# Agent Bain vs. Agent McKinsey: A New Text-to-SQL Benchmark for the Business Domain

Anonymous ACL submission


## Abstract

Text-to-SQL benchmarks have traditionally only tested simple data access as a translation task of natural language to SQL queries. But in reality, users tend to ask diverse questions that require more complex responses including data-driven predictions or recommendations. Using the business domain as a motivating example, we introduce CORGI, a new benchmark that expands text-to-SQL to reflect practical database queries encountered by end users. CORGI is composed of synthetic databases inspired by enterprises such as DoorDash, Airbnb, and Lululemon. It provides questions across four increasingly complicated categories of business queries: descriptive, explanatory, predictive, and recommendational. This challenge calls for causal reasoning, temporal forecasting, and strategic recommendation, reflecting multi-level and multi-step agentic intelligence. We find that LLM performance degrades on higher-level questions as question complexity increases. CORGI also introduces and encourages the text-to-SQL community to consider new automatic methods for evaluating open-ended, qualitative responses in data access tasks. Our experiments show that LLMs exhibit an average 33.12% lower success execution rate (SER) on CORGI compared to existing benchmarks such as BIRD, highlighting the substantially higher complexity of real-world business needs. We release the CORGI dataset, an evaluation framework, and a submission website to support future research.

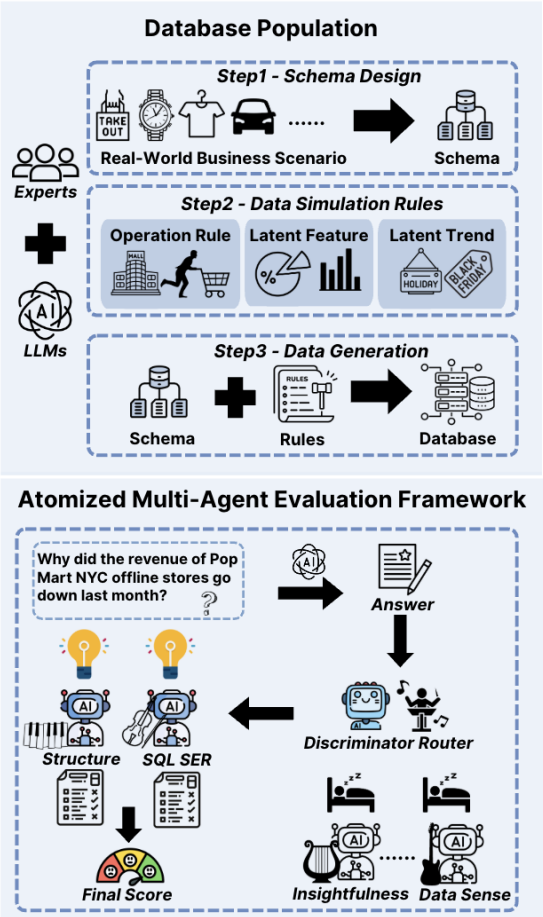




Figure 1: We incorporate business logic into the construction of CORGI and introduce a multi-agent evaluation framework with one discriminator and six scoring agents, each corresponding to a distinct scoring dimension for four types of business queries.

## 1 Introduction

Text-to-SQL is a long standing machine learning task of translating natural language questions (NLQ) into executable SQL queries (Zhong et al., 2017; Yu et al., 2018a). Most recently, LLMs have dominated state-of-the-art text-to-SQL approaches (Ma et al., 2025; Dönder et al., 2025; Yao et al., 2025) with commercial systems such as ChatGPT and Gemini making simple text-to-SQL functionality widely accessible. This calls into question the continued utility of text-to-SQL benchmarks, a task with a long tradition dating back to the 1970s, a much different era. Despite increasing complexity in joins and nested query structures, existing text-to-SQL benchmarks such as BIRD23-train-filtered (Li et al., 2023b), Wik-

iSQL (Zhong et al., 2017), and Spider2.0 (Team, 2023) primarily focus on simple past data retrieval – a translation task of mapping natural language input to SQL programming language.

In practice, end users require more than querying past records; they also seek trend analysis, forecasting, and strategic recommendations, which calls for a broader conception of text-to-SQL that goes beyond pure translation and incorporates identifying relevant data and building data-driven responses (Shmueli, 2010; Fildes et al., 2008; Armstrong, 2001; Davenport and Ronanki, 2018). Focusing on business domain, we propose a more expansive paradigm for text-to-SQL benchmarking in the LLM era, emphasizing diverse user query types, LLM-generated qualitative responses, and automatic evaluation.

Our main contributions are as follows:

First, we present **CORGI**[1], a business domain-specific text-to-SQL benchmark. CORGI provides simulated databases from verticals such as food delivery, rental, and advertising, as shown in Table 1, together with **four categories of business queries** including descriptive, explanatory, predictive, and recommendational. This captures real-world retrieval practices while enabling the evaluation of higher-level decision-making.

Second, we introduce a **business literature inspired evaluation framework** and an **atomized multi-agent evaluation mechanism**. Figure 1 illustrates the multi-agent evaluation mechanism. Based on the query type, the discriminator selects relevant scoring metrics and routes the input to specialized scoring agents, ensuring scalability and contextual precision. Table 3 indicates that the scores given by LLMs under our framework are consistent with those of human experts.

Third, through preliminary experiments, we show that leading open-source and closed-source LLMs exhibit **performance drops** on explanatory, predictive or recommendational tasks, highlighting open challenges for future research.

