Appendix for UniDM: A Unified Framework for Data Manipulation with Large Language Models

We update the revision and provide more details and experiments. The revised contents are summarized as follows:

- Experiments of UniDM on different LLMs (Table A).
- Additional ablation study of UniDM on various datasets (Table B) and tasks (Table C).
- Additional fine-tuning experiments of UniDM on LLaMA2-7B (Table D).
- Overhead in the format of token consumption compared with FM (Table E).)
- Detailed prompt templates (Appendix B).)
- Case study of our retrieval-based UniDM compared with FM (Appendix C).)

A Additional Experimental Results

We additionally evaluate UniDM over 5 LLMs variants: GPT-4-Turbo (OpenAI, 2021), Claude2 (about 100B) (Claude2, 2023), LLaMA2 (7B and 70B) (LLaMA2, 2023), Qwen-7B (Qwen, 2023) on data imputation task. We observe consistent high performance using UniDM across different base LLMs, evidencing its adaptability and robustness.

We also conduct more ablation experiments on more tasks and datasets (as shown in Table B and Table C). For the data transformation task, it does not require to extract context data and this step is not included in the ablation study.

Table A. UniDM results on data imputation with LLMs variants.

Model	Data Imputation Acc (%)	
	Restaurant	Buy
GPT-3-175B	93.0	98.5
GPT-4-Turbo	96.5	98.5
Claude2	89.5	96.9
LLaMA2-7B	86.0	95.4
LLaMA2-70B	88.4	96.9
Qwen-7B	86.0	93.8

We extend the fine-tuning experiment on an additional LLM (open weight LLM LLaMA2 (LLaMA2, 2023)). As shown in Table D, the results further validate adaptability and robustness of our approach across various LLMs.

In Table E, we compare the LLMs' tokens between the FM and our method. It is evident that our method incurs greater token consumption than the FM; however, our method automates the processes of context retrieval and target prompt construction, which significantly reduces the human labor.

Table B. Ablation study of UniDM on data imputation task (Buy dataset).

Instance-wise Retrieval	Meta-wise Retrieval	Target Prompt Construction	Context Data Parsing	Acc (%)
				90.8
✓				92.3
	✓			90.8
✓	✓			92.3
✓	✓	✓		96.9
✓	✓	✓	✓	98.5

Table C. Ablation study of UniDM on data transformation task.

Target Prompt Construction	Context Data Parsing	Data Transformation Acc (%)	
		Stack Overflow	Bing QueryLogs
		63.3	52.0
	✓	65.3	52.0
✓		65.3	54.0
✓	✓	67.4	56.0

Table D. Fine-tuning experiments: F1-score of UniDM on entity resolution task (Walmart-Amazon dataset).

LLM	F1-Score (%)	
	FM	UniDM
GPT-J-6B	17.6	17.8
GPT-J-6B (fine-tune)	84.2	86.6
LLaMA2-7B	NA	40.6
LLaMA2-7B (fine-tune)	NA	89.4
GPT-3-175B	87.0	88.2

Table E. Token consumption (per-query) comparison with FM.

Method	Token Consumption	
	Restaurant	Buy
FM	174	246
UniDM (w/o retrieval)	325	384
UniDM	6860	7323

B Prompt Template

We provide a running example to explain how our method automatically generates the desired cloze question of target task based on the in-context learning of LLMs. The prompt example is as follows:

(Input to LLMs):

Write the claim as a cloze question.
Claim:
The task is data imputation which produces the missing data with some value to retain most of the data. The context is Wenham, Marysville, and Westmont are cities in the United States, identified by the ISO3 code USA. The target is city:New Cassel, iso3:USA, country:?
Cloze question:
Wenham, Marysville, and Westmont are cities in the United States, identified by the ISO3 code USA. New Cassel is the name of a city whose ISO3 country code is USA. New Kassel belongs to the country _.
Claim:
The task is data transformation which is the process of converting data from one format to another required format within a record. The context is data before transformation: 20000101 data after transformation: 2000-01-01. The target is 19990415:?
Cloze question:
20000101 can be transformed to 2000-01-01, and 19990415 can be transformed to _.
Claim:
The task is error detection which detect attribute error within a record in a data cleaning system. The context is the address of 2505 u s highway 431 north is not an error, the county name of mxrshxll is an error. The target is whether there is an error in city:sheffxeld.
Cloze question:
The address ”2505 U.S. Highway 431 North” has no error, whereas the county name ”mxrshxll” contains an error. It is required to identify if there is an error in the city name ”sheffxeld”. Is there an error in the city name? Yes or No. _
Claim:
The task is entity resolution which is the process of predicting whether two records are referencing the same real-world thing. The context is A is the Punch! Home Design Architectural Series 4000 v10, manufactured by Punch! Software, is priced at $199.99. B is The Punch Software 41100 Punch! Home Design Architectural Series 18, manufactured by Punch Software, is priced at $18.99. The target is are A and B the same?
Cloze question:

