

VerifAI: Verified Generative AI

Nan Tang
HKUST (GZ) / HKUST
China
nantang@hkust-gz.edu.cn

Chenyu Yang*
HKUST (GZ)
China
cyang662@connect.hkust-gz.edu.cn

Ju Fan†
Renmin University
China
fanj@ruc.edu.cn

Lei Cao
University of Arizona / MIT
USA
lcao@csail.mit.edu

Yuyu Luo
HKUST (GZ) / HKUST
China
yuyuluo@hkust-gz.edu.cn

Alon Halevy
Meta AI
USA
ayh@fb.com

ABSTRACT

Generative AI has made significant strides, yet concerns about the accuracy and reliability of its outputs continue to grow. Such inaccuracies can have serious consequences such as inaccurate decision-making, the spread of false information, privacy violations, legal liabilities, and more. Although efforts to address these risks are underway, including explainable AI and AI regulation practices such as transparency, privacy protection, bias mitigation, and social and environmental responsibility, misinformation caused by generative AI will remain a significant challenge. We propose that verifying the outputs of generative AI from a data management perspective is an emerging issue for generative AI. This involves analyzing the underlying data from multi-modal data lakes, including text files, tables, and knowledge graphs, and assessing its quality and consistency. By doing so, we can establish a stronger foundation for evaluating the outputs of generative AI models. Such an approach can ensure the correctness of generative AI, promote transparency, and enable decision-making with greater confidence. Our vision is to promote the development of verifiable generative AI and contribute to a more trustworthy and responsible use of AI.

1 INTRODUCTION

Making effective decisions based on data requires data of high quality. The exact meaning of high quality varies from one application to another, but typically it means that the data has undergone verification, validation, and evaluation to ensure its reliability and accuracy for a specific use case or application. Furthermore, we may also require that the appropriate levels of security and privacy protection be in place. This is particularly crucial in fields such as finance, healthcare, and government, where the decisions based on the data can affect individuals and society as a whole. Organizations that rely on good data for making important decisions must take steps to ensure that the data they use is trustworthy and reliable.

Regrettably, real-world data is often incomplete, inconsistent, or inaccurate [7, 16]. Researchers across different fields have invested significant effort in addressing these issues. For instance,

the database community has developed methods for detecting data errors [2, 6, 17, 23, 24, 26, 27, 34–36, 40], while the natural language community has developed techniques for identifying fake news and misinformation [5, 38].

The issue of data quality becomes more important in the era of generative AI, where advanced generative models like GPT-4, Midjourney, and SlidesAI can generate intricate outputs. This technology has immense potential to transform various fields, including relational data synthesis, natural language question answering, image generation for advertising and marketing, and many others.

More specifically, numerous companies have audacious and transformative plans for integrating generative AI into a wide range of their products, including Microsoft (Excel, Word, PowerPoint, Bing Search), Google, Meta, Alibaba, and Baidu, and it is reasonable to expect this trend to proliferate beyond the big technology companies. Consequently, the data produced by these systems will be utilized for multiple purposes, such as decision-making, knowledge acquisition, and presentations. However, it is crucial to acknowledge that the accuracy and reliability of the generated data cannot be guaranteed. Therefore, it is vital to exercise caution when utilizing such data to make important decisions or draw conclusions.

There are significant risks associated with using “bad” data generated by generative AI, such as the following.

- (1) *Inaccurate decision-making*: Decisions based on bad data can lead to poor outcomes, e.g., financial losses.
- (2) *Spreading misinformation*: Bad data can be utilized to disseminate false information, and individuals may rely on the knowledge from generative AI irrespective of its accuracy.
- (3) *Privacy violations*: It can lead to privacy violations if personal information is shared or used inappropriately.
- (4) *Legal liabilities*: Organizations that make decisions based on bad data can face legal liabilities if they cause harm to individuals or violate laws and regulations.
- (5) *Damage to trust*: The use of bad data can erode trust in organizations, leading to long-term reputational damage.

Although the risks discussed above are already prevalent when using low quality data before the era of generative AI, they are significantly amplified with the widespread adoption of generative models. In addition to the many discussions about regulating generative AI, from a technical perspective there have been two main bodies of work on the topic, explainable AI and AI regulations.

Explainable AI (XAI) [32, 33, 42] seeks to provide explanations for the outputs produced by AI systems. While XAI can be a valuable tool for improving the trustworthiness of generative AI, it

*Chenyu Yang is a co-first author.

†Ju Fan is the corresponding author.