We release the CORGI dataset, code, evaluation questions, and evaluation framework on GitHub[2]. We further provide a development set including gold labels for descriptive queries and human-annotated evaluation results for higher-level reasoning questions to facilitate community engagement.

[1]In this paper, we use SQLite as the only SQL dialect.
[2]github.com/corgibenchmark/CORGI

## 2 Related Work

### 2.1 Toward Business Intelligence (BI): Expanding the Scope of Text-to-SQL

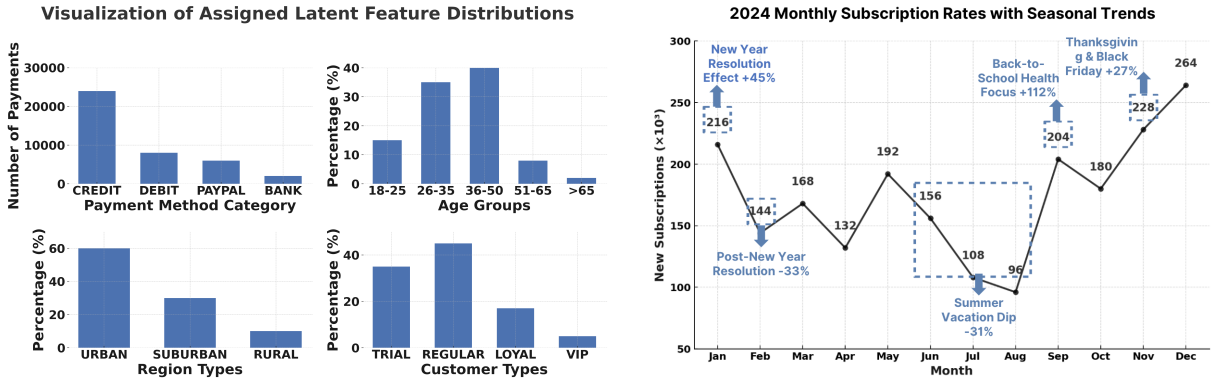
Earlier benchmarks such as WikiSQL (Zhong et al., 2017) and Spider (Yu et al., 2018b) predate the LLM era and primarily frame text-to-SQL as a translation task, often also called intent-to-SQL or NL-to-SQL. This has been primarily a simple task about accessing past data from databases. More recent benchmarks, including BIRD (Li et al., 2023b) and its extensions BIRD-CRITIC (Li et al., 2025), BIRD-Interact (Huo et al., 2025), and LiveSQL-Bench (BIRD Team, 2025) have been newly introduced, expanding evaluation to complex queries, SQL debugging, and agentic interaction, but still remain focused on text-to-SQL translation.

However, real-world users require more than past data access. In business scenarios, database queries often demand forecasting and data-informed advice on topics such as pricing strategies and capacity planning (Bughin et al., 2018; Gathani and at et al., 2022). In healthcare, clinicians query patient records to predict risk and support treatment decisions (Topol, 2019; Shortliffe and Sepúlveda, 2021).

### 2.2 Holistic Evaluation Framework

Traditional text-to-SQL evaluation emphasizes the correctness of generated SQL queries and their execution results, typically measured by metrics such as SER and execution accuracy (EA) (Wang et al., 2019; Xu et al., 2021). However, as text-to-SQL is extended to BI settings, evaluation becomes increasingly subjective, since outputs often involve open-ended qualitative explanations rather than exact matches. Metrics like EA fail to capture key interpretability aspects, such as identifying and grounding relevant data, which are essential for downstream reasoning quality and decision relevance (Bommasani et al., 2021; Zhou et al., 2023). Prior work in business contexts further highlights that effective evaluation must go beyond technical correctness to account for strategic alignment, creativity, and judgment under uncertainty (Perga, 2024; Jain, 2024; Malberg et al., 2025).

Assessing such qualitative dimensions traditionally relies on human judgment, which is costly, inconsistent, and difficult to scale. While prior studies report mixed evidence on LLMs as evaluators (Zheng et al., 2023), other work shows that LLM-based judgments can closely align with




(a) Latent feature distribution (partial)

(b) Latent seasonal trends

Figure 2: Illustrations of data simulation rules without compounding effects, using the *Persona Nutrition* database as a case study. Additional details are provided in Appendix B.

human scoring when guided by explicit rubrics and applied to open-ended, reasoning-intensive tasks (Liu et al., 2023). Comparable performance has been observed in domains such as law, where LLMs evaluate complex reasoning tasks at near-expert levels (Shui et al., 2023). These advantages in scalability and consistency have driven the growing adoption of LLM-based evaluation, including in settings such as LLM alignment, where LLMs are used to assess the outputs of other models (Anthropic, 2025).

## 3 Benchmark Construction

### 3.1 Database Population

We first construct database schemas and then populate them with simulated data. We design the entity–relationship diagrams and schemas from publicly available sources, technical documentation, and application interfaces to emulate real-world business operations.

We generate data using three types of rules [3]: (i) business operational constraints (e.g., the follow-up assessments conducted every six months), (ii) latent feature distributions (e.g., the customer type distribution illustrated in Figure 2a) , and (iii) seasonal trends and fluctuations (e.g., the 45% New Year resolution effect illustrated in Figure 2b). Appendix B provides detailed explanations and examples of these rules.

### 3.2 Question Generation

Queries can be broadly categorized into two classes: those concerning the *past* and those oriented toward the *future* (Shmueli, 2010; Fildes et al., 2008). Building on this distinction, we categorize queries into four types, as illustrated in Figure 3. Overall, type 1 questions correspond to the classic data retrieval task, while the remaining types extend beyond pure translation and require data retrieval to support higher-level reasoning.

