
ORCA: ORchestrating Causal Agent

Joanie Hayoun Chung^{1,2*}, Chaemyung Lim^{2*}, Sumin Lee¹, Sungbin Lim^{1,3†}

¹Department of Statistics, Korea University

²Business School, Korea University

³LG AI Research

{jchung02, xoaud14, sungbin}@korea.ac.kr

Abstract

Causal inference is essential for decision-making science while the complexity of the data analysis workflow, ranging from data wrangling to causal analysis, increases substantially as the scale of data grows in complicated business environments. Especially, the execution of the workflow in relational databases by non-experts can result in repetitive bottlenecks which impede timely and responsible business insights. To address this challenge, we propose **ORCA (Orchestrating Causal Agent)**, an LLM agentic system that can automate routine workflows in RDBMS while preserving expert oversight via human-AI interactions. ORCA orchestrates the full data analysis pipeline: interpreting natural language queries, navigating tables from DB servers, generating proper SQL codes, preprocessing data, and configuring modeling processes using causal inference libraries. Domain experts still can control the automation through iterative interactions with ORCA, enabling robust data-driven decision making with less technical expertise in statistical computing. Empirical evaluations on benchmark and synthetic e-commerce datasets demonstrate competitive performance of ORCA in table understanding, query generation, and cause-effect estimation—achieving over $7\times$ improvement in estimating average treatment compared to GPT-4o mini.

1 Introduction

Understanding causality is increasingly critical to decision-making situations in medical [12, 25], political [5, 10], and business domains [4, 11], which require responsibility and safety. However, valid causal analysis demands domain knowledge to search a suitable data generating process, as well as statistical expertise to choose appropriate identification and estimation strategies. As the scale of data increases, the complexity of causal analysis grows exponentially since causal discovery and cause-effect estimation pose substantial computational and operational challenges [2, 7, 18]. Contrary to advancements in data infrastructure and management system, user-friendly tools to perform scalable causal analysis remain underdeveloped, especially in relational databases. In high-dimensional relational databases, the end-to-end pipeline from navigating complex schemas to executing estimations is not only time-consuming, but also hinders reliable causal analysis.

Recent developments in large language models (LLMs) have enabled agentic systems that aim to automate the data analysis workflow [13, 17, 21]. Early works have focused on individual tasks, such as translating natural language queries into SQL code (Text2SQL) or summarizing data tables. More recently, multi-agent frameworks, including LangGraph, AutoGen [26], and CrewAI, has enabled agents to interpret analytical goals, identify relevant data, and coordinate modular tasks toward end-to-end analysis pipelines [8, 16, 22, 23]. However, current agent-based systems remain fragmented

*Equal Contribution.

†Corresponding Author. E-mail: sungbin@korea.ac.kr.

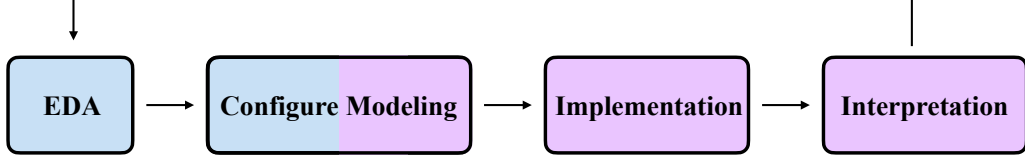


Figure 1: An abstraction of repetitive data analysis process depicted in two high-level phases: data examining ■ and causal analysis ■. While the former involves understanding and preparing the data through exploration, the latter focuses on implementing statistical methods to uncover cause-effect relationships and draw insights. Especially for configure modeling, it requires orchestration of both processes, as it is needed to consider the statistical model while examining the data.

in both scope and capability so that they fail to move beyond surface-level query generation to support advanced statistical reasoning. These systems often treat schema interpretation, data retrieval, modeling, and explanation as isolated components, lacking a unified design.

To implement an agentic system for a unified framework, we propose **ORCA (ORchestrating Causal Agent)**, that covers the entire data analysis pipeline, from data wrangling and SQL generation to cause-effect estimation and interpretation. The data analysis process can be grouped into two high-level phases; examining data and causal analysis (Figure 1). Correspondingly, ORCA is composed of two agents: the **Data Wrangler** and the **Causal Analyzer**, each responsible for one of these phases. By integrating LLM reasoning, tool orchestration, and causal modeling, ORCA supports complex data analysis through intuitive natural language interaction. ORCA leverages LLM-based agents to interpret user intent, retrieve and process data, apply causal inference techniques, and present interpretable results. Through iterative human-agent interaction, ORCA balances automation with expert oversight to address advanced analytical tasks. The main contributions of this paper are:

- We propose ORCA, an LLM agentic system that operates on relational databases and automates both data engineering and advanced statistical modeling pipelines, enabling end-to-end data analysis through natural language interfaces.
- We design an interactive framework by actively engaging with users throughout the iterative process to clarify user feedback and refine outputs to enhance analytical accuracy.
- We present REEF (Relational E-commerce Evaluation Framework), a multi-table, semi-synthetic dataset, to facilitate robust evaluation in relational databases, that reflects the structural complexity, noise, and incompleteness of real-world business environments. REEF provides a challenging yet controlled setting for assessing the full analytical pipeline, including the estimation of causal effects in realistic relational data environments.

2 Related Works

LLM Agents The capabilities of LLMs have evolved from static text generation to dynamic, tool-augmented reasoning. Early systems combined LLMs with external tools such as ChatGPT [19] Code Interpreter, APIs, and retrieval-augmented generation (RAG), enabling them to perform structured tasks like table manipulation, computation, and data querying. These tool-augmented agents demonstrated that LLMs could serve as flexible interfaces for interacting with structured environments. Building on these foundations, researchers have proposed multi-agent systems in which several LLM-based agents, each assigned a specialized role and toolset, collaborate to solve complex tasks [14, 26]. Frameworks such as CAMEL [14] and AutoGen [26] coordinate agents through message-passing, role-based prompting, and memory sharing, allowing agents to collectively plan, execute, and revise actions toward a shared objective [24]. As tasks become more complex, agent orchestration emerges as a critical challenge. Orchestration refers to the dynamic management of agent roles, communication flow, and task dependencies [9]. Effective orchestration is key to reducing redundancy, ensuring coherent reasoning, and enabling scalability. ORCA builds on these advances by combining tool-augmented reasoning, multi-agent collaboration, and human-in-the-loop orchestration into a unified framework for end-to-end causal data analysis.