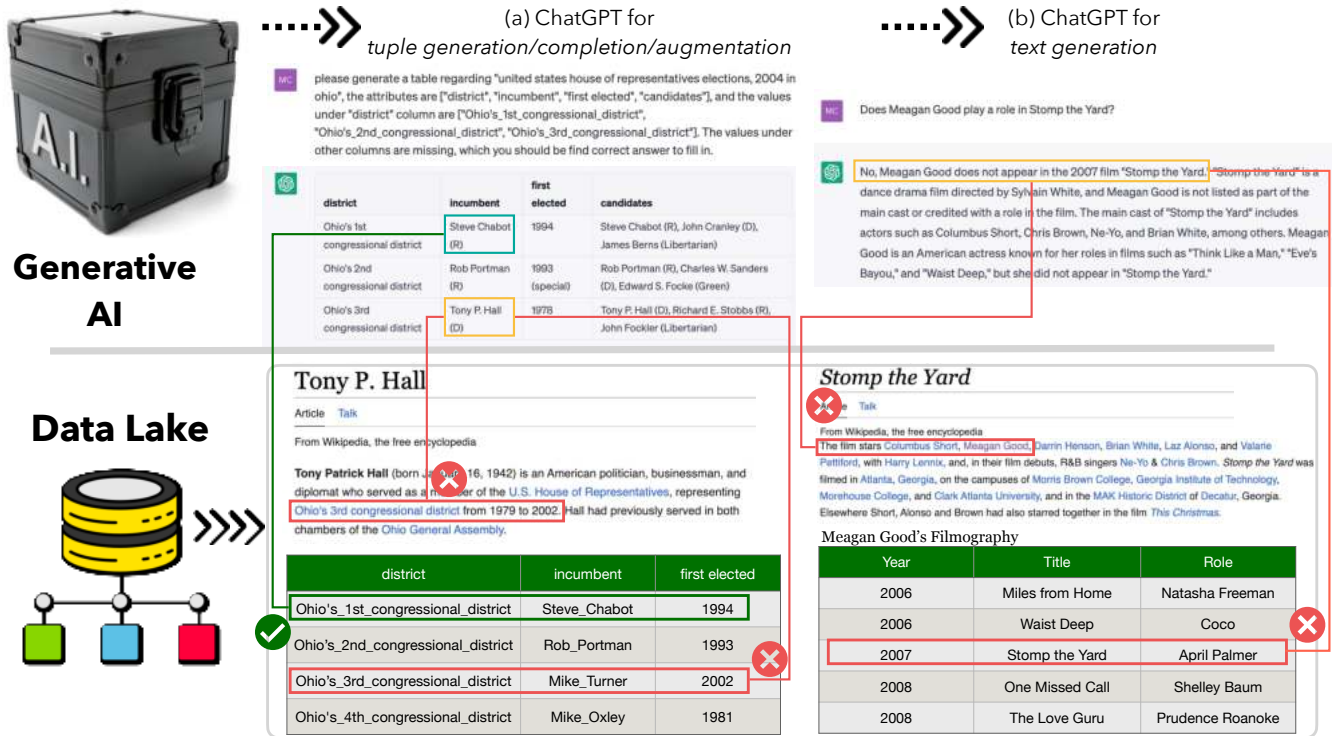


Figure 1: Generative AI can generate (a) values in tuples and (b) text. Our system, VerifAI, tries to either verify or refute generated value, by reasoning the (generated data, evidence) pair where the evidence is discovered from data lakes.

cannot completely eliminate concerns regarding the accuracy and reliability of generated outputs. This is due in part to the fact that generative AI models can be highly complex, which makes it difficult to provide a complete explanation of how a given output was produced.

AI regulations involve the ethical and accountable development and use of AI systems, which encompasses a range of practices such as transparency, privacy protection, bias mitigation, and social and environmental responsibility. Although model providers like OpenAI and Google are making efforts to regulate the development of generative AI models, the issue of flawed generative data is expected to continue to be a significant challenge.

Our Vision. We advocate for the establishment of a more reliable framework to evaluate the accuracy and reliability of generative AI outputs. We propose a data management perspective for verifying generative AI outputs that is based on identifying supporting data from data lakes and reasoning with it in order to assess the quality and consistency of the generated data. We believe that a verification approach can complement the aforementioned approaches in order to ensure that systems using generative AI are deployed responsibly, effectively, and in a trustworthy manner. It is important to clarify that our proposed approach focuses on verifying generative data that has ground truth, as opposed to subjective data [22].

We illustrate our approach with the following examples based on running our proposed framework VerifAI.

EXAMPLE 1. [Tuple generation/completion/augmentation (Figure 1(a)).] Consider the prompt that provides a serialized table with missing values in attribute **incumbent** and asks ChatGPT to complete these tuples. The new table with completed tuples by ChatGPT is shown below the prompt.

For the first tuple, VerifAI can “search” a tuple in the data lake that verifies the imputed and generative value that the **incumbent** value is correct.

For the third tuple, VerifAI can search a tuple and a text file in the data lake that both validate the imputed value to be incorrect.

[Text generation (Figure 1(b)).] We asked ChatGPT a question “Does Meagan Good play a role in Stomp the Yard” and got an answer as shown in the figure.

VerifAI can search a text file and a tuple in the data lake that both validate the generated text to be incorrect. □

The examples above only scratch the surface of generated (bad) data. These data can be leveraged for downstream applications such as analytical (e.g., OLAP) queries and data visualization, as well as query acquisition. However, if not properly managed, generative data can have negative and even disastrous consequences.

Despite efforts from model providers (e.g., OpenAI) and retrieval-enhanced methods [3, 9, 15] for improving the accuracy of generative AI during the data generation process, the spread of misinformation from generative AI will remain a significant problem. Our

proposed post-generation verification approach can also complement these methods by enhancing the overall reliability of generative AI. By combining our verification approach with those existing and rapidly evolving generative methods, we can create a synergistic effect that further improves the accuracy and reliability of generative AI.

Challenges. Generative AI can generate data in various contexts and domains, utilizing the *world knowledge* that the model has learned [15]. However, when the generated data is dirty or inaccurate, traditional data cleaning [13, 25, 30, 37, 46] and data integration [11, 14, 31, 39, 41, 43, 44] methods that rely on the data at hand may not be sufficient. Moreover, when generative AI is used in the context of a specific enterprise, we may need to consult data that is specific to the enterprise in order to verify the correctness of the generated data, raising several new challenges.

- (C1) *Indexing and searching multi-modal data lakes.* Although there have been efforts to manage data lakes with relational data [18] and textual data [29], indexing multi-modal data lakes at scale and effectively retrieving top- k data instances cross data modalities remains an unsolved problem.
- (C2) *Cross-modal data verification.* Data matching is a key concept in data integration and data cleaning [11]. However, matching and reasoning across different data modalities are not well-addressed.
- (C3) *Trust of heterogeneous datasets in multi-modal data lakes.* Evaluating the trustworthiness of web sources for knowledge fusion has been well studied [12]. However, evaluating the trustworthiness of different datasets in data lakes, particularly when they are not well curated, remains an open problem.
- (C4) *Provenance of the verification process.* Fully automated verification of generative AI outputs can be challenging. It is important to store the lineage of the end-to-end verification process, in case the retrieved data from data lakes is flawed or incomplete, or the verification process itself makes mistakes. This allows for later human checks or debugging.

