

PNEUMA: Leveraging LLMs for Tabular Data Representation and Retrieval in an End-to-End System

MUHAMMAD IMAM LUTHFI BALAKA, University of Indonesia, Indonesia

DAVID ALEXANDER, University of Indonesia, Indonesia

QIMING WANG, The University of Chicago, USA

YUE GONG, The University of Chicago, USA

ADILA KRISNADHI, University of Indonesia, Indonesia

RAUL CASTRO FERNANDEZ, The University of Chicago, USA

Finding relevant tables among databases, lakes, and repositories is the first step in extracting value from data. Such a task remains difficult because assessing whether a table is relevant to a problem does not always depend only on its content but also on the context, which is usually tribal knowledge known to the individual or team. While tools like data catalogs and academic data discovery systems target this problem, they rely on keyword search or more complex interfaces, limiting non-technical users' ability to find relevant data. The advent of large language models (LLMs) offers a unique opportunity for users to ask questions directly in natural language, making dataset discovery more intuitive, accessible, and efficient.

In this paper, we introduce PNEUMA, a retrieval-augmented generation (RAG) system designed to efficiently and effectively discover tabular data. PNEUMA leverages large language models (LLMs) for both table representation and table retrieval. For table representation, PNEUMA preserves schema and row-level information to ensure comprehensive data understanding. For table retrieval, PNEUMA augments LLMs with traditional information retrieval techniques, such as full-text and vector search, harnessing the strengths of both to improve retrieval performance. To evaluate PNEUMA, we generate comprehensive benchmarks that simulate table discovery workload on six real-world datasets including enterprise data, scientific databases, warehousing data, and open data. Our results demonstrate that PNEUMA outperforms widely used table search systems (such as full-text search and state-of-the-art RAG systems) in accuracy and resource efficiency.

CCS Concepts: • **Information systems** → **Document representation; Evaluation of retrieval results; Specialized information retrieval.**

Additional Key Words and Phrases: Data Discovery, Large Language Models, Natural-Language Questions

ACM Reference Format:

Muhammad Imam Luthfi Balaka, David Alexander, Qiming Wang, Yue Gong, Adila Krisnadhi, and Raul Castro Fernandez. 2025. PNEUMA: Leveraging LLMs for Tabular Data Representation and Retrieval in an End-to-End System. *Proc. ACM Manag. Data* 3, 3 (SIGMOD), Article 200 (June 2025), 28 pages. <https://doi.org/10.1145/3725337>

1 Introduction

Identifying *relevant* data is a prerequisite to solving data problems and creating value. Data may be relevant to a data problem because of its *content*, i.e., columns and rows in tabular data. For

Authors' Contact Information: Muhammad Imam Luthfi Balaka, muhammad.imam07@ui.ac.id, University of Indonesia, Depok, Indonesia; David Alexander, david.alexander01@ui.ac.id, University of Indonesia, Depok, Indonesia; Qiming Wang, qmwang@uchicago.edu, The University of Chicago, Chicago, USA; Yue Gong, yuegong@uchicago.edu, The University of Chicago, Chicago, USA; Adila Krisnadhi, adila@cs.ui.ac.id, University of Indonesia, Depok, Indonesia; Raul Castro Fernandez, raulcf@uchicago.edu, The University of Chicago, Chicago, USA.



This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

© 2025 Copyright held by the owner/author(s).

ACM 2836-6573/2025/6-ART200

<https://doi.org/10.1145/3725337>

instance, to calculate total sales for a product, a relevant table must include columns related to sales and products, with rows containing individual sales records. Data may also be relevant to a data problem because of its *context* [24], e.g., an enterprise data catalog that contains the origin, purpose and usage of the datasets. For example, to fill in missing values, it is crucial to understand the mechanism that explains the missing data [14]. Since context is often not explicitly included in table content, both content and context are commonly considered when searching for data [12].

Consider a user who asks for temperature data from a room that was sampled uniformly, e.g., because they need to know the sampling method for some downstream analysis. It is possible that a schema explicitly indicates the sampling procedure for the data. Still, it is also plausible that this information is instead incorporated into some form of documentation, e.g., a PDF, associated with the table. As an example, of 300k datasets in data.gov, $\approx 150k$ contain some form of associated context (in PDF, Text, or HTML). Our experience with organizations reveals that internal data is similarly documented in external repositories such as wikis, catalogs, and others [3, 43]. Ideally, the user articulates their need without thinking whether satisfying that need requires looking at the content of the table—a schema that models sampling method—or context, the PDF. Today’s solutions consider content and context when computing relevance. Data discovery systems largely overlook context [18, 27, 54, 55] while data catalogs index context. Furthermore, keyword search remains the primary interface in many dataset search platforms, with few offering natural language interfaces to address these types of data discovery tasks [55], yet natural language remains the most accessible method for non-technical users to identify relevant data.

In this paper, we present PNEUMA,¹ an open-source system that retrieves tables from table repositories based on their content, context, or both. After indexing a collection of tables, PNEUMA takes natural language questions as input and produces a ranking of relevant tables as output. A main challenge PNEUMA faces is that content and context are represented fundamentally differently. While the content of tables is highly structured, the context representation may range from free text to different degrees of semi-structuredness. To address this, PNEUMA leverages large language models (LLMs) within a retrieval-augmented generation (RAG) architecture [36], enabling both content and context to be represented as vectors. This paper contributes to the fields of data discovery and data catalogs by exploring two primary questions that arise from PNEUMA’s approach:

1. How do we represent content and context as vectors? There are numerous tabular representation techniques [27, 54, 55, 58, 60], with many focused on generating vectors to achieve high retrieval accuracy. While accuracy is crucial, we find that these methods often produce indices with storage footprints several orders of magnitude larger than the original data, severely impacting scalability and making them impractical for large data collections.

Instead, PNEUMA introduces a novel table representation method that leverages LLMs to *narrate* table schema. LLMs, equipped with extensive knowledge, can provide meaningful column descriptions, even for abbreviations or domain-specific terms that may be challenging for humans or smaller models to interpret. For example, one table in our benchmarks has the schema <Player | AB | AVG | BABIP>, which might be ambiguous to a user unfamiliar with baseball statistics, but an LLM recognizes these as Major League Baseball (MLB) statistics. AB refers to “the number of at-bats for a batter” and BABIP to “Batting Average on Balls In Play”. Schema narrations are further complemented by selected portions of the table’s content, helping resolve semantic ambiguities and providing a more comprehensive table representation. PNEUMA’s table representation method handles both content and context by transforming them into free text as an intermediate step. This text is then encoded into vectors using state-of-the-art embedding techniques [45].

¹code is available at <https://github.com/TheDataStation/pneuma>

2. How do we ensure high retrieval accuracy? To return relevant tables, the retrieval method must be compatible with the underlying vector representation. Retrieval techniques like full-text search [51] excel when string matching² suffices but suffer whenever there is semantic ambiguity, like when the input natural question uses a language different than the tables. Vector search handles semantic ambiguity better [32], but alone, it rarely offers high-quality answers if test data is beyond the domain of training data. RAG architectures treat vector search as a process to generate candidates which are then processed by a downstream LLM [22], i.e., the vector search results are incorporated into the LLM’s context window.

PNEUMA introduces a retrieval method that improves over the above. It combines full-text and vector search, *and* leverages an LLM as a mechanism to *refine* the suggested ranking, rather than relying on the LLM to directly rank or answer the question. In other words, PNEUMA’s retriever takes on more responsibility during candidate generation, reducing the workload for the downstream LLM. The LLM’s role is simplified to identifying and filtering out irrelevant entries from the final ranking. Since the LLM is tasked with a more focused, manageable job—but one that would still be difficult for traditional methods to handle—it performs much better than with more open-ended tasks. We find that this combination of retrieval and LLM evaluation significantly improves retrieval accuracy for both content and context questions.

In addition to these two primary technical contributions, this paper contributes two artifacts, a benchmark generator and an end-to-end system that implements the above techniques. Although data contexts must be supplied by external agents in practice, determining how to incentivize individuals to provide these contexts is beyond the scope of this paper. Instead, we simulate data contexts in our benchmark generator using an LLM. The benchmark generator takes table collections as input and produces content and context questions. It leverages LLMs to generate meaningful content questions and to create table contexts by answering the questions introduced in [24], which are designed to describe and document datasets. We use this benchmark generator to produce the benchmarks used in the evaluation. The end-to-end system, PNEUMA, indexes content and context, accepts natural language questions and produces ranked tables as outputs.

Evaluation Results. PNEUMA surpasses state-of-the-art information retrieval systems, including full-text search with BM25 [51], the RAG system implemented by LlamaIndex [39], and the advanced table retrieval system Solo [55], in retrieval quality and efficiency. PNEUMA achieves up to a **22.95% higher** relevant table hit rate compared to these baselines. PNEUMA serves user queries up to **31× faster** than the baselines while requiring **orders of magnitude less storage** than baselines that produce numerous embeddings.

Outline. We begin by introducing the preliminaries, including the definitions of context and content questions (Section 2). We then provide an overview of PNEUMA’s architecture (Section 3), followed by a detailed explanation of its two key components: table representation (Section 4) and table retrieval (Section 5). In Section 6, we describe the process of generating the table discovery benchmark, and in Section 7, we present the evaluation results. Lastly, we present related work and conclusions.

2 Content and Context Questions

Data discovery is the problem of identifying and retrieving documents that satisfy an information need [10]. In the paper, we focus on tabular datasets: a table D consists of a relational schema $\mathcal{R}(A_1, \dots, A_n)$ over n columns, and a set of tuples $\mathcal{T}$, which are instances of the schema $\mathcal{R}$. Solving data discovery is challenging when there is a large volume of tables. For example, in large organizations, data is often distributed across multiple independently maintained databases. This

²or simple approximate string matching

creates information silos and requires extensive tribal knowledge to locate and access the relevant data. Additionally, many organizations use data lakes to store vast quantities of raw data, further complicating data discovery due to the sheer volume and lack of comprehensive metadata and documentation. Furthermore, scientific repositories, which archive experimental results and observational data, also present significant discovery challenges. Researchers need efficient methods to locate datasets relevant to their specific scientific questions.

While substantial research has explored discovering combinations of tables based on an initial table provided by the user [61], i.e., based on *content*, there is significantly less work that identifies relevant tables based on their *content* or *context*.

Definitions. A table’s content consists of column names and row values, which can be directly provided to a system. However, its context can only be generated by external agents, such as the person who collected the data or previously used the table for specific tasks. While table content is inherently structured, table context can range from free text to semi-structured documents. In this work, we treat contexts as free-form text that provides additional information about the table, which cannot be directly derived from its content.

Table Content			Metadata.json
Year	Product Category	Revenue (USD)	<pre>{ "Table Description" : "This table shows the annual revenue for a technology company's two main product categories. In 2021, the release of a foldable smartphone boosted smartphone sales." }</pre>
2020	Smartphones	120 millions	
2020	Laptops	85 millions	
2021	Smartphones	135 millions	
2021	Laptops	110 millions	
...	...	...	

Fig. 1. Example table content and context

Figure 1 presents an example of table content and context. On the left is a table showing the revenues of electronic products for a technology company. On the right is the table’s context, the table description stored in a metadata file from an enterprise catalog.

A content question is answered by the table content, e.g., “Which dataset contains information about annual revenue for smartphones and laptops?”. The table content is crucial for finding the table to answer this question, as it asks for specific entities within the table.

