

LINGUA MANGA : A Generic Large Language Model Centric System for Data Curation

Zui Chen
IIIS, Tsinghua University
Beijing, China
chenzui19@mails.tsinghua.edu.cn

Lei Cao
U of Arizona/MIT
Tucson, Arizona
lcao@csail.mit.edu

Sam Madden
MIT
Cambridge, Massachusetts
madden@csail.mit.edu

ABSTRACT

Data curation is a wide-ranging area which contains many critical but time-consuming data processing tasks. However, the diversity of such tasks makes it challenging to develop a general-purpose data curation system. To address this issue, we present LINGUA MANGA, a user-friendly and versatile system that utilizes pre-trained large language models. LINGUA MANGA offers automatic optimization for achieving high performance and label efficiency while facilitating flexible and rapid development. Through three example applications with distinct objectives and users of varying levels of technical proficiency, we demonstrate that LINGUA MANGA can effectively assist both skilled programmers and low-code or even no-code users in addressing data curation challenges.

PVLDB Reference Format:

Zui Chen, Lei Cao, and Sam Madden. LINGUA MANGA : A Generic Large Language Model Centric System for Data Curation. PVLDB, 16(12): 4074 - 4077, 2023.
doi:10.14778/3611540.3611624

1 INTRODUCTION

In the era of big data, organizations are collecting vast amounts of data from various sources. However, the data is often messy, incomplete, or contains errors. To effectively perform data analytics or any other applications, it is often necessary to engage in a data curation process [8, 15], which includes tasks like data discovery, integration, and cleaning. Nonetheless, devising a data curation solution for a particular scenario can be a cumbersome and time-consuming process. It entails extensive communication of requirements, a collaboration between domain experts and programmers, multiple rounds of debugging and testing, and active maintenance to accommodate new use cases. This underscores the need for a *generic* system that enables users to efficiently and effectively address diverse data curation challenges and apply solutions to various applications.

However, building a system that addresses diverse data curation problems and applications is challenging. Data curation involves a range of distinct tasks, such as data discovery through table search, schema matching, entity resolution during data integration, and data imputation in data cleaning. Moreover, in real applications, various extra tasks are often involved in the data curation processes,

This work is licensed under the Creative Commons BY-NC-ND 4.0 International License. Visit <https://creativecommons.org/licenses/by-nc-nd/4.0/> to view a copy of this license. For any use beyond those covered by this license, obtain permission by emailing info@vldb.org. Copyright is held by the owner/author(s). Publication rights licensed to the VLDB Endowment.
Proceedings of the VLDB Endowment, Vol. 16, No. 12 ISSN 2150-8097.
doi:10.14778/3611540.3611624

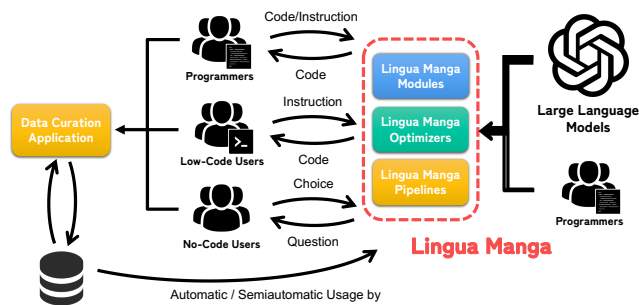


Figure 1: An illustration of the LINGUA MANGA System. such as event extraction, name extraction, anomaly detection, data summarization, visualization, table semantic parsing, etc. Worst yet, applications in different domains have diverse requirements. Effectively resolving data curation problems often relies heavily on users' domain knowledge, making developing a *general-purpose* solution improbable. For example, entity resolution over *books* focuses on text understanding, while entity resolution over *images* should rather pay special attention to image processing. As a result, existing data curation systems typically focus on only a few tasks on specific data formats. For example, Data Tamer [8] solves schema matching and entity resolution problems; ZenCrowd [4] targets entity linking; while CrowdDB [7] addresses fuzzy SQL semantics.

Recent advances in Large Language Models (LLMs) have shed light on this highly challenging yet critical data curation area. Pre-trained LLMs such as GPT-3 [10], Codex [6], ChatGPT [17], and LLaMA [5] have not only shown the ability to understand human needs and instructions but also to generate code like a programmer. Such ability could be leveraged to solve specific data curation problems like data imputation or entity resolution [2].