Contributions. In this paper, we present **VerifAI**, a framework for verified generative AI, which offers the following contributions:

- (1) **A modularized framework.** We propose a modular framework for verifying generative data that is extendable to different types of data sources and generative data. The framework comprises three main components: an **Indexer**, a **Reranker**, and a **Verifier**, as illustrated in Figure 2 (more details can be found in Section 3).
 - The **Indexer** module serves the purpose of indexing datasets from diverse modalities including but not limited to tuples, tables, text, and knowledge graphs. It comprises generic and coarse-grained indexes that aid in this task.
 - The **Reranker** module is responsible for reranking the top- k data sources from the **Indexer** with respect to a specific generated data object. This step is more fine-grained and task-specific in nature, in order to further optimize the ranking of the retrieved data sources.

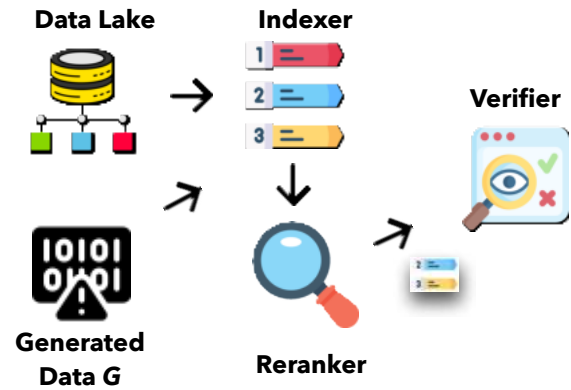


Figure 2: An overview of VerifAI.

For each retrieved data instance, the **Verifier** module will determine whether it can verify or invalidate the generated data object.

- (2) **Experiments.** We show that **VerifAI** achieves high accuracy for generated tables and text (Section 4), which demonstrates the feasibility of using multi-modal data lakes to verify generative AI.
- (3) **Open problems.** Contribution (1) highlighted the potential of **VerifAI** in addressing challenges C1 and C2. However, there are still open problems for these challenges, as well as for other challenges such as C3 and C4, which will be discussed in Section 5.

2 VERIFIED GENERATIVE AI: THE PROBLEM

Generative Data. Generative data generally refers to data that is created or generated by a model or algorithm, rather than being directly observed or collected from the real world.

In the context of this work, we specifically focus on data generated by large language models, such as ChatGPT, using natural language generation techniques. This includes text data such as paragraphs and sentences, or tabular data such as tuples and tables, but not other data modalities such as image data or graph data.

Multi-Modal Data Lakes. A multi-modal data lake provides a single repository or data store for storing and managing multiple types of data, including structured, semi-structured, and unstructured data. This can include tables, text, knowledge graphs, images, audio, and other forms of digital content.

This work focuses on showcasing data lakes that are designed to handle and integrate tabular data and text data.

In the rest of this paper, we will use the term *data object* for generated data and *data instance* for a specific unit of data within a data lake, which can take the form of a tuple, table, or text.

We will briefly discuss potential strategies for supporting and integrating other data modalities (e.g., knowledge graphs) into a multi-modal data lake in Section 5.

Verified Generative AI. Given a generated data object g , and a multi-modal data lake L , the problem of *verified generative AI* involves discovering a set of data instances L^g from L relevant to g ,

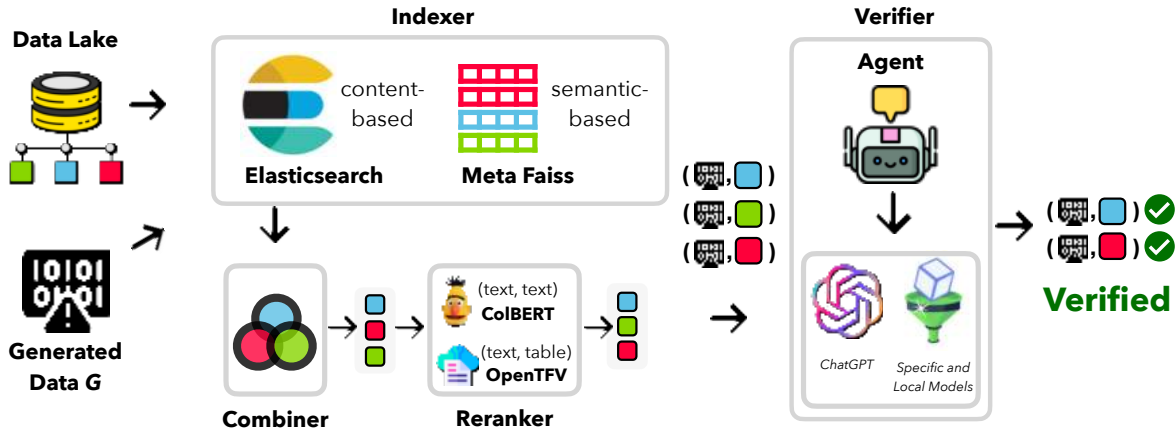


Figure 3: Modules of VeriFAI.

and verifying each data instance $\mathbf{x} \in \mathbf{L}^g$ as a mapping from (g, \mathbf{x}) to a ternary value as $\text{verify}(g, \mathbf{x}) \rightarrow 0|1|2$, indicating **verified** | **refuted** | **not related**, respectively.

Remark. The verification process is specific to the application and requires additional metadata from the user. For instance, when given a tuple, the verification requirement could be either on the entire tuple or on a specific column e.g., only verifying the attribute **incumbent** in Figure 1(a).