A context question is answered using the table’s context rather than the table content. The table context encompasses all information that informs the understanding and use of the data [26], such as metadata like column descriptions, how the data was generated, and its intended purpose. A crucial aspect of table context is that it cannot be inferred directly from the dataset. For instance, questions like “For what purpose was the dataset created?” and “What sampling method was used to generate this dataset?”, which are important for practical data science [24], can only be answered by external sources, such as the data creator. Datasheets for datasets [24] introduces a structured approach to eliciting dataset context through a series of questions aimed at data creators. An example context question is “Which dataset reflects the market impact of cutting-edge flexible display technology in the mobile industry?”. This question requires considering the table’s context, as the context reveals that the company released a groundbreaking foldable smartphone in 2021.

Where do data contexts come from? This depends on the scenario. Sometimes, table contexts are readily available, such as in online repositories (ICPSR, Dataverse, or HuggingFace). HuggingFace’ datasets page contains “data cards”, which represent “context”. In enterprise scenarios, table contexts are sometimes maintained in data catalogs. Additionally, alternative research lines provide

approaches to generating such context, e.g., internal data markets to incentivize the creation of metadata/context [19]. PNEUMA is designed to leverage data context when it is available. In the absence of data context, PNEUMA can still perform table searches based solely on content.

Today’s Landscape and Problem Statement. Academic indexes, such as PubMed [50] and JSTOR [31], support table discovery via basic metadata such as titles, authors, or abstract descriptions. However, they often lack detailed data context, such as the dataset’s purpose or the methods used for data collection. Additionally, these systems typically rely on keyword-based searches over metadata, offering a limited semantic understanding of user queries and search capabilities over the dataset content itself. Industry catalogs such as Google Dataset Search [7] and Amundsen [3], face similar issues, relying heavily on complete and accurate metadata while neglecting the need to search through the data content and context. Internal resources, such as wikis or Slack channels, are often disorganized and reliant on tribal knowledge, further complicating access.

The Retrieval-Augmented Generation (RAG) architecture [36] is a promising solution to the table discovery problem because it permits *normalizing* the representation of content and context and searches over vectors. Furthermore, adding LLMs to the retrieval pipeline may improve ranking relevancy and, most importantly, enable querying tables using natural language.

PROBLEM 2.1 (TABLE DISCOVERY). *Given a natural language question Q , a collection of tables $\mathcal{D}$, a corpora of table context $X_{\mathcal{D}}$ and a table discovery system S , find the relevant table $D_Q \in \mathcal{D}$ that is relevant to answer Q , i.e., $D_Q = S(Q, \mathcal{D}, X_{\mathcal{D}})$.*

3 PNEUMA Architecture

In this section, we present the design of PNEUMA and explain how it addresses the table discovery problem using content and context. PNEUMA’s offline stage focuses on table representation and its online stage on table retrieval. Figure 2 illustrates the system architecture of PNEUMA.

During the offline phase, users register tables and associated contexts via the Data Register component, which offers APIs to represent the context and to ingest tables from different sources such as databases and CSV files. The volume of table collections can be large, thus PNEUMA employs the Content Summarizer to represent large tables into smaller documents, referred to as “content summaries”, while preserving schema and row-level information. These summaries, along with table contexts, are indexed by the Discovery Index Builder into both full-text and vector indices, enabling efficient table retrieval. At this stage, PNEUMA treats both data content and context uniformly as text documents.

During the online phase, PNEUMA retrieves tables based on the user query Q by integrating three signals: lexical (BM25), semantic (vector search), and a signal based on LLM judgment. The lexical and semantic retrievers complement each other—BM25 handles exact lexical matches, while vector search captures semantic similarity, even without exact matches. To harness both strengths, PNEUMA scores document relevance using both retrievers. These scored candidates are then passed to LLM Judge, which determines whether each candidate document is relevant to the query. Based on these judgments, PNEUMA reorders the documents, and returns the top-ranked tables associated with the retrieved documents.

Support Content and Context Questions. PNEUMA seamlessly supports table discovery queries based on table content, context or both via its two core mechanisms: *table representation* (offline) and *table retrieval* (online). PNEUMA is designed to leverage table context when it is available but table contexts need not be available for PNEUMA to work. Users can point PNEUMA to a folder containing tables and their associated contexts, which are then transformed into normalized representations and indexed. Once the indexing process is complete, users can query the system using natural language questions. The table retrieval mechanism ensures that relevant documents—whether

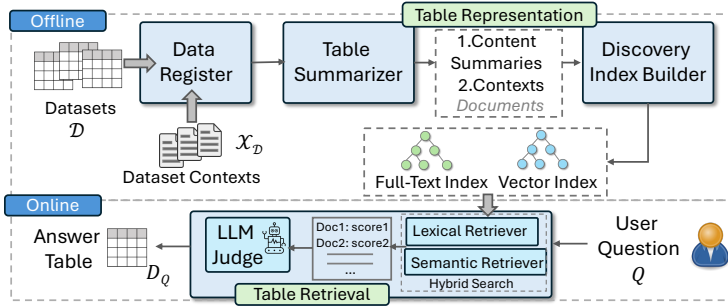


Fig. 2. System Overview of PNEUMA

content or context—are retrieved in response to user queries. For example, content-based questions like "What were the total sales for a specific product in the last quarter?" are answered by retrieving relevant table content summaries. Similarly, context-based questions, such as "Find datasets created using randomized controlled trials to evaluate the effectiveness of new treatments", are addressed by retrieving the appropriate context documents. This approach eliminates the need for manual search and allows users to discover relevant tables simply by asking questions in natural language.

The next two sections introduce table representation and table retrieval in more detail.

4 Table Representation

Effectively representing tables is essential for accurate and efficient table retrieval. Some approaches [27, 58] represent tables by embedding them into vector spaces, typically trained for specific tasks like table question-answering [30]. However, these approaches often lose row-level details, which are critical for precise data retrieval, especially for queries targeting specific facts or values. Additionally, these methods are resource-intensive to train and do not generalize well across new table corpora or tasks [55]. Other approaches [39, 55] split tables into smaller text snippets, typically at the row level, treating each snippet as a document. While this preserves row-level information, it leads to scalability issues, as storage requirements grow linearly with the number of rows, creating significant overhead. This approach also slows down the indexing process for large datasets containing millions of rows. For instance, LLAMAINDEX [39], a popular RAG library, produces a 44 GB index for the Chicago Open Data (829 MB) and a 26 GB index for a subset—500 rows max per table—of BIRD [37] (218 MB). These indexes are much larger than the original datasets and hence harder to justify their storage costs. On the other hand, PNEUMA produces 458 MB and 23.9 MB indexes for the Chicago Open Data and the BIRD dataset, respectively.

Table-level embedding approaches struggle with questions that require keyword matching at the row level, as the raw text information often gets lost in the embedding process. On the other hand, indexing every row in a table leads to high storage overhead and neglects the schema's structural context. Indexing only the table schema might save space, but it becomes ineffective when many tables share similar schemas, making it hard to differentiate between them. Moreover, real-world datasets often contain ambiguous or uninformative column names, so merely concatenating column names to represent the schema is not sufficient. Thus, a well-rounded table representation must integrate both schema and row-level information while retaining some raw textual information.

To address these challenges, PNEUMA creates table representations by generating both schema summaries and a smaller number of row summaries. Schema summaries describe the table's columns, preserving structural information crucial for understanding the relational context of the data. Row

summaries, on the other hand, combine row values, providing a detailed representation of the table's content. PNEUMA's Table Summarizer leverages an LLM to generate schema summaries. Additionally, PNEUMA optimizes the generation process by dynamically adjusting the inference batch size, filling each batch size with tasks of similar computational workload for balanced and efficient processing. We next introduce how PNEUMA generates schema and row summaries.

Generate schema summaries. Concatenating column names alone is insufficient for effective schema summaries, as column names in many tables are often cryptic or abbreviated. For example, the columns "AB" and "BABIP" are unclear. To address this, PNEUMA uses an LLM to generate schema summaries by providing a meaningful description of each column. Here, the LLM recognizes the "AB" column as the number of at-bats and the column "BABIP" as an abbreviation of Batting Average on Balls In Play, then narrates each with more descriptions. These narrations are then concatenated to form schema summaries, as illustrated in Figure 3. To ensure accuracy and coherence, PNEUMA provides the full schema as context to the LLM when prompting it to describe individual columns.

Generate row summaries. Row summaries provide more detailed information about the table content and differentiate tables with similar schemas. For instance, in our evaluation using the FetaQA [46] where many tables share the same schema, row information was critical for improving the hit rate. To address this, PNEUMA randomly samples r rows, as it avoids making any assumptions about the table content. For each sampled row, PNEUMA concatenates its values with their respective column names to generate r summaries, as illustrated in Figure 4. Including column names is essential, as it provides context for each value, ensuring clarity in the data's meaning. Without the column names, values like "76" could be ambiguous, potentially representing a variety of attributes, such as age, quantity, or price. By pairing values with their respective columns, PNEUMA maintains clarity and structure, allowing for more accurate and contextually aware table retrieval.

AB: This represents the number of at-bats for a batter. | BABIP: This represents Batting Average on Balls In Play, a statistic that measures a batter's success rate in converting balls put into play into hits. | ...

Fig. 3. An example schema summary

AB: 76 | BABIP: 0.317 | ...
AB: 11 | BABIP: 0.111 | ...
...

Fig. 4. An example row summary

Data ingestion via dynamic batch-size selection. PNEUMA employs an LLM to generate schema summaries for each column in a table collection, requiring one inference per column. For large collections, this leads to a large number of LLM calls, making LLM inference the primary computational bottleneck during the ingestion process. Thus, optimizing LLM inference time is crucial for the practical scalability of PNEUMA on large datasets. For instance, on the FeTaQA dataset, which contains over 10K tables, LLM inference without optimization requires approximately 15 hours. With PNEUMA's optimization techniques, this time is reduced to just 2 hours. We detail how PNEUMA achieves this improvement below.

Batch processing offers a way to optimize LLM inference by allowing multiple inputs to be processed in parallel during a single forward pass through the LLM [2]. By grouping several requests into a batch, the system leverages parallel computation on GPUs, reducing processing time and improving resource utilization. However, determining the optimal batch size for LLM inference is not straightforward. We cannot simply choose a batch size based on available GPU memory because the memory usage varies depending on the complexity and token length of the prompt. Larger batches improve throughput but demand more memory, potentially causing out-of-memory errors, while smaller batches underutilize resources and reduce efficiency. To address this challenge,

PNEUMA dynamically adjusts batch size during inference, automatically determining the optimal batch size for efficient processing.

PNEUMA begins by pooling prompts, where each prompt instructs the LLM to generate a schema description for a table column. It then employs a binary search algorithm to identify the optimal batch size within a predefined range. If a batch size exceeds memory limits and triggers an out-of-memory error, PNEUMA reduces the batch size; conversely, if a batch size runs without errors for several iterations, PNEUMA increases it. To further optimize memory usage, PNEUMA sorts the prompts by size and assigns them to each batch in a way that ensures each batch contains a similar number of tokens, balancing memory requirements across batches for maximum efficiency. For example, when processing the Adventure Works dataset, we get a batch size of 50 instead of 14 using this balancing approach. For reference, this reduces the summarization time by 78% (from 288 to 63 seconds).

5 Table Retrieval

We present the table retrieval method in Section 5.1 and indexes in Section 5.2.