(Output of LLMs):
Punch! Home Design Architectural Series 4000 v10, manufactured by Punch! Software, is priced at $199.99, whereas Punch Software 41100 Punch! Home Design Architectural Series 18, also manufactured by Punch Software, is priced at $18.99. Are these two products the same? Yes or No. _

C Case Study

For case study, we present the final results of FM (random setting), FM (manual setting), and our retrieval-based UniDM. For FM method (random setting), we have the final prompt and the output of LLMs as follows:

(Input to LLMs):
name: anthonys. addr: 3109 piedmont rd. just south of peachtree rd.. phone: 404/262-7379. type: american. What is the city? atlanta
name: rose pistola. addr: 532 columbus ave.. phone: 415/399-0499. type: italian. What is the city? san francisco
name: american place. addr: 2 park ave. at 32nd st.. phone: 212/684-2122. type: american. What is the city? new york
name: ruth’s chris steak house (los angeles). addr: 224 s. beverly dr.. phone: 310-859-8744. type: steakhouses. What is the city?

(Output of LLMs): los angeles
Ground Truth: Beverly Hills

For FM method (manual setting), we have the final prompt and the output of LLMs as follows:

(Input to LLMs):
name: oceana. addr: 55 e. 54th st.. phone: 212/759-5941. type: seafood. What is the city? new york
name: oceana. addr: 55 e. 54th st.. phone: 212-759-5941. type: seafood. What is the city? new york city
name: ruth’s chris steak house (los angeles). addr: 224 s. beverly dr.. phone: 310-859-8744. type: steakhouses. What is the city?

(Output of LLMs): los angeles
Ground Truth: Beverly Hills

For our retrieval-based UniDM, we have the final prompt and the output of LLMs as follows:

(Input to LLMs):
The name of the place is Belvedere. The address is 9882 Little Santa Monica Blvd. The city is Beverly Hills.
The name of the grill is Jack Sprat’s Grill and its address is 10668 W. Pico Blvd. in the city of Los Angeles.
The name of the establishment is Border Grill, located on 4th Street in Los Angeles.
Ruth’s Chris Steak House (Los Angeles) is located at 224 S. Beverly Dr. Ruth’s Chris Steak House (Los Angeles) is located in the city of _.

(Output of LLMs): Beverly Hills
Ground Truth: Beverly Hills

In the appendix, we first provide more complex tasks supported by our method, including TabelQA (Appendix D), join discovery (Appendix E), and information extraction (Appendix F). In Appendix G, we also represent the main function, executed by UniDM, in pseudo-code.

D Explaining the Table Question Answer Task

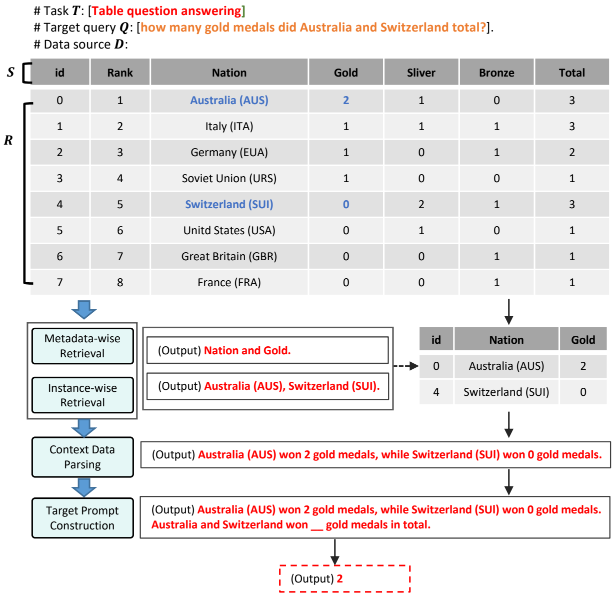
To show the generality of our UniDM solution, we apply it on the more complex table question answer (TableQA) task. This is a task to ask a question to retrieve answers from a data table. Figure 1 gives an illustrative example on WikiTableQuestions dataset (Pasupat & Liang, 2015). Here we have a question Q: "how many gold medals did Australia and Switzerland total?" and the answer on the number of gold medals could be obtained from the table by finding the relevant information.