**Type 1: Past, Descriptive.** Type 1 questions focus on direct factual retrieval by translating NLQ into SQL queries (Zhong et al., 2017).

**Type 2: Past, Explanatory.** Type 2 queries require explaining historical patterns rather than merely retrieving data, necessitating appropriate SQL queries to identify relevant data as well as reasoning over underlying patterns (Shmueli, 2010).

Consider the question: "Why did the revenue of Pop Mart's offline stores in New York City decline over the past 90 days?" Answering it requires analyzing *retrieved sales and inventory data* to identify potential factors, such as supply chain disruptions.

**Type 3: Future, Predictive.** Type 3 questions involve forecasting future outcomes from historical data (Armstrong, 2001; Fildes et al., 2008), requiring trend identification and predictive model selection (e.g., time-series models (Faw et al., 2025)) based on SQL-retrieved data.

For instance, the question "What are the expected Labubu sales for the North America online store in November 2026?" requires forecasting demand by combining *historical sales trends* with upcoming promotions such as Black Friday.

**Type 4: Future, Recommendational.** Type 4 queries require LLMs to generate actionable recommendations by identifying relevant historical records, evaluating alternative strategies, and

[3] We apply the three rule types independently as a simplifying assumption.

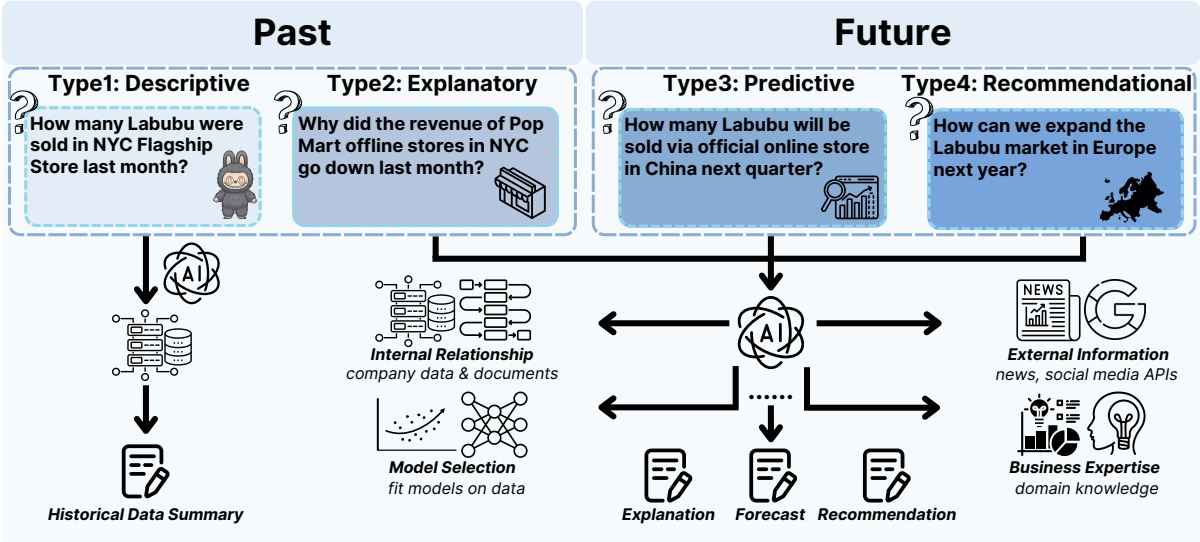



Figure 3: Queries range in complexity from fact retrieval to speculative business strategy.

proposing interventions aligned with business objectives (Davenport and Ronanki, 2018; Provost and Fawcett, 2013).

Consider the query: "How can the Labubu market in Europe be expanded next year?" A plausible answer can be derived by analyzing *past offline revenue and conversion data* to suggest partnerships with high-performing retailers.

### 3.3 CORGI Statistics

CORGI covers 10 real-world industry verticals,[4] spanning three broad categories: (i) consumer platforms (e.g., DoorDash), (ii) retail and commerce (e.g., Lululemon), and (iii) digital services (e.g., Apple App Store), as summarized in Table 1. Appendix A provides the detailed descriptions.

Table 2 compares the overall and question-type-level structural complexity of CORGI with existing benchmarks, highlighting the substantially greater structural demands and answer-generation challenges posed by CORGI. Compared with BIRD, CORGI exhibits a markedly higher number of JOINs per question (7.48 vs. 0.93), more queries per question (2.31 vs. 1.0), and significantly longer answers (1002 vs. 161 characters). In addition, while the average SQL SER on the BIRD dev set reaches 88.0%, the average SQL SER for type 2–4 questions in CORGI drops to 54.88%, indicating increased difficulty in generating executable queries. Within CORGI itself, type 1 questions are markedly simpler than higher-level question types. Specifically, the generated SQL queries for type 2–4 questions are substantially longer on average (1159.33 vs. 255 characters) and involve more JOIN operations (3.80 vs. 0.55), further underscoring the challenges introduced by advanced BI-oriented tasks. Appendix I provides additional statistics of CORGI.

