

ERATTA: Extreme RAG for enterprise-Table To Answers with Large Language Models

Sohini Roychowdhury*, Marko Krema*, Anvar Mahammad*, Brian Moore*, Arijit Mukherjee §, Punit Prakashchandra §

*Corporate Data and Analytics Office (CDAO), Accenture, USA

§CDAO, Accenture, India

Abstract—Large language models (LLMs) with retrieval augmented-generation (RAG) have been the optimal choice for scalable generative AI solutions in the recent past. Although RAG implemented with AI agents (agentic-RAG) has been recently popularized, it suffers from unstable cost and unreliable performances for Enterprise-level data-practices. Most existing use-cases that incorporate RAG with LLMs have been either generic or extremely domain specific, thereby questioning the scalability and generalizability of RAG-LLM approaches. In this work, we propose a unique LLM-based system where multiple LLMs can be invoked to enable data authentication, user-query routing, data-retrieval and custom prompting for question-answering capabilities from Enterprise-level data tables on sustainability. The source tables here are highly fluctuating and large in size storing carbon footprint, energy and water usage at buildings in regional levels globally and the proposed framework enables structured responses in under 10 seconds per query. Additionally, we propose a five metric scoring module that detects and reports hallucinations in the LLM responses. Our proposed system and scoring metrics achieve > 90% confidence scores across hundreds of user queries in the sustainability, financial health and social media domains. Extensions to the proposed extreme RAG architectures can enable heterogeneous source querying using LLMs.

Index Terms—Retrieval Augmented Generation (RAG), authentication, sustainability, hallucination, Unicorn, TextBison

I. INTRODUCTION

Generative AI solutions using Large Language Models (LLMs) have been extremely popular since the launch of Open AI's chatGPT-3 in late 2022 [1]. Most LLM solutions today can typically be categorized as Gen-AI applications, such as a chatbot-like experience with a virtual assistant for code, language and content generation [2]; or Gen-AI agents that are capable of automated orchestration to perform specific tasks such as booking tickets, writing blogs, articles and generating software automatically [3]. Most such solutions implement the retrieval augmented-generation (RAG) approach, wherein the most relevant data sources are first identified followed by the knowledge-graph (KG) method to isolate only relevant data entities that are collated with natural language instructions and answering guidelines and sent to the LLM to receive the required responses [4]. Although agentic-RAG approaches have been found to enhance the "knowledge-quality" of retrieved content instead of relying on LLM memory, these approaches can be costly and time-intensive owing to the web-search and KG method. Thus, for *Enterprise-level* data products involving large numbers of tabular rows and columns, agentic-RAG can

prove impractical for question-answering for a large number of users/groups. In this paper, we present a novel system architecture that maximizes the RAG impact through multiple LLM calls for tasks such as user authentication, query routing and natural language question-answering while ensuring security/access requirements and accuracy standards by flagging hallucinations for critical domains such as sustainability.

So far, towards use-cases that require question-answering from large volumes of tabular data at *Enterprise* level, language-based RAG approaches have been accurate as presented in [2]. The proposed framework enhances system scalability by compartmentalizing the RAG process for specific tasks (extreme RAG) and by relying on automated code generation as an intermediate step for tables-to-answers use-cases. The various RAG tasks are as follows: 1) user-access authentication for tabular data based on minimal-user-profile (MUP) data, 2) user-query understanding and routing based on intention 3) SQL code generation from natural language to retrieve specific sub-tabular data (also known as seq2sql [5]) 4) response generation using the retrieved sub-tabular context and natural language. The proposed extreme RAG and hallucination control metrics presented here are scalable and efficient for knowledge retrieval with lower maintenance and operating costs when compared to the existing works [2] [4]. The process of generating code (SQL, Python, C/C++) using LLMs to query tabular-data has several advantages. First, LLMs can accept data in tabular format, which allows better manageability for *relevant data* search to each user query when compared to semantic retrieval on textual data from vector databases as presented in the work [2]. Structured data tables eliminate the need to manage a complex retrieval engine and results ranking system. Second, passing data in smaller sub-tables to LLMs enables scalability on new data tables without the need for the an offline table to text generation and vector embedding process in [2]. For tabular data sets that have high refresh rates such as stock market prices, inflation indices, carbon footprint etc., passing the relevant sub-tables ensures most accurate data-to-answers. Third, tabular data is structured and concise, which minimizes the scope for data misrepresentation and hallucinations. It is noteworthy that LLMs have been found to be more accurate and stable for text-to-code generation than for text-to-text generation in terms of *fake responses* or *hallucinations* [1]. Thus, our novel extreme-RAG system shown in Fig. 1 incorporates text-to-SQL code

