Fin-DBQA Shared-task: Database Querying and Reasoning

Rungsiman Nararatwong,[1] Natthawut Kertkeidkachorn,[2]Hiroya Takamura,[1] Ryutaro Ichise[3,1]
[1]Artificial Intelligence Research Center, AIST, Japan
[2]Japan Advanced Institute of Science and Technology
[3]Institute of Science Tokyo
{r.nararatwong, takamura.hiroya}@aist.go.jp
natt@jaist.ac.jp, ichise@iee.e.titech.ac.jp

Abstract

This paper presents the results of the Fin-DBQA shared task based on a question-answering dataset, focusing on database querying and reasoning. The dataset, consisting of 400 questions grouped into 40 conversations, evaluates language models' abilities to answer sequential questions with complex reasoning and multi-hop queries in a multi-turn conversational question-answering setting. Each sample includes the question, answer, database queries, querying result (tables), and a program (series of operations) that produces the answer from the result. We received 52 submissions from three participants, with scores significantly surpassing the baselines. One participant submitted a paper detailing a prompt-based solution using large language models with additional data preprocessing that helps improve the overall performance.

1 Introduction

While earlier research on question answering has predominantly focused on text-based QA systems (Rajpurkar et al., 2016; Chen et al., 2021a; Gaim et al., 2023), recent efforts have expanded to include tabular QA (Zhang et al., 2020; Pal et al., 2023), and hybrid QA approaches (Chen et al., 2020; Zhu et al., 2021; Chen et al., 2022). These advancements, however, typically assume that all required tables or datasets are provided as inputs during experimentation. In contrast, real-world scenarios often involve more complex requirements. For example, answering a question like "What is the difference in net profit between Amazon and Microsoft in 2023?" (Q1) necessitates a two-step process: querying relevant data and performing reasoning. Specifically, models must retrieve the revenues of the two companies for 2023 and subsequently apply mathematical reasoning to compute the difference.

In a conversational question setting, users build on previous queries. A user might ask, "Did that figure increase from the previous year?" (Q2). To answer Q2, a model must first resolve the coreference ("that" refers to the revenue difference from Q1), then retrieve the relevant data for 2022, compute the difference for that year, and compare it to the result from Q1. Alternatively, a follow-up question might be unrelated to Q1 yet require complex reasoning, such as, "Which company had the highest revenue in the technology sector in 2023?" Answering this involves multi-hop querying: the model must first identify the technology sector, then locate the relevant companies, and finally compare their revenues. These examples highlight the challenges of sequential and multi-hop question answering, where models must integrate reasoning, coreference resolution, and data navigation to provide accurate answers.

To address the limitations of previous studies concerning the querying step in question answering, we introduce the Fin-DBQA shared task based on the DBQR-QA dataset (Nararatwong et al., 2024). This task is built around a novel question-answering dataset designed to evaluate database querying and reasoning capabilities. The dataset comprises 400 questions organized into 40 conversations, enabling the assessment of language models in handling sequential, multi-hop queries within a multi-turn conversational setting. Each data sample includes the question, its answer, corresponding database queries, the querying results (tables), and a program detailing the operations required to derive the answer from the results. The task attracted 52 submissions from three participants, with performance metrics significantly surpassing the established baselines. One participant proposed a prompt-based approach leveraging large language models, complemented by additional data preprocessing techniques, which further enhanced overall performance.

Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal, pages 385–391
January 19–20, 2025.