5.1 PNEUMA's Hybrid Retrieval

In the online stage, PNEUMA retrieves a ranked list of tables for a user's query by integrating three signals: lexical (BM25 [51]), semantic (vector search over embeddings of context and content summaries), and the LLM judgment signal. These signals are combined to ensure robust and accurate retrieval across a wide range of user queries. First, lexical and semantic retrieval methods are used to generate candidate tables and their initial relevance scores. The LLM then evaluates these candidates to refine the relevance ranking. While traditional retrieval methods like BM25 and vector search are effective for identifying relevant documents, they may miss subtle nuances in the query or context. The LLM receives a ranked list of candidate tables provided by the two retrievers, and refines the results for better precision, especially in complex or ambiguous queries where context and meaning are critical.

The novelty of PNEUMA's approach lies not in any single signal, but in the integration of all signals. During table retrieval, PNEUMA grounds the LLM with factual information—relevance scores—from two reliable retrievers. Unlike many approaches that rely on the LLM for re-ranking, our method assigns the LLM a simpler, well-defined task: to judge whether a candidate document (content/context) is relevant to the question. In other words, we decompose an otherwise difficult task for the LLM, improving its performance.

Hybrid Search. PNEUMA adopts a hybrid search strategy that utilizes both full-text and vector retrievers to retrieve and rank relevant documents. PNEUMA retrieves $n \cdot k$ documents independently from both the full-text and vector retrievers. Here, k represents the number of documents the user wants to inspect, which is explicitly set by a user. Meanwhile, n is a multiplicative factor that determines the number of documents to retrieve initially for each retriever. n is transparent to the user. n ensures that the LLM Judge has a larger pool of candidates to evaluate.

Next, each unique document, from a set of up to $2nk$ documents, is assigned a combined relevance score. This score is calculated by weighting the relevance scores from both retrievers. If a document is retrieved by only one retriever, its relevance score for the other retriever is also computed. The final combined score for a document, $s(d)$, is evaluated as follows:

$$s(d) = \alpha \cdot s_{\text{lexical}}(d) + (1 - \alpha) \cdot s_{\text{semantic}}(d)$$

Where $s_{\text{lexical}}(d)$ is the normalized relevance score (using min-max scaling) from the full-text retriever, $s_{\text{semantic}}(d)$ is the normalized score from the vector retriever, α is a weighting factor (between 0 and 1) that balances the influence of both retrievers.

This scoring function allows PNEUMA to integrate the strengths of both retrieval methods, ensuring a more accurate ranking of documents based on both lexical and semantic relevance.

LLM Judge. The performance of table retrieval can be further improved by LLM Judge. Formally, its task is to classify pairs of documents and questions as either relevant or not. We reorder the documents based on the LLM’s output. As we iterate through the list, any document marked as irrelevant by the LLM is moved to the end, ensuring the most relevant documents are prioritized at the top. At the same time, the relative order between documents deemed relevant by the LLM is preserved based on their scores from the previous step.

The output of LLM-Judge is the final list of relevant documents to a user’s question. The tables associated with the retrieved documents are then returned to the user.

Illustration of PNEUMA’s Retrieval Mechanism We use a real example to demonstrate the benefits of PNEUMA’s retrieval mechanism. Consider this context question from ChEMBL (a chemical database) *“Develop a comprehensive database that catalogs and organizes the fundamental characteristics and attributes of tiny organic substances, encompassing their structural compositions, mass, and other relevant details.”*. The correct table context to address this question is *“The dataset was created to provide a comprehensive repository of chemical compounds, specifically small molecules, with their associated properties and information. The primary task in mind was to fill a gap in the availability of structured data on approved and investigational small molecule drugs, including their chemical structures, therapeutic uses, and regulatory information.”*

At $k = 1$, neither the full-text or vector-based retriever retrieved this document or its associated table. The full-text retriever failed due to limitations in lexical matching, while the vector-based retriever struggled because the pre-trained embedding model does not fully capture the nuanced semantics in specialized domains like chemistry. In contrast, PNEUMA’s hybrid search successfully retrieved this context using its vector retriever at a lower rank. It then elevated the context to the top rank by weighting the combined score. LLM Judge validated the relevance of this context to the question, keeping it in the first rank. Consequently, Hybrid Search + LLM Judge correctly retrieved the relevant context and associated table for the question.

In summary, PNEUMA achieves robust table retrieval by leveraging the complementary strengths of lexical and semantic retrieval, combined with the advanced semantic understanding of LLMs, to handle a wide range of user queries effectively.

5.2 Efficient Index Construction

To support the table retrieval in PNEUMA, we need to build efficient data discovery indices during the offline stage. PNEUMA needs to index two types of documents: content summaries and table contexts. Unlike content summaries, table contexts are already in text form and can be treated directly as documents without passing through the Table Summarizer. At this stage, PNEUMA treats both content and context uniformly, viewing them as documents for indexing.

To support table retrieval, PNEUMA builds two types of indices: full-text and vector indices. Each of these indices processes all documents, associating them with their respective table and document type—either as a content summary or a table context.

Block-based Embedding Generation. Building the vector index requires generating embeddings for documents in advance. The challenge is that in PNEUMA, each document is small, and there are a large number of documents to process, making the indexing process time-consuming and storage-intensive. To improve the indexing time and storage overhead, we group documents from the same table and of the same type into fixed-size blocks determined by the context window size of the encoding model. For instance, if there are three row summaries for Table A, rather than embedding each summary individually, we combine them into a single document and then generate

a single embedding for the block. This reduces the number of embeddings required, accelerates the indexing process, and lowers storage overhead, while still preserving all relevant information for retrieval. It also avoids scenarios where a given document is larger than an LLM's context window because the document size is bounded by the window size of the embedding model which is typically much smaller than the LLM's context window.

Incrementally Maintaining the Index. When there is an update to the schema, PNEUMA can generate new schema summaries for the updated data by calling Table Summarizer and then insert the new summaries into the index. For row updates, one could similarly call Table Summarizer to do the re-sampling, either eagerly, or only after a defined threshold is exceeded, e.g., when more than 20% of the rows are updated. PNEUMA does not have a built-in mechanism to detect those changes because it makes no assumptions of the storage system that keeps those tables. For instance, they may be stored in S3. But PNEUMA is engineered to expose an API for updating that can be invoked by external clients.

6 Table Discovery Benchmark

Existing data discovery benchmarks are limited. NQ_tables [27], FeTaQA [46], and QATCH [47], which are designed for table question answering, often reveal the correct answer table in the question itself. CMDBench [17] targets table discovery but suffers from high false-negative rates due to incomplete ground truth. Additionally, BIRD [37] and Spider [59], which were developed for text-to-SQL tasks, focus on table content and overlook table context, limiting their ability to evaluate systems that leverage content and context.

To address this gap, we introduce a benchmark generator that produces benchmarks specifically designed for table discovery, covering both content and context questions. By incorporating both types of queries, a benchmark simulates a diverse range of real-world scenarios, allowing for a more comprehensive evaluation of a system's capabilities. A benchmark consists of two components: the content benchmark and the context benchmark. Each benchmark maps a question to a list of tables, where each individual table within the list can answer the question. The content benchmark focuses on content questions, represented as $(Q_c, [T])$, while the context benchmark addresses context questions, represented as $(Q_x, [T])$. In the following sections, we detail the process of generating both content and context benchmarks.

6.1 Content Benchmark

For the content benchmark, our goal is to generate questions that can be directly answered by the content of a table. A straightforward approach is to prompt powerful LLMs, such as GPT-4 [1], to generate questions based on several sample rows from a table. However, this approach is insufficient for two reasons. First, LLMs are prone to hallucination, often generating overly simplistic or irrelevant questions that may not reflect the complexity of the data. Second, there is little control over the diversity of the questions generated—such as ensuring they cover different aspects of the table (e.g., addressing various columns)—which is critical for evaluating table discovery systems comprehensively.

Therefore, we use SQL queries as an intermediate proxy for generating natural language questions. This approach complements LLMs by leveraging the structured nature of SQL to ensure both relevance and diversity. SQL queries inherently target specific columns and operations, controlling the scope of questions and reducing the likelihood of hallucinations. Given a table, we first generate SQL queries based on the table's structure. These SQL queries are then converted into natural language questions. We introduce these two stages below.

6.1.1 SQL Generation. Recent work, SOLO [55], has demonstrated that using a SQL template is an effective approach for generating meaningful questions. While SOLO follows the simple template defined in WIKISQL [64], we extend it to include more complex SQL operators such as Group By, Having and Order By. In our template, which is shown in Figure 5, we exclude JOIN operations, focusing on questions that can be answered by a single table. Additionally, the FROM clause is omitted because identifying the table identifier is the goal of table discovery systems, and thus explicitly including the table name is forbidden in question generation. By omitting the FROM clause, we guarantee that the table name will not appear in the generated question.

We generate SQL queries by filling in the SQL template through random sampling. Specifically, we randomly select columns and corresponding values within the selected columns. Additionally, we assign probabilities for including aggregation functions, GROUP BY, HAVING, and ORDER BY clauses, and include them based on those probabilities. We ensure that the generated SQL queries maintain valid syntax even if certain clauses are not selected (with the exception of the FROM clause, which is omitted by design).

```
SELECT column_1 , MAX(column_2)
WHERE column_3 = value_1 AND column_4 < value_2
GROUP BY column_1 HAVING MAX(column_2) > value_3
ORDER BY column_1 LIMIT 3;
```

Fig. 5. Example SQL Query Template

6.1.2 Question Generation. Next, we use an LLM to convert SQL into a natural language question. The LLM receives the table schema, a few sampled rows, and the SQL query. However, the initial questions often suffer from quality issues, such as overusing keywords from cell values or inconsistencies due to LLM hallucinations. We address these issues and enhance question quality through the following two stages.

Question Rephrasing. To better simulate real-world user queries, we ensure LLM-generated questions do not directly replicate column names or long cell values, as users often cannot recall exact details in practice. To implement this, we identify questions that contain long cell texts matching the associated table data. We then instruct the LLM to paraphrase the questions by rewording them to avoid exact matches while maintaining original meaning.

Consistency Check. Even when the LLM is guided to generate questions based on SQL queries, it may still produce hallucinations due to its inherent limitations. This can result in questions that are inconsistent with the original SQL query. For example, the SQL may be designed to return two columns, but the generated question might only reference one of them. Additionally, the LLM can confuse column names that look similar, leading to inaccuracies in the question. To ensure consistency between the SQL query and the generated question, we apply the concept of *Cycle Consistency*, originally introduced in machine translation [34]. Specifically, we ask the LLM to translate the generated question back into SQL (referred to as the Back SQL) and compare it with the original SQL. We use the SQL exact match metric from [59] to evaluate SQL equivalence, which requires the two SQL queries to have the exact same components. If the Back SQL does not match the original SQL exactly, we consider the question inconsistent and discard it.

6.1.3 Annotating other answer tables. There may be more than one table capable of answering a given question. This introduces a challenge when designing a benchmark, as it could lead to false negatives. Specifically, a system might identify a valid table that answers the question, but it could

be unfairly penalized for not retrieving the specific table on which the question was originally generated. To address this, we extend our search to identify other tables that can also answer the generated question. Each question corresponds to a SQL query during content question generation. Thus, a table is identified as an alternative answer table for a question if it contains all the columns and values referenced in the corresponding SQL query of the question.