However, the current use of LLMs and LLM plugins is still far from the ultimate solution. Repeatedly calling LLM services can be costly and may lead to data privacy issues. For instance, consider a large e-commerce organization with thousands of documents and tables containing billions of rows – it would be neither affordable nor secure to let an LLM access the entire data lake to complete the tasks. LLMs tend to be less effective without access to enterprise data due to the gap between the enterprise data and the public data LLMs trained on. Thus, there is an urgent need to harness the power of LLMs systematically, efficiently, and securely. Properly integrating LLMs and traditional data curation methods could be a promising path toward a general-purpose data curation system.

In this demonstration, we present LINGUA MANGA, an LLM-centric system that enables the swift development of effective yet efficient data curation solutions. It benefits programmers as well as low-code or even no-code users. Specifically, LINGUA MANGA shows the following key properties:

- **User-friendly:** Users can quickly develop a data curation solution by utilizing the templates, built-in modules, and, most importantly, by effectively interacting with LLMs through LINGUA MANGA as a middleware.

- **Flexible:** Users can effortlessly communicate with LINGUA MANGA through natural language (NL) to inject domain knowledge or instructions into the data curation solution such that it can meet the needs of the specific applications or domains.

- **Intelligent:** LINGUA MANGA automatically optimizes the solution to fix errors, improves performance, continuously learns from the data, etc.

- **Highly Performant:** LINGUA MANGA minimizes the frequency of calling the LLM service, which incurs heavy computation and monetary costs, thus resulting in an economical, efficient, and privacy-preserving solution.

- **Label Efficient:** with our system, users can develop a data curation solution with no or only a few labeled examples from the specific application while still achieving accuracy comparable to the SOTA ML-based methods trained with thousands of labels.

2 PRELIMINARIES: PROMPT-TUNING AND CODE GENERATION IN LLMs

Starting from GPT-3 [10], LLMs’ size and computational cost have grown so much that users can no longer afford to fine-tune them for every downstream task. In the meantime, LLMs have also gained significant knowledge through training on large-scale corpora, leading to the emergence of a new paradigm known as prompt-tuning [9]. Using prompt-tuning, users can generalize LLMs to new tasks and data domains without modifying the LLM itself. This is achieved by incorporating prompts, such as descriptions, tags, and learnable components, into the input data.

Moreover, cutting-edge LLMs like ChatGPT [17] have demonstrated their effectiveness in interacting with humans using natural-language-based prompts [3]. These prompts are interpretable by humans. For instance, a straightforward prompt like "Please determine if the following entities are equivalent" can enable LLMs to carry out entity resolution. This capability is especially useful in LINGUA MANGA since it facilitates LLMs to interactively execute functions vaguely specified by users in natural language.

Furthermore, prompts like "Please write a code for entity resolution" trigger LLMs to generate custom software modules via code generation. GPT-3 [10], ChatGPT [17], Codex, and Copilot [6] are the mainstream LLMs that support code generation. However, existing methods that use these models to generate code have limitations [19]. They are typically inadequate for producing code with complex logic and building large-scale software. This calls for a new methodology to effectively utilize LLM Generated Code (LLMGC).

3 LINGUA MANGA

LINGUA MANGA comprises a Domain-Specific Language (DSL), a Compiler, an Optimizer, and a set of templates.

LINGUA MANGA is a **user-friendly** workflow system that enables users to quickly build data curation solutions by composing pipelines of logical operators. The system features a DSL to simplify the workflow-building process.

LINGUA MANGA is **automatic**. Like a relational database, it auto-compiles each logical operator into a physical, executable *module*.

LINGUA MANGA is **extensible**. LINGUA MANGA allows the programmers to implement their own physical modules, which can be seamlessly plugged into the system as long as they follow the standard interfaces of LINGUA MANGA .

LINGUA MANGA is **flexible**. To meet the specific need of different applications, the LINGUA MANGA optimizer interacts with users via LLMs to generate customized physical modules.

LINGUA MANGA is **intelligent**. The LINGUA MANGA optimizer uses LLMs to enhance the modules’ capability.

In addition, to further facilitate users in building pipelines, LINGUA MANGA offers templates, essentially groups of pre-built pipelines. Rather than creating a pipeline from scratch, LINGUA MANGA allows users to start with a pre-defined, well-optimized pipeline that the target application can directly use.

The fundamental observation underlying this research is that while current LLMs may not yet be able to produce exceedingly intricate and large software systems, they are proficient at generating modular code on a smaller scale. Moreover, LLMs possess learning and logical reasoning ability that enable them to complete data curation tasks directly, indicating that LLMs could accelerate the process of developing a data curation solution by automating parts of it and, thus, reduce the workload of software developers.