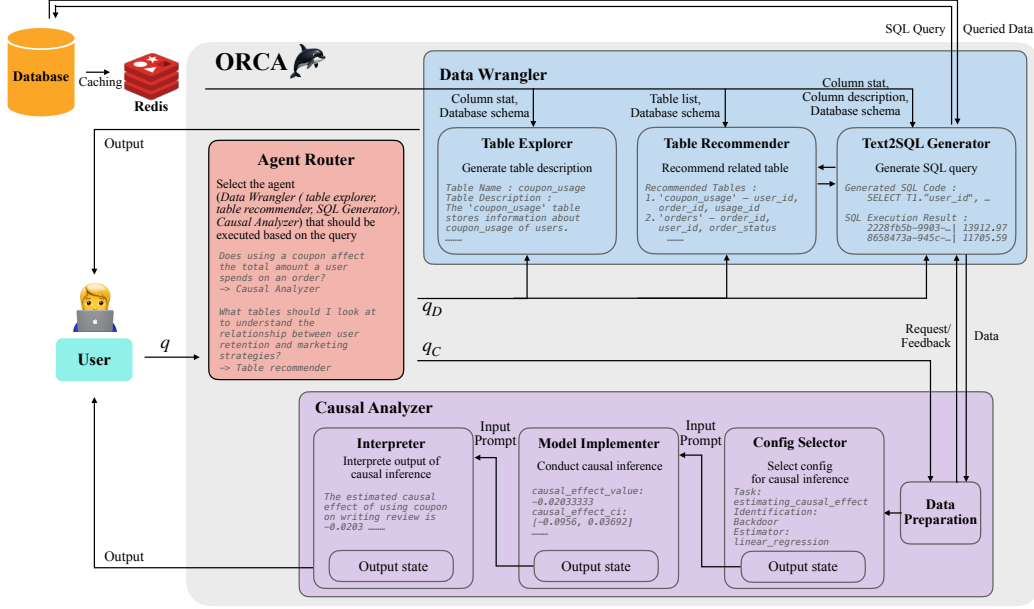


Figure 2: Overall diagram of ORCA framework. For given relational database and cache server, ORCA interacts with a user by receiving a query q through **Agent Router**, which classifies the query to route data query q_D and causal query q_C , and automates data analysis workflow by operating two modular agents, **Data Wrangler** and **Causal Analyzer**.

Causal Inference Softwares Causal inference libraries such as DoWhy [3, 20], EconML [1], and CausalML [6] offer robust and principled pipelines for estimating treatment effects under various assumptions. DoWhy, in particular, emphasizes modeling of causal assumptions and organizes inference into four stages, identification, estimation, refutation, and interpretation. This modular design supports transparency and flexibility for analysts. However, these tools assume access to clean, structured data and require users to manually select causal variables and specify identification and estimation strategies, each of which demands understanding of causal methodology. This requirement poses a barrier in real-world scenarios where data is distributed across complex relational databases, with noise or missing values. ORCA fills this gap by unifying data engineering, LLM-based reasoning, and causal estimation within an interactive agent framework. Unlike existing causal libraries, ORCA automatically extracts relevant variables from databases, constructs plausible causal models, and communicates findings through natural language. This enables non-expert users to conduct sophisticated causal analyses without manual data preparation or statistical programming.

3 ORCA: Orchestrating Causal Agent

3.1 Framework Overview

ORCA is an LLM-based agent system designed to reflect the end-to-end workflow of a data analyst. Given a natural language query, ORCA interprets the analytical intent and dynamically routes it to the appropriate agent. ORCA integrates two modular agents: the **Data Wrangler**, which automates data discovery and retrieval by interacting with external systems such as relational databases and cache servers; and the **Causal Analyzer**, which executes causal inference tasks (see Figure 2). The **Causal Analyzer** operates on the data provided by the **Data Wrangler**, and the two modules interact iteratively to refine the analysis. Consequently, two agents form a full pipeline for causal analysis. Furthermore, ORCA supports interactive execution. Users can provide clarification and feedback at each step, so the agent can request additional queries when uncertainty arises. This design enables robust, adaptable analysis workflows without requiring deep technical expertise from the user.

3.2 Data Wrangler

Understanding the structure and role of each table in large and complex databases is a well-known bottleneck for data analysts. In particular, identifying relevant tables and columns for a given analytical goal and writing the correct SQL codes to extract data, can be both time-consuming and error-prone. The **Data Wrangler** receives the data query q_D and performs three key functions: (1) table description, implemented by the *Table Explorer*; (2) table retrieval, handled by the *Table Recommender*; and (3) SQL code generation, carried out by the *Text2SQL Generator*.

3.2.1 Table Explorer

The *Table Explorer* automatically generates a detailed overview of each table by integrating column statistics and metadata into clear, analysis-ready insights and recommendations. This enables users to efficiently grasp the table’s structure, supporting informed interaction and seamless progression to downstream analytical tasks.

Input Prompt When a query requests information about a specific table, the *Table Explorer* uses Prompt B.1 to generate a detailed description of the table, including descriptive statistics such as high null ratios, skewed distributions, and potential outliers for each column. The prompt is supplemented with context such as basic statistics for each column, such as data types, mean, median, null counts, unique value counts, and representative example values, which are retrieved from the cache server.

Output State With the generated description, the *Table Explorer* outputs possible analyses, such as user-level aggregation, time-series trend analysis, or cohort comparison, by utilizing foreign key relationships retrieved from the cache server. Sample output can be found in Figure D.1.

3.2.2 Table Recommender

The *Table Recommender* serves as an essential component in the initial stages of analysis by automatically selecting table and columns that correspond to the user’s analytical intent. By intent-aware selection and visual schema representation, the module improves accessibility and planning efficiency, particularly in collaborative or enterprise-scale analytical workflows.

Input Prompt A user query is given either in natural language or as a document. The *Table Recommender* is instructed by Prompt B.2 to extract or summarize the underlying analytical objective and recommend the most relevant tables and key columns. Metadata such as table and column names, and foreign key relationships, pre-stored in the cache server, are provided as context. When dealing with large-scale database, it is infeasible to include all table and column information directly in the prompt. To address this scalability challenge, the *Table Recommender* first performs an embedding-based similarity search, comparing the embedding of the extracted analytical objective with the pre-computed embeddings of table and column metadata in the vector database.

Output State Along with the recommended tables/columns, the module generates an Entity-Relationship Diagram (ERD) to enhance schema interpretability. This visual representation, exported as a PNG file, provides an intuitive overview of the associations among the recommended tables and columns, thereby supporting more structured and efficient analytical planning. See Figure D.2 for detailed output.

3.2.3 Text2SQL Generator

The *Text2SQL Generator* interacts directly with the database for code execution and data retrieval. To accurately translate natural language queries into SQL codes, the module follows a three-stage pipeline that ensures both syntactic validity and semantic alignment with user intent. Before generating the SQL query, it is necessary to select only the tables and columns necessary to the analysis, rather than using all available tables in the database. The *Text2SQL Generator* leverages the *Table Recommender* to identify the most relevant tables and columns for the query.

Generation Prompt The *Text2SQL Generator* uses Prompt B.3 to break down the user query into sub-questions, using the metadata of selected tables/columns and foreign key relationships as

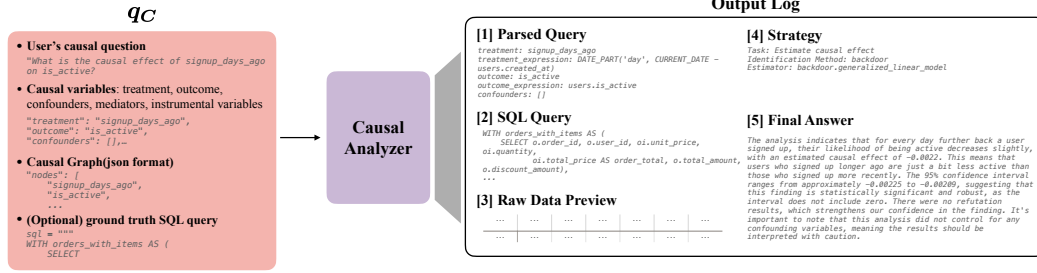


Figure 3: Overview of the **Causal Analyzer** workflow. Given a causal query q_C from the **Agent Router**, the agent executes a causal data analysis pipeline and outputs a result log.

context. The prompt includes few-shot examples as well as instructions for Chain-of-Thought (CoT) decomposition, guiding the LLM to explicitly generate intermediate reasoning steps.