6.2 Context Benchmark

Context questions are those that can only be answered by considering the broader context of a table, rather than its content alone. Therefore, generating context questions requires first establishing the contexts associated with each table.

6.2.1 Context Generation. A table's context is typically provided by external agents, such as the data creator or individuals familiar with its use. However, manually creating contexts for our table collection, which includes 11,501 tables in total, would be prohibitively expensive. To overcome this issue, we leverage large language models (LLMs), which have shown significant potential in role-playing scenarios [52]. We instruct an LLM to assume the role of the table creator and automatically generate context for each table.

But how do we prompt the LLM to elicit relevant table context? Datasheets for Datasets [24] propose 51 questions aimed at data creators to gather comprehensive data contexts. These questions cover topics such as the reasons for creating the dataset, data collection methods, and usage considerations.

For each table, we prompt the LLM (acting as the data creator) to answer this set of questions, thereby generating detailed context automatically. In the prompt, we emphasize two key characteristics—completeness and relevance—to ensure the LLM produces high-quality responses. Completeness means that the answer definitively and comprehensively addresses all parts of the question, particularly important for long and compounded questions. Relevance focuses on providing only the information requested, suppressing the LLM's tendency to include extraneous details. These criteria are crucial for generating high-quality, focused answers that provide the necessary context for each table while avoiding unnecessary or off-topic information. For reference, the LLM we use for context generation, Meta-Llama-3-8B-Instruct, takes approximately 4 minutes to generate contexts for each table (batch size $k = 1$), which are essentially answers to the 51 context-elicitation questions.

6.2.2 Context Benchmark Generation. Next, we generate the context benchmark by prompting an LLM to create questions based on table contexts. We use stratified sampling to select 1,020 contexts, with 20 contexts for each of the 51 context-elicitation questions defined in [24]. For each selected context, we prompt the LLM to generate a question that asks for a table based on that context. The table associated with the context becomes the answer table for the generated question. However, many of the generated questions contain keywords directly taken from the context, which does not accurately reflect real-world user queries. To address this, we have the LLM rephrase the questions to make them more natural and realistic, which is similar to the process we follow for content.

6.2.3 Annotate other answer tables. Other tables may have relevant contexts that can also answer a question in the context benchmark. Similar to what we explained in Section 6.1.3, it is important to consider these tables to avoid unfairly penalizing a table discovery system during evaluation. To identify additional answer tables, we prompt an LLM to check which other contexts are capable of answering a given question.

7 Evaluation

In this section, we answer the following research questions:

- **RQ1 (End-to-end evaluation):** Does PNEUMA effectively identify the relevant tables for the given questions?
- **RQ2 (Efficiency and Scalability):** Is PNEUMA efficient and scalable in answering user questions during the online stage? Additionally, does it optimize storage footprint and data preparation time in the offline stage?
- **RQ3 (Ablation study):** Is each component within PNEUMA useful? Is the design of each PNEUMA component well-justified?
- **RQ4 (Microbenchmarks):** How do PNEUMA's hyperparameters affect its performance? We also investigate whether PNEUMA is sensitive to the choice of LLMs and evaluate its performance on a simpler benchmark where the questions have significant keyword overlap with the data.

Datasets. We evaluate PNEUMA using six real-world datasets (Table 1), spanning public, scientific, business, and enterprise data environments. This variety ensures comprehensive testing across different workloads and scenarios. Specifically, we apply our content and context benchmark generators to ChEMBL, Adventure Works, Public BI, Chicago Open Data, and FeTaQA, resulting in 5 content benchmarks and 5 context benchmarks. In addition, we leverage an external benchmark BIRD [37], which provides content questions. Table 1 shows the statistics of these datasets.

Table 1. Datasets Characteristics

Name	#Tables	Avg. #Rows	Avg. #Attributes	Size
ChEMBL	78	5,161	7	67 MB
Adventure Works	88	9,127	8	102 MB
Public BI	203	20	66	2.2 MB
Chicago Open Data	802	2,812	17	829 MB
FeTaQA	10,330	14	6	42 MB
BIRD	597	614,450	7	17 GB

- **ChEMBL** [23]: a scientific database containing bioactivity data on drug-like molecules, used in pharmaceutical research.
- **Adventure Works** [44]: simulates an enterprise database, with tables related to business operations, sales, and manufacturing.
- **Public BI** [5]: datasets used for business intelligence and data warehousing tasks. It contains data typically found in reports, dashboards, and business decision-making systems.
- **Chicago Open Data** [49]: contains publicly available data from the City of Chicago, representing various aspects of civic operations, including transportation, crime, and public services.
- **FeTaQA** [46]: a dataset designed for factual table question answering (QA) tasks, where users pose questions to retrieve factual information from tables.
- **BIRD** [37]: a cross-domain dataset designed for evaluating text-to-SQL tasks, covering more than 37 professional domains, such as healthcare and blockchain.

PNEUMA Implementation. We build PNEUMA using Python 3.12.2. The LLM inside Table Summarizer and LLM Judge is Qwen2.5-7B-Instruct [57].³ To generate row summaries, we set $r = 5$, meaning we sample five rows per table. The lexical retriever is implemented using BM25s [41], a

³<https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>

leading Python implementation of BM25. For the vector index, we employ a SOTA vector database, ChromaDB [13], which uses the HNSW index [42] to support efficient vector search.

System Setup. All experiments are conducted on a DGX A100 server, which has 1 TB RAM and dual AMD EPYC 7742 CPUs (128 cores in total). In terms of software, this server has Python 3.12.2, CUDA 12.4, and Nvidia driver 550.90.07. For RQ1 and RQ2, we use these hyperparameters for PNEUMA: $\alpha = 0.5$ and $n = 5$. We study the sensitivity of these hyperparameters in Section 7.4.

Baselines. We compare PNEUMA against the following baselines, which are state-of-the-art systems for table discovery.

- **FULL-TEXT SEARCH:** Treats tables as plain text documents and builds a full-text search index over them. It then applies the BM25 algorithm to retrieve tables that best match a user's query. We implement this using a state-of-the-art BM25 library [41].
- **LLAMAINDEX:** We implement a RAG system using LlamaIndex [39], a widely-used open-source library for retrieval-augmented generation. The system utilizes the DBREADER API provided by LlamaIndex, which converts each row in a table into a vector representation, allowing the LLM to retrieve relevant table information based on user queries.
- **SOLO:** Solo [55] is a state-of-the-art system for table retrieval that employs a self-supervised learning approach. Given natural-language questions, it returns the most relevant table as output.

7.1 RQ1: End-to-End Evaluation

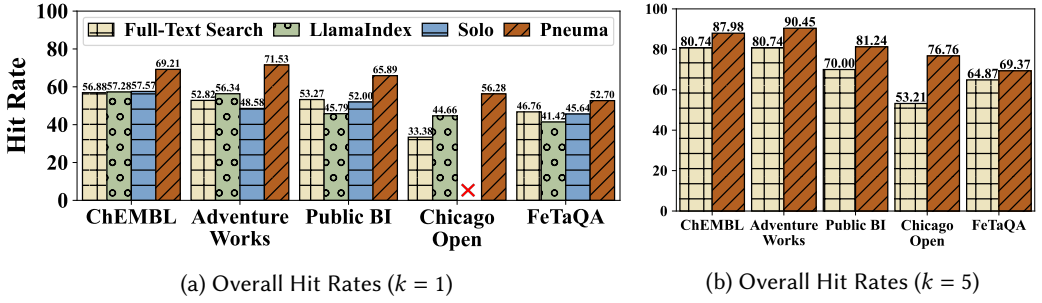
We use **hit rate** to evaluate PNEUMA's performance in identifying tables relevant to the given questions. In our benchmarks, while a question may have multiple relevant tables, any one of these tables is sufficient to answer the question fully. Therefore, the hit rate is defined as the proportion of questions for which *at least* one relevant table appears in the top- k retrieved documents. Next, we show the overall hit rates across context and content benchmarks, and we break down the hit rates over each benchmark.

7.1.1 PNEUMA's Overall Hit Rate. Figure 6 shows the overall hit rate, over both content and context questions (2k questions in total), of PNEUMA and the baselines at $k \in \{1, 5\}$. When $k = 1$, all systems must rank the relevant documents, associated with one of the correct answer tables, at the top. On the other hand, $k = 5$ represents a scenario where users want to inspect more documents.

At $k = 1$, PNEUMA outperforms all baselines in every scenario. For example, on the Adventure Works variant, PNEUMA obtains a hit rate of 71.53%, which is 18.71, 15.19, and 22.95 percentage points more than FULL-TEXT SEARCH, LLAMAINDEX, and SOLO, respectively. SOLO cannot be run on the Chicago dataset due to its excessive vector generation (730 million vectors, ~1 TB storage). Additionally, SOLO is excluded at $k = 5$ because its hyperparameter k refers to the number of tables, making it incomparable to other baselines. LLAMAINDEX is also excluded at $k = 5$ due to its overly slow performance at $k = 1$.

At $k = 5$, PNEUMA still outperforms FULL-TEXT SEARCH, but the difference is smaller compared to $k = 1$. This is because as k increases, all approaches can include more tables and achieve higher hit rates. Nevertheless, it is still considerable because even a 10% difference corresponds to approximately 202 questions.

7.1.2 PNEUMA's Hit Rate in Answering Content Questions. We compare PNEUMA with the baselines on content benchmarks only (1k questions) across different datasets. As shown in Figure 8(a), PNEUMA has higher hit rates than all baselines in every scenario. Other baselines have certain scenarios in which they perform worse compared to others. FULL-TEXT SEARCH performs worst on the Chicago dataset, LLAMAINDEX performs worst on the Public BI dataset, and SOLO performs worst on the Adventure Works dataset.

Fig. 6. Overall Hit Rates on Context and Content Benchmarks ($k \in \{1, 5\}$)

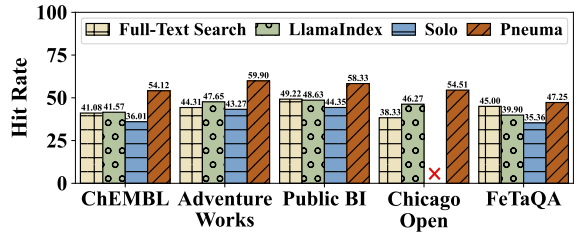
We also evaluate PNEUMA on BIRD where 1595 questions meet our criteria of being fully answerable by a single table. We first do not tamper with the original BIRD benchmark, using the same 1595 questions and their tables to form BIRD (original). To account for questions that could potentially be answered by alternative tables with similar information, we created BIRD (annotated) by annotating all possible answer tables using PNEUMA's benchmark generator, identifying 120 (7.5%) questions with two or more valid tables.