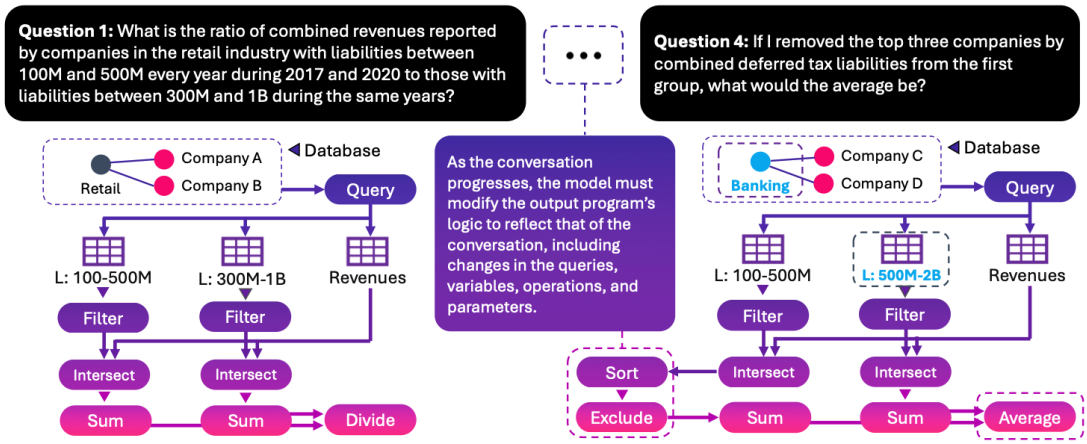

Figure 1: Example in DBQR-QA.

2 Related Work

Large language models (LLMs) have achieved significant advancements in reasoning-based question answering (QA). This progress is evident across diverse benchmarks, including the DROP dataset (Dua et al., 2019) for reading comprehension and arithmetic QA, the GSM-8K dataset (Cobbe et al., 2021) for solving grade-school math word problems, the MMLU benchmark (Hendrycks et al., 2021) for multi-domain multiple-choice questions, and the NumHG dataset (Huang et al., 2024) for number-focused headline generation.

Tabular QA is another domain that demands advanced reasoning skills. Key datasets in this area include TAT-QA (Zhu et al., 2021), which focuses on hybrid financial tabular and textual data; FinQA (Chen et al., 2021b), designed for numerical reasoning in finance; and FeTaQA (Nan et al., 2022), which supports free-form table question answering. Building upon the foundations of TAT-QA and FinQA, our dataset extends the scope of reasoning by integrating querying and reasoning into the problem-solving process.

Numerous conversational QA datasets focus on large knowledge bases, enabling diverse multi-hop questions. Notable examples include SQA (Iyyer et al., 2017) for Wikipedia tables, CSQA (Saha et al., 2018) for reasoning over knowledge bases, and ConvQuestions (Christmann et al., 2019), which spans five domains. Non-knowledge-base QA datasets also present significant challenges, such as CoQA (Reddy et al., 2019) for machine comprehension and QuAC (Choi et al., 2018) for dialog-based contexts. Despite years of extensive research in conversational QA, its tabular and reasoning aspects remain underexplored. ConvFinQA (Chen et al., 2022) addresses this gap by focusing on numerical reasoning chains within single-table conversational QA.

3 Dataset Construction

3.1 Dataset Overview

Figure 1 illustrates an example from the DBQR-QA dataset. This dataset, developed for the Fin-DBQA shared task, features questions that require a combination of database querying and complex multi-step table manipulations. The tasks are further complicated by a multi-branch chain of reasoning, where each question in the sequence introduces, modifies, or removes queries, variables, operations, and parameters. This progressive complexity challenges models not only to memorize information but also to dynamically adapt and refine their reasoning throughout the conversation.

The questions in the proposed DBQR-QA dataset were derived from the TAT-QA and FinQA datasets, both of which were manually crafted and annotated by financial experts. However, the limited variety of reasoning operations in these datasets led to many questions exhibiting similarities. To address this, similar questions were grouped into a template-based representation. Using BART (Lewis et al., 2020), these elements were extracted to generate generalized templates. For example, the question "What was the net revenue in 2019?" was abstracted to "What was the [concept] in [year]?" This abstraction process involved calculating string similarity scores, grouping templates by similarity, and refining them to align with the graph database context, extending beyond simple tabular data.