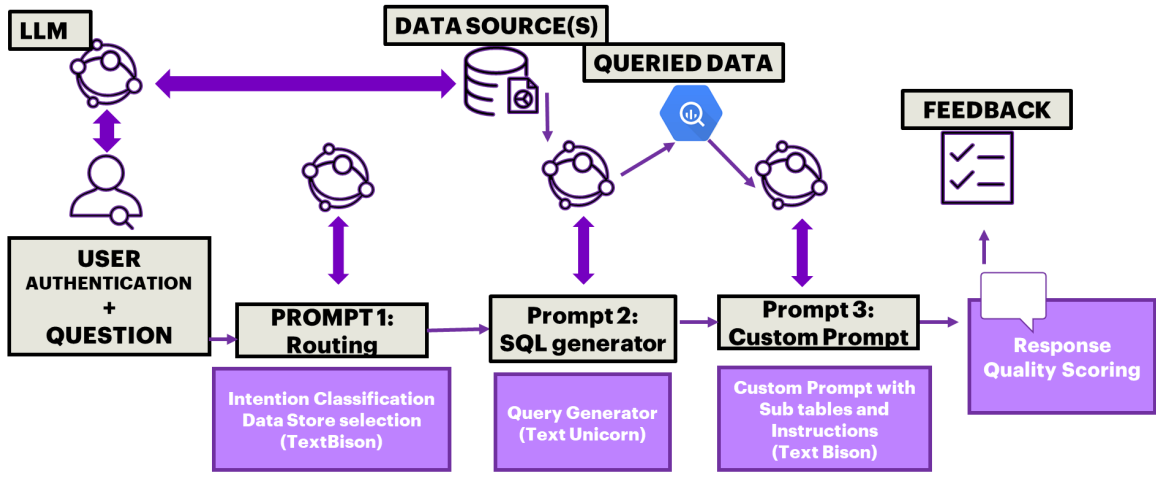


Fig. 1. The proposed extreme-RAG system architecture that invokes 3-4 LLM prompts per user query to authenticate, route, extract smaller tabular data and fetch the required response from fast-varying and large tabular sustainability data sources.

generation followed by passing shorter data tables to LLM prompts to enable a variety of table-to-answers that are tested on *Sustainability domain* datasets.

This paper makes three key contributions. First, we present a serial framework that implements RAG at multiple levels to complete specific tasks per step to separate the user-query understanding, data access and response generation processes. Second, we present non-LLM hallucination detection metrics that evaluate the accuracy and reliability of each query response. Third, we present extensions to the extreme-RAG framework towards answering predictive and prescriptive user-queries from Enterprise-tabular data. We compare the performance of the extreme-RAG with agentic-RAG systems and observe higher reliability and accuracy for the extreme RAG approach. This is primarily due to the intermediate stage of (Prompt 2 in Fig. 1) where SQL queries are automatically generated to retrieve the most relevant data from the enterprise tables. By focusing on smaller, sub-tables with labeled information, our system capitalizes on the LLM's strengths in handling concise and well-defined datasets.

II. RELATED WORK

Generating responses from tabular data sources has been an area of focus for the past few decades. Early works such as [6] relied on the structure of tabular data in xml/html documents to parse and extract the required answers. Prior to the advent of LLMs, some major efforts in this domain included the development of algorithms known as 'seq2sql' models that can accept a sequence of input words and convert it into a corresponding SQL query as shown in Table I. In this work, we extend the state-of-the-art by breaking down the tasks of data fetching from an input sequence of words into three stages, 1) query routing, 2) multiple-SQL generation, 3) SQL and text combination to yield accurate solutions to complex, multi-tabular queries from enterprise data products.