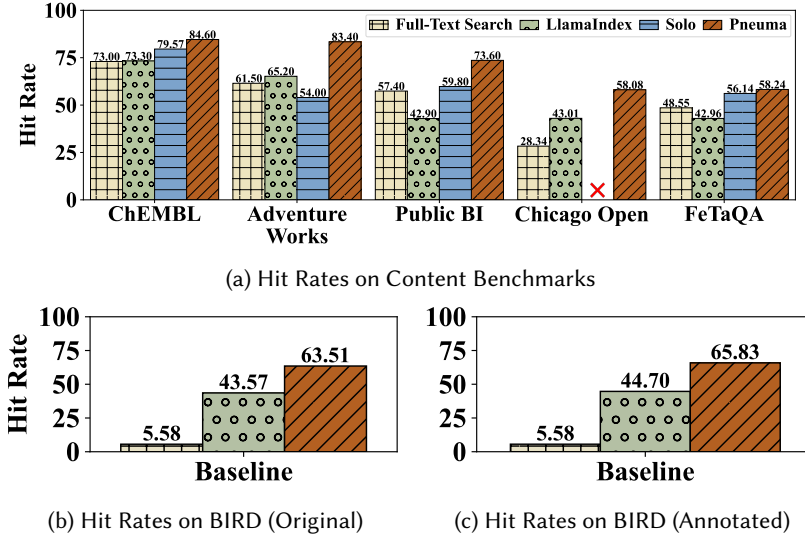
We exclude SOLO because it generates an excessive number of vectors (8.88 billion vectors, ~12TB storage). We limit LLAMAINDEX to only consider the first 5000 rows from each table in BIRD due to computational limitations: our instance only has access to approximately 90 GB of RAM, which is insufficient for LLAMAINDEX's processing of the whole dataset because it needs to load all documents to the memory at once before converting them to embeddings. Nevertheless, LLAMAINDEX indexed 1.3 million vectors and took 34 hours to answer 1595 questions. In contrast, PNEUMA did the same task in less than 13 minutes.

As shown in Figure 8(b), PNEUMA outperforms other baselines significantly on BIRD (original). PNEUMA achieves a hit rate of 63.51%, which is 57.93 and 19.94 percentage points higher than FULL-TEXT SEARCH and LLAMAINDEX, respectively. On BIRD (annotated), PNEUMA achieves an even higher hit rate of 65.83%, exceeding FULL-TEXT SEARCH and LLAMAINDEX by 60.25 and 21.13 percentage points, respectively. FULL-TEXT SEARCH performs badly mainly because lexical matching on row information is insufficient to answer the questions. LLAMAINDEX also utilizes row information, but the embedding model encodes semantic information, helping RAG better answer the questions.

7.1.3 PNEUMA's Hit Rate in Answering Context Questions.

In this section, we evaluate PNEUMA and the baselines on context benchmarks only across different datasets (1k questions). As shown in Figure 7, PNEUMA performs better than the three other baselines, highlighting PNEUMA's ability to handle a variety of context questions. For reference, SOLO cannot index dataset context, but it is specifically trained on each dataset. Thus, it can point to the relevant tables for context questions such as the type of instances available in the tables and what each instance is about.

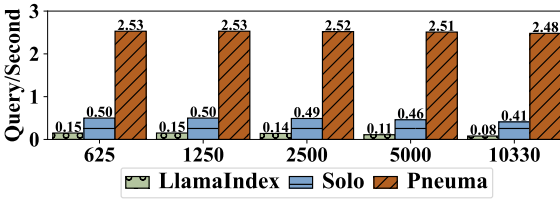
Fig. 7. Hit Rates on Context Benchmarks ($k = 1$)

Fig. 8. Hit Rates on Content Benchmarks and BIRD ($k = 1$)

7.1.4 Summary of Results. Overall, we see that PNEUMA is capable of handling a variety of context and content questions, outperforming all baselines across all datasets. On BIRD, the performance gap between PNEUMA and other baselines is even more pronounced.

7.2 RQ2: Efficiency of PNEUMA

We compare the efficiency of PNEUMA with LLAMAINDEX and SOLO. FULL-TEXT SEARCH is excluded from this comparison because its performance suffers when there is insufficient lexical overlap between the question and the tables, and it lacks semantic understanding, which is often the performance bottleneck. The efficiency is evaluated from two perspectives: **1. Online stage:** How quickly can PNEUMA serve queries? This is measured in query throughput. **2. Offline stage:** How much time does PNEUMA take to prepare data and how much is its storage footprint? We use the largest dataset, FeTaQa, to answer this research question. We evaluated performance using varying table counts: 625, 1250, 2500, 5000, and 10330 (the full FeTaQa dataset).

Fig. 9. Query Throughput of PNEUMA, LLAMAINDEX, and SOLO ($k = 1$)

16.6× faster than LLAMAINDEX and **5×** faster than SOLO. At 10,330 datasets, PNEUMA outperforms

7.2.1 Querying Throughput (Online). We use **query throughput** to assess online efficiency. We compiled a list of 100 questions from our benchmarks and recorded the total time required to answer them. To ensure we did not capture noise in the results and that the results were consistent, we ran the performance benchmark for each setting 10 times and averaged the results. As shown in Figure 9, PNEUMA significantly outperforms both LLAMAINDEX and SOLO. With 625 datasets, PNEUMA is

LLAMAINDEX by **29.6×** and Solo by **6×**. PNEUMA is significantly more efficient and requires far less storage compared to LLAMAINDEX. This is because PNEUMA generates concise table representations, indexing fewer documents and resulting in a smaller search space, whereas LLAMAINDEX must search through a larger number of vectors, increasing both computational and storage overhead.

7.2.2 Data Preparation Time vs. Storage Footprint (Offline). The input dataset must be prepared before the system can answer questions. The preparation process for PNEUMA and LLAMAINDEX includes ingesting the data, representing tables, and creating the index. For Solo, this process includes indexing the data and training a model. We call this offline time. Figure 10 shows LLAMAINDEX has the fastest runtime but the largest storage footprint. PNEUMA balances these factors, with linear growth in both.

The LLAMAINDEX system has very low offline runtime because the only data preparation needed is ingesting and directly generating embeddings. PNEUMA, on the other hand, in addition to ingesting the data and generating embeddings, has an additional step of generating schema narrations. PNEUMA takes considerably more preparation time because generating schema narrations requires prompting an LLM. However, this means PNEUMA does not keep all row data from the dataset; thus, the storage footprint is smaller than LLAMAINDEX. With 625 tables, PNEUMA requires **14.6×** longer but requires only **0.68×** the storage space. At 10330 tables, PNEUMA requires **18.6×** longer but only **0.6×** the storage space. Compared to SOLO, however, PNEUMA is always faster.

The FeTaQA dataset predominantly consists of small tables, averaging 14 rows each. PNEUMA creates a summary of 5 rows and an LLM-generated schema, while LLAMAINDEX indexes entire tables. Consequently, PNEUMA's summaries only halve the storage size for FeTaQA. However, on the Adventure Works dataset (88 tables, averaging 9,127 rows), PNEUMA outperforms LLAMAINDEX significantly in both storage footprint and offline time. With this dataset, LLAMAINDEX requires 3357.2s offline time and 14.5GB of storage space, while PNEUMA only requires 97.7s offline and 30.3MB. LLAMAINDEX takes **34.3×** longer and **477.7×** the storage footprint.

For reference, SOLO's data is irregular for 2 reasons. First, the training time, which consists of constructing training data and model training, varies. These steps depend on the number of questions in the training data instead of the number of tables in a dataset. Second, SOLO's index size can be different depending on the number of vectors it generates. When it is < 1 million, an exact exhaustive search is used and vectors are stored without compression. Otherwise, an approximate

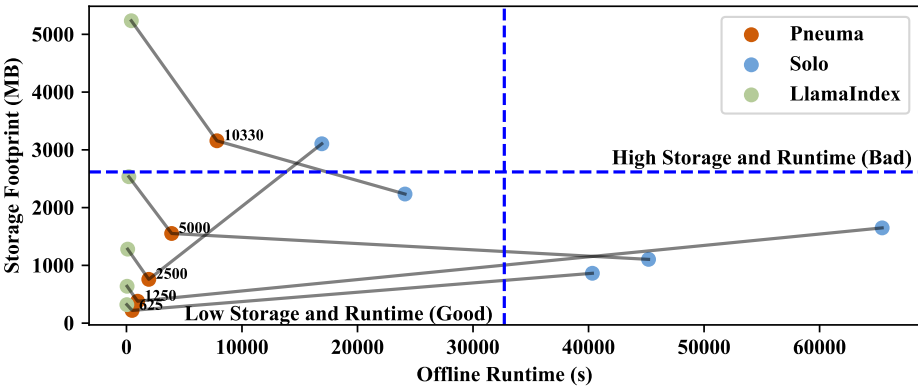


Fig. 10. Comparison of Offline Runtime vs. Storage Footprint of PNEUMA, LLAMAINDEX, and SOLO

search is used and vectors are compressed with product quantization. This causes SOLO's storage footprints for 5,000 and 10,330 tables to be smaller than in other scenarios.

7.2.3 Schema Summary Generation Throughput. In the event of changes in table schema or the addition of new tables, schema summaries need to be (re)generated. This process involves prompting the LLM to describe each table column that has changed. We report schema summary generation throughput in columns per second. A single process of PNEUMA's summarizer can process **8 columns/second**. Whether processing 8 columns/second is sufficient depends entirely on the rate of schema changes in the dataset. If the rate of changes exceeds this throughput, it is easy to increase throughput by adding more processors. Further improvements to a processor's performance will also bring benefits.

7.3 RQ3: Ablation Study

We evaluate the impact of PNEUMA's components on its hit rates. Specifically, we investigate the contributions of Table Summarizer, Hybrid Retrieval, and LLM Judge by considering alternative variations for each of them.

7.3.1 Impact of Table Summarizer. To justify the design of Table Summarizer, we index variants of content summaries for Hybrid Retrieval ($k = 1, n = 5, \alpha = 0.5$) and run it on the content benchmarks. The variants include (1) **SAMPLEROWS**: A random sample of $r = 5$ rows from each table. For each row, we concatenate its values along with the corresponding column names. (2) **SCHEMACONCAT**: A concatenation of the column names. (3) **SCHEMANARRATIONS**: Similar to SCHEMACONCAT, but each column name is appended with LLM-narrated descriptions. (4) **DBREADER**: Similar to SAMPLEROWS, but includes all rows.

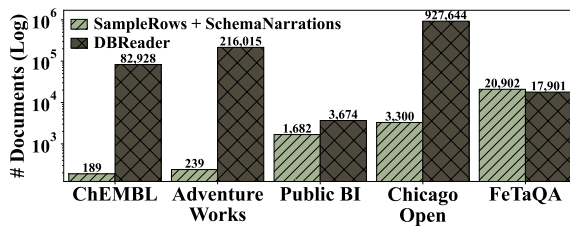


Fig. 12. Comparison of Total Number of Documents (Log Scale) Between Table Summarizer and DBReader

because the FeTaQA dataset has many tables with similar schemas. For example, there are 420 tables

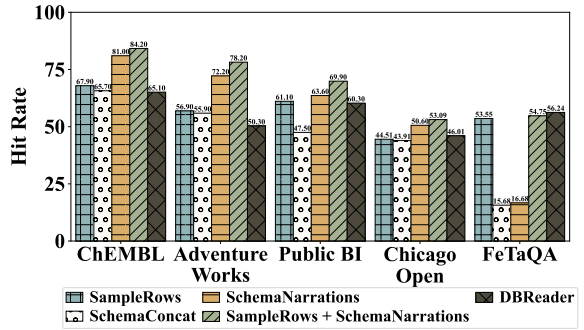


Fig. 11. Hit Rates of Hybrid Retrieval on Content Benchmarks as its Content Summaries Vary

SCHEMANARRATIONS is better than SCHEMACONCAT. As shown in Figure 11, SCHEMANARRATIONS leads to better hit rates than SCHEMACONCAT in every scenario. This demonstrates the effectiveness of narrating the schemas of the tables using an LLM. However, both of them obtain low hit rates on the FeTaQA dataset.