Similar to ConvFinQA, DBQR-QA converts

questions from FinQA into a conversational format, but it differentiates itself by incorporating table manipulation throughout the reasoning process. Unlike ConvFinQA, which relies on only six basic arithmetic operators—such as addition, subtraction, multiplication, and division, DBQR-QA includes 26 operators within the Pandas DataFrame. This expanded set of operators facilitates more complex and expressive queries compared to previous datasets.

After establishing the question templates, we populated them with entities (e.g., companies), financial concepts, and numerical data, ensuring alignment with the US-GAAP taxonomy. We prioritized terms based on their frequency of occurrence in the questions, selecting those represented in the graph to guarantee the accuracy of the generated answers. Next, we defined a set of operations and combined them to create a program for each question, marking the initial stage of the annotation process.

3.2 Automatic-Answer Annotation

To leverage the responses annotated by financial experts in TAT-QA and FinQA, we developed a knowledge graph derived from financial report tables formatted as XBRL documents. This integration enables the handling of complex tasks requiring extensive data interlinking by storing the relevant information within the graph. The graph's querying mechanism facilitates the transformation of results into tables that can be further manipulated during reasoning steps. Through the knowledge graph, automatic-answer annotations for generated questions become readily accessible. For instance, a question from TAT-QA, such as "How much revenue came from LinkedIn in 2018?" is adapted to "How much profit came from Apple in 2023?" in our dataset. In TAT-QA, the annotation process involves extracting the triple (revenue, LinkedIn, 2018) to answer the question. In our context, the corresponding automatic-answer annotation consists of the triple (profit, Apple, 2023), providing a preliminary answer.

3.3 Answer Verification

We utilized Amazon Mechanical Turk workers to validate our automatic-answer annotations. Their task involved reading the questions and constructing a program (a sequence of tabular operations) based on data queried from the database. The system subsequently compared their program's output with our own. In cases of discrepancies, the workers were required to identify which program, or whether the question itself, was incorrect. This method reduced the potential bias of our interpretation influencing theirs, a concern that would have arisen if we had asked them to verify our programs directly.

To ensure the quality of annotations, only workers who achieved a minimum score of 70% on three qualification tests were considered eligible. Furthermore, they provided sufficient explanations for any discrepancies in their answers, demonstrating their ability to identify and address potential issues. A question was deemed valid if it received a majority consensus. We reviewed the workers' feedback and identified questions that were flagged as incorrect, such as those involving the possibility of a negative value when measuring the "difference" between two quantities (e.g., the difference between A and B). These issues were addressed with additional clarification.

4 Dataset Statistics

The DBQR-QA dataset is divided into five distinct subsets, each categorized according to question type and complexity. This classification introduces a diverse range of question types designed to assess querying and reasoning abilities. These categories are specifically structured to explore the intricate aspects of financial datasets, addressing various objectives and levels of complexity. An overview of the five unique question types within the dataset is presented below.

Type 1: Simple Query with Specific Companies (Simple)

This type involves direct questions concerning specific companies, requiring the extraction of data and the application of basic arithmetic to derive solutions. A typical example might involve financial metrics over a designated period, such as determining which year to exclude in order to maximize the average deferred revenues of a particular company.

Type 2: Complex Query with Unspecified Companies (Complex)

The complexity in this context arises from the lack of specification regarding the companies of interest, as well as the incorporation of conditional thresholds for financial metrics. The objective is to select criteria that fit a specific financial parameter across a set of companies. For example, this

could involve identifying the year with the highest average contractual obligations, based on varying minimum thresholds for purchase obligations.

Type 3: Reasoning Steps Requiring Multiple Tables (Multi-Table)

This category involves synthesizing data from multiple tables to address questions that require comparative analysis or the aggregation of financial metrics across different periods or conditions. It evaluates the ability to navigate and interpret interconnected datasets, such as comparing average earnings per share across various years, while accounting for differences between basic and diluted shares.