## 4 Evaluation Framework

### 4.1 Motivation

We introduce the concept of automatic qualitative evaluation for text-to-SQL tasks, using LLMs as evaluators. Although the traditional formulation of text-to-SQL often assumes that gold labels are unambiguous, even seemingly simple fact-retrieval queries can admit multiple valid interpretations. For example, a query such as "how many students do I have?" may refer to (i) the number of students currently enrolled in a school, (ii) the total number of students previously taught, or (iii) the number of students in a specific class. Similarly, the query "what were my sales this month?" could correspond to (i) total revenue, (ii) total items sold, (iii) total profit (revenue minus cost), or (iv) profit after tax. Prior text-to-SQL benchmarks, such as BIRD, explicitly provide additional context or constraints to resolve such ambiguities and enforce a single intended interpretation (Li et al., 2023b). In contrast, our expanded text-to-SQL benchmark produces qualitative, open-ended responses that cannot be adequately evaluated using gold labels alone. To

[4]We consider only 10 industry domains in this paper, with additional ones left for future work.

[5]We select BI questions in LiveSQLBench-Base-Lite and generate answers using GPT-5.

| Business Scenario | Representative Platform | Business Model | Tables | #C/T | #R |
|---|---|---|---|---|---|
| Food delivery | DoorDash | B2C/B2B2C | 30 | 4.0 | 25 |
| Home rental | Airbnb | C2C | 18 | 4.3 | 25 |
| Clothing e-commerce | Lululemon | B2C | 34 | 3.9 | 40 |
| Freelancing marketplace | Upwork | C2C/C2B | 25 | 3.8 | 33 |
| App store | Apple App Store | B2C/B2B2C | 18 | 4.3 | 22 |
| Car rental | Turo | C2C | 30 | 6.7 | 68 |
| Luxury consignment | The RealReal | C2C | 30 | 6.5 | 65 |
| User analytics | Amplitude | B2B | 20 | 7.5 | 25 |
| Personalized product | Persona Nutrition | B2C | 28 | 7.8 | 38 |
| E-commerce enablement | Shopify | B2B | 23 | 7.0 | 60 |

Table 1: Overview of the ten industry verticals in CORGI. #C/T and #R denote the average number of columns per table and the number of foreign key relationships.

| Type | #Q | J/Q | Q/Qs | J/Qr | C/Qr | G/Qr | S/Qr | Len |
|---|---|---|---|---|---|---|---|---|
| **CORGI** | | | | | | | | |
| CORGI-Type1 | 300 | 0.55 | 1.00 | 0.55 | 1.78 | 0.16 | 0.05 | 255 |
| CORGI-Type2 | 300 | 11.07 | 2.83 | 3.99 | 1.66 | 1.54 | 0.71 | 1152 |
| CORGI-Type3 | 300 | 8.96 | 3.37 | 2.65 | 1.52 | 1.78 | 0.28 | 1036 |
| CORGI-Type4 | 300 | 16.26 | 3.34 | 4.76 | 1.07 | 2.25 | 0.50 | 1290 |
| **CORGI-Overall** | 1200 | 7.48 | 2.31 | 3.24 | 1.47 | 1.58 | 0.41 | 1002 |
| **Other Benchmarks** | | | | | | | | |
| BIRD-Simple | – | 0.75 | 1.00 | 0.75 | 1.27 | 0.05 | 0.00 | 129 |
| BIRD-Moderate | – | 1.16 | 1.00 | 1.16 | 1.73 | 0.12 | 0.01 | 193 |
| BIRD-Challenging | – | 1.35 | 1.00 | 1.35 | 1.75 | 0.13 | 0.07 | 271 |
| BIRD-Overall (Li et al., 2023b) | 12751 | 0.93 | 1.00 | 0.93 | 1.46 | 0.08 | 0.01 | 161 |
| Spider (Yu et al., 2018a) | 1034 | 0.50 | 1.00 | 0.50 | 0.57 | 0.27 | 0.00 | 107 |
| WikiSQL (Zhong et al., 2017) | 80654 | 0.00 | 1.00 | 0.00 | 1.38 | 0.00 | 0.00 | 70 |
| Advising (Finegan-Dollak et al., 2018) | 4570 | 0.79 | 1.00 | 0.79 | 4.15 | 0.06 | 0.00 | 488 |
| LiveSQLBench-Base-Lite[5] (BIRD Team, 2025) | 270 | 1.17 | 1.00 | 1.17 | 0.43 | 0.57 | 0.10 | 527 |

Table 2: Structural complexity comparison between CORGI and the development sets of existing benchmarks. #Q denotes the number of questions; J/Q denotes JOINs per question; Q/Qs denotes queries per question; J/Qr denotes JOINs per query; C/Qr denotes conditions per query; G/Qr denotes GROUP BY clauses per query; S/Qr denotes subqueries per query; and Len denotes query length in characters.

address this challenge, we propose evaluating BI-related qualitative text-to-SQL outputs with LLMs that approximate human judgment, drawing inspiration from how responses are assessed in real-world management consulting settings.

### 4.2 Strategic Management Consulting–Inspired Evaluation Framework

SQL-based data access is central to business management consulting (Brynjolfsson et al., 2011; Bughin et al., 2018). To evaluate LLMs in text-to-SQL-centric business decision-making contexts, we integrate this criteria for assessing professional strategic consulting with recent work on LLM-based decision-making agents (Greiner and Poulfelt, 2005; Christensen, 2013; Park et al., 2023; Yao et al., 2023). This framework is collaboratively developed with business school faculty, professional consultants, and corporate practitioners.