Next, we discuss the modules and the LINGUA MANGA optimizer, which are the critical components of LINGUA MANGA .

3.1 Modules in LINGUA MANGA

A module is a function $f : X \rightarrow Y$ that takes X as input and returns Y . Modules are usually viewed as black boxes, and their behavior is determined by their expected output. In LINGUA MANGA , modules can be classified into four types:

- **Custom Module:** As a basic module, a custom module is implemented with manually written code. It can be created by users with programming skills or provided by LINGUA MANGA as a default built-in module.

- **LLM Module:** An LLM itself can be a module. An LLM module tends to be more powerful than a traditional module because of the common sense knowledge and logical inference abilities possessed by an LLM. However, an LLM module requires a good task description as input; and LLM outputs typically need proper validation, as textual responses generated by the LLMs could be diverse and unstable. For instance, numerical validation that ensures the LLM’s output is within a reasonable value range can identify and correct potential errors in LLM output.

- **LLMGC Module:** An LLM can dynamically generate code to implement an LLMGC module, replacing the role of programmers. LINGUA MANGA allows LLMGC to call other modules in the system or use external tools such as a calculator or another pre-trained LLM. Although an LLMGC module might be less expressive than custom modules written by skilled programmers, it enables no-code or low-code users to implement or customize a module effortlessly.

- **Decorated Module:** As the most advanced module in LINGUA MANGA , a decorated module can comprise multiple basic modules and be enhanced by the optimizer that leverages the common sense knowledge encoded in the LLMs and their learning ability.

3.2 LINGUA MANGA Optimizer

The optimizer in LINGUA MANGA can improve the performance of physical modules within a pipeline by enhancing their efficiency and effectiveness, e.g., avoiding repeatedly calling the LLM service to minimize the execution costs and the potential data privacy leak. Unlike conventional database optimizers, the optimizer in LINGUA MANGA is modularized and can be selectively composed by users to serve specific applications best. The key modules in LINGUA MANGA include:

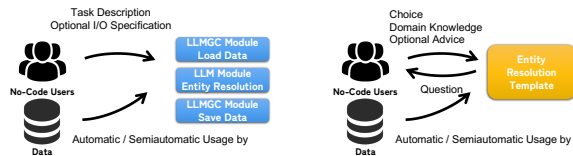
- **Validator:** It checks whether the target module behaves correctly on a few example test cases. It then uses the failed test cases to trigger the LLM to improve the target module and fix the errors. Specifically, the validator first calls an LLM to generate the suggestion by reading the code and the failure cases. Then, the code, failure cases, and the generated suggestion are sent to another LLM to generate a new version of the code. This validation cycle repeats until either all test cases are executed successfully, or a timeout ensues, leading to a re-generation of the LLMGC module until an additional timeout.

- **Simulator:** A simulator automatically generates a more efficient and equally effective alternative to a given module that already functions well. For example, a module that frequently calls LLM service can be expensive. Potentially, a supervised learning counterpart could replace it. This is because LLMs are versatile, while a data curation task typically only needs part of its functionality. It is thus possible to produce a module proficient in simulating the specific LLM functionality required by a task.

Because each module is treated as a black-box function, an ML-based simulator can replicate the target module through supervised learning. The target module will function as intended during initialization, and a control logic will decide when the simulated version should take over, such as after achieving the desired accuracy or reaching a certain level of confidence.

A simulator could improve accuracy as well. At first sight, this is counter-intuitive. Like the teacher-student model in machine learning, traditionally, the student model, which learns from a teacher model, is expected to be upper bounded by the teacher model on performance. However, studies have shown that self-training with filters [12, 16, 18, 20] can produce a superior student model because of better generalization. Moreover, the simulator will continuously monitor the real data flow in the original pipeline. It can thus constantly learn to adapt to the data distribution, potentially outperforming the original static module.

- **Connector:** Concerning efficiency and data privacy, it is crucial for applications to reduce the amount of data exposed to LLMs, while still ensuring the effectiveness. For instance, in an NL2SQL task, granting LLMs access to the entire table contents is neither secure nor cost-effective. However, relying solely on the table schema is not adequate to generate accurate predicates. To address this issue, a locally-running connector can be employed to manage the selective data upload to LLMs. A pre-defined connector for tabular data enables LLMs to execute SQL commands in local databases and obtain the resulting data while ensuring that the execution is limited to the queries specified by the user. In multi-modal data scenarios, appropriate pre-defined connectors are provided, such as connectors designed for handling extensive textual data.