Type 4: Multi-hop Query (Multi-Hop)

Multi-hop queries require a series of logical steps and inferences to reach a conclusion. These types of questions typically involve complex, industry-specific analyses, such as comparing averages or trends across various criteria or time periods. For instance, a question might inquire which industry-level factor leads to a higher average net cash provided by operating activities, necessitating an understanding of temporal trends and the unique characteristics of different industries.

Type 5: Instruction QA (Instruction)

Instruction-based questions involve intricate scenarios that direct the analyst through a sequence of data retrieval and analytical tasks across multiple dimensions, such as time, industry, and financial metrics. These questions simulate real-world data analysis challenges, necessitating a deep understanding and the capacity to follow multi-step instructions in order to compare and contrast averages or identify trends within specific groups of companies.

5 Evaluation

5.1 Manual Evaluation

There are two primary types of answers: textual and numerical. An answer can consist of a single value or a set, which may include either texts or numbers. Textual answers may take the form of comparisons (such as "higher," "lower," or "equal") or references to entities, including financial terms defined in the taxonomy, companies, individuals, industries, and countries. No other types of textual answers are permitted. Human evaluators are required to focus solely on the answer itself, disregarding any additional contextual information or other details, regardless of their accuracy. In the case of an answer being a set of values, the predicted and actual sets must match exactly, with the order of elements being irrelevant. That is, all values must be present in the answer, and no extraneous values should be included. When the set consists of specific years or entities—such as company revenues within a certain period—the predicted values must clearly identify all the correct years or entities.

5.2 Automatic Evaluation

5.2.1 Heuristic Evaluator

The heuristic evaluator is less flexible in handling the model's output, especially for a prompt-based approach. For example, the model may output "higher" or "greater," possibly with an explanation, for a question asking whether something is more or less than another. Even so, it offers a quick preliminary evaluation that works well with numbers, covering most answers. The evaluator refers to the label to determine the answer type, then applies the following rules to process the answers:

1. *Integer*: Convert the numeric answer into an integer using int(answer).

2. *Float*: Convert the numeric answer into a string with two-digit floating point using "

3. *Set*: All items in the prediction and label sets must match. Otherwise, the algorithm flags the answer as incorrect.

4. *Dictionary*: All keys and values must match. The label uses the entity/concept names, not their mentions, e.g., "CATERPILLAR INC" not "Caterpillar" and "us-gaap: Revenues" not "total revenue."

5.2.2 GPT-4 Evaluator

We instructed GPT-4 to compare the generated response with human annotations (refer to Appendix A for the prompt). In the DBQR-QA dataset's experiment, we created two evaluation prompts: Binary and scoring. The binary prompt asks the model to determine whether the answer is correct. The scoring prompt asks the model to grade the answer from 0 to 10, 0 being no match and 10 being an exact match. However, we only use the binary prompt in this shared task for simplicity and cost management.

	Grader	GPT	Human
Practice (50 questions)			
Jan Strich	.54	.54	.56
Training (200 questions)			
Jan Strich	.33	.31	.37
Test (150 questions)			
Dunamu-ML	.64	.63	.64
Jan Strich	.52	.51	.55
Jonathan Zhou	.26	.21	.30

Table 1: Evaluation scores of all participants.

6 Results

A total of 5 submissions were received for the practice set (50 questions), 2 for the training set (200 questions), and 45 for the test set (150 questions). Each set included all five types of questions. Table 1 presents a summary of the best scores achieved by each participant. The scores across evaluators are generally similar. Based on the assessments of human evaluators, the highest scores for the practice, training, and test sets were 0.56, 0.37, and 0.64, respectively.