Self-correction Prompt Once an initial SQL code is generated, it is executed against the database to detect potential errors. If execution fails, the module is prompted to analyze the error and revise the query. Specifically, the prompt includes the faulty SQL statement, the corresponding error message, the metadata of the involved tables and columns, and the original user query (see Prompt B.4). The refinement is performed iteratively, with up to three attempts to balance accuracy and latency.

Validation Prompt Even if the generated SQL code executes successfully, the result may not always reflect the intent of user request. Hence The *Text2SQL Generator* uses Prompt B.5 to evaluate whether the returned result satisfactorily answers the initial user request, with both the original user query and the result provided as context. The prompt additionally includes the executed SQL query and the relevant table and column metadata, which enables the agent to determine whether the outcome stems from correct but overly restrictive query logic or from flaws in query construction. If the output is deemed insufficient or inconsistent with the intent of user, the module undergoes a refinement process, in which feedback is provided and the output is improved accordingly.

Output State The module returns both the generated SQL code and its execution results. Additionally, LLM-based feedback on the code is provided. Sample output can be found in Figure D.3.

3.3 Causal Analyzer

The **Data Wrangler** processes user queries q_D on the database schema. ORCA extends beyond this phase to the entire causal analysis pipeline, which typically requires statistical expertise. The **Causal Analyzer** operates causal inference modules accurately and transparently to prevent procedural errors, hence it enables reliable data analysis and ensures interpretability at each step in the pipeline. The **Causal Analyzer**, also built using the LangGraph framework, comprises three core modules—*Config Selector*, *Model Implementer*, and *Interpreter*—preceded by a Data Preparation submodule that extracts and cleans the dataset based on the user’s causal query q_C (see Figure 3). This process reflects the labor-intensive preprocessing phase typically performed by data analysts. The sample output can be found in Figure D.4.

Data Preparation By receiving the user query q_C which includes the identified variables and causal graph, the Data Preparation sub-module prompts the *Text2SQL Generator* to handle schema mappings, joins, and filtering conditions automatically, when constructing SQL queries to extract all the necessary variables. The retrieved dataset is subsequently preprocessed as needed. For example, object-type variables are encoded into numeric or categorical formats to ensure compatibility with the causal inference library. Additional normalization steps can optionally be applied.

3.3.1 Config Selector

The *Config Selector* module identifies an appropriate causal inference configuration tailored to the user’s question and the dataset characteristics. Once the dataset is prepared, it reviews the input ques-

tion, variables, and a data sample to determine a suitable strategy. The selection corresponds to the core configurations of the textttDoWhy framework: causal task, identification strategy, estimation method, and optionally, a refutation technique for robustness checks. To generate this configuration, the agent employs Prompt B.6 that lists valid options with concise descriptions. This automation alleviates the need for users to understand causal theory or API-specific syntax.

3.3.2 Model Implementer

The *Model Implementer* module executes the selected strategy by estimating the causal effect based on the full dataset and causal graph. It receives four key inputs: the strategy selected from the previous step, a parsed query specifying causal variables, a dataset, and the causal graph. This module follows the four-stage procedure of DoWhy [3, 20]: (1) constructing a causal model, (2) identifying the estimand based on the graph and variable configuration, (3) estimating the causal effect using the chosen method, and (4) optionally running refutation to assess robustness. Internally, estimation methods require additional configurations depending on the estimator type and outcome variable. The resulting model object, estimand, and estimate are stored in the system state. Additionally, outputs including the estimated effect and its confidence interval are saved for downstream use.

3.3.3 Interpreter

The *Interpreter* module translates the causal inference results into natural language. It uses an LLM with Prompt B.7 to produce a concise but understandable summary to users without a statistical background. Based on the treatment effect estimate, confidence interval, and variable descriptions, the module explains what the effect means, whether it is statistically significant, and the robustness of the findings. The output is a 3–6 sentence summary including key insights, any assumptions or limitations.

4 Experiments

To evaluate the effectiveness of the proposed framework across its core modules, we conduct a series of four experiments, each aligned with a specific sub-task. These experiments are designed to measure how well the system understands structured data, selects relevant tables, generates executable SQL queries, and performs causal inference under realistic database conditions. For fair comparison, all components and the baseline use GPT-4o-mini, with schema information explicitly included in prompts. The source code is available at our GitHub repository³.

Dataset We evaluate our system using two datasets. First, we use BIRD [15], a text-to-SQL benchmark grounded in real-world domains, designed to test models under challenging conditions such as external knowledge reasoning, SQL efficiency over large databases, and complex query structures including implicit joins and aggregations. For reproducibility and efficient error analysis, we adopt the official BIRD mini-dev subset, which includes 500 natural language queries across 12 diverse databases⁴. In addition, we construct a custom e-commerce database, **REEF**(Relational E-commerce Evaluation Framework). Existing benchmarks reflect less explicit causal relationships across relational data. REEF consists of 18 interrelated tables (e.g., products, orders, users) to validate our system in realistic environments. We design the underlying data generating process to encode specific causal relationships among variables, enabling the construction of ground-truth causal graphs. Details are provided in the Appendix A.

4.1 Task 1: Table Description

We evaluate the ability of *Table Explorer* to generate informative and concise natural language descriptions of database tables. For each table, both the baseline and ORCA are tasked with producing a description that conveys the core semantics and structure of the table. The quality of these descriptions is then assessed using a standardized rubric. The baseline model is prompted to take a table as input and to create a description. Due to the large size of the tables, only the first 100 rows are used as input for baseline model. The evaluation is conducted using a 5-point scale, which measures

³<https://github.com/ChaemyungLim/ORCA>

⁴https://github.com/bird-bench/mini_dev

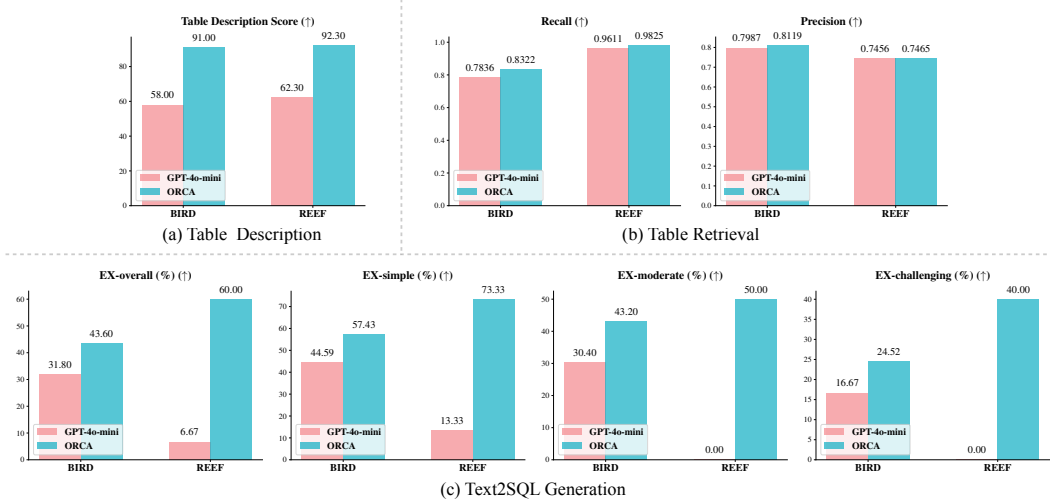


Figure 4: Barplots of tasks (a) Table Description, (b) Table Retrieval, (c) Text2SQL Generation.

how informative and meaningful each description is. GPT-4o mini serves as the evaluator scoring the output from both models. To ensure fairness, GPT-4o-mini is prompted with clear and consistent grading instructions. Detailed grading criteria are provided in the Appendix C.