3 THE DESIGN OF VERIFAI

An overview of different modules of VeriFAI is given in Figure 3.

3.1 Indexer

The **Indexer** module is designed based on two key principles:

- *Task-agnostic:* It can support a wide range of tasks and use cases, making it versatile and adaptable to different needs.
- *Support for both content- and semantic-based search:* It can handle traditional string similarity search as well as vector-based search, enabling more advanced search and retrieval capabilities.

Based on the aforementioned principles, VeriFAI currently consists of two indexes.

- Elasticsearch [19] supports content-based search, where tables or text files are serialized as strings and then indexed by Elasticsearch.
- Meta Faiss [28] is a library for efficient similarity search and clustering of dense vectors. We first use embedding techniques (e.g., tuple-to-vec or text-to-vec using BERT) to convert tuples or chunked text files into vectors, which are then indexed by Faiss.

The above two types of indexes are commonly used for indexing data lakes. For example, Aurum [18] utilizes Elasticsearch, while Microsoft Azure Cognitive Search (<https://azure.microsoft.com/en-us/products/search>) indexes embeddings (or vectors) of textual data. Combining these two approaches can enhance recall and serve as a foundation for indexing data lakes more effectively.

Remark. As these indexes are task-agnostic, the retrieved top- k results may have low accuracy. To ensure the retrieval of correct data instances, k is typically set to a large number (e.g., 100 to 1000 [29]). However, in order to achieve good accuracy with a much smaller number of results after reranking, it is desired to perform task-aware reranking over the retrieved results.

Combiner. While different indexes use different techniques (e.g., content- or semantic-based), their retrieved results typically overlap. The **Combiner** simply combines these retrieved results from multiple indexes and removes duplicates. Due to its simplicity, we did not depict and discuss it in Figure 2.

3.2 Reranker

The retrieved results from multiple indexes are initially ranked in a coarse-grained fashion using a task-agnostic similarity measure. However, to achieve a more fine-grained and task-specific ranking of the results, the **Reranker** module is utilized. The primary idea is that, following the reranking process, we only need to focus on a limited number of top- k' retrieved results (e.g., $k' = 5$).

Next, we describe two different approaches, one for reranking (text, text) pairs, and the other for reranking (text, table) pairs.

- **(text, text) Reranker:** We adopt ColBERT [29], a holistic approach to compare each token of a query and each token of a retrieved text file, resulting in a more precise score.
- **(text, table) Reranker:** We have devised a semantic-based reranking method between a natural language statement and a table, called OpenTFV [21], for open domain table-based fact verification.

Remark. We are currently working on expanding our support for different types of fine-grained **Rerankers**, such as those used for reranking (tuple, tuple) or (text, tuple) pairs.

3.3 Verifier

This reasoning module in our system determines whether generated data is verified or refuted, based on a retrieved data instance. It utilizes multiple **Verifiers**, each tailored to a specific task. An **Agent** decides which **Verifier** to use for a given task.

In **VerifAI**, we utilize two types of **Verifiers**. The first type is a one-size-fits-all model such as ChatGPT. The second type is composed of specific and localized models designed for different tasks, such as the table fact verification model [20] for (text, table) verification, and a fine-tuned RoBERTa model for (tuple, tuple) verification.

While ChatGPT can be used by default for simplicity, there are two main reasons why we support specific and localized models.

- (1) **Data privacy.** Many applications (e.g., healthcare, government) contain highly sensitive information. One legitimate concern when using externally hosted models such as ChatGPT is data privacy. The model could potentially learn and retain sensitive information from the data it has seen [1].
- (2) **Better accuracy.** Specific and local models, when being fine-tuned on specific datasets and tasks, can oftentimes outperform generic models such as ChatGPT, as will be shown later in Section 4.

Remark. When retrieving data for a generated data instance, it's possible that we may retrieve multiple data instances that either verify or refute the generated data object. This can occur for several reasons, e.g., incorrect data instances being retrieved. Therefore, understanding the trustworthiness of different data sources and maintaining the provenance of verification are important.

4 PRELIMINARY RESULTS

In this section, we showcase preliminary experimental results that highlight the initial achievements of **VerifAI** in facilitating the verification of generative AI.

4.1 Setting

We consider three verification tasks.

Generative AI Task	Retrieved Data Modalities
tuple completion	tuples
tuple completion	textual files
text generation	relational tables

Note that the (text, text) verification problem is essentially equivalent to the standard fact-checking problem in the natural language processing community [29], which has already been demonstrated to be viable. Therefore, for the sake of this discussion, we will focus primarily on scenarios that involve tuples or tables.

Tuples in need of verification. We collected 100 tuples from web tables. For each tuple, we randomly removed a non-key attribute cell value and then asked ChatGPT to infer the missing value by utilizing the given template provided below. If multiple tuples share the same schema, we can handle them together as a batch.

Prompt template of tuple completion with ChatGPT			
Question:			
Table name			
column 1	column 2	...	column n
a_1	NaN	...	z_1
a_2	b_2	...	NaN
...
Please fill the missing values, annotated by NaN			

ChatGPT will produce a complete table that doesn't have any missing values. An example of this is shown in Figure 1(a). However, each tuple that has had missing values imputed must be verified.

Textual claims in need of verification. We perform a controlled study to assess textual claims, employing 1,300 textual claims from the TabFact [8] benchmark, which is currently the most advanced benchmark for verifying the credibility of textual hypotheses by utilizing a given table. This benchmark is suitable for our aim of verifying textual claims. Nevertheless, it's worth mentioning that although TabFact offers a (textual claim, table) pair, our task demands us to search the data lakes for table(s), which is an exceedingly more difficult task.