For the TableQA task, we directly set the task query Q to be the question contained in the task description. The set R of records and set S of attributes are set to contain all records and attributes in the table D_i, respectively. When applying UniDM to solve TableQA, in the first context information retrieval, we set the set of candidate attributes $S' = S$. By applying prompts p_{rm} and p_{ri}, UniDM first automatically retrieves a content snapshot $\mathcal{C}$ from the data table D_i. This snapshot contains a selection of columns ('Nation' and 'Gold' in our example) and rows ('Australia (AUS)' and 'Switzerland (SUI)' in our example) that summarize the information most relevant to the task and the query ('how many gold medals did Australia and Switzerland total?').

In the second step, the context snapshot $\mathcal{C}$ is serialized and then parsed into a natural text representation $\mathcal{C}'$. In our example, we now know that "Australia (AUS) won 2 goal medals, while Switzerland (SUI) won 0 goal medals". To facilitate open-ended cloze question generation centered around the target query and contextual information drawn from the data, the prompt engineering module is employed. Ultimately, the resulting cloze question is fed into the LLM to yield an answer. UniDM could correctly output "2" as the answer. This exhibit that UniDM is good at not only processing instance-level tasks such as data imputation and error correction, but also can be applied to retrieve table-level information.

E Join Discovery Task

In data lakes, the join relations across tables are not specified. Join discovery task is a major challenge in data analysis. It aims at finding semantically joinable columns across different tables. This task could be subsumed and solved by our UniDM framework.



# Task T: [Table question answering]
# Target query Q: [how many gold medals did Australia and Switzerland total?].
# Data source D:

id	Rank	Nation	Gold	Sliver	Bronze	Total
0	1	Australia (AUS)	2	1	0	3
1	2	Italy (ITA)	1	1	1	3
2	3	Germany (EUA)	1	0	1	2
3	4	Soviet Union (URS)	1	0	0	1
4	5	Switzerland (SUI)	0	2	1	3
5	6	Unitd States (USA)	0	1	0	1
6	7	Great Britain (GBR)	0	0	1	1
7	8	France (FRA)	0	0	1	1

(Output) Nation and Gold.

(Output) Australia (AUS), Switzerland (SUI).

id	Nation	Gold
0	Australia (AUS)	2
4	Switzerland (SUI)	0

(Output) Australia (AUS) won 2 gold medals, while Switzerland (SUI) won 0 gold medals.

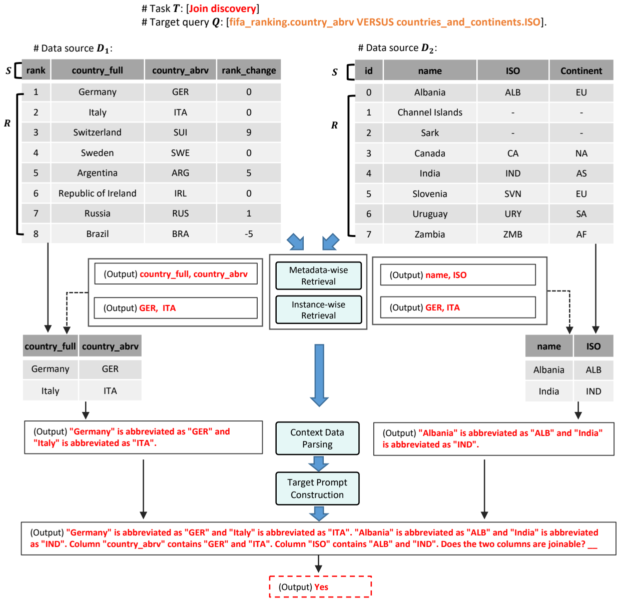
(Output) Australia (AUS) won 2 gold medals, while Switzerland (SUI) won 0 gold medals. Australia and Switzerland won __ gold medals in total.