Results As shown in Figure 4 (a), ORCA significantly outperforms the baseline in both datasets, achieving average scores of 91.0 and 92.3 compared to 58.0 and 62.3, respectively. This demonstrates ORCA’s strong ability to interpret and summarize structured data effectively.

4.2 Task 2: Table Retrieval

We evaluate the ability of the *Table Recommender* to recommend relevant tables given a natural language question. The goal of this task is to identify which tables from a database are necessary to answer a user query. The baseline model (GPT-4o mini) is prompted with a list of all table names in the database, followed by the user question. It is instructed to select which tables would be helpful for answering the question.

Metric Recall denotes the average proportion of ground-truth tables that were included in the system’s recommendation. Precision means the average proportion of tables recommended by the system that are in the ground-truth set. The metrics are computed as follows:

$$\text{Recall} = \frac{1}{N} \sum_{n=1}^N \frac{|G_n \cap R_n|}{|G_n|}, \quad \text{Precision} = \frac{1}{N} \sum_{n=1}^N \frac{|G_n \cap R_n|}{|R_n|}, \quad (1)$$

where N is the total number of evaluation examples, G_n denotes the set of ground-truth tables for the n -th example, and R_n is the set of tables recommended by the system.

Results As shown in Figure 4 (b), ORCA consistently outperforms baseline on both datasets, achieving lower AMR and AER scores. This indicates that ORCA is more effective at identifying relevant tables while minimizing the inclusion of unnecessary ones.

4.3 Task 3: Text2SQL Generation

We evaluate the ability of *Text2SQL Generator* to generate a valid SQL query for the user’s request. Each SQL prediction is executed against the corresponding database, and the output is compared with that of the ground-truth SQL to compute execution accuracy.

Metric We adopt Execution Accuracy (EX) as the primary evaluation metric. EX measures the percentage of evaluation examples where the predicted SQL and the ground-truth SQL produce the

same execution result. Considering the result set as V_n executed by the n -th ground-truth SQL Y_n , and the result set \hat{V}_n executed by the predicted SQL \hat{Y}_n , EX is computed by:

$$\text{EX} = \frac{1}{N} \sum_{n=1}^N \mathbb{1}(V_n = \hat{V}_n). \quad (2)$$

Results As shown in Figure 4 (c), ORCA consistently outperforms GPT-4o mini across all difficulty levels on both BIRD and REEF datasets. Notably, while the REEF dataset reflects a more realistic and structurally complex database environment, ORCA shows a strong execution accuracy of 60.00%, whereas GPT-4o mini’s performance significantly drops to 6.67%.

4.4 Task 4: Treatment Effect Estimation

We evaluate causal analysis performance on the REEF by evaluating treatment effect estimation. While ORCA retrieves and processes relevant data and schema based on the given treatment and outcome variables along with the causal graph, the GPT-4o mini is explicitly provided with the same information. To evaluate the data access accuracy on the final causal estimate, we report two experimental results. ORCA (oracle) is given the ground-truth SQL query to directly retrieve the correct dataframe while ORCA (agentic) performs the full pipeline autonomously, starting from SQL generation based on inputs such as the causal graph and variable names.

Metric We measure the confidence interval (CI) coverage rate, the proportion of queries for which the predicted Average Treatment Effect (ATE) falls within the ground-truth CI, and the Absolute Error between predicted and ground-truth ATE values. These metrics jointly reflect the correctness and stability of causal effect estimation.

Model	CI Coverage (%)	MAE	MSE	Max Abs. Error
GPT-4o mini	13.0 ± 2.99	14.2 ± 5.0	2500.0 ± 1400.0	140.0 ± 47.0
ORCA (agentic)	74.8 ± 5.3	1.6 ± 1.2	4.4 ± 3.6	1.2 ± 9.4
ORCA (oracle)	89.9 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0

Table 1: Causal estimation performance (mean ± 95% CI) over 10 runs on the REEF dataset.

Results As shown in Table 1, ORCA significantly outperforms the GPT-4o mini baseline across all evaluation metrics. While GPT-4o mini achieves only 13.0% CI coverage, ORCA (agentic) achieves 74.8%, and ORCA (oracle) reaches 89.9%, demonstrating over a $7\times$ improvement. ORCA (oracle) achieves accurate estimates, confirming that the **Causal Analyzer** can produce highly precise results when provided with exact data access. Nevertheless, ORCA (agentic) achieves competitive performance compared to GPT-4o mini, showing that the agentic system can reach practical levels.

5 Conclusion and Limitation

We introduce ORCA, an LLM-based system designed to enable intuitive, end-to-end causal analysis over relational databases. ORCA allows non-expert users to inspect schemas, generate executable SQL queries, and perform robust causal inference with minimal manual intervention. The system serves as a practical blueprint for combining LLMs with structured reasoning to support trustworthy, scalable data analysis workflows.

However, ORCA does not claim to fully replace human expertise. In particular, it assumes access to a causal graph and does not yet support data-driven causal discovery, which typically requires domain knowledge. Our goal is not full automation, but rather safe delegation: automating routine components while preserving opportunities for expert oversight. Future work includes tighter integration of feedback loops and extending ORCA to open-ended discovery tasks.