SAMPLEROWS and DBREADER improve hit rates on FeTaQA. Using row information leads to 3× better hit rate on the FeTaQA dataset compared to SCHEMANARRATIONS and SCHEMACONCAT. This is because the FeTaQA dataset has many tables with similar schemas. For example, there are 420 tables

about elections with exactly the same schema. Thus, Hybrid Retrieval needs row-level information to differentiate tables with similar schemas.

Between SAMPLEROWS and DBREADER, the differences in hit rates are minimal, but the latter corresponds to many more documents. For example, DBREADER has 927,644 documents in the Chicago variant, which is 459× more documents than SAMPLEROWS. However, it only achieves 3.37% better hit rate compared to SAMPLEROWS.

Table Summarizer combines schema and row information. Although SAMPLEROWS and DBREADER achieve much better hit rates on the FeTaQA dataset, SCHEMANARRATIONS is better on the other datasets. For instance, in the ChEMBL variant, Hybrid Retrieval obtains a hit rate of 81%, which is 13.1 and 15.9 percentage points higher than SAMPLEROWS and DBREADER, respectively.

Therefore, Table Summarizer combines both schema narrations and sample rows as content summaries. In most scenarios, this approach leads to better hit rates and lower number of summaries—the latter is shown in Figure 12. In the Chicago dataset, for instance, DBREADER has 281× more documents than Table Summarizer, but Table Summarizer’s hit rate is 5.29 percentage points higher.

7.3.2 Impact of Hybrid Retrieval. In this section, we want to show that Hybrid Retrieval in PNEUMA’s retriever ($k = 1, n = 5, \alpha = 0.5$), is better than vector-search-only (PNEUMA-VS) or full-text-search-only (PNEUMA-FTS). All retrievers index the same documents: content summaries and table contexts. As shown in Table 2, Hybrid Retrieval obtains the highest hit rates in 70% of the scenarios and competitive otherwise. For example, in the Adventure Works-based content benchmark, Hybrid Retrieval obtains a hit rate of 81.40%, which is 7 and 13.7 percentage points higher than PNEUMA-FTS and PNEUMA-VS, respectively.

PNEUMA-FTS outperforms PNEUMA-VS on content questions, while the opposite is true for context questions. Context questions are closely tied to their original contexts, so rephrasing has a greater impact on hit rates for these questions. Hybrid Retrieval leverages the fact that the two retrievers answer different sets of questions incorrectly, but with some overlap. It combines both retrievers and harnesses their respective strengths, resulting in a compounding improvement in performance.

Table 2. Hit Rate Comparison Between PNEUMA-FTS, PNEUMA-VS, and Hybrid Retrieval ($k = 1, n = 5, \alpha = 0.5$)

Dataset	Benchmark Type	Retriever		
		PNEUMA-FTS	PNEUMA-VS	Hybrid Retrieval
ChEMBL	Content	75.40	73.00	83.00
	Context	44.22	53.24	50.88
Adventure Works	Content	74.40	67.70	81.40
	Context	48.24	57.55	56.08
Public BI	Content	72.10	42.90	71.10
	Context	49.80	57.65	57.75
Chicago Open	Content	50.10	40.92	55.09
	Context	41.18	57.16	52.25
FeTaQA	Content	54.05	25.17	56.14
	Context	40.10	44.90	46.96

7.3.3 Impact of LLM Judge. To gain insight into LLM Judge’s impact on Hybrid Retrieval, we compare it with two pre-trained re-ranker models: BAAI/bge-reranker-v2-m3 (BGE) [11] and dunzhang/stella_en_1.5B_v5 (Stella) [33]. The former is a widely-used re-ranker model, while the latter is the best embedding model for re-ranking purposes for its size (as of 2024-10-17) according

to the MTEB Leaderboard [45]. (It is overall ranked second, slightly below a model 7 times its size.) At the same time, we try a different model for LLM Judge: HuggingFaceH4/zephyr-7b-beta (Zephyr) [53].

We compare the hit rates of Hybrid Retrieval ($k = 1, n = 5, \alpha = 0.5$), with and without judges, in Figure 13. We observe that LLM Judge, both Zephyr and Qwen variants, perform better than the re-ranker models. For example, in the FeTaQA-based content benchmark, LLM Judge (Qwen) obtains a hit rate of 58.24%, which is 22.08 and 20.18 percentage points higher than BGE and Stella, respectively. In almost all scenarios, both Stella and BGE reduce hit rates. This is undesirable because users have traded query time for better hit rates, not worse. Between Zephyr and Qwen, the differences are subtle, with Qwen obtaining slightly better hit rates in more scenarios. Overall, LLM Judge’s performance is not sensitive to the selection of LLMs of similar size.

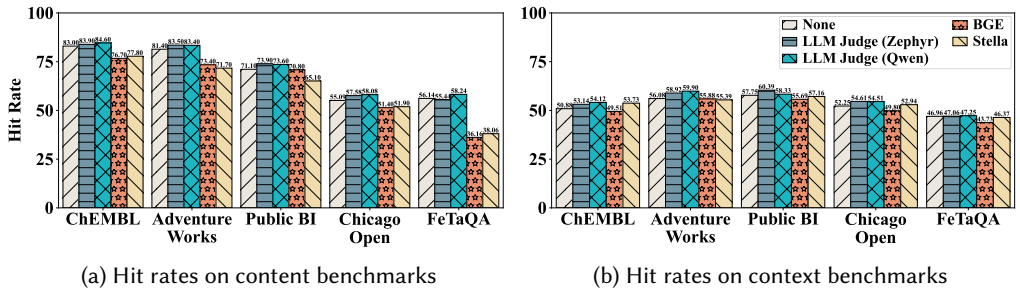


Fig. 13. Hit Rates of Hybrid Retrieval ($k = 1, n = 5, \alpha = 0.5$) with Different Judges

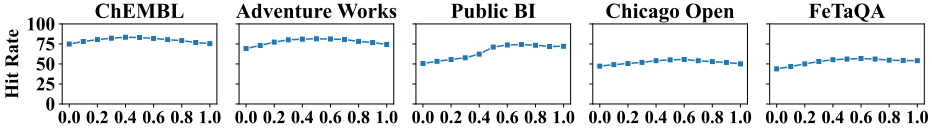
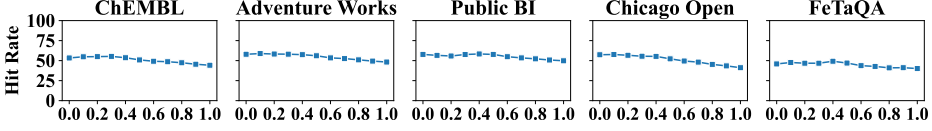
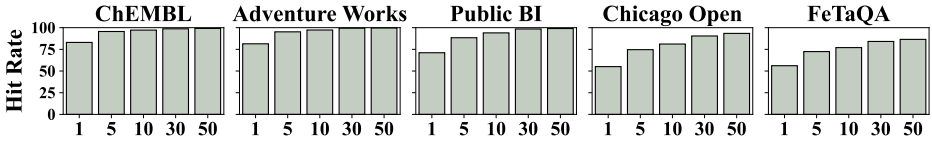
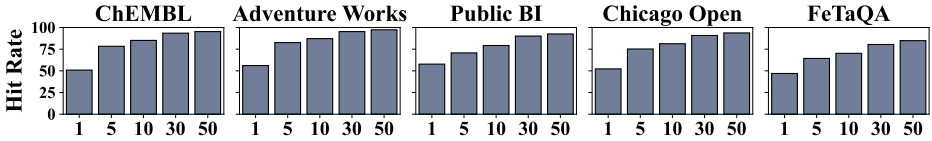
7.4 RQ4: Microbenchmarks

We evaluate PNEUMA’s retrieval performance by varying its hyperparameters (α , k , and n) and testing different LLMs as schema summarizers for robustness. We also explore the baselines’ behavior when questions share many keywords with dataset contexts and row values, and investigate the impact of hallucinations.

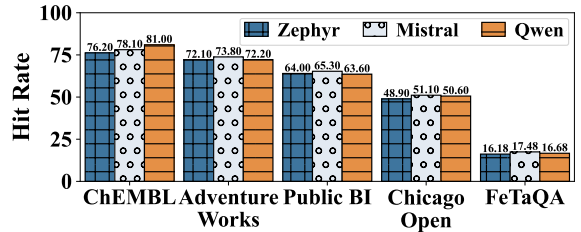
7.4.1 Changing α . This hyperparameter linearly combines the normalized relevance scores given by vector and full-text retrievers. Higher α gives more weight to the latter. To understand the impact, we experiment with $\alpha \in \{0.0, 0.1, \dots, 1.0\}$, $k = 1, n = 5$. As shown in Figure 14, the hit rates on both benchmarks peak at around 0.4 – 0.6. This suggests the effectiveness of weighting both retrievers at approximately the same weight.

7.4.2 Changing k . k determines the number of documents retrieved. Higher k values increase the chance of retrieving relevant tables, potentially improving hit rates. We tested $k \in \{1, 5, 10, 30, 50\}$ with $n = 5, \alpha = 0.5$. Figure 15 confirms this: hit rates improve for both benchmarks as k increases, with the largest gain from $k = 1$ to $k = 5$, followed by smaller but continued improvements.

7.4.3 Changing n . The last hyperparameter is the multiplicative factor of k for retrieving documents from both vector and full-text indices. By retrieving nk instead of k documents, we gather more relevant ones. We study the effect of changing the value of n on hit rates. We do so by experimenting with these hyperparameters: $k = 1, n \in \{1, 5, 10, 15, 20\}, \alpha = 0.5$. As shown in Figure 16, the differences in hit rates occur mostly from $n = 1$ to $n = 5$, and are minuscule otherwise.

(a) Hit rates on content benchmarks with a variety of α values(b) Hit rates on context benchmarks with a variety of α valuesFig. 14. Hit Rates of Hybrid Retrieval ($k = 1, n = 5$) on Content and Context Benchmarks as α Varies(a) Hit rates on content benchmarks with a variety of k values(b) Hit rates on context benchmarks with a variety of k valuesFig. 15. Hit Rates of Hybrid Retrieval ($n = 5, \alpha = 0.5$) on Content and Context Benchmarks as k Varies

7.4.4 Changing LLMs for Schema Summarizer. We tested Mistral-7B-Instruct-v0.3 and Zephyr-7b-beta as alternatives to Qwen for schema summarization, using Hybrid Retrieval ($k = 1, n = 5, \alpha = 0.5$). Figure 17 shows minimal differences in hit rates, with Mistral slightly outperforming others. We retain Qwen for its conciseness, enabling faster summarization. These results demonstrate that our system's performance is not significantly model-dependent.