Multi-Modal Data Lakes. We utilized 19,498 tables that contain 269,622 tuples, and 13,796 textual files. Of these, 16,573 tables were taken from TabFact [8], while the others were sourced from WikiTable-TURL [10]. Furthermore, since some cells contain references to entities, such as links to Wikipedia pages, we gathered 13,796 entities and retrieved their corresponding text from the associated Wikipedia pages to obtain the textual files.

Verifier. ChatGPT can serve as the default **Verifier** for both data types. The template employed to interact with ChatGPT during the verification process is presented below.

Prompt template of verified with ChatGPT

Please use the evidence below to validate the generative data.

Evidence: [Use the retrieved tuple/table/text]

Generative Data: [Data object to be verified]

Result: Verified/Refuted/Not Related + Further explanation

It is worth mentioning that we have also developed specific and localized models for certain scenarios. In the case of evaluating (tuple, tuple) pairs, the local model's accuracy is comparable to ChatGPT [4]; therefore, we only present ChatGPT's results. For (text, table) verification, we employed our previous work PASTA [20] as an alternative to ChatGPT. Later in the section, we will show that PASTA outperforms ChatGPT for (text, table) verification.

Evaluation Metrics for Retrieval. We use Elasticsearch [19] to retrieve the top-3 tuples and top-3 text files for each tuple with an imputed missing value, top-5 tables for each textual claim. As we have a small number of relevant data, we evaluate the retrieval process using only the recall metric.

However, defining "relevance" presents a challenge in evaluating the performance of retrieval. Ideally, we would consider all data instances in the data lakes that can support or refute the generative data to be relevant. However, manually labeling relevant data for each tuple or claim would be time-consuming and impractical. Therefore, we take a different approach in this paper. As each tuple has an original, complete counterpart in the data lake, we consider it to be relevant evidence, while the remaining tuples are considered irrelevant. Moreover, we consider the textual files about entities present in a tuple to be relevant evidence as well. Additionally, each textual claim is associated with a corresponding table in the original dataset. Thus, we consider the corresponding table to be relevant evidence, while the remaining tables are considered irrelevant.

Due to space limitations, we will not report the effectiveness of reranking in our study. However, it's worth noting that reranking

Generated data type	retrieved data type	recall
tuple	tuple	0.99
	text	0.58
textual claim	table	0.88

Table 1: Recall on retrieved data instances.

(text, text) pairs has been proven effective in ColBERT [29], and reranking (tuple, tuple) pairs has been discussed in RetClean [4].

Evaluation Metric for Verifier. For evaluating the Verifier’s performance, we use accuracy as the measure. A Verifier’s decision is considered correct in one of the following three cases:

- (1) When the retrieved data instance supports the imputed tuple or claim, the Verifier outputs “true”;
- (2) When the retrieved table refutes the imputed tuple or claim, the Verifier outputs “false”;
- (3) When the retrieved data instance can neither support nor refute it, the Verifier outputs “not related”. To compare ChatGPT with PASTA that only offers two different answers: “true” or “false”, in this case, we consider it’s also correct when PASTA outputs “false”.

4.2 Preliminary Results

The accuracy of ChatGPT in imputing missing values for tuples and determining the correctness of claims is only **0.52** and **0.54**, respectively, in the absence of additional data. These findings emphasize the significance of verifying generative data to guarantee accuracy and dependability.

Table 1 displays the retrieval outcomes, revealing that the retrieval module performs well for (tuple, tuple) and (textual claim, table). However, it doesn’t retrieve the associated text files effectively based on a tuple. This is due to the fact that we simply utilized Elasticsearch as the Indexer and only retrieved three text files. We anticipate that the retrieval performance will improve when we expand the number of retrieved files and conduct further experiments by adding the Reranker.

Table 2 reveals that as the Verifier, ChatGPT can accurately determine whether the imputed value is correct or not, with a high accuracy of **0.88**. Regarding textual claims, we present results in two settings. When a relevant table is retrieved and provided as evidence to the Verifier in the form of a (text, relevant table) pair, PASTA is able to verify the textual claim with higher accuracy than ChatGPT based on the table. However, many retrieved tables are irrelevant to the claim. In all (text, retrieved table) pairs, PASTA’s accuracy drops to 0.72 because it hasn’t encountered this scenario during training. On the other hand, ChatGPT has superior generalization capabilities and performs better than PASTA when dealing with many irrelevant tables.

In Figure 4, we present a case of verifying a textual claim based on retrieved tables using ChatGPT. VerifAI retrieves two tables E_1 and E_2 , where E_1 can be used with an aggregation query to refute the claim while E_2 is not related because it is for the year 1959. The red boxes in Figure 4 show that ChatGPT can provide not only a verification result but also some explanation. Hence, when the retrieved data is highly related to the generative data,

	ChatGPT	PASTA
(tuple, tuple+text)	0.88	NA
(text, relevant table)	0.75	0.89
(text, retrieved table)	0.91	0.72

Table 2: Evaluation on Verifier.

Claim: In 1954 u.s. open (golf), the cash prize for tommy bolt, fred haas, and ben hogan was 960 in total.
(Ground Truth: a false claim that should be **Refuted**)

Retrieved Evidence and Verification

Table E1: 1954 u.s. open (golf)

```

I place | player | country | score | to par | money |
| t1 | ed furgol | united states | 71 + 70 + 71 + 72 = 284 | + 4 | 1 6000 |
| t2 | gene littler | united states | 70 + 69 + 76 + 70 = 285 | + 5 | 3600 |
| t3 | lloyd mangrum | united states | 72 + 71 + 72 + 71 = 286 | + 6 | 1500 |
| t3 | dick mayer | united states | 72 + 71 + 70 + 73 = 286 | + 6 | 1500 |
| t5 | bobby locke | south africa | 74 + 70 + 74 + 70 = 288 | + 8 | 960 |
| t6 | tommy bolt | united states | 72 + 72 + 73 + 72 = 289 | + 9 | 570 |
| t6 | fred haas | united states | 73 + 73 + 71 + 72 = 289 | + 9 | 570 |
| t6 | ben hogan | united states | 71 + 70 + 76 + 72 = 289 | + 9 | 570 |
| t6 | shelley mayfield | united states | 73 + 75 + 72 + 69 = 289 | + 9 | 570 |
| t6 | billy joe patton (a) | united states | 69 + 76 + 71 + 73 = 289 | + 9 | 1 0 1
                    
```

Verification result: **Refuted**. **Explanation:** The cash prize for Tommy Bolt, Fred Haas, and Ben Hogan was \$570 each, totaling \$1710.