III. DATA AND METHODS

For our experiments, we utilize enterprise-scale sustainability-specific datasets containing information regarding carbon footprint, water/electricity consumption and renewable energy usage across countries and continents measured at daily, monthly and yearly levels. The dataset comprises of 7 tables of sizes 50 MB to 1.2 GB each with over 1000 rows and 50 columns of data in each table. Based on these tables, the user-queries can be categorized into the following 7 major categories and their combinations as shown in Table II. The descriptions for the extreme-RAG system components are presented below.

A. Extreme-RAG System Components

1) *Authentication RAG*: This process is a scalable extension to an otherwise rule-based look-up that enables/maps tables to users based on the user-access restrictions. As shown in Fig. 1, upon each user login, there is a need to enable a database lookup to retain access only to the tables that the user can retrieve from and to withhold access to all remaining tables. For example, a user from 'North America' can only have access to power-consumption information from North America. So only the tables with user-access are pre-loaded to the Postgres DB or BigQuery for to set up a question-answering session while all remaining geo-location tables are skipped. Next, a comprehensive table-access profile (or MUP) needs to be generated to accommodate super-users who may need to access multiple tables across continents. For instance, a user specializing in renewable energy may require access to specific data tables across continents. For authenticating access for hundreds and thousands of users on thousands of data tables with varying schema, a custom prompt that outputs the minimal access constraints (or MUP) in json/xml formats is the first step for a chat session set-up.

2) *Prompt 1, Routing a User-query*: Once authenticated, each user-query needs to be routed for its intention and appropriate data tables. This query understanding and routing is an essential step since the instructions to respond to the different

TABLE I
A SUMMARY OF SEQ2SQL ALGORITHMS PRIOR TO LLMs FOR TABLES-TO-ANSWERS USE-CASES AND THEIR CAPABILITIES.

| Model Name | Year | Core Capabilities |
|--|------------|---|
| Bidirectional Attention for SQL Generation [7] | June, 2018 | <ul style="list-style-type: none"> - Three sub-modules were designed with deep learning models for inference. - Bi-directional attention mechanisms and character embeddings were implemented with convolutional neural networks (CNNs). - Experimental evaluations were presented for the Wiki SQL dataset (87,726 hand annotated text and SQL). |
| Tranx [8] | 2018 | <ul style="list-style-type: none"> - A transition-based neural semantic parser that mapped natural language (NL) utterances into formal meaning representations. - Highly generalizable system that can be applied to new representations by just writing a new abstract syntax description. - Experimental verification performed for 4 semantic parsing and code generation tasks. |
| STAMP+RL [9] | 2018 | <ul style="list-style-type: none"> - For typical seq2SQL models, inaccuracies occur due to mismatch between question words and table contents. - Quality of generated SQL query was improved through content replication by column name, cells or SQL keywords. - Generation of 'WHERE' clause was further improved by using column-cell relations. |
| PT-MAML [10] | 2018 | <ul style="list-style-type: none"> - Adapted meta learning to solve the situation where the number of conditions in a text-to-SQL query vary significantly. - This method decoded a text sequence guided by a fixed syntax pattern by tagging for schema and constant values. |

TABLE II
CATEGORIES OF QUESTIONS ROUTED BY PROMPT 1

| Intention | Type | Example user-queries |
|-----------|---------|---|
| 0 | Percent | What % of our offices are at 100% renewable electricity? |
| 1 | Change | What is the annual reduction of emissions globally? |
| 2 | Rank | Which country has the highest Emissions type 1 emissions? |
| 3 | Level | What is scope 1 emission levels for offices in Argentina? |
| 4 | Rank | Which city had the highest water consumption for Dec 2022? |
| 5 | Multi | Which countries reduced scope 3 emissions consistently in the last 2 years and increased renewable electricity? |
| 6 | FAQ | What is included in business travel? |

categories of questions in Table I vary significantly. Query extension and query rewriting is then performed by converting the user-query into its high-dimensional embedded format and matching it to at-most five prior query samples in the vector space. The output of this prompt 1 is a comprehensive list of data sources relevant to the user-query.