(Output) 2

Figure 1. Explanation example for table question answering by the UniDM.

Figure 2 gives an illustrative example of join discovery task. The query Q is set to be the textual name of the two tables, e.g., "fifa_ranking.country_abrv" and "countries_and_continents.ISO". The set R of records and the set S of attributes are set to contain all records and attributes in the two tables D_1 and D_2, respectively. $F_T(R, S, \mathcal{D})$ outputs all pairs of attributes where $s \in S_i, s' \in S_j$ that could be joined with each other. In the first context information retrieval, UniDM extracts joinable attributes and records $\mathcal{C}_1$ and $\mathcal{C}_2$ between the two tables. The context snapshots $\mathcal{C}_1$ and $\mathcal{C}_2$ are serialized and then parsed into natural text representations $\mathcal{C}'_1$ and $\mathcal{C}'_2$ separately. By using the prompt engineering module, UniDM constructs a cloze question with the target query and contextual information from the two tables. Ultimately, the resulting cloze question is fed into the LLM to yield an answer "Yes" that indicates the two tables are joinable.

For experiment on join discovery task, we use NextiaJD (Flores Herrera et al., 2021) that composes four splits according to their file size. The dataset labels the join quality of attribute pairs based on a measure that considers both containment and cardinality proportion with empirically determined thresholds. In experiments, we use a subset with 4404 pairs (2239 positive and 2164 negative) of attributes whose quality is labeled as Good and High. For the baseline, we select an embedding-based solution WarpGate (Cong et al., 2022), the SOTA method. As shown in Figure 3, we report precision, recall and F1-score under various threshold values. We find that UniDM consistently obtains higher F1-score compared with WarpGate (Cong et al., 2022) under various threshold values. This exhibits that UniDM's



# Task T: [Join discovery]
# Target query Q: [fifa_ranking.country_abrv VERSUS countries_and_continents.ISO].
# Data source D_1:
# Data source D_2:
S
R
Metadata-wise Retrieval
Instance-wise Retrieval
Context Data Parsing
Target Prompt Construction

rank	country_full	country_abrv	rank_change
1	Germany	GER	0
2	Italy	ITA	0
3	Switzerland	SUI	9
4	Sweden	SWE	0
5	Argentina	ARG	5
6	Republic of Ireland	IRL	0
7	Russia	RUS	1
8	Brazil	BRA	-5

id	name	ISO	Continent
0	Albania	ALB	EU
1	Channel Islands	-	-
2	Sark	-	-
3	Canada	CA	NA
4	India	IND	AS
5	Slovenia	SVN	EU
6	Uruguay	URY	SA
7	Zambia	ZMB	AF

(Output) country_full, country_abrv

(Output) GER, ITA

(Output) name, ISO

(Output) GER, ITA

country_full	country_abrv
Germany	GER
Italy	ITA

name	ISO
Albania	ALB
India	IND

(Output) "Germany" is abbreviated as "GER" and "Italy" is abbreviated as "ITA".

(Output) "Albania" is abbreviated as "ALB" and "India" is abbreviated as "IND".

(Output) "Germany" is abbreviated as "GER" and "Italy" is abbreviated as "ITA". "Albania" is abbreviated as "ALB" and "India" is abbreviated as "IND". Column "country_abrv" contains "GER" and "ITA". Column "ISO" contains "ALB" and "IND". Does the two columns are joinable? __

(Output) Yes

Figure 2. Explanation example for join discovery by the UniDM.