Fig. 17. Hybrid Retrieval Hit Rates ($k = 1, n = 5, \alpha = 0.5$) on Content Benchmarks as Schema Summaries Vary

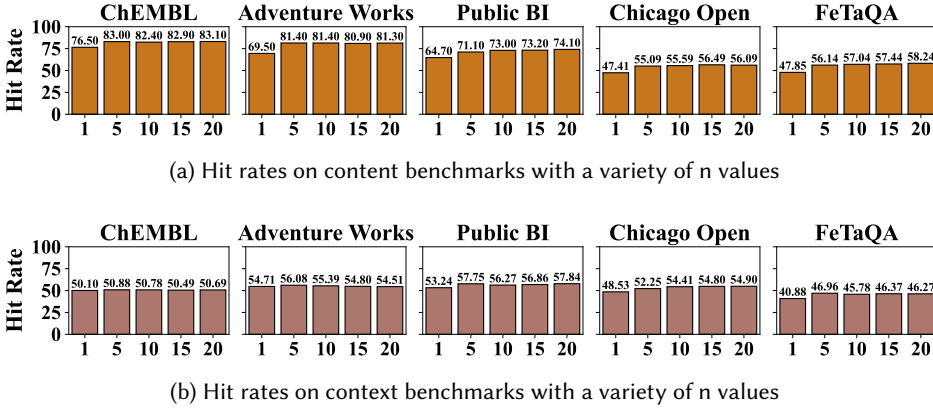


Fig. 16. Hit Rates of Hybrid Retrieval ($k = 1, \alpha = 0.5$) on Content and Context Benchmarks as n Varies

7.4.5 Investigating the impact of hallucinations. We have made efforts to minimize hallucinations by prompting the LLM to avoid generating descriptions when uncertain and using greedy decoding. Despite our efforts, certain LLMs still hallucinate. While we expect this to improve as LLM technology advances, we conducted experiments to assess how hallucination-prone summaries affect hit rates. Specifically, we compared the hit rates of Hybrid Retrieval ($k = 1, n = 5, \alpha = 0.5$) when indexing schema summaries generated with three different LLM configurations: (1) **TEMP-0** (default: greedy search), (2) **TEMP-1.5** (sampling enabled with temperature set to 1.5), and (3) **NON-INSTRUCT-TEMP-1.5** (same as (2) but using the non-instruct variant of Qwen).

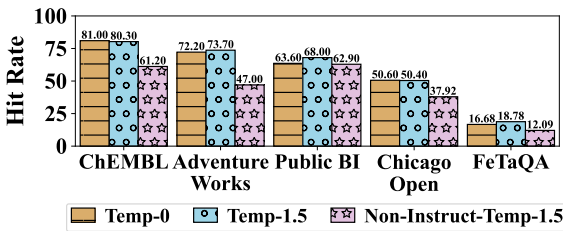


Fig. 18. Hit Rates of Hybrid Retrieval ($k = 1, n = 5, \alpha = 0.5$) on Content Benchmarks for different LLMs

racies are observed across other summaries generated by NON-INSTRUCT-TEMP-1.5, leading to significantly lower hit rates.

To assess whether the LLM summarizer is prone to hallucinations, PNEUMA's benchmark generator can be used to create a table discovery benchmark for a sample of tables. The benchmark serves as a test dataset to evaluate the quality of the generated summaries. A very low hit rate on this test set indicates that the LLM is prone to hallucinations and fails to produce useful summaries.

7.4.6 Using Keyword-Heavy Benchmarks. We evaluated PNEUMA ($k = 1, n = 5, \alpha = 0.5$) and other baselines on keyword-heavy variants of context and content benchmarks, where questions share many keywords with contexts and row values. Figure 19 shows PNEUMA outperforming all baselines

Figure 18 shows that TEMP-0 performs on par with TEMP-1.5 because increasing the temperature results in summaries with equivalent meanings but a broader vocabulary. In contrast, NON-INSTRUCT-TEMP-1.5 produces hallucinated summaries. For example, the 'molregno' column in the ChEMBL dataset represents an internal ID for compounds. Both TEMP-0 and TEMP-1.5 correctly identify it as IDs for molecule or compound, but NON-INSTRUCT-TEMP-1.5 incorrectly describes it as representing alphanumeric molecules. Similar inaccuracies are observed across other summaries generated by NON-INSTRUCT-TEMP-1.5, leading to significantly lower hit rates.

on content benchmarks and competing closely with FULL-TEXT SEARCH on context benchmarks. Due to keyword overlap, FULL-TEXT SEARCH generally outperforms LLAMAINDEX, contrary to more realistic scenarios (Section 7.1). These results highlight PNEUMA's effectiveness in handling keyword-heavy questions, leveraging its consideration of lexical information.

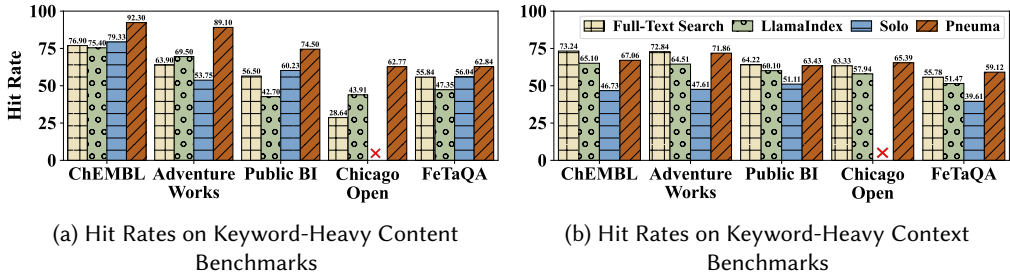


Fig. 19. Hit Rates on Keyword-Heavy Content and Context Benchmarks ($k = 1$)

8 Related Work

Information Retrieval. Information retrieval methods such as BM25 [51] match queries with documents based on term overlap, while vector search captures semantic similarity by embedding queries and documents in high-dimensional spaces [32]. Hybrid approaches combine both methods to balance precision and semantic understanding [8]. PNEUMA extends these techniques with an LLM-judge which aids in reranking results. Unlike other LLM reranking methods [21], which rely on LLMs to directly output a ranking, PNEUMA assigns the LLM a simpler task: binary classification of the relevance of documents to questions.

Tabular Data Discovery. Data discovery investigates identifying and retrieving relevant data—either an individual dataset [9, 10, 55] or a combination of datasets [16, 20, 25]—to satisfy the user's information need. PNEUMA focuses on retrieving relevant datasets individually, without performing any combination or integration across them. Keyword search [9, 60] is a traditional interface to discover tabular data. However, keywords are not specific enough to express user intent, and it is difficult for users to choose the right keywords. Aurum [18] proposed programming APIs for more advanced data discovery features but requires users to have technical backgrounds and hence less accessible. Recently natural-language question (NLQ) [27, 54, 55] has become a more popular interface for tabular data discovery, thanks to the quick development of LLMs. Systems such as OpenDTR [27] and GTR [54] require expensive human-annotated data for training on a new table corpus to perform well. Although Solo [55] proposed a self-supervised approach, it suffers from long training time as shown in Section 7.2.2. PNEUMA uses NLQ as an interface and does not need training by exploiting the combination of pre-trained embedding model, LLM, and hybrid search.

Table Representation for Data Discovery. There are generally two approaches to table representation for data discovery. The first approach learns representations on tabular data [27, 40, 54, 58, 63]. However, this approach may not generalize well to different datasets [55]. The second approach converts tables into multiple text snippets and uses off-the-shelf models [15] to represent them. Among text-based approaches, Liu [39] converts each row to a snippet and then encodes each row snippet into a vector. Wang and Castro Fernandez [55] treats each pair of columns in a row as a text snippet. Both Liu [39] and Wang and Castro Fernandez [55] generate many vectors, impacting storage and indexing scalability. PNEUMA produces fewer vectors without trading search quality,

and even performs better by representing tables with LLM-generated schema descriptions and sample rows.

Benchmarks for Tabular Data Discovery. QATCH [47], BIRD [37], and Spider [35], are designed for table question answering and text-to-SQL tasks. QATCH reveals the correct table in the question, making it unsuitable for table discovery. BIRD and Spider can be adapted for table discovery tasks but involve fewer tables, offering a less challenging setting compared to PNEUMA's benchmark. CMDBench [17] aligns better with PNEUMA's goals but lacks comprehensive ground truth, which is critical for accurately interpreting hit rates. PNEUMA offers a more comprehensive and challenging benchmark generator for table discovery tasks.

Table Question Answering. Table question answering (QA) is a prominent downstream application of tabular data. Prior work has explored both specialized and general-purpose approaches. TaPaS [28] and OmniTab [29], for instance, train/fine-tune specialized models for table QA, while others use large language models with tools (e.g., SQL engines), as in Chain-of-Table [56] and ReAcTable [62]. LOTUS [48] introduces semantic operators for transforming tables using natural-language criteria, which has been demonstrated to be useful for table QA [6]. Palimpzest [38] tackles analytical queries over input tables or tabulated unstructured data. This generalization is also adopted by Aryn [4], which handles tables, among other types of modality, in documents. Overall, these systems could serve as downstream users of the tables retrieved by PNEUMA.

9 Conclusions

We introduce PNEUMA, an end-to-end system that helps users discover desired tables in large data repositories using natural language queries based on both data content and context. PNEUMA enables efficient table discovery through two key components: table representation and table retrieval. Table representation generates schema and row summaries by leveraging an LLM to narrate schemas, while table retrieval combines the LLM Judge with two traditional information retrieval methods, effectively utilizing the strengths of both approaches. PNEUMA outperforms state-of-the-art table retrieval systems in both response quality and efficiency, while requiring orders of magnitude less storage.

