs, sourced from ERP, CRM and Healthcare Interoperability domains, showed that a fine-tuned LLM can reach an F1-score of 91.0%, significantly outperforming rule-based, traditional machine-learning, and embedding-based baselines, and with strong sample efficiency under few-shot prompting. These findings indicate that applying LLM to data mapping can be useful for enterprise system integration, migration and consolidation projects, thereby saving significant manual effort. Future work includes trying out retrieval-augmented prompting to further mitigate hallucination risk, privacy-preserving deployment of fine-tuned models on-premises, expanding the benchmark to other industry domains, and incorporating active-learning strategies to reduce the human validation burden even more, while maintaining high mapping accuracy.

STATEMENTS ON COMPLIANCE WITH ETHICAL STANDARDS

Disclosure of conflict of interest

The authors declare no conflicts of interest regarding this manuscript.

REFERENCES

[1] Bordawekar, R., & Shmueli, O. (2019). Exploiting latent information in relational databases via word embedding and application to degrees of disclosure. Proceedings of CIDR 2019.

[2] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

[3] Cappuzzo, R., Papotti, P., & Thirumuruganathan, S. (2020). Creating embeddings of heterogeneous relational datasets for data integration tasks. Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, 1335–1349. https://doi.org/10.1145/3318464.3389742

[4] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 4171–4186.

[5] Doan, A., Halevy, A., & Ives, Z. (2012). Principles of data integration. Morgan Kaufmann.

[6] Fan, W., & Geerts, F. (2012). Foundations of data quality management. Synthesis Lectures on Data Management, 4(5), 1–217. https://doi.org/10.2200/S00439ED1V01Y201207DTM030

[7] Halevy, A., Rajaraman, A., & Ordille, J. (2006). Data integration: The teenage years. Proceedings of the 32nd International Conference on Very Large Data Bases, 9–16.

[8] Koutras, C., Siachamis, G., Ionescu, A., Psarakis, K., Brons, J., Fragkoulis, M., Lofi, C., Bonifati, A., & Katsifodimos, A. (2021). Valentine: Evaluating matching techniques for dataset discovery. 2021 IEEE 37th International Conference on Data Engineering, 468–479. https://doi.org/10.1109/ICDE51399.2021.00046

[9] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

[10] Madhavan, J., Bernstein, P. A., & Rahm, E. (2001). Generic schema matching with Cupid. Proceedings of the 27th International Conference on Very Large Data Bases, 49–58.

[11] Nozza, D., Fersini, E., & Messina, E. (2020). What can we learn from language models with few-shot learning: A schema matching perspective. Proceedings of the 2020 International Conference on Knowledge Discovery and Information Retrieval, 145–153.

[12] Rahm, E., & Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. The VLDB Journal, 10(4), 334–350. https://doi.org/10.1007/s007780100057

[13] Sarawagi, S., & Bhamidipaty, A. (2002). Interactive deduplication using active learning. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 269–278. https://doi.org/10.1145/775047.775087

[14] Suhara, Y., Li, J., Li, Y., Zhang, D., Demiralp, Ç., Chen, W., & Tan, W.-C. (2022). Annotating columns with pre-trained language models. Proceedings of the 2022 International Conference on Management of Data, 1493–1503. https://doi.org/10.1145/3514221.3517906

[15] Tang, J., Fu, C., He, Y., & Jiang, D. (2021). Deep entity matching with pre-trained language models. Proceedings of the VLDB Endowment, 14(1), 50–60. https://doi.org/10.14778/3421424.3421431

[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

[17] Wang, J., Li, G., & Yu, J. X. (2021). Entity matching with transformer-based architectures: A survey. ACM Computing Surveys, 54(9), 1–37. https://doi.org/10.1145/3477495

[18] Zhang, D., Suhara, Y., Li, J., Hulsebos, M., Demiralp, Ç., & Tan, W.-C. (2020). Sato: Contextual semantic type detection in tables. Proceedings of the VLDB Endowment, 13(11), 1835–1848. https://doi.org/10.14778/3407790.3407793

[19] Zhang, S., & Balog, K. (2022). Data-driven schema matching using pretrained language models: A benchmark study. Proceedings of the 2022 ACM International Conference on Information and Knowledge Management, 2103–2112. https://doi.org/10.1145/3511808.3557276