We organize the framework into **six** main dimensions in **four** aspects: **(i) Communication quality (how you say it):** (1) Structure (Minto, 2009), **(ii) Analytical reasoning (how you think):** (2) Data Sense (Chambers et al., 1971; Binder and Watkins, 2024), (3) Insightfulness (Binder and Watkins, 2024; Backler et al., 2025), **(iii) Strategic decision quality (how you decide):** (4) Operational Implementability (Chambers et al., 1971; Garner, 2013), (5) Purpose Alignment (Porter and Kramer, 2011), and **(iv) Governance&risk control (how you govern):** (6) Compliance (Cowgill et al., 2020; Guo et al., 2025). Appendix D elaborates each dimension via sub-metrics.

Among the six dimensions, some are universal, while others are question-type specific. For example, *Structure* applies to all questions, requiring clear organization and coherent argumentation. In contrast, *Compliance* is specific to type 4 questions, where risk management, regulatory, and ethical

considerations are essential for recommendations.

We apply the proposed evaluation framework to higher-level type 2–4 questions. We use SQL SER to evaluate performance across all question types, and additionally report EA for type 1 questions. Table 4 reports the values of SER and EA.

### 4.3 Atomized Multi-Agent Evaluation Mechanism

Consistent with the evaluation framework, our multi-agent evaluation mechanism consists of one discriminator and six scoring agents, as illustrated in Figure 1. The discriminator look-up serves as a routing component for evaluation. By allowing agents to specialize in distinct evaluation dimensions, this architecture mirrors human committee-based assessment, where diverse perspectives improve reliability and validity (Chan et al., 2023; Rasheed et al., 2024).

The evaluation proceeds as follows:

**(1) Input**: Each input consists of the question type, question, and corresponding answer. **(2) Discriminator Routing**: The discriminator identifies relevant scoring dimensions based on the question type and content. **(3) Scoring**: General and specialized scoring agents assign scores along relevant dimensions. **(4) Aggregation**: The final score is the average of the dimension scores.

Table 5 indicates that LLMs do not exhibit a clear self-preference bias, i.e., a tendency to assign systematically higher scores to outputs generated by the same model. For instance, on type 2 questions, GPT-5 assigns an average score of 3.87 to its own outputs, while Gemini-2.5-Pro and Llama 4 assign 4.16 when evaluating GPT-5's responses.

### 4.4 Automatic and Human Evaluation Consistency

In addition to automatic LLM-based evaluation, we conduct a human expert assessment on a randomly selected subset of the benchmark by sampling 3 of the 10 databases (AppStore, Clothing, and Shopify) and manually scoring the corresponding LLM-generated answers using the same evaluation framework.

Table 3 reports the consistency between automatic evaluation and human annotations using Kendall's $\tau$[5]. Overall, the agreement is moderate and decreases with question complexity, from 0.52–0.71 (type 2) to 0.43–0.64 (type 3) and 0.31–0.57 (type 4), reflecting the increased difficulty of evaluating qualitative, judgment-intensive responses.

[5]An introduction to Kendall's $\tau$ is provided in Appendix H.

**Algorithm 1:** Query Generation and Answer Synthesis for a Single NLQ

**Initialize** $S \leftarrow [\,]$, $R \leftarrow [\,]$
$S \leftarrow h(Q, D, K, P \mid \theta)$
**for** $s \in S$ **do**
  $r \leftarrow e(s)$
  **if** $r$ *is invalid* **then**
    $r \leftarrow \emptyset$
  Append $r$ to $R$
$A \leftarrow f(Q, R, K \mid \theta)$
**return** $A, S, R$

## 5 Methodology

### 5.1 Problem Formulation

Given an NLQ $Q$, a database schema $D$, and a knowledge set $K = \{K_I, K_E\}$, where $K_I$ denotes internal knowledge (e.g., database descriptions) and $K_E$ denotes external knowledge (e.g., information retrieved via web search APIs), an LLM parameterized by $\theta$ consumes prompts $P$ to generate a set of SQL queries $S$, their execution results $R$, and a final answer $A$. Algorithm 1 summarizes the procedure, and we illustrate the process below.

**Query generation.** $h(\cdot \mid \theta)$ maps $(Q, D, K, P)$ to a list of SQL queries $S$. For type 1 questions, the query generator produces a single SQL query. For the others, it generates up to five queries to support complex reasoning:

$$S = h(Q, D, K, P \mid \theta).$$

**Query execution.** $e(\cdot)$ runs the generated queries on the database defined by $D$ to obtain execution results $R$:

$$R = e(S).$$

**Answer composition.** $f(\cdot \mid \theta)$ produces the final answer from the question $Q$, execution results $R$, and knowledge set $K$:

$$A = f(Q, R, K \mid \theta).$$

**Output.** The system returns the answer $A$ with index-aligned $(S, R)$ pairs, discarding erroneous queries or replacing them with empty results.

| Type | Generator\Evaluator | Gemini-2.5-Pro | GPT5 | Llama4 |
|---|---|---|---|---|
| Type 2 | Gemini-2.5-flash-lite | 0.62 | 0.66 | 0.57 |
| | Gemini-2.5-Pro | 0.67 | 0.68 | 0.58 |
| | GPT5 | 0.71 | 0.69 | 0.62 |
| | Llama4 | 0.53 | 0.61 | 0.63 |
| | Qwen3 | 0.57 | 0.65 | 0.52 |
| Type 3 | Gemini-2.5-flash-lite | 0.58 | 0.52 | 0.48 |
| | Gemini-2.5-Pro | 0.63 | 0.59 | 0.54 |
| | GPT5 | 0.51 | 0.64 | 0.46 |
| | Llama4 | 0.54 | 0.47 | 0.55 |
| | Qwen3 | 0.56 | 0.43 | 0.41 |
| Type 4 | Gemini-2.5-flash-lite | 0.45 | 0.43 | 0.35 |
| | Gemini-2.5-Pro | 0.41 | 0.43 | 0.37 |
| | GPT5 | 0.46 | 0.41 | 0.31 |
| | Llama4 | 0.57 | 0.38 | 0.43 |
| | Qwen3 | 0.38 | 0.36 | 0.32 |

Table 3: Kendall's tau consistency scores between generators (rows) and evaluators (columns) on CORGI.