6.1 Participant's Solution

Of the three participating teams, one submitted a paper describing their methodology and experimental results. In their study, the authors introduced a prompt-based approach incorporating a preprocessing step that converts tables into a "tidy data" format (Wickham, 2014), wherein each column corresponds to a variable and each row represents an observation. As presented in Table 2, their experiments conducted on four large language models demonstrate consistent and significant improvements compared to the baseline approach employed by DBQR-QA.

7 Conclusion

The Fin-DBQA shared task highlights the challenges associated with addressing multi-turn conversational question-answering that involves complex reasoning and multi-hop queries. While the solutions proposed by participants achieved performance metrics significantly surpassing the baseline, considerable scope for improvement remains, offering opportunities for further advancements in future research.

Model	Pass	Fail	Crash
DBQR-QA baseline	.18	.52	.27
Llama 3.1 8B + *tidy data* + *5-shot*	.20	.61	.20
Llama 3.1 70B (FP8) + *tidy data* + *5-shot*	.22	.61	.17
GPT-4o-mini + *tidy data* + *5-shot*	.39	.53	.08
GPT-4o + *tidy data*	.51	.46	.04

Table 2: Evaluation scores submitted by Jan Strich (participant). We only reported the experimental condition for each model that yielded the highest pass rate. **5-shot**: With 5-shot examples.

Acknowledgement

This shared task is partially based on the results obtained from a project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021a. Nquad: 70,000+ questions for machine comprehension of the numerals in text. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2925–2929.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020. Hybridqa: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021b. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6292, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettle-

moyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Philipp Christmann, Rishiraj Saha Roy, Abdalghani Abujabal, Jyotsna Singh, and Gerhard Weikum. 2019. Look before you hop: Conversational question answering over knowledge graphs using judicious context expansion. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 729–738, New York, NY, USA. Association for Computing Machinery.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Fitsum Gaim, Wonsuk Yang, Hancheol Park, and Jong Park. 2023. Question-answering in a low-resourced language: Benchmark dataset and models for Tigrinya. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11857–11870, Toronto, Canada. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Jian-Tao Huang, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2024. NumHG: A dataset for number-focused headline generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12323–12329, Torino, Italia. ELRA and ICCL.

Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. 2017. Search-based neural structured learning for sequential question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1821–1831, Vancouver, Canada. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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*, 10:35–49.

Rungsiman Nararatwong, Chung-Chi Chen, Natthawut Kertkeidkachorn, Hiroya Takamura, and Ryutaro Ichise. 2024. DBQR-QA: A question answering dataset on a hybrid of database querying and reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15169–15182, Bangkok, Thailand. Association for Computational Linguistics.

Vaishali Pal, Andrew Yates, Evangelos Kanoulas, and Maarten de Rijke. 2023. MultiTabQA: Generating tabular answers for multi-table question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6322–6334, Toronto, Canada. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Amrita Saha, Vardaan Pahuja, Mitesh M. Khapra, Karthik Sankaranarayanan, and Sarath Chandar. 2018. Complex sequential question answering: Towards learning to converse over linked question answer pairs with a knowledge graph. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Hadley Wickham. 2014. Tidy data. *Journal of statistical software*, 59:1–23.

Shuo Zhang, Zhuyun Dai, Krisztian Balog, and Jamie Callan. 2020. Summarizing and exploring tabular data in conversational search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*,

System prompt
You are an evaluator. Given a series of conversational questions, your task is to compare an answer to the last question predicted by an AI to an answer labeled by a human.
Binary evaluator
Question: ... AI's answer: ... Human's answer: ... Are the two answers to the last question the same? Answer "yes" or "no" in the following JSON format: ‘‘‘ { "result": "yes" or "no" } ‘‘‘ Do not explain or output anything else.

Table 3: Evaluation prompt for GPT-4.

SIGIR '20, page 1537–1540, New York, NY, USA. Association for Computing Machinery.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287, Online. Association for Computational Linguistics.

A Evaluation Prompt

We use OpenAI's GPT models for evaluation. Table 3 shows the prompt we used for evaluation.