(a) Custom pipeline.

(b) Built-in template.

Figure 2: An illustration of two possible Entity Resolution workflows using LINGUA MANGA .

4 DEMONSTRATION

We use three applications to demonstrate LINGUA MANGA . Fig. 5 shows its user interfaces with name extraction as an example task.

4.1 Entity Resolution: Effortless to the Novices

Consider the scenario where a technical novice wants to perform entity resolution on a dataset but lacks programming experience.

LINGUA MANGA effectively fulfills these requirements by providing a user-friendly solution. To begin, users can easily search for existing templates within the system. If none are available, they can create a simple pipeline that includes data loading, entity resolution, and data-saving operators, as depicted in Figure 2a. It is worth noting that the user does not have to write any code. Instead, they can simply describe the task to the LLMs using the suggested prompt templates and provide optional input and output specifications through examples. The LINGUA MANGA optimizer will then improve the performance automatically.

We evaluated the F1 scores of the solution produced by LINGUA MANGA on three entity resolution datasets: BeerAdvo-RateBeer, Fodors-Zagats, and iTunes-Amazon [1]. As shown in Table 1¹, it significantly outperforms the previous LLM-based method [2], while achieving performance comparable to supervised learning algorithms trained with hundreds or even thousands of labeled examples [13].

Table 1: Quantitative Experiment on Entity Resolution.

Dataset	Magellan [1]	Ditto [13]	FMs [2]	LINGUA MANGA
BeerAdvo-RateBeer	78.8	94.37	78.6	89.66
Fodors-Zagats	100.0	100.00	87.2	95.65
iTunes-Amazon	91.2	97.06	65.9	92.00

4.2 Name Extraction: Flexible for the Adeptes

Consider a scenario where a domain expert, as a low-code user with basic programming skills, aims to get a high-accuracy and efficient name extraction solution, i.e., find all person names in a text passage. In this case, the domain expert understands that name extraction typically involves three steps: tokenization, noun-phrase extraction, and tagging. As a result, the domain expert can create a pipeline of three individual operators (ignoring data loading and saving for ease of presentation), as shown in Fig. 3.

The domain expert recognizes that tagging is a complex process and is often the performance bottleneck. Therefore, he or she chooses to utilize the LLM module within LINGUA MANGA along with an example-based validator. The other two operators are relatively simple and thus are realized as LLMGC modules – using LLM to generate the code. Suppose the resulting LLMGC modules, such as the noun-phrase extraction module, are not precise enough. The

¹The numbers of the baselines are obtained from [2].

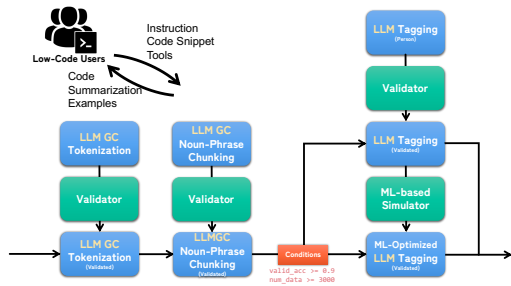


Figure 3: A possible Name Extraction pipeline.

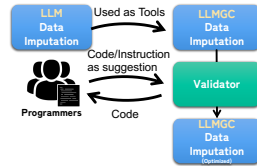


Figure 4: A possible Data Imputation pipeline.

domain expert can further enhance them by providing external tool APIs, domain knowledge instructions, or code snippets to optimize the code generation process. Additionally, since repeatedly using LLM to tag each data record can be costly, the domain expert may use the simulator to create an ML-based alternative solution that simulates LLM tagging with significantly lower expenses.

To showcase the versatility of LINGUA MANGA, we experimented on a real-world name extraction dataset obtained from a startup company. This task is unique in that it has to handle multi-lingual data, which significantly degrades the accuracy of name extraction. LINGUA MANGA quickly resolves this issue by incorporating an LLM language detection module and providing multi-lingual tools to the LLMGC module. This shows that LINGUA MANGA can swiftly adapt to specific needs, being flexible and efficient.

4.3 Data Imputation: Excellency with the Experts

Consider a scenario where an expert programmer needs to optimize a specific data imputation solution at all costs. This demo assumes the expert is dealing with the Buy dataset [2, 14], which comprises three attributes: products with names, descriptions, and manufacturers. The manufacturers’ names are missing.