### 5.2 Scoring Strategy

We adopts the following two-stage scoring strategy:

**Step 1: Dimension scoring.** Each answer is evaluated along the relevant dimensions, using a discrete 0–3–5 scale where 5 denotes excellent performance meeting professional consulting standards and 0 indicates critical failure.

**Step 2: Score aggregation.** The final score is the average over all applicable dimensions:

$$\text{Final Score} = \frac{1}{n}\sum_{i=1}^{n} \text{Dimension Score}_i,$$

where $n$ is the number of relevant dimensions. Alternative aggregation strategies are discussed in Appendix G.

## 6 Experiments

### 6.1 Experimental Setup

For the benchmark experiments, we evaluate a mix of popular open-source and closed-source LLMs. For answer generation, we use *Gemini-2.5-flash-lite* (DeepMind, 2025), *Gemini-2.5-Pro* (Gemini Team, Google DeepMind, 2025), *GPT-5* (OpenAI, 2025), *Llama 4* (model ID: `meta-llama/Llama-4-Scout-17B-16E-Instruct`) (Meta AI, 2025), and *Qwen 3* (model ID: `Qwen/Qwen3-235B-A22B-Instruct-2507`) (Yang et al., 2025). For automatic evaluation, we use *Gemini-2.5-Pro*, *GPT-5*, and *Llama 4*. We set the *temperature* to 0 for all models to reduce sampling variance and improve reproducibility (Li et al., 2023a). For LLM-based automatic evaluation, we batch inputs by question type. Appendix C presents prompts used for answer generation and evaluation.

We randomly sample 3 out of the 10 databases (AppStore, Clothing, and Shopify). We provide gold labels for type 1 queries and conduct LLM-based automatic evaluation as well as human expert annotation on type 2–4 questions using the framework introduced in Section 4.

### 6.2 Experimental Results and Discussion

For type 1 questions, we evaluate models using SQL SER and EA. For type 2–4 questions, we evaluate models using SQL SER together with the strategic management consulting–inspired evaluation framework. Tables 4 and 5 present the experimental results.

**Finding 1: Closed-source LLMs perform well on type 1 questions.** Closed-source models demonstrate consistently stronger performance on type 1 questions, achieving EA scores of 88.89% (GPT-5), 81.48% (Gemini-2.5-Pro), and 80.62% (Gemini-2.5-Flash-Lite), alongside SER scores of 92.59% (GPT-5), 88.89% (Gemini-2.5-Pro), and 92.59% (Gemini-2.5-Flash-Lite). In contrast, open-source models exhibit noticeably lower performance, with EA scores of 66.67% (Llama 4) and 29.63% (Qwen 3), and SER scores of 85.19% (Llama 4) and 40.74% (Qwen 3). Overall, these results suggest that state-of-the-art closed-source LLMs are better equipped to generate correct and executable SQL queries for simple factual retrieval tasks.

**Finding 2: LLMs exhibit a monotonic decrease in average SER from type 1 to type 4 questions.** Table 4 reports SER values across all question types

| Type | Metric | G5 | G2.5P | G2.5F | L4 | Q3 | Overall |
|---|---|---|---|---|---|---|---|
| Type 1 | EA | 88.89% | 81.48% | 80.62% | 66.67% | 29.63% | 69.46% |
| Type 1 | SER | 92.59% | 88.89% | 92.59% | 85.19% | 40.74% | 79.47% |
| Type 2 | SER | 83.33% | 71.11% | 65.30% | 27.66% | 48.39% | 59.16% |
| Type 3 | SER | 63.79% | 53.49% | 67.91% | 56.52% | 48.98% | 58.14% |
| Type 4 | SER | 58.62% | 50.00% | 70.18% | 34.78% | 23.08% | 47.33% |

Table 4: EA is reported for type 1 questions, while SER is reported across all question types and models. G5 denotes GPT-5; G2.5P denotes Gemini-2.5-Pro; G2.5F denotes Gemini-2.5-Flash-Lite; L4 denotes Llama 4; and Q3 denotes Qwen 3. *Overall* reports the average values across all models.