References

- [1] Keith Battocchi, Eleanor Dillon, Maggie Hei, Greg Lewis, Paul Oka, Miruna Oprescu, and Vasilis Syrgkanis. EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. <https://github.com/py-why/EconML>, 2019. Version 0.x.
- [2] Zewude A Berkessa, Esa Läärä, and Patrik Waldmann. A review of causal methods for high-dimensional data. *IEEE Access*, 2024.
- [3] Patrick Blöbaum, Peter Götze, Kailash Budhathoki, Atalanti A Mastakouri, and Dominik Janzing. Dowhy-gcm: An extension of dowhy for causal inference in graphical causal models. *Journal of Machine Learning Research*, 25(147):1–7, 2024.
- [4] Iavor I Bojinov, Albert Chen, and Min Liu. The importance of being causal. *Harvard Data Science Review*, 2(3), 2020.
- [5] Pedro Carneiro, Karsten T Hansen, and James J Heckman. Estimating distributions of treatment effects with an application to the returns to schooling and measurement of the effects of uncertainty on college, 2003.
- [6] Huigang Chen, Totte Harinen, Jeong-Yoon Lee, Mike Yung, and Zhenyu Zhao. Causalm: Python package for causal machine learning. *arXiv preprint arXiv:2002.11631*, 2020.
- [7] Shuyu Dong, Kento Uemura, Akito Fujii, Shuang Chang, Yusuke Koyanagi, Koji Maruhashi, and Michèle Sebag. High-dimensional causal discovery: Learning from inverse covariance via independence-based decomposition. *arXiv preprint arXiv:2211.14221*, 2022.
- [8] Siyuan Guo, Cheng Deng, Ying Wen, Hechang Chen, Yi Chang, and Jun Wang. DS-agent: Automated data science by empowering large language models with case-based reasoning. *arXiv preprint arXiv:2402.17453*, 2024.
- [9] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024.
- [10] James J Heckman and Rodrigo Pinto. The econometric model for causal policy analysis. *Annual review of economics*, 14(1):893–923, 2022.
- [11] Hoje Jo and Maretno A Harjoto. The causal effect of corporate governance on corporate social responsibility. *Journal of business ethics*, 106(1):53–72, 2012.
- [12] F Kühne, M Schomaker, I Stojkov, B Jahn, A Conrads-Frank, S Siebert, G Sroczynski, S Puntcher, D Schmid, P Schnell-Inderst, and U Siebert. Causal evidence in health decision making: methodological approaches of causal inference and health decision science. *Ger Med Sci*, 20:Doc12, 2022.
- [13] Mandar Kulkarni. Agent-s: Llm agentic workflow to automate standard operating procedures. *arXiv preprint arXiv:2503.15520*, 2025.
- [14] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008, 2023.
- [15] Jinyang Li, Binyuan Hui, Ge Qu, Binhua Li, Jiaxi Yang, Bowen Li, Bailin Wang, Bowen Qin, Rongyu Cao, Ruiying Geng, Nan Huo, Chenhao Ma, Kevin C. Chang, Fei Huang, Reynold Cheng, and Yongbin Li. Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. *ArXiv*, abs/2305.03111, 2023.
- [16] Ziming Li, Qianbo Zang, David Ma, Jiawei Guo, Tuney Zheng, Minghao Liu, Xinyao Niu, Yue Wang, Jian Yang, Jiaheng Liu, et al. Autokaggle: A multi-agent framework for autonomous data science competitions. *arXiv preprint arXiv:2410.20424*, 2024.
- [17] Sun Maojun, Ruijian Han, Binyan Jiang, Houduo Qi, Defeng Sun, Yancheng Yuan, and Jian Huang. Lambda: A large model based data agent. *Journal of the American Statistical Association*, pages 1–20, 2025.

- [18] Sakib A Mondal, Prashanth Rv, and Sagar Rao. A fast algorithm for high-dimensional causal discovery. *International Journal of Data Science and Analytics*, pages 1–10, 2025.
- [19] OpenAI. Introducing ChatGPT. <https://openai.com/index/chatgpt/>, 2022.
- [20] Amit Sharma and Emre Kiciman. Dowhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*, 2020.
- [21] Irene Testini, José Hernández-Orallo, and Lorenzo Pacchiardi. Measuring data science automation: A survey of evaluation tools for ai assistants and agents. *arXiv preprint arXiv:2506.08800*, 2025.
- [22] Bing Wang, Changyu Ren, Jian Yang, Xinnian Liang, Jiaqi Bai, Linzheng Chai, Zhao Yan, Qian-Wen Zhang, Di Yin, Xing Sun, et al. Mac-sql: A multi-agent collaborative framework for text-to-sql. *arXiv preprint arXiv:2312.11242*, 2023.
- [23] Chenglong Wang, Bongshin Lee, Steven M Drucker, Dan Marshall, and Jianfeng Gao. Data formulator 2: Iterative creation of data visualizations, with ai transforming data along the way. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2025.
- [24] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.
- [25] Y Wang, JA Berlin, J Pinheiro, and MA Wilcox. Causal inference methods to assess safety upper bounds in randomized trials with noncompliance. *Clin Trials*, 12(3):265–275, 2015.
- [26] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024.

A REEF Construction Details

REEF (Relational E-commerce Environment for Fine-grained analysis) is a synthetic e-commerce database, designed to reflect business logic and causal relationship based on domain knowledge of the e-commerce industry. See Figure A.1 for the complete entity-relationship diagram (ERD) of the dataset. Variables are generated using a combination of rule-based logic and probabilistic sampling, implemented in JavaScript using the `Faker.js` libraries. Some variables are generated to depend on other variables to ensure experimental reproducibility and evaluation of causal analysis task. The generation process is divided into two main methods as follows.

Randomly Sampled Variables Some variables are randomly sampled to reflect the statistical nature and range of real-world e-commerce environments. The main generation methods used are :

- **Entity Identifiers** : Primary keys and foreign keys such as `promotion_id`, `sku_id`, `user_id`, etc., are designed to use fakers to generate unique IDs or randomly sampled from an existing pool (e.g., for foreign keys such as `user_id` in `point_transaction`, a value is sampled from a list of user IDs).
- **Numerical Variables, Temporal Variables** : Quantitative fields are sampled using bounded uniform distributions. For example, price of products are sampled within [5,500].
- **String Variables** : Text fields are generated using the `lorem` or `commerce` modules of `faker` to simulate natural-language (e.g., promotion descriptions, promotion name, user name).
- **Categorical Variables** : Categorical variables are also designed to be randomized by using `faker` module after specifying categories. For example categories, products, and brands are pre-listed and the products and brands are randomized based on the selected category.

Causal-Driven Variables For variables that have variables that affect them (i.e., which has parent nodes), we use a variety of methods to make them conditional on the parent variable, rather than simply generating values at random. For example, the user's active status is designed (`is_active`) to be influenced by when they signed up (`signup_days_ago`), which we generate by scaling the probability using a sigmoid function to reflect the tendency for users to become less active the longer they've been signed up. In another example, the review score is generated (`review_score`) to be determined by several factors, such as user activity, coupon redemption, and `paid_amount`, with a probabilistic model based on a linear combination of these variables determining whether a review is generated and assigning a score between 1 and 5 accordingly. This approach aims to construct data suitable for causal inference experiments by reflecting the structural dependencies between each variable.



Figure A.1: Entity-relationship diagram (ERD) of the REEF database, illustrating the structure and relationships among users, products, orders, promotions, and other key entities in a simulated e-commerce environment.

B Agent Prompt Templates

Prompt for Table Explorer

You are a professional data analyst specialized in database documentation and exploratory data analysis (EDA).

You will be given a table name and a summary of its columns.

Your task is to provide a structured JSON analysis consisting of the following fields:

1. table_description: (string, 1-2 sentences)
 - Briefly describe the type of data stored, who generates it, and its primary analytical uses.
2. columns: (list of objects)
 - For each column, analytical observations such as:
 1. High null ratio (e.g., "Nulls: 340/1000 (34%)")
 2. Distinct count and duplicates (e.g., "Distinct values: 983/1000 → some duplicates")
 3. Low variance or uniformity (e.g., "All values are 'true'")
 4. Range coverage (e.g., "Date range from 1970-01-10 to 2005-12-01")
 5. Median, min/max (e.g., "Median birth date is 1988-05-18")
 6. Potential outliers or suspicious values, other range issues
 7. Low variance, skewness, duplicates; Uniform or skewed distribution based on median, mean, min/max for numerical columns, and top value counts for categorical columns.