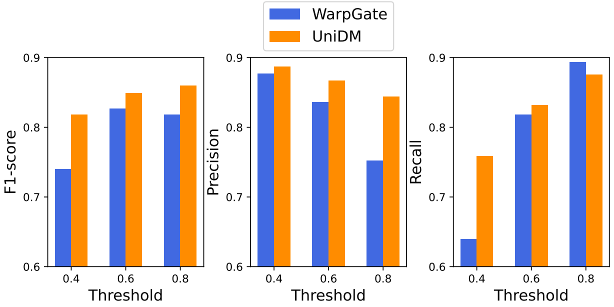


Figure 3. F1-score, precision and recall on join discovery task.

potential to manipulate data across tables.

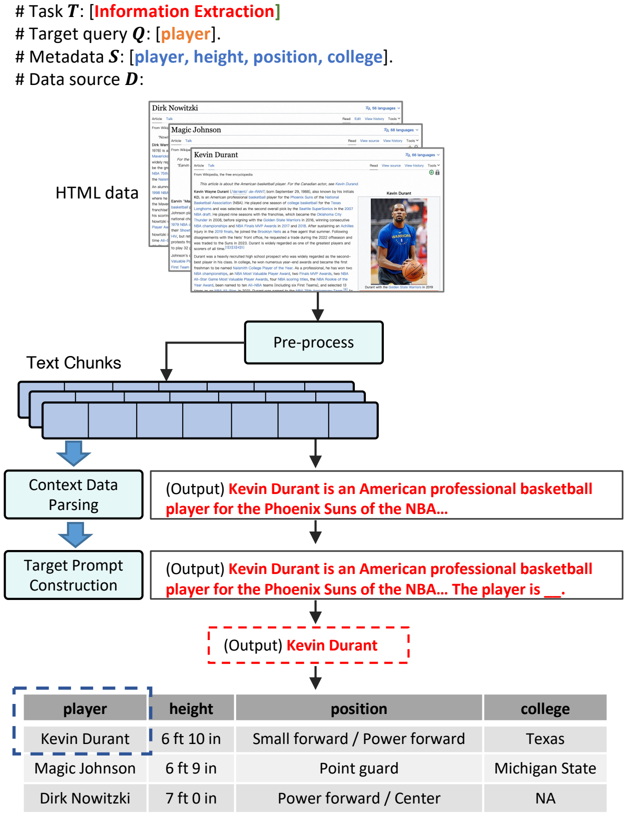
F Information Extraction Task

Table F. Text F1-score on information extraction task.

Method	Information Extraction NBA player
Evaporate-code	40.6
Evaporate-code+	84.6
UniDM	70.1

In order to demonstrate the performance of UniDM in handling more complex data manipulation tasks, we conduct experiments on an information extraction task. For the information extraction task, we aim to construct a structured view (i.e., tabular) of a set of semi-structured documents (e.g., HTML).

Figure 4 gives an illustrative example of information extraction task. D_i is a semi-structured document and $R = D_i$. S represents a set of pre-defined attributes to be extracted from D_i. $F_T(R, S, \mathcal{D})$ outputs a set of attribute values from the document D_i. The query Q is set to be the target attribute S, e.g., "player". Note that we temporarily removed the context retrieval module because the attributes and the instance is pre-defined by users. We slightly modified the context parsing prompt by adding the query Q. It aims to extract and transform the context information ("Kevin Durant is an American professional basketball player...") according to the query Q. We then construct the target cloze question and yield the final result (i.e., "Kevin Durant").



# Task T: [Information Extraction]
# Target query Q: [player].
# Metadata S: [player, height, position, college].
# Data source D:
HTML data
Dirk Nowitzki
Magic Johnson
Kevin Durant
Pre-process
Text Chunks
Context Data Parsing
Target Prompt Construction

(Output) Kevin Durant is an American professional basketball player for the Phoenix Suns of the NBA...

(Output) Kevin Durant is an American professional basketball player for the Phoenix Suns of the NBA... The player is __.

(Output) Kevin Durant

player	height	position	college
Kevin Durant	6 ft 10 in	Small forward / Power forward	Texas
Magic Johnson	6 ft 9 in	Point guard	Michigan State
Dirk Nowitzki	7 ft 0 in	Power forward / Center	NA

Figure 4. Explanation example for information extraction by the UniDM.