This routing prompt is trained on a specific template to identify table relevance. The result is a list of table names in Python list format, which can be easily integrated with other data-processing tasks. Although this can be envisioned as a standard classification task as in [2], LLMs are found to be more generalizable and scalable across a wide range of use queries and source tables with minimal training than standard supervised learning models.

Following the identification of *relevant* data sources, the proposed system proceeds to retrieve a list of data source configurations for each identified source. These data source configurations represent tabular metadata [10] and are important to isolate the appropriate SQL queries in the next step.

3) *Prompt 2, Data Retrieval*: Once the appropriate data tables are mapped to a user query, the data retrieval prompt converts standard language to SQL code. This prompt takes in 3 inputs: 1) the rewritten and extended sub-queries, 2) a list of data source configurations (metadata) obtained by semantic matching of the queries with prototype questions, and 3) a sample question and its answer. Utilizing these inputs, Prompt 2 generates a complex, nested, SQL query. A query runner process runs this SQL query against the data sources specified by the previous step to fetch the required information in tabular form. Only the *relevant* rows and fields from the pre-loaded tables are fetched and loaded into a specific BigQuery Table.

The metadata corresponding to the mapped tables, can be expressed as “Table Configurations” and the example questions can be specified as “Prototype Question configurations” as shown in Fig. 2. Here, *Table Configurations* represents an

encapsulation (class) with properties for the table name, a list of relevant table fields, and a list of sample field values. *Prototype Question Query* similarly is another class that maps prior user-queries to data source names. The primary advantage of using these data structures in Prompt 2 is that it eliminates the need for code changes whenever the data domains change. For each new data source, the Table Configuration List is automatically updated, and with each new user-query category, the Prototype Question Query list is updated. Thus, these data structures enable generalizability and scalability.

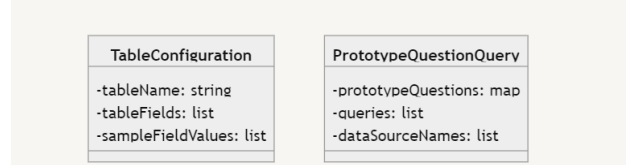


Fig. 2. Examples of class definitions to extract tabular metadata and enable scaling prompt 2 across source tables and user query types.

4) *Prompt 3, Answer Retrieval*: Finally, the third prompt per user-query utilizes the re-written and extended questions and the tabular data loaded at the end of prompt 2 in BigQuery to generate a customized prompt that is then sent to the LLM for a natural language response. This step is an extension to the classical *seq2sql* models as it collects the SQL query outputs combines them with standard instructional guardrails and a sample question and answer for response generation. The style and format are specified as part of the prompt. Additional instructions are provided to display relevant error messages related to lack of tabular-data access, lack of data and irrelevant user-queries. It is noteworthy that per user-query the requirement is for all 3 prompts to complete under 10 seconds of run time to ensure real-time qualities. An example of the three prompts per user query is shown in Fig. 3.

B. Quality Assurance

LLMs have been well known to be riddled with *hallucinations* [2] even after setting model parameters like temperature and top-P to zero. We propose the following factual checks (flags) $[s_{1,i}, s_{2,i}, s_{3,i}, s_{4,i}, s_{5,i}]$ for each user-query (i) in combination with the outcomes in prompts 2 and 3 to detect possible hallucinations for the extreme-RAG responses. For these checks, *spacy*-based modules are used to extract and compare the named entities in the user-query (Q_{Ni}), with those