| Type | Generator \ Evaluator | S | | | DS | | | I | | | OI | | | PA | | | C | | | Avg. | | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | G2.5P | G5 | L4 | G2.5P | G5 | L4 | G2.5P | G5 | L4 | G2.5P | G5 | L4 | G2.5P | G5 | L4 | G2.5P | G5 | L4 | G2.5P | G5 | L4 | |
| **Type 2** | | | | | | | | | | | | | | | | | | | | | | | |
| | Gemini-2.5-flash-lite | 3.85 | 3.80 | 3.55 | 4.05 | 3.97 | 4.10 | 2.95 | 3.39 | 3.45 | — | — | — | — | — | — | — | — | — | 3.65 | 2.95 | 3.70 | 3.43 |
| | Gemini-2.5-Pro | 4.02 | 3.31 | 3.82 | 4.23 | 3.23 | 4.77 | 2.97 | 3.17 | 4.20 | — | — | — | — | — | — | — | — | — | 3.74 | 3.24 | 4.26 | 3.75 |
| | GPT-5 | 4.26 | 3.36 | 3.75 | 4.80 | 4.33 | 4.60 | 3.43 | 3.90 | 4.13 | — | — | — | — | — | — | — | — | — | 4.17 | 3.86 | 4.16 | 4.06 |
| | Llama 4 | 1.71 | 1.53 | 3.58 | 2.20 | 1.57 | 3.77 | 1.63 | 1.30 | 2.80 | — | — | — | — | — | — | — | — | — | 1.85 | 1.47 | 3.38 | 2.23 |
| | Qwen 3 | 1.96 | 1.60 | 2.44 | 1.50 | 0.87 | 1.87 | 0.67 | 0.60 | 1.20 | — | — | — | — | — | — | — | — | — | 1.37 | 1.02 | 1.84 | 1.41 |
| | *Closed-source Overall* | | 3.75 | | | 4.23 | | | 3.51 | | | — | | | — | | | — | | | 3.75 | | **3.75** |
| | *Open-source Overall* | | 2.14 | | | 1.96 | | | 1.37 | | | — | | | — | | | — | | | 1.82 | | **1.82** |
| **Type 3** | | | | | | | | | | | | | | | | | | | | | | | |
| | Gemini-2.5-flash-lite | 3.55 | 3.10 | 3.60 | 3.95 | 3.10 | 4.10 | 2.65 | 2.30 | 3.10 | — | — | — | — | — | — | — | — | — | 3.38 | 2.83 | 3.60 | 3.27 |
| | Gemini-2.5-Pro | 4.31 | 3.69 | 4.04 | 4.89 | 3.47 | 4.57 | 3.07 | 2.83 | 4.17 | — | — | — | — | — | — | — | — | — | 4.09 | 3.33 | 4.26 | 3.89 |
| | GPT-5 | 4.11 | 3.80 | 4.02 | 5.00 | 4.64 | 4.78 | 3.37 | 3.67 | 3.93 | — | — | — | — | — | — | — | — | — | 4.16 | 4.04 | 4.24 | 4.15 |
| | Llama 4 | 2.58 | 2.31 | 3.95 | 3.23 | 2.24 | 4.16 | 2.13 | 1.87 | 3.50 | — | — | — | — | — | — | — | — | — | 2.65 | 2.14 | 3.87 | 2.89 |
| | Qwen 3 | 2.29 | 2.11 | 3.09 | 1.84 | 1.69 | 2.62 | 0.93 | 1.33 | 2.13 | — | — | — | — | — | — | — | — | — | 1.69 | 1.71 | 2.61 | 2.00 |
| | *Closed-source Overall* | | 3.80 | | | 4.28 | | | 3.23 | | | — | | | — | | | — | | | 3.77 | | **3.77** |
| | *Open-source Overall* | | 2.72 | | | 2.63 | | | 1.98 | | | — | | | — | | | — | | | 2.45 | | **2.45** |
| **Type 4** | | | | | | | | | | | | | | | | | | | | | | | |
| | Gemini-2.5-flash-lite | 3.20 | 2.75 | 3.30 | 4.10 | 3.20 | 4.30 | 2.70 | 2.30 | 3.10 | 3.30 | 2.80 | 3.50 | 3.55 | 2.90 | 3.70 | 1.90 | 1.60 | 2.10 | 3.13 | 2.76 | 3.33 | 3.12 |
| | Gemini-2.5-Pro | 3.95 | 3.04 | 3.77 | 5.00 | 3.49 | 5.00 | 3.60 | 3.00 | 4.87 | 4.03 | 3.60 | 4.53 | 4.60 | 3.27 | 4.70 | 3.27 | 2.44 | 4.73 | 4.08 | 3.14 | 4.60 | 3.94 |
| | GPT-5 | 4.33 | 3.49 | 3.44 | 5.00 | 4.80 | 5.00 | 3.53 | 3.00 | 4.27 | 4.57 | 4.33 | 4.93 | 4.67 | 3.70 | 4.67 | 3.38 | 2.73 | 4.69 | 4.25 | 3.68 | 4.50 | 4.14 |
| | Llama 4 | 2.76 | 2.65 | 4.29 | 4.17 | 2.97 | 4.93 | 0.01 | 1.20 | 3.00 | 2.07 | 1.80 | 2.97 | 4.20 | 3.10 | 4.73 | 1.40 | 1.33 | 3.44 | 2.43 | 2.17 | 3.89 | 2.83 |
| | Qwen 3 | 1.69 | 1.78 | 2.35 | 2.49 | 2.10 | 2.67 | 0.01 | 1.33 | 1.60 | 1.37 | 1.03 | 2.27 | 2.33 | 1.70 | 2.60 | 1.82 | 1.42 | 2.91 | 1.62 | 1.56 | 2.40 | 1.86 |
| | *Closed-source Overall* | | 3.47 | | | 4.43 | | | 3.37 | | | 3.95 | | | 3.97 | | | 2.98 | | | 3.72 | | **3.72** |
| | *Open-source Overall* | | 2.59 | | | 3.22 | | | 1.19 | | | 1.92 | | | 3.11 | | | 2.05 | | | 2.35 | | **2.35** |

Table 5: Evaluation scores for type 2–4 questions under the proposed framework. S denotes Structure; DS denotes Data Sense; I denotes Insightfulness; OI denotes Operational Implementability; PA denotes Purpose Alignment; C denotes Compliance; G2.5P denotes Gemini-2.5-Pro; G5 denotes GPT-5; and L4 denotes Llama 4. Dashes indicate metrics not applicable to a given question type. *Overall* reports the average values across all models.

and models. For type 2–4 questions, we allow up to five generated queries to support complex reasoning, whereas type 1 questions use a single SQL query. We observe an overall decline in average SER as task difficulty increases, decreasing from 79.47% for type 1 to 59.16% for type 2, 58.14% for type 3, and 47.33% for type 4. This monotonic downward trend culminates in a pronounced drop for type 4 questions, highlighting the greater challenge of generating multiple executable SQL queries for higher-level reasoning tasks. By our evaluation framework, we observe no similar trend.