For information extraction, we use SWDE benchmark in (Hao et al., 2011) and choose NBAplayer dataset. This dataset includes complex HTML from the NBA Wikipedia pages for NBA players. Following the previous method, we consult closed information extraction setting, where a pre-defined schema is provided and UniDM is used to populate the table. Ground truth value is available for the target attributes. As shown in Table F, UniDM outperforms the baseline methods Evaporate-code in terms of the F1-score. Evaporate-code+ achieves better results due to its utilization of ensemble methods. This exhibits that UniDM's potential to manipulate semi-structured data.

Algorithm 1 Unified Framework for Data Manipulation

```
Input: a data lake $\mathcal{D}$, a subset of records $R \subseteq D_i$ extracted from a
    table $D_i \in \mathcal{D}$, a subset of attributes $S \subseteq S_i$ under the schema
    $S_i$, a target task $T$ and a query $Q$
if Context Retrieval then
    // Meta-wise retrieval
    $p_{rm} \leftarrow$ prompt_meta$(\mathcal{D}, T, Q, S)$
    $S^t \leftarrow$ LLM$(p_{rm})$
    // Instance-wise retrieval
    $p_{ri} \leftarrow$ prompt_instance$(\mathcal{D}, T, Q, R)$
    $\{score_i\}_{i=1}^{m} \leftarrow$ LLM$(p_{ri})$
    $R^t \leftarrow$ top-k$(\{score_i\})$
    // Select cells based on retrieved results
    $\mathcal{C} \leftarrow$ data_select$(\mathcal{D}, S^t, R^t)$
else
    // randomly sample context from the data
    $\mathcal{C} \leftarrow$ data_sample$(\mathcal{D})$
end if
$\mathcal{V} \leftarrow$ serialize$(C) = \{(s : r[s]) | \forall r[s] \in \mathcal{C}\}$
if Data Parsing then
    // Parse data into a natural text representation
    $p_{dp} \leftarrow$ prompt_parse$(\mathcal{V})$
    $\mathcal{C}' \leftarrow$ LLM$(p_{dp})$
else
    $\mathcal{C}' = \mathcal{V}$
end if
// Recursively uses the LLM to reformat the data task.
$p_{cq} \leftarrow$ prompt_construction$(T, Q, \mathcal{C}')$
$p_{as} \leftarrow$ LLM$(p_{cq})$
$Y \leftarrow$ LLM$(p_{as})$
Return $Y$
```

G Algorithm

Algorithm 1 represents the main function for data manipulation tasks, executed by UniDM, in pseudo-code. The input integrates a data lake for a target data manipulation task, with user-provided parameters. These parameters include a subset of schema, a subset of records, a task description, and a target query. First, we automatically retrieve context information according to the task and the input query. This module uses LLMs to select valuable attributes for the task and the target attribute, and then perform a fine-grained filtering on all records to identify relevant ones w.r.t. target records. Next, the context information $\mathcal{C}$, represented in a tabular form, is transformed into a more effective format C' for LLMs. The target prompt construction is to find an effective prompt to organize the task description T, the context information C' and the query Q. Finally, this prompt is fed into LLMs to yield the final answer of our task.

References

Claude2. Claude2, 2023. URL https://claude.ai/.

Cong, T., Gale, J., Frantz, J., Jagadish, H., and Demiralp, Ç. Warpgate: A semantic join discovery system for cloud data warehouse. *arXiv preprint arXiv:2212.14155*, 2022.

Flores Herrera, J. d. J., Nadal Francesch, S., and Romero Moral, Ó. Towards scalable data discovery. In *Advances in Database Technology: EDBT 2021, 24th International Conference on Extending Database Technology: Nicosia, Cyprus, March 23-26, 2021: proceedings*, pp. 433–438. OpenProceedings, 2021.

Hao, Q., Cai, R., Pang, Y., and Zhang, L. From one tree to a forest: a unified solution for structured web data extraction. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011)*. Association for Computing Machinery, Inc., July 2011.

LLaMA2. Llama2, 2023. URL https://ai.meta.com/llama/.

OpenAI. Openai api, 2021. URL https://openai.com/api/.

Pasupat, P. and Liang, P. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1470–1480, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1142. URL https://aclanthology.org/P15-1142.

Qwen. Qwen, 2023. URL https://tongyi.aliyun.com/qianwen/.