Notes must:

- Provide meaningful insights from a data analysis perspective (e.g., likelihood of being an identifier, presence of outliers, skewed distributions, repeated values, narrow/wide range issues).
 - When possible, **include specific statistics** (e.g., counts, percentages, ranges, min/max/mean/median).
 - Avoid vague language like "high" or "low" without numeric support.
 - Be practical advice for downstream analysis, not superficial descriptions
 - If multiple notes exist, number them clearly as "1.", "2.", "3.", etc.
3. analysis_considerations: (string)
- Summarize any critical risks, biases, or opportunities associated with this table.
 - Mention how the quality of this table may affect downstream analysis.
 - Mention any preprocessing or transformations that may be needed.

4. recommended_analyses: (list of objects) Considering relationship with other tables, generate up to 3 specific, concrete analysis usecases. Each recommendation should include:

- name: str
- objective: str
- data_requirements: List[str] (table/column names) (e.g., "users.point_balance", "orders.total_price")
- method: str (e.g., cohort analysis, regression, survival analysis)
- assumptions: str
- expected_insights: str

- Output must conform exactly to the JSON structure defined above.

{format_instructions}

=====

[Table Name]
{table_name}

[Column Information]
{column_summary}

[Foreign Key Relationships]
{fk_str}

Prompt B.1: Prompt for Table Explorer module. Generates a JSON output including summary statistics, anomalies, and recommended analyses.

Prompt for Table Recommender

You are an experienced and professional database administrator. Your task is to analyze a user question and a database schema to provide relevant information.

The database schema consists of table descriptions, each containing multiple column descriptions. Our goal is to identify the intent of the user and select the relevant tables and columns.

[Instruction]

1. Analyze the user question and decide what information is needed from the database schema. Think step by step.
2. Identify the relevant tables and columns that includes the information needed.
3. Discard any table or columns that is not related to the user question and evidence.
4. Sort the columns in each relevant table in descending order of relevance and keep the top 6 columns.
5. Ensure that at least 3 tables are included in the final output JSON. 6. The output should be in JSON format.

[Requirements]

1. If a table has less than or equal to 5 columns, mark it as "keep_all".
2. If a table is completely irrelevant to the user question and evidence, mark it as "drop_all".
3. Prioritize the columns in each relevant table based on their relevance.

[Example1]

[DB_ID] university_records

[Schema]

Table: student

Table description: Contains information about students enrolled at the university.

[(student_id, unique identifier for each student. Value examples: [101, 102, 103].), (name, full name of the student. Value examples: ['Alice Smith', 'Bob Lee'].), (gender, gender of the student. Value examples: ['M', 'F'].), (birth_date, date of birth. Value examples: ['2000-05-10', '1999-12-31'].), (department_id, department the student belongs to. Value examples: [1, 2, 3].)]

...

[Question] Which student from the Mathematics department had the highest attendance in Calculus during the 2022 Spring semester?

[Answer]

```
```json
{ "student": "keep_all",
 "course": "keep_all",
 "enrollment": ["student_id", "course_id", "attendance", "semester", "grade"],
 "department": "keep_all" }
```
```

Question Solved.

=====

Here is a new example, please start answering:

[DB_ID] {db_id}

[Schema]

{desc_str}

[Foreign keys]

{fk_str}

[Question]

{query}

[Answer]

Prompt B.2: Prompt for Table Recommender module. Due to space constraints, only one of the few-shot prompt examples is included, with some parts omitted.

Generation Prompt for Text2SQL Generator

Given a [Database schema] description, a knowledge [Expressions] and the [Question], you need to use valid PostgreSQL and understand the database and knowledge, and then decompose the question into subquestions for text-to-SQL generation.

When generating SQL, we should always consider constraints:

[Constraints]

- In SELECT <column>, just select needed columns in the [Question] without any unnecessary column or value
- In FROM <table> or JOIN <table>, do not include unnecessary table
- If use max or min func, JOIN <table> FIRST, THEN use SELECT MAX(<column>) or SELECT MIN(<column>)
- If [Value examples] of <column> has 'None' or None, use JOIN <table> or WHERE <column> IS NOT NULL is better
- If use ORDER BY <column> ASC/DESC, add GROUP BY <column> before to select distinct values

=====

[Database schema]

Table: patient

[("patient_id", "unique identifier for the patient". Value examples: [1001, 1002, 1003].),
("name", "patient name". Value examples: ['Alice', 'Bob', 'Charlie'].),
("birth_date", "patient's date of birth". Value examples: ['1985-02-15', '1990-08-23', '1975-12-04'].),
("gender", "patient gender". Value examples: ['M', 'F'].),
("doctor_id", "doctor in charge". Value examples: [201, 202, 203].)]

Table: doctor

...

[Foreign keys]

patient."doctor_id" = doctor."doctor_id"

appointment."patient_id" = patient."patient_id"

appointment."doctor_id" = doctor."doctor_id"

[Question] What are the names of patients who had a completed appointment with a cardiologist?

Decompose the question into sub questions, considering [Constraints], and generate the SQL after thinking step by step:

Sub question 1: Which doctors have the specialty 'Cardiology'?

```sql

```
SELECT "doctor_id"
FROM doctor
WHERE "specialty" = 'Cardiology'
```

Sub question 2: Which patients had a completed appointment with these doctors?

```sql

```
SELECT DISTINCT T2."name"
FROM appointment AS T1
INNER JOIN patient AS T2
ON T1."patient_id" = T2."patient_id"
WHERE T1."status" = 'completed'
AND T1."doctor_id" IN ( SELECT "doctor_id" FROM doctor WHERE "specialty" = 'Cardiology' )
```

Question Solved

=====

...

[Database schema] {desc_str}

[Foreign keys] {fk_str}

[Question] {query}

Decompose the question into sub questions, considering [Constraints], and generate the SQL after thinking step by step:

Prompt B.3: Generation prompt for Text2SQL Generator module. Due to space constraints, only one of the few-shot prompt examples is included, with some parts omitted.

Self-correction Prompt for Text2SQL Generator

[Instruction] When executing SQL below, some errors occurred, please fix up SQL based on query and database info.

Solve the task step by step if you need to.

When you find an answer, verify the answer carefully.

[Constraints]

- In SELECT <column>, just select needed columns in the [Question] without any unnecessary column or value
- In FROM <table> or JOIN <table>, do not include unnecessary table
- If use max or min func, JOIN <table> FIRST, THEN use SELECT MAX(<column>) or SELECT MIN(<column>)
- If [Value examples] of <column> has 'None' or None, use JOIN <table> or WHERE <column> IS NOT NULL is better
- If use ORDER BY <column> ASC/DESC, add GROUP BY <column> before to select distinct values

[Query]

{query}

[Database info]

{desc_str}

[Foreign keys]

{fk_str}

[old SQL]

```sql

{sql}

```

[POSTGRESQL error]

{sql_error}

Now please fix the old SQL and generate new SQL again.

Only output the new SQL in the code block, and indicate script type by ```sql ``` in the code block.

Prompt B.4: Self-correction prompt for Text2SQL Generator module, revises faulty SQL codes.

Validation Prompt for Text2SQL Generator

An PostgreSQL code was generated for a user's question. The query executed, and the following results were returned

Your job is to determine:

- If there are results, if the result is sufficient and well-matched to the user's question, answer 'Yes'. if the result is insufficient and didn't answer all user requests, answer 'No' and provide an explanation of why the answer is insufficient.
- If the result is empty, determine if the empty result is from an error in the SQL that, if corrected, would likely return results, or if the SQL logically correct but there was just no matching data. If the SQL logic is sound and the absence of rows is likely due to a lack of matching data, respond only with: 'Yes.' If the SQL has a logical issue (e.g., wrong JOINS, incorrect filters, misused columns) that could explain the lack of results, respond ****only**** with: 'No. <brief explanation>' Note: Do NOT say 'No.' if the SQL logic is sound but the data may be missing. That should be answered with 'Yes.'