Table E2: 1959 u.s. open (golf)

```

I player | country | year (s) won | total | to par | finish |
| ben hogan | united states | 1948, 1950, 1951 , 1953 | 287 | + 7 | t8 |
| cary middlecoff | united states | 1949, 1956 | 294 | + 14 | t19 |
| liack fleck | united states | 1955 | 294 | + 14 | t19 |
| liulus boros | united states | 1952 | 297 | + 17 | t28 |
| tommy bolt | united states | 1958 | 301 | + 21 | t38 | v2:
                    
```

Verification result: **Not related**.

Figure 4: Verifying a textual claim using retrieved tables.

local models like PASTA have higher accuracy while protecting privacy. In contrast, ChatGPT is better at generalizing and providing explanations for further judgments. Users can select the appropriate model based on their requirements.

In summary, we have evaluated the usefulness of VerifAI through various experiments and demonstrated several use cases in Figures 1 and 4.

5 OPEN PROBLEMS

In this paper, we propose a framework called VerifAI that addresses the growing concern about the reliability of generative AI, which is leading to the spread of misinformation at an alarming rate. The modular design of VerifAI enables the verification of generated data using multi-modal data lakes, paving the way for research activities that will produce practical solutions for enhancing the reliability of generative AI. In addition, we have identified important open problems that can make significant advances in this field and improve the reliability of generative AI.

- **Cross-Modal Data Discovery.** Data discovery is a challenging problem in data preparation [18], particularly in data lakes that contain heterogeneous data stored in various formats, including

structured (e.g., tables), semi-structured (e.g., graphs), and unstructured data (e.g., images and videos).

Unlike data lakes that contain only relational tables, discovering data from different modalities requires addressing the heterogeneity of the data.

To address this issue, a promising direction is to explore *cross-modal representation learning*, which involves encoding data from different modalities into a homogeneous vector space. This approach can facilitate a unified data discovery process, such as using a semantic-based indexer, as illustrated in Figure 3.

• **Cross-Modal Verification.** In addition to textual and relational data, datasets in other modalities, such as knowledge graph entities (or small sub-graphs), can contain valuable information for verifying generative AI. As discussed earlier, generic models like ChatGPT may not provide a comprehensive solution for reasoning due to challenges such as privacy and accuracy.

Therefore, a promising direction is to develop local models that are specifically trained for certain use cases, such as (text, knowledge graph entity) or (tuple, text). By focusing on specific cases, these local models can provide more accurate and effective solutions for verifying generative AI.

• **Trustworthiness of Data Sources.** The accuracy of discovering and verifying data across different modalities in a data lake can be influenced by the quality and reliability of the underlying data sources.

Therefore, it is crucial to assess the trustworthiness of different sources accurately to enhance the overall accuracy and reliability of the entire verification process.

• **Provenance Management.** It is important to maintain a record of the provenance of data instances or sources used in verification to facilitate further human checking or debugging.

• **Managing Data Generated by Generative AI.** Generative AI solution providers, including OpenAI, are continuously collecting prompts and generated data. While this information is valuable for improving generative AI models, it can also be useful for end-users, particularly enterprise users.

Therefore, a promising direction is to explore how to manage the (conversational) prompts and data generated by generative AI to enable better prompt engineering and data lineage tracking, similar to ModelDB [45] which is used for managing machine learning models. Such a solution would enable end-users to better facilitate effective collaboration between different stakeholders.

ACKNOWLEDGMENTS

This work was partly supported by NSF (2327954), the NSF of China (62122090, 62072461 and 62072458), the Fund for Building World-Class Universities (Disciplines) of Renmin University of China, the Beijing Natural Science Foundation (L222006), and the Research Funds of Renmin University of China and University of Arizona.