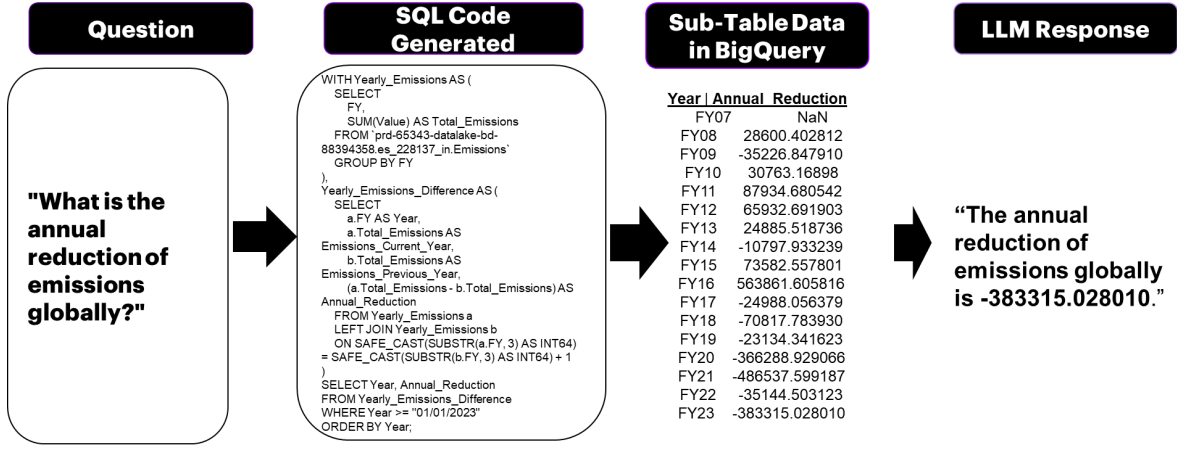


Fig. 3. Example of the steps in Prompts 1, 2 and 3 to retrieve an appropriate response from the sustainability data tables.

in the response (R_{N_i}), in the SQL queries (Σ_{N_i}) and the data loaded in BigQuery by Prompt 2 (D_{N_i}) as follows.

- 1) **Number check:** This flag in prompt 3 ensures that all numerical values mentioned in the answers are directly derived from the underlying data sources without any discrepancies or fabrications by the end of Prompt 2. This check is crucial for maintaining the factual accuracy of numerical data presented in responses as shown in (1).
- 2) **Entity Check:** This flag verifies that all entities mentioned in the user-query are accurately reflected in the response, ensuring the response's relevance and completeness. This step is critical for confirming that the answer addresses all aspects of the query comprehensively, thereby enhancing user satisfaction and trust in the system. For example, in a user-query regarding "water consumption rates in the USA", this check ensures that the named entity "USA" is included in the response (R_{N_i}) and query (Q_{N_i}) as shown in (2).
- 3) **Query Check:** This flag confirms that all essential keywords and conditions stated in the user question are included in the SQL commands that are executed in Prompt 2. This step verifies the presence of specific filter words used in the user question (Q_{N_i}) within the SQL queries (Σ_{N_i}). This flag is configured using column metadata for each table and each filter check is added individually per data schema. This step ensures that the query accurately reflects the criteria and constraints specified in the question and can be envisioned by equation (2) wherein R_{N_i} is changed to Σ_{N_i} .
- 4) **Regurgitation check:** This warning flag identifies if the response merely replicates information from the prompt 3 without paraphrasing. This step checks for word sequences where ten or more consecutive words from the prompt 3 are repeated. This step detects responses that fail to expand upon the query, which is essential for providing value-added insights to the user [2].
- 5) **Increase/Decrease Modifier check:** This flag checks for consistency between the directional changes mentioned in the user query and that described in the prompt 3

response. This step ensures that the response accurately reflects the dynamics or trends requested by the user, critical for analyses involving changes over time or comparisons.

$$s_{1,i} = \begin{cases} 1 & \text{if } \text{numbers}(R_i) \subset \text{numbers}(D_i) \\ 0 & \text{otherwise} \end{cases}, \quad (1)$$

$$s_{2,i} = \begin{cases} 1 & \text{if } Q_{N_i} == R_{N_i} \\ 0 & \text{otherwise} \end{cases}. \quad (2)$$

IV. EXPERIMENTS AND RESULTS

For validation of the proposed system we have deployed the proposed extreme-RAG system to answer specific questions across domains and databases using the Vertex AI platform in Google cloud platform. Prompt 1 and 3 are built using TextBison002 while Prompt 2 is built using the Unicorn LLM. The proposed system is scalable and can be generalized to other cloud-based and on-prem LLMs as well.