Because LINGUA MANGA effectively integrates LLMs, it is advantageous over traditional rule-based or learning-based approaches. For example, without having the relevant knowledge, it is almost impossible for the traditional methods to deduce that the product “PlayStation 2 Memory Card 8MB” is produced by “Sony”. This is where the LLM module could help.

Experts can offer comprehensive guidelines or codes to create a better LLMGC module, as illustrated in Figure 4. These guidelines facilitate the validation procedure, enabling the LLMGC module to undergo iterative refinement. As a result, it can effectively use the LLM as an external tool to resolve complex cases while performing more efficiently than a pure LLM module on easy cases.

The LINGUA MANGA produced solution results in an accuracy of 94.48%, which is a remarkable improvement compared to the 16.2% accuracy by HoloClean [11]. It is comparable to IMP [14], which trains a Transformer model on thousands of training examples and achieves an accuracy of 96.5%. It is noteworthy that our optimized

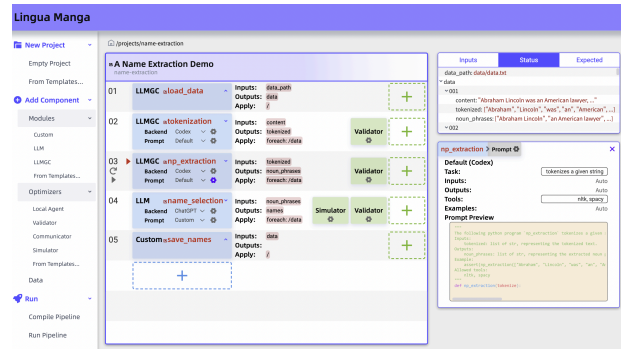


Figure 5: UI of the LINGUA MANGA System.

version, which combines the LLM module with the LLMGC module, is cost-effective, using only 1/6 LLM calls to achieve higher accuracy, compared to the version that only uses the LLM module to call LLM service repeatedly (accuracy 93.92%). Previous LLM-based research achieved an accuracy of 84.6% [2].

5 CONCLUSION

We introduced LINGUA MANGA, a user-friendly and versatile system designed to facilitate the development of data curation applications, leveraging the capabilities of LLMs. We demonstrated the efficiency, effectiveness, and user-friendliness of LINGUA MANGA through three data curation tasks. Nonetheless, it should be noted that LINGUA MANGA is currently in the prototype stage, which presents numerous opportunities for enhancements. These improvements encompass pipeline optimizations, supporting multi-modal applications, and implementing robust mitigation strategies to tackle LLM-induced hallucinations, etc.

REFERENCES

- [1] Sanjib et al. Das. [n.d.]. The Magellan Data Repository.
- [2] Avaniika Narayan et al. 2022. Can Foundation Models Wrangle Your Data? *VLDB* (2022).
- [3] Fatih Kadir Akin et al. 2023. Awesome ChatGPT Prompts.
- [4] Gianluca Demartini et al. 2012. ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *WWW*.
- [5] Hugo Touvron et al. 2023. LLaMA: Open and Efficient Foundation Language Models. (2023).
- [6] Mark Chen et al. 2021. Evaluating Large Language Models Trained on Code. (2021).
- [7] Michael J. Franklin et al. 2011. CrowdDB: answering queries with crowdsourcing. In *SIGMOD*.
- [8] Michael Stonebraker et al. 2013. Data Curation at Scale: The Data Tamer System. In *CIDR*.
- [9] Pengfei Liu et al. 2023. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.* (2023).
- [10] Tom B. Brown et al. 2020. Language Models are Few-Shot Learners. In *NeurIPS*.
- [11] Theodoros Rekatsinas et al. 2017. HoloClean: Holistic Data Repairs with Probabilistic Inference. *VLDB* (2017).
- [12] Timo Schick et al. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. (2023).
- [13] Yuliang Li et al. 2020. Deep Entity Matching with Pre-Trained Language Models. *VLDB* (2020).
- [14] Yinan Mei et al. 2021. Capturing Semantics for Imputation with Pre-trained Language Models. In *ICDE*.
- [15] André Freitas and Edward Curry. 2016. Big data curation. (2016).
- [16] Gautier Izacard and Edouard Grave. 2021. Distilling Knowledge from Reader to Retriever for Question Answering. In *ICLR*.
- [17] OpenAI. 2023. ChatGPT.
- [18] Timo Schick and Hinrich Schütze. 2021. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In *EACL*.
- [19] Yichen Xu and Yanqiao Zhu. 2022. A Survey on Pretrained Language Models for Neural Code Intelligence. (2022).
- [20] David Yarowsky. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *ACL*.