10 Acknowledgements

We thank the reviewers for their constructive feedback. We thank the Indonesia-US Research Collaboration Open Digital Technology program, the Data Science Institute's Research Initiative on Data Ecology, and the National Science Foundation (CAREER Award 2340034) for supporting this project.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
- [2] Amey Agrawal, Nitin Kedia, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav S. Gulavani, Alexey Tumanov, and Ramachandran Ramjee. 2024. Taming Throughput-Latency Tradeoff in LLM Inference with Sarathi-Serve. *arXiv:2403.02310 [cs.LG]* <https://arxiv.org/abs/2403.02310>
- [3] amundsen. [n. d.]. <https://www.amundsen.io/>.
- [4] Eric Anderson, Jonathan Fritz, Austin Lee, Bohou Li, Mark Lindblad, Henry Lindeman, Alex Meyer, Parth Parmar, Tanvi Ranade, Mehul A. Shah, Benjamin Sowell, Dan Tecuci, Vinayak Thapliyal, and Matt Welsh. 2024. The Design of an LLM-powered Unstructured Analytics System. *arXiv:2409.00847 [cs.DB]* <https://arxiv.org/abs/2409.00847>
- [5] Public BI Benchmark. [n. d.]. Public BI Benchmark. https://github.com/cwida/public_bi_benchmark
- [6] Asim Biswal, Liana Patel, Siddharth Jha, Amog Kamsetty, Shu Liu, Joseph E. Gonzalez, Carlos Guestrin, and Matei Zaharia. 2024. Text2SQL is Not Enough: Unifying AI and Databases with TAG. *arXiv:2408.14717 [cs.DB]* <https://arxiv.org/abs/2408.14717>
- [7] Dan Brickley, Matthew Burgess, and Natasha Noy. 2019. Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. In *The world wide web conference*. 1365–1375.
- [8] Sebastian Bruch, Siyu Gai, and Amir Ingber. 2023. An analysis of fusion functions for hybrid retrieval. *ACM Transactions on Information Systems* 42, 1 (2023), 1–35.
- [9] Sonia Castelo, Rémi Rampin, Aécio Santos, Aline Bessa, Fernando Chirigati, and Juliana Freire. 2021. Auctus: A dataset search engine for data discovery and augmentation. *Proceedings of the VLDB Endowment* 14, 12 (2021), 2791–2794.
- [10] R. Castro Fernandez, Z. Abedjan, F. Koko, G. Yuan, S. Madden, and M. Stonebraker. 2018. Aurum: A Data Discovery System. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*. 1001–1012. doi:10.1109/ICDE.2018.00094
- [11] Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In *Findings of the Association for Computational Linguistics ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, 2318–2335. doi:10.18653/v1/2024.findings-acl.137
- [12] Wenhui Chen, Ming-Wei Chang, Eva Schlinger, William Yang Wang, and William W Cohen. 2020. Open Question Answering over Tables and Text. In *International Conference on Learning Representations*.
- [13] chromadb. [n. d.]. <https://www.trychroma.com/>.
- [14] Tlameo Emmanuel, Thabiso Maupong, Dimane Mpoeleng, Thabo Semong, Banyatsang Mphago, and Oteng Tabona. 2021. A survey on missing data in machine learning. *Journal of Big data* 8 (2021), 1–37.
- [15] Hugging Face. [n. d.]. <https://huggingface.co/models>.
- [16] Grace Fan, Roei Shraga, and Renee J. Miller. 2024. Gen-T: Table Reclamation in Data Lakes. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE Computer Society, Los Alamitos, CA, USA, 3532–3545. doi:10.1109/ICDE60146.2024.00272
- [17] Yanlin Feng, Sajjadur Rahman, Aaron Feng, Vincent Chen, and Eser Kandogan. 2024. CMDBench: A Benchmark for Coarse-to-fine Multimodal Data Discovery in Compound AI Systems. In *Proceedings of the Conference on Governance, Understanding and Integration of Data for Effective and Responsible AI* (Santiago, AA, Chile) (GUIDE-AI '24). Association for Computing Machinery, New York, NY, USA, 16–25. doi:10.1145/3665601.3669846
- [18] Raul Castro Fernandez, Ziawasch Abedjan, Famiem Koko, Gina Yuan, Samuel Madden, and Michael Stonebraker. 2018. Aurum: A data discovery system. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*. IEEE, 1001–1012.
- [19] Raul Castro Fernandez, Pranav Subramaniam, and Michael J Franklin. 2020. Data market platforms. *Proceedings of the VLDB Endowment* 13, 12 (2020), 1933–1947.
- [20] Sainyam Galhotra, Yue Gong, and Raul Castro Fernandez. 2023. Metam: Goal-oriented data discovery. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*. IEEE, 2780–2793.
- [21] Jingtong Gao, Bo Chen, Xiangyu Zhao, Weiwen Liu, Xiangyang Li, Yichao Wang, Zijian Zhang, Wanyu Wang, Yuyang Ye, Shanru Lin, et al. 2024. LLM-enhanced Reranking in Recommender Systems. *arXiv preprint arXiv:2406.12433* (2024).
- [22] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997* (2023).
- [23] Anna Gaulton, Louisa J. Bellis, A. Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, Shaun McGlinchey, David Michalovich, Bissan Al-Lazikani, and John P. Overington. 2011. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* 40, D1 (09 2011), D1100–D1107. doi:10.1093/nar/gkr777 [arXiv:https://academic.oup.com/nar/article-pdf/40/D1/D1100/16955876/gkr777.pdf](https://academic.oup.com/nar/article-pdf/40/D1/D1100/16955876/gkr777.pdf)

- [24] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.
- [25] Yue Gong, Zhiru Zhu, Sainyam Ghalotra, and Raul Castro Fernandez. 2023. Ver: View Discovery in the Wild. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*. 503–516. doi:10.1109/ICDE55515.2023.00045
- [26] Joseph M. Hellerstein, Vikram Sreekanti, Joseph E. Gonzalez, James Dalton, Akon Dey, Sreyashi Nag, Krishna Ramachandran, Sudhanshu Arora, Arka Bhattacharyya, Shirshanka Das, Mark Donsky, Gabriel Fierro, Chang She, Carl Steinbach, Venkat Subramanian, and Eric Sun. 2017. Ground: A Data Context Service. In *8th Biennial Conference on Innovative Data Systems Research, CIDR 2017, Chaminade, CA, USA, January 8-11, 2017, Online Proceedings*. www.cidrdb.org. <http://cidrdb.org/cidr2017/papers/p111-hellerstein-cidr17.pdf>
- [27] Jonathan Herzig, Thomas Mueller, Syrine Krichene, and Julian Eisenschlos. 2021. Open Domain Question Answering over Tables via Dense Retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 512–519.
- [28] Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly Supervised Table Parsing via Pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 4320–4333. doi:10.18653/v1/2020.acl-main.398
- [29] Zhengbao Jiang, Yi Mao, Pengcheng He, Graham Neubig, and Weizhu Chen. 2022. OmniTab: Pretraining with Natural and Synthetic Data for Few-shot Table-based Question Answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 932–942. doi:10.18653/v1/2022.naacl-main.68
- [30] Nengzheng Jin, Joanna Siebert, Dongfang Li, and Qingcai Chen. 2022. A survey on table question answering: recent advances. In *China Conference on Knowledge Graph and Semantic Computing*. Springer, 174–186.
- [31] jstor. [n. d.]. <https://www.jstor.org/>.
- [32] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 6769–6781.
- [33] Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. 2022. Matryoshka Representation Learning. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 30233–30249. https://proceedings.neurips.cc/paper_files/paper/2022/file/c32319f4868da7613d78af9993100e42-Paper-Conference.pdf
- [34] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043* (2017).
- [35] Fangyu Lei, Fangyu Lei, Jixuan Chen, Yuxiao Ye, Ruisheng Cao, Dongchan Shin, Hongjin Su, Zhaoqing Suo, Hongcheng Gao, Wenjing Hu, Pengcheng Yin, Victor Zhong, Caiming Xiong, Ruoxi Sun, Qian Liu, Sida Wang, and Tao Yu. 2024. Spider 2.0: Evaluating Language Models on Real-World Enterprise Text-to-SQL Workflows. <https://api.semanticscholar.org/CorpusID:273970164>
- [36] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.
- [37] Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, Xuanhe Zhou, Chenhao Ma, Guoliang Li, Kevin C.C. Chang, Fei Huang, Reynold Cheng, and Yongbin Li. 2024. Can LLM already serve as a database interface? a big bench for large-scale database grounded text-to-SQLs. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (New Orleans, LA, USA) (NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 1835, 28 pages.
- [38] Chunwei Liu, Matthew Russo, Michael Cafarella, Lei Cao, Peter Baile Chen, Zui Chen, Michael Franklin, Tim Kraska, Samuel Madden, Rana Shahout, and Gerardo Vitagliano. 2025. Palimpsest: Optimizing AI-Powered Analytics with Declarative Query Processing. In *Proceedings of the Conference on Innovative Database Research (CIDR)* (2025).
- [39] Jerry Liu. 2022. *LlamaIndex*. https://github.com/jerryliu/llama_index
- [40] Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian guang Lou. 2021. TAPEX: Table Pre-training via Learning a Neural SQL Executor. *arXiv:2107.07653 [cs.CL]*
- [41] Xing Han Lu. 2024. BM25S: Orders of magnitude faster lexical search via eager sparse scoring. *arXiv e-prints*, Article arXiv:2407.03618 (July 2024), arXiv:2407.03618 pages. doi:10.48550/arXiv.2407.03618 arXiv:2407.03618 [cs.IR]
- [42] Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence* 42, 4 (2018), 824–836.

- [43] AirBnB Metis. [n.d.]. <https://medium.com/airbnb-engineering/metis-building-airbnbs-next-generation-data-management-platform-d2c5219edf19>.
- [44] Microsoft. 2024. Adventure Works. <https://learn.microsoft.com/en-us/sql/samples/adventureworks-install-configure?view=sql-server-ver16&tabs=ssms>
- [45] Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 2014–2037. doi:10.18653/v1/2023.eacl-main.148
- [46] Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Hailey Schoelkopf, Riley Kong, Xiangru Tang, Mutethia Mutuma, Ben Rosand, Isabel Trindade, Renusree Bandaru, Jacob Cunningham, Caiming Xiong, Dragomir Radev, and Dragomir Radev. 2022. FeTaQA: Free-form Table Question Answering. *Transactions of the Association for Computational Linguistics* 10 (2022), 35–49. doi:10.1162/tacl_a_00446
- [47] Simone Papicchio, Paolo Papotti, and Luca Cagliero. 2023. QATCH: Benchmarking SQL-centric tasks with Table Representation Learning Models on Your Data. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 30898–30917. https://proceedings.neurips.cc/paper_files/paper/2023/file/62a24b69b820d30e9e5ad4f15ff7bf72-Paper-Datasets_and_Benchmarks.pdf
- [48] Liana Patel, Siddharth Jha, Parth Asawa, Melissa Pan, Carlos Guestrin, and Matei Zaharia. 2024. Semantic Operators: A Declarative Model for Rich, AI-based Analytics Over Text Data. arXiv:2407.11418 [cs.DB] <https://arxiv.org/abs/2407.11418>
- [49] Chicago Data Portal. [n.d.]. *Chicago Data Portal*. <https://data.cityofchicago.org/>
- [50] pubmed. [n.d.]. <https://pubmed.ncbi.nlm.nih.gov/>.
- [51] Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval* 3, 4 (2009), 333–389.
- [52] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A Trainable Agent for Role-Playing. arXiv:2310.10158 [cs.CL] <https://arxiv.org/abs/2310.10158>
- [53] Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct Distillation of LM Alignment. arXiv:2310.16944 [cs.LG] <https://arxiv.org/abs/2310.16944>
- [54] Fei Wang, Kexuan Sun, Muhao Chen, Jay Pujara, and Pedro Szekely. 2021. Retrieving complex tables with multi-granular graph representation learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1472–1482.
- [55] Qiming Wang and Raul Castro Fernandez. 2023. Solo: Data Discovery Using Natural Language Questions Via A Self-Supervised Approach. *Proc. ACM Manag. Data* 1, 4, Article 262 (dec 2023), 27 pages. doi:10.1145/3626756
- [56] Zilong Zhang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, and Tomas Pfister. 2024. Chain-of-Table: Evolving Tables in the Reasoning Chain for Table Understanding. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=4L0xnS4GQM>
- [57] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yeqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 Technical Report. arXiv:2407.10671 [cs.CL] <https://arxiv.org/abs/2407.10671>
- [58] Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 8413–8426.
- [59] Tao Yu, Rui Zhang, Kai-Chou Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Z Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir R. Radev. 2018. Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task. *ArXiv abs/1809.08887* (2018). <https://api.semanticscholar.org/CorpusID:52815560>
- [60] Shuo Zhang and Krisztian Balog. 2018. Ad hoc table retrieval using semantic similarity. In *Proceedings of the 2018 world wide web conference*. 1553–1562.

- [61] Shuo Zhang and Krisztian Balog. 2020. Web table extraction, retrieval, and augmentation: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)* 11, 2 (2020), 1–35.
- [62] Yunjia Zhang, Jordan Henkel, Avriella Floratou, Joyce Cahoon, Shaleen Deep, and Jignesh M. Patel. 2024. ReAcTable: Enhancing ReAct for Table Question Answering. *Proc. VLDB Endow.* 17, 8 (April 2024), 1981–1994. doi:10.14778/3659437.3659452
- [63] Zixuan Zhao and Raul Castro Fernandez. 2022. Leva: Boosting machine learning performance with relational embedding data augmentation. In *Proceedings of the 2022 International Conference on Management of Data*. 1504–1517.
- [64] Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning. *ArXiv abs/1709.00103* (2017). <https://api.semanticscholar.org/CorpusID:25156106>

Received October 2024; revised January 2025; accepted February 2025