A. Qualitative Scalability Analysis

We assess the scalability of our extreme RAG-system on 100 random queries from multiple domains such as financial reporting, healthcare and social-media. Sample user-queries along with the generated SQL and the quality check scores are shown in Table III. Here, we observe that although the prompt 3 contains instructions based on the sustainability dataset, using a new-domain for validation dataset results in similar response performances. For financial domain data tables and queries, the proposed system achieves highest confidence scores (all scores are 1). For healthcare related queries, we observe some numerical hallucinations in the patient ID numbers that leads to inaccurate numerical predictions (80% accuracy for s_1 metric). For the social-media use case, the proposed system hallucinates numbers (that can be flagged by our scoring module) and the modifier flag returns a not applicable value of -1.

TABLE III
SAMPLE QUERIES, GENERATED SQL AND RESPONSE SCORES

| Question | SQL command | s_1, s_2, s_3, s_4, s_5 |
|--|--|---------------------------|
| What was the revenue growth for the Northeast region in last quarter compared to the previous quarter? | SELECT region, SUM(revenue) AS revenue FROM financials WHERE region = 'Northeast' AND (quarter = 'Last' OR quarter = 'Previous') GROUP BY quarter | [1,1,1,1,1] |
| How many new patients were admitted to the oncology department this month and how does this compare to last month? | SELECT department, COUNT (patient-id) AS new-patients FROM hospital-admissions WHERE department = 'Oncology' AND (month = 'This' OR month = 'Last') GROUP BY month | [0.8,1,1,1,1] |
| What was the click-through rate (CTR) for the digital marketing campaign on social media platforms in January? | SELECT platform, AVG (click-through-rate) AS CTR FROM marketing-data WHERE campaign = 'Digital Marketing' AND platform = 'Social Media' AND month= 'January' | [0,1,1,1,-1] |

B. Quantitative Response Analysis

The quantitative responses on 500 sample variants of 60 queries generated on the sustainability datasets are shown in Fig. 4. We observe that all responses have average scores around 90%. The number check score is lowest at 89.09% followed by entity check and modifier checks at 94.55% each. Averaged query-check and regurgitation-check are around 92.73%. This shows the impact of the proposed Extreme-RAG system towards hallucination control and scalability.

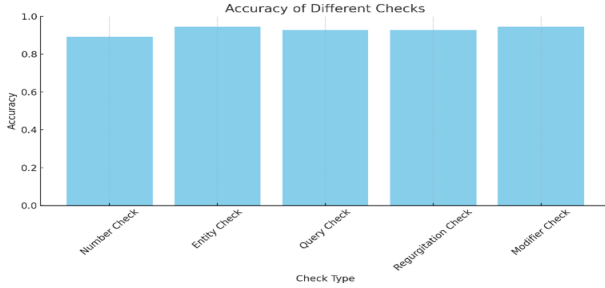


Fig. 4. Averaged response quality scores on sustainability dataset $[s_1$ to $s_5]$

C. Extreme-RAG Extensions

The extreme-RAG architecture can further be extended to scenario planning and predictive capabilities using LLMs. Fig. 5 shows a sample flow where a real-time optimization/predictive algorithm is invoked based on the named entities extracted from prompt 1, and the result returned from the algorithm can then be saved into BigQuery followed by a prompt 3 that furnishes the response in natural language. Such extensions/modifications to the code generation LLM prompt (prompt 2) enhance the scalability of the proposed system.

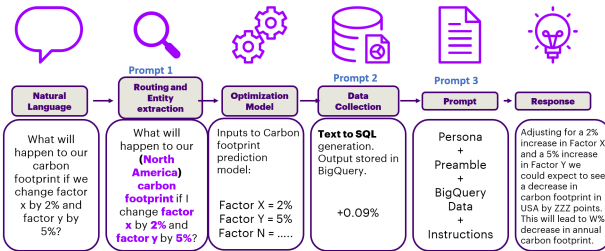


Fig. 5. Example of an extension of the proposed Extreme RAG system to solve for predictive and prescriptive trend questions using back-end optimization algorithms that can be invoked by LLM prompts (prompt 2).