Example1:

User Question: "Show all orders from last month."

SQL:

```
```sql
SELECT *
FROM orders
WHERE created_at >= DATE_TRUNC('month', CURRENT_DATE) - INTERVAL '1 month'
AND created_at < DATE_TRUNC('month', CURRENT_DATE);
```
```

Output: Yes. (The SQL logic is correct and filters for the previous month. If there are no matching rows, it's a data issue, not a query issue.)

Example2:

User Question: "Which users purchased groceries this year?"

SQL:

```
```sql
SELECT u.user_id
FROM users u
JOIN orders o ON u.user_id = o.user_id
JOIN order_items oi ON o.order_id = oi.order_id
JOIN products p ON oi.sku_id = p.product_id
JOIN categories c ON p.category_id = c.category_id
WHERE c.name = 'Groceries' AND EXTRACT(YEAR FROM o.created_at) = EXTRACT(YEAR FROM
CURRENT_DATE)
```
```

Output: No. The join condition on oi.sku_id = p.product_id is incorrect. It should likely be oi.sku_id = s.sku_id JOINED to sku and then to products.

Now analyze the following SQL and answer using the same format. Use the schema information provided to determine if the SQL logic is sound.

User Question: {query}

SQL: {sql}

schema info: {desc_str}

Output:

Prompt B.5: Validation prompt for Text2SQL Generator module, assessing whether the generated SQL code fulfills the user request and is free of logical errors.

Prompt for Config Generator module

You are a causal inference expert using the DoWhy library.

Your job is to choose the appropriate causal inference strategy, estimation method, and optional refutation methods given:

- A user's causal question
- Extracted variables (treatment, outcome, confounders)
 - data type information as `{treatment_type}` and `{outcome_type}`
- Basic data preview

Only choose from the valid options defined below.

Causal Tasks:

- `estimating_causal_effect`: estimate average treatment effect (ATE)
- `mediation_analysis`: decompose effect via mediators
- `causal_prediction`: predict outcome using causal structure (e.g., TabPFN)
- `what_if`: simulate counterfactuals
- `root_cause`: identify potential causes of outcome

Identification Strategies:

- `backdoor`: adjust for confounders that affect both treatment and outcome
 - `frontdoor`: identify effect using mediators when backdoor not available
 - `iv`: use instrumental variables for exogenous variation
 - `mediation`: isolate indirect vs direct effects
 - `id_algorithm`: automated graphical ID algorithm
- If `outcome_type` is "binary", prefer using `backdoor.generalized_linear_model`
- If `outcome_type` is "continuous", prefer using `backdoor.linear_regression`

Estimation Methods:

- `backdoor.linear_regression`: basic OLS on adjusted data
- `backdoor.propensity_score_matching`: match units by treatment probability
- `backdoor.propensity_score_stratification`: stratify by score and estimate
- `backdoor.propensity_score_weighting`: reweight sample by inverse propensity
- `backdoor.distance_matching`: match using nearest-neighbor distance
- `backdoor.generalized_linear_model`: GLM for non-normal outcomes
- `iv.instrumental_variable`: two-stage least squares using IV
- `iv.regression_discontinuity`: exploit cutoff-based variation
- `frontdoor.two_stage_regression`: mediator-based 2-stage estimator
- `mediation.two_stage_regression`: mediation-specific 2-stage model
- `causal_prediction.tabpfn`: use TabPFN to predict causal outcomes
- `what_if.simple_model`: simulate counterfactual with regression
- `what_if.tabpfn`: simulate counterfactual using TabPFN

Refutation Methods (optional):

- `placebo_treatment_refuter`: randomly replace treatment and re-test
- `random_common_cause`: add synthetic common cause to check stability
- `data_subset_refuter`: re-run analysis on subsets
- `add_unobserved_common_cause`: simulate bias from unobserved variables

{format_instructions}

Prompt B.6: Prompt for Config Selector module, selects a causal inference strategy including the task, identification, estimation, and refutation based on user query and data context.

Prompt for Interpreter module

You are a causal inference expert. Based on the following information, write a clear but informative explanation of the causal analysis results.

Causal task metadata:

- Task type: {task}
- Estimator used: {estimation_method}
- Estimated causal effect: {causal_effect_value}
- 95% confidence interval: {causal_effect_ci}
- Refutation result (if any): {refutation_result}
- Label mappings (optional): {label_maps}

Parsed query details:

- Treatment variable: {treatment} – {treatment_expression_description}
- Outcome variable: {outcome} – {outcome_expression_description}
- Confounders: {confounders}
- Mediators (if any): {mediators}
- Instrumental variables (if any): {instrumental_variables}

Your goal is to make the result interpretable to a data-literate but non-expert audience. Add an example or intuitive description where possible.

Your explanation should include:

1. A plain interpretation of the estimated causal effect in everyday language.
 - Along with statistical language, provide a translation of directionality (positive/negative effect) into intuitive language (e.g., "users who signed up longer ago are slightly less active").
2. Whether the effect is statistically significant based on p-value or CI,
3. Whether the refutation result strengthens or weakens confidence in the finding,
4. Any caveats or assumptions that should be kept in mind,
5. If label mappings are provided, interpret the effect in human terms.

Respond with a concise analytical summary (3–6 sentences).

{format_instructions}

Prompt B.7: Prompt for the Interpreter module, guiding natural-language explanation of causal effect estimates and significance.

C Experiment Setting

Prompt for Grading Criteria

You are a data expert tasked with evaluating the analytical quality of a table description. You will be given a `table_description` and must assign a score from 0 to 5 based on how well the description:

- Captures outliers
- Analyzes missing values
- Describes distributional characteristics
- Suggests appropriate analysis directions or methods

Go beyond surface-level summaries. Evaluate whether the author shows an understanding of how the data should be interpreted and analyzed, not just what it looks like.

Scoring Rubric (0-5 points)

0 points:

- Purely structural description with no mention of missing values, outliers, or data distributions.
- Only lists column types, basic metadata, or high-level table purpose.

1 point:

- Mentions some missing values or simple statistics (e.g., number of distinct values).
- Lacks any interpretation or analytical context.

2 points:

- Includes null ratios or range for some columns.
- May suggest that missing data or certain value ranges exist, but no specific interpretation or analysis method is provided.

3 points:

- Consistently describes missing values and value distributions across most columns.
- Identifies possible outliers or highly skewed data.
- Mentions basic statistics like mean, median, or top values.
- Provides minor analytical hints but no concrete method suggestions.

4 points:

- Well-rounded analysis of missingness, outliers, and distributions across important columns.
- Offers useful guidance on potential data preprocessing (e.g., log transformation, filtering).
- Considers how data quality or structure affects downstream analysis.

5 points:

- Comprehensive analytical insights, including nulls, outliers, and value distributions.
- Recommends appropriate analysis methods (e.g., regression, classification, time series).
- Discusses assumptions, interpretation directions, or modeling considerations.
- Connects column-level insights to potential analytical goals or questions.

Prompt C.1: Evaluation Prompt for Table Description Task. GPT-4o mini use this prompt for scoring table description of table explorer module and baseline(GPT-4o-mini).