REFERENCES

- [1] [n.d.]. ChatGPT tied to Samsung's alleged data leak. <https://cybernews.com/news/chatgpt-samsung-data-leak/>. Accessed: 2023-04-06.
- [2] Ziawasch Abedjan, Xu Chu, Dong Deng, Raul Castro Fernandez, Ihab F. Ilyas, Mourad Ouzzani, Paolo Papotti, Michael Stonebraker, and Nan Tang. 2016. Detecting Data Errors: Where are we and what needs to be done? *Proc. VLDB Endow.* (2016).
- [3] Mohammad Shahmeer Ahmad, Zan Ahmad Naeem, Mohamed Y. Eltabakh, Mourad Ouzzani, and Nan Tang. 2023. RetClean: Retrieval-Based Data Cleaning Using Foundation Models and Data Lakes. *CoRR abs/2303.16909* (2023). <https://doi.org/10.48550/ARXIV.2303.16909> arXiv:2303.16909
- [4] Mohammad Shahmeer Ahmad, Zan Ahmad Naeem, Mohamed Y. Eltabakh, Mourad Ouzzani, and Nan Tang. 2023. RetClean: Retrieval-Based Data Cleaning Using Foundation Models and Data Lakes. *CoRR abs/2303.16909* (2023).
- [5] Fatma Arslan, Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2020. A Benchmark Dataset of Check-Worthy Factual Claims. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, Munmun De Choudhury, Rumi Chunara, Aron Culotta, and Brooke Foucault Welles (Eds.). AAAI Press, 821–829. <https://ojs.aaai.org/index.php/ICWSM/article/view/7346>
- [6] Chengliang Chai, Lei Cao, Guoliang Li, Jian Li, Yuyu Luo, and Samuel Madden. 2020. Human-in-the-loop Outlier Detection. In *SIGMOD Conference*. ACM, 19–33.
- [7] Chengliang Chai, Jiabin Liu, Nan Tang, Ju Fan, Dongjing Miao, Jiayi Wang, Yuyu Luo, and Guoliang Li. 2023. GoodCore: Data-effective and Data-efficient Machine Learning through Coreset Selection over Incomplete Data. *Proc. ACM Manag. Data* 1, 2 (2023), 157:1–157:27.
- [8] Wenhui Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyu Zhou, and William Yang Wang. 2019. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164* (2019).
- [9] Zui Chen, Zihui Gu, Lei Cao, Ju Fan, Samuel Madden, and Nan Tang. 2023. Symphony: Towards Natural Language Query Answering over Multi-modal Data Lakes. In *13th Conference on Innovative Data Systems Research, CIDR 2023, Amsterdam, The Netherlands, January 8-11, 2023*. www.cidrdb.org. <https://www.cidrdb.org/cidr2023/papers/p51-chen.pdf>
- [10] Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. 2022. Turl: Table understanding through representation learning. *ACM SIGMOD Record* 51, 1 (2022), 33–40.
- [11] AnHai Doan, Alon Y. Halevy, and Zachary G. Ives. 2012. *Principles of Data Integration*. Morgan Kaufmann. <http://research.cs.wisc.edu/dibook/>
- [12] Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. 2015. Knowledge-Based Trust: Estimating the Trustworthiness of Web Sources. *Proc. VLDB Endow.* 8, 9 (2015), 938–949.
- [13] Amr Ebaid, Ahmed K. Elmagarmid, Ihab F. Ilyas, Mourad Ouzzani, Jorge-Arnulfo Quian -Ruiz, Nan Tang, and Si Yin. 2013. NADEEF: A Generalized Data Cleaning System. *Proc. VLDB Endow.* 6, 12 (2013), 1218–1221. <https://doi.org/10.14778/2536274.2536280>
- [14] Muhammad Ebraheem, Saravanan Thirumuruganathan, Shafiq R. Joty, Mourad Ouzzani, and Nan Tang. 2018. Distributed Representations of Tuples for Entity Resolution. *Proc. VLDB Endow.* 11, 11 (2018), 1454–1467. <https://doi.org/10.14778/3236187.3236198>
- [15] Sebastian Borgeaud et al. 2021. Improving language models by retrieving from trillions of tokens. *CoRR abs/2112.04426* (2021).
- [16] Wenfei Fan and Floris Geerts. 2012. *Foundations of Data Quality Management*. Morgan & Claypool Publishers.
- [17] Wenfei Fan, Jianzhong Li, Nan Tang, and Wenyuan Yu. 2014. Incremental Detection of Inconsistencies in Distributed Data. *IEEE Trans. Knowl. Data Eng.* 26, 6 (2014), 1367–1383. <https://doi.org/10.1109/TKDE.2012.138>
- [18] Raul Castro Fernandez, Ziawasch Abedjan, Famiem Koko, Gina Yuan, Samuel Madden, and Michael Stonebraker. 2018. Aurum: A Data Discovery System. In *ICDE*.
- [19] Clinton Gormley and Zachary Tong. 2015. *Elasticsearch: The Definitive Guide* (1st ed.). O'Reilly Media, Inc.
- [20] Zihui Gu, Ju Fan, Nan Tang, Preslav Nakov, Xiaoman Zhao, and Xiaoyong Du. 2022. PASTA: Table-Operations Aware Fact Verification via Sentence-Table Cloze Pre-training. In *EMNLP*.
- [21] Zihui Gu, Ruixue Fan, Xiaoman Zhao, Meihui Zhang, Ju Fan, and Xiaoyong Du. 2022. OpenTFV: An Open Domain Table-Based Fact Verification System. In *SIGMOD*.
- [22] Alon Y. Halevy. 2020. Subjective Databases. In *ENASE*.
- [23] Shuang Hao, Chengliang Chai, Guoliang Li, Nan Tang, Ning Wang, and Xiang Yu. 2023. HOFD: An Outdated Fact Detector for Knowledge Bases. *IEEE Trans. Knowl. Data Eng.* 35, 10 (2023), 10775–10789. <https://doi.org/10.1109/TKDE.2023.3248223>
- [24] Shuang Hao, Nan Tang, Guoliang Li, and Jianhua Feng. 2018. Discovering Mis-Categorized Entities. In *34th IEEE International Conference on Data Engineering, ICDE 2018, Paris, France, April 16-19, 2018*. IEEE Computer Society, 413–424. <https://doi.org/10.1109/ICDE.2018.00045>
- [25] Jian He, Enzo Veltri, Donatello Santoro, Guoliang Li, Giansalvatore Mecca, Paolo Papotti, and Nan Tang. 2016. Interactive and Deterministic Data Cleaning. In *Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 - July 01, 2016*, Fatma Özcan, Georgia Koutrika, and Sam Madden (Eds.). ACM, 893–907. <https://doi.org/10.1145/2882903.2915242>
- [26] Alireza Heidari, Joshua McGrath, Ihab F. Ilyas, and Theodoros Rekatsinas. 2019. HoloDetect: Few-Shot Learning for Error Detection. In *Proceedings of*