D. Comparison with Agentic-RAG

Langchain (LangGraph) and LlamaIndex have been utilized heavily by several custom products till date to enable LLMs with orchestration capabilities to build intelligent AI agents [11]. For any user-query where the underlying data is relatively large to skim through or highly varying in nature, AI agents can perform a web-search, knowledge identification followed by RAG and automated verification (using Trulens [12]). In Table IV, we compare the performance of an agentic-RAG incorporated using Langchain after prompt 1 to the proposed Extreme-RAG solution. It is noteworthy that agentic-RAG involves prior vector-embedding of the enterprise Tables as an offline process and an additional LLM-based evaluation step.

TABLE IV
COMPARISON WITH AGENTIC RAG AFTER PROMPT 1

| Method | # LLM calls | avg time (s) | Evaluation metrics | Avg score |
|-----------|-------------|--------------|----------------------------------|-----------|
| Proposed | 2 | 8.5 | $[s_1-s_5]$ | 89-95% |
| Langchain | 4-5 | 40 | [Relevance, Toxicity, Grounding] | 40-98% |

V. CONCLUSION

In this work, we propose a novel multi-LLM system that enables scalable tables-to-answers for Enterprise-level sustainability data tables with an averaged accuracy of $> 90\%$ in under 10 seconds per user-query. The proposed process of dividing a user-query for query authentication, query understanding and routing, sub-tabular retrieval followed by answer generation is more stable and faster when compared to agentic-RAG approaches, which makes it a better fit for numeric-intensive use cases such as Sales, Financial and Sustainability domains. Future works can be directed towards extending RAG approaches to heterogeneous data sources for automation and insight generation tasks.

REFERENCES

- [1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
- [2] S. Roychowdhury, A. Alvarez, B. Moore, M. Crema, M. P. Gelpi, P. Agrawal, F. M. Rodríguez, Á. Rodríguez, J. R. Cabrejas, P. M. Serrano *et al.*, "Hallucination-minimized data-to-answer framework for financial decision-makers," in *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 2023, pp. 4693–4702.
- [3] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023, pp. 1–22.
- [4] T. Bui, O. Tran, P. Nguyen, B. Ho, L. Nguyen, T. Bui, and T. Quan, "Cross-data knowledge graph construction for llm-enabled educational question-answering system: A case study at hcmut," *arXiv preprint arXiv:2404.09296*, 2024.

- [5] X. Zhang, F. Yin, G. Ma, B. Ge, and W. Xiao, "M-sql: Multi-task representation learning for single-table text2sql generation," *IEEE Access*, vol. 8, pp. 43 156–43 167, 2020.
- [6] X. Wei, B. Croft, and A. McCallum, "Table extraction for answer retrieval," *Information retrieval*, vol. 9, pp. 589–611, 2006.
- [7] G. Huilin, G. Tong, W. Fan, and M. Chao, "Bidirectional attention for sql generation," in *2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*. IEEE, 2019, pp. 676–682.
- [8] P. Yin and G. Neubig, "Tranx: A transition-based neural abstract syntax parser for semantic parsing and code generation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (Demo Track)*, 2018.
- [9] Y. Sun, D. Tang, N. Duan, J. Ji, G. Cao, X. Feng, B. Qin, T. Liu, and M. Zhou, "Semantic parsing with syntax-and table-aware sql generation," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 361–372.
- [10] P.-S. Huang, C. Wang, R. Singh, W.-t. Yih, and X. He, "Natural language to structured query generation via meta-learning," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, 2018.
- [11] J. Pinheiro, W. Victorio, E. Nascimento, A. Seabra, Y. Izquierdo, G. García, G. Coelho, M. Lemos, L. A. P. P. Leme, A. Furtado *et al.*, "On the construction of database interfaces based on large language models," in *WEBIST*, 2023, pp. 373–380.
- [12] A. Datta, M. Fredrikson, K. Leino, K. Lu, S. Sen, R. Shih, and Z. Wang, "Exploring conceptual soundness with trulens," in *NeurIPS 2021 Competitions and Demonstrations Track*. PMLR, 2022, pp. 302–307.