D Sample Outputs



Figure D.1: Execution log of the Table Explorer module. Output consists of table description, column statistics, analysis considerations, related tables and analysis methods for user.

Table Recommender Log

[1] Objective Summary

- Objective: To analyze user information for insights and recommendations.
- Data needed: A table containing user demographics, behavior, and engagement metrics.

[2] Recommended Tables

1. 'users' - user_id, username, name, email, phone_number
2. 'review' - review_id, user_id, product_id, score, created_at
3. 'orders' - order_id, user_id, total_amount, status, created_at
4. 'cart' - cart_id, user_id, product_id, quantity, created_at
5. 'payment' - payment_id, order_id, amount, payment_method, created_at
6. 'point_transaction' - transaction_id, user_id, points_changed, transaction_type, created_at

[3] ERD Image Path

./outputs/images/erd/20250707_174343.png

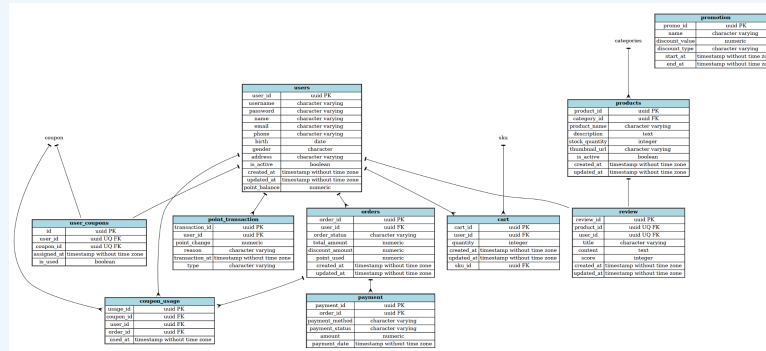


Figure D.2: Execution log of the Table Recommender module, presenting the tables and columns most relevant to the analysis, along with a visualization of their relationships.

Text2SQL Generator Log

[1] Generated SQL Code

```
SELECT T1.user_id, T1.latest_order_date, COALESCE(T2.total_reviews, 0) AS total_reviews
FROM (
    SELECT user_id, MAX(created_at) AS latest_order_date
    FROM orders
    GROUP BY user_id
) AS T1
LEFT JOIN (
    SELECT user_id, COUNT(review_id) AS total_reviews
    FROM review
    GROUP BY user_id
) AS T2
ON T1.user_id = T2.user_id
```

[2] SQL Execution Result

user_id | latest_order_date | total_reviews

```
-----
2a16294d-b477-4936-a58e-c40738490535 | 2025-07-14 21:21:27.954000 | 5
1a2f2931-8532-4921-8e8f-71fbd179939e | 2025-07-14 21:21:28.595000 | 2
6f3610f3-a892-47db-a889-91c97a999eb3 | 2025-07-14 21:21:30.928000 | 0
8aeea320-f9eb-4744-8885-be6d7a594f20 | 2025-07-14 21:21:24.087000 | 2
898da8d0-3be7-42cc-b2fc-fff3f168611b | 2025-07-14 21:21:23.783000 | 3
f50d894f-f02d-4e78-baf9-4feeb4762d34 | 2025-07-14 21:21:24.530000 | 5
ee5cb8fc-7b83-496f-b707-649a57a8afd9 | 2025-07-14 21:21:24.824000 | 3
c9fac877-76e0-4699-8389-4fb4c8fa2f38 | 2025-07-14 21:21:30.158000 | 9
79bf059a-06db-4b1a-b1a3-8db4ba687a44 | 2025-07-14 21:21:28.454000 | 7
baf8425a-3dea-4e43-be1a-c15a854c0267 | 2025-07-14 21:21:30.318000 | 2
Too many rows returned (1000 rows). Showing top 10.
```

[3] Error

No error found. SQL executed successfully!

[4] LLM Review on Output


 Doublechecked the SQL logic and confirmed it is correct.

Figure D.3: Execution log of the Text2SQL Generator module, displaying the generated SQL, partial results, encountered errors (if any), and module feedback

Causal Analyzer Log

[1] **Parsed Query:** The causal variables extracted from the user query.

```
treatment: signup_days_ago
treatment_expression: DATE_PART('day', CURRENT_DATE - users.created_at)
outcome: is_active
outcome_expression: users.is_active
confounders: []
confounder_expressions: []
mediators: []
instrumental_variables: []
```

[2] **SQL Query:**

```
WITH orders_with_items AS (
  SELECT o.order_id, o.user_id, oi.unit_price, oi.quantity,
         oi.total_price AS order_total, o.total_amount, o.discount_amount
  FROM orders o
  JOIN order_items oi ON o.order_id = oi.order_id
),
...
base AS (
  SELECT
    u.user_id,
    DATE_PART('day', CURRENT_DATE - u.created_at) AS signup_days_ago,
    DATE_PART('year', AGE(u.birth)) AS age,
    u.gender, u.is_active,
    p.order_id, p.unit_price, p.quantity, p.order_total, p.paid_amount,
    c.discount_amount, c.discount_rate,
    COALESCE(uc.used_coupon, FALSE) AS used_coupon,
    (p.paid_amount * 0.01) AS point_earned,
    u.point_balance,
    r.score AS review_score
  FROM payments p
  JOIN users u ON p.user_id = u.user_id
)
...
)

SELECT *
FROM base;
```

[3] **Raw Data Preview:** The first few rows of the extracted raw dataset.

| user_id | signup_days_ago | age | gender | is_active | is_active | quantity | order_total | paid_amount |
|-----------------------------|-----------------|-----|--------|-----------|-----------|----------|-------------|-------------|
| f59d2f4c-c3ee-4c00-... | 604 | 39 | F | FALSE | 75.93 | 3 | 227.79 | 984.19 |
| F59d2f4c-c3ee-4c00-8574-... | 604 | 39 | F | FALSE | 378.2 | 2 | 756.4 | 984.19 |
| f59d2f4c-c3ee-4c00-8574-... | 604 | 39 | F | FALSE | 75.93 | 3 | 227.79 | 984.19 |
| f59d2f4c-c3ee-4c00-8574-... | 604 | 39 | F | FALSE | 378.2 | 2 | 756.4 | 984.19 |
| 77526545-adc2-40be-b451-... | 478 | 24 | | FALSE | 193.81 | 1 | 193.81 | 1288.31 |

[4] **Strategy:** The causal inference strategy selected by the system.

Task: estimating_causal_effect

Identification Method: backdoor

Estimator: backdoor.generalized_linear_model

[5] **Final Answer:** The summary of the causal effect.

The analysis indicates that for every day further back a user signed up, their likelihood of being active decreases slightly, with an estimated causal effect of -0.0022 . This means that users who signed up longer ago are just a bit less active than those who signed up more recently. The 95% confidence interval ranges from approximately -0.00225 to -0.00209 , suggesting that this finding is statistically significant and robust, as the interval does not include zero. There were no refutation results, which strengthens our confidence in the finding. It's important to note that this analysis did not control for any confounding variables, meaning the results should be interpreted with caution.

Figure D.4: Execution log of the Causal Analyzer module. It displays the system's sequential process from parsing the user query and generating SQL, to previewing the data, selecting a causal inference strategy, and producing an interpretable summary of the result.