- the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019. ACM, 829–846. <https://doi.org/10.1145/3299869.3319888>
- [27] Ihab F. Ilyas and Felix Naumann. 2022. Data Errors: Symptoms, Causes and Origins. *IEEE Data Eng. Bull.* 45, 1 (2022), 4–9. <http://sites.computer.org/debull/A22mar/p4.pdf>
- [28] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547.
- [29] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. *CoRR* (2020).
- [30] Zuhair Khayyat, Ihab F. Ilyas, Alekh Jindal, Samuel Madden, Mourad Ouzzani, Paolo Papotti, Jorge-Arnulfo Quiané-Ruiz, Nan Tang, and Si Yin. 2015. BigDancing: A System for Big Data Cleansing. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, Victoria, Australia, May 31 - June 4, 2015*, Timos K. Sellis, Susan B. Davidson, and Zachary G. Ives (Eds.). ACM, 1215–1230. <https://doi.org/10.1145/2723372.2747646>
- [31] Yuliang Li, Jinfeng Li, Yoshihiko Suhara, AnHai Doan, and Wang-Chiew Tan. 2020. Deep Entity Matching with Pre-Trained Language Models. *Proc. VLDB Endow.* 14, 1 (2020), 50–60. <https://doi.org/10.14778/3421424.3421431>
- [32] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2021. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* (2021).
- [33] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2021. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* 23, 1 (2021), 18. <https://doi.org/10.3390/E23010018>
- [34] Yuyu Luo, Chengliang Chai, Xuedi Qin, Nan Tang, and Guoliang Li. 2020. Interactive Cleaning for Progressive Visualization through Composite Questions. In *ICDE*. IEEE, 733–744.
- [35] Mohammad Mahdavi, Ziawasch Abedjan, Raul Castro Fernandez, Samuel Madden, Mourad Ouzzani, Michael Stonebraker, and Nan Tang. 2019. Raha: A Configuration-Free Error Detection System. In *Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019*, Peter A. Boncz, Stefan Manegold, Anastasia Ailamaki, Amol Deshpande, and Tim Kraska (Eds.). ACM, 865–882. <https://doi.org/10.1145/3299869.3324956>
- [36] Abdulhakim Ali Qahtan, Ahmed K. Elmagarmid, Mourad Ouzzani, and Nan Tang. 2018. FAHES: Detecting Disguised Missing Values. In *34th IEEE International Conference on Data Engineering, ICDE 2018, Paris, France, April 16-19, 2018*. IEEE Computer Society, 1609–1612. <https://doi.org/10.1109/ICDE.2018.00188>
- [37] Theodoros Rekatsinas, Xu Chu, Ihab F. Ilyas, and Christopher Ré. 2017. HoloClean: Holistic Data Repairs with Probabilistic Inference. *Proc. VLDB Endow.* 10, 11 (2017), 1190–1201. <https://doi.org/10.14778/3137628.3137631>
- [38] Shivam Sharma, Firoj Alam, Md. Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Y. Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022. Detecting and Understanding Harmful Memes: A Survey. *CoRR* abs/2205.04274 (2022).
- [39] Michael Stonebraker and Ihab F. Ilyas. 2018. Data Integration: The Current Status and the Way Forward. *IEEE Data Eng. Bull.* 41, 2 (2018), 3–9. <http://sites.computer.org/debull/A18june/p3.pdf>
- [40] Saravanan Thirumuruganathan, Laure Berti-Équille, Mourad Ouzzani, Jorge-Arnulfo Quiané-Ruiz, and Nan Tang. 2017. UGuide: User-Guided Discovery of FD-Detectable Errors. In *Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD Conference 2017, Chicago, IL, USA, May 14-19, 2017*, Semih Salihoglu, Wenchao Zhou, Rada Chirkova, Jun Yang, and Dan Suciu (Eds.). ACM, 1385–1397. <https://doi.org/10.1145/3035918.3064024>
- [41] Saravanan Thirumuruganathan, Han Li, Nan Tang, Mourad Ouzzani, Yash Govind, Derek Paulsen, Glenn Fung, and AnHai Doan. 2021. Deep learning for blocking in entity matching: a design space exploration. *Proceedings of the VLDB Endowment* 14, 11 (2021), 2459–2472.
- [42] Saravanan Thirumuruganathan, Mourad Ouzzani, and Nan Tang. 2019. Explaining Entity Resolution Predictions: Where are we and What needs to be done?. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics, HILDA@SIGMOD 2019, Amsterdam, The Netherlands, July 5, 2019*. ACM, 10:1–10:6. <https://doi.org/10.1145/3328519.3329130>
- [43] Jianhong Tu, Ju Fan, Nan Tang, Peng Wang, Chengliang Chai, Guoliang Li, Ruixue Fan, and Xiaoyong Du. 2022. Domain Adaptation for Deep Entity Resolution. In *SIGMOD '22: International Conference on Management of Data, Philadelphia, PA, USA, June 12 - 17, 2022*, Zachary G. Ives, Angela Bonifati, and Amr El Abbadi (Eds.). ACM, 443–457. <https://doi.org/10.1145/3514221.3517870>
- [44] Jianhong Tu, Ju Fan, Nan Tang, Peng Wang, Guoliang Li, Xiaoyong Du, Xiaofeng Jia, and Song Gao. 2023. Unicorn: A Unified Multi-tasking Model for Supporting Matching Tasks in Data Integration. *Proc. ACM Manag. Data* 1, 1 (2023), 84:1–84:26. <https://doi.org/10.1145/3588938>
- [45] Manasi Vartak and Samuel Madden. 2018. MODELDB: Opportunities and Challenges in Managing Machine Learning Models. *IEEE Data Eng. Bull.* (2018).
- [46] Jiannan Wang and Nan Tang. 2014. Towards dependable data repairing with fixing rules. In *International Conference on Management of Data, SIGMOD 2014, Snowbird, UT, USA, June 22-27, 2014*, Curtis E. Dyreson, Feifei Li, and M. Tamer Özsu (Eds.). ACM, 457–468. <https://doi.org/10.1145/2588555.2610494>