**Finding 3: Open-source LLMs particularly struggle with type 2 to type 4 questions.** Table 5 shows that open-source LLMs consistently struggle with type 2–4 questions. Llama 4 and Qwen 3 achieve substantially lower performance than closed-source LLMs across type 2–4 questions. In particular, the average scores for open-source models reach 1.82 on type 2 questions, 2.45 on type 3 questions, and 2.35 on type 4 questions, indicating the substantially greater difficulty of generating SQL queries for open-ended BI tasks.

## 7 Conclusion

We introduce CORGI, a new text-to-SQL benchmark featuring 10 hand-curated industry-inspired databases, a new business literature-informed evaluation framework, and two new evaluation mechanisms. We show that popular LLMs struggle with increasingly difficult business insight questions such as those requiring recommendations. While LLMs are good at generating structured, MBA-trained responses, they fail to give sound implementable action plans. We also show that on average open-source models trail behind close-source models in SQL generation. We release CORGI as a public benchmark for researchers to contribute to and help to further our understanding of LLMs as management consultants.

## Limitations

While CORGI advances schema realism and task relevance for business applications, several limi-

tations remain. First, all database contents and SQL queries are synthesized rather than drawn from production systems; this ensures privacy and scalability but omits natural artifacts such as noise, redundancy, and schema drift. Second, CORGI focuses on single-turn queries, leaving out multi-turn conversational refinement that is common in real-world systems. Third, each query is paired with only one natural language formulation, limiting the evaluation of model robustness to paraphrasing or informal inputs. Finally, this study covers only 10 business verticals. These limitations highlight promising directions for extending CORGI in future work.

## Ethical Considerations

Ethical considerations play an important role in our research. First, our evaluation framework incorporates an *Ethical Responsibility* metric to assess whether generated analyses and recommendations adhere to established principles of AI ethics. Second, we publicly release the dataset, evaluation questions, and evaluation framework on GitHub, enabling other researchers to replicate, scrutinize, and extend our work, thereby promoting transparency and collective progress in the text-to-SQL community. With respect to data privacy, all data used to construct the dataset are sourced exclusively from publicly available materials and do not contain confidential or sensitive information. Through these measures, we aim to introduce a new SQL evaluation benchmark while ensuring its responsible, transparent, and equitable use.

## Acknowledgments

We thank the anonymous reviewers and area chairs for their valuable feedback.

## References

Anthropic. 2025. Bloom: An open-source framework for automated behavioural evaluations of frontier ai models. https://www.anthropic.com/research/bloom.

J. Scott Armstrong. 2001. *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Springer.

W. Backler, A. Iny, N. D'Intino, E. Parker, and S. Hirashita. 2025. Navigating the future with strategic foresight.

Jonathan Binder and Michael D. Watkins. 2024. To solve a tough problem, reframe it. *Harvard Business Review*, 102(1-2):80–89.

BIRD Team. 2025. Livesqlbench: A dynamic and contamination-free benchmark for evaluating llms on real-world text-to-sql tasks. https://livesqlbench.ai. Benchmark and dataset release. No official paper available at the time of writing.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, and et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Erik Brynjolfsson, Lorin M. Hitt, and Heekyung Hellen Kim. 2011. Strengthening the links between decision-making and performance through data-driven decision-making. *MIS Quarterly*, 35(2):243–260.

Jacques Bughin, Jeongmin Seong, James Manyika, Michael Chui, and Raoul Joshi. 2018. Notes from the ai frontier: Applications and value of deep learning. *McKinsey Global Institute*.

J. C. Chambers, S. K. Mullick, and D. D. Smith. 1971. How to choose the right forecasting technique. *Harvard Business Review*.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *Preprint*, arXiv:2308.07201.

Laurence J. Christensen. 2013. *The Business of Consulting*. Wiley.

Bo Cowgill, Francesca Dell'Acqua, Senyu Deng, Daniel Hsu, Nitin Verma, and Augustin Chaintreau. 2020. Biased programmers? or biased data? a field experiment in operationalizing ai ethics. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 679–681.

Thomas H. Davenport and Rajeev Ronanki. 2018. *Artificial Intelligence for the Real World*. Harvard Business Publishing.

Google DeepMind. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Yusuf Denizay Dönder, Derek Hommel, Andrea W. Wen-Yi, David Mimno, and Eun Seo Jo. 2025. Cheaper, better, faster, stronger: Robust text-to-sql without chain-of-thought or fine-tuning. *arXiv preprint*, arXiv:2505.14174.

Matthew Faw, Rajat Sen, Yichen Zhou, and Abhimanyu Das. 2025. In-context fine-tuning for time-series foundation models. In *Forty-second International Conference on Machine Learning*.

Robert Fildes, Paul Goodwin, Michael Lawrence, and Konstantinos Nikolopoulos. 2008. Forecasting and operational research: A review. *Journal of the Operational Research Society*, 59(9):1150–1172.

Catherine Finegan-Dollak, Jonathan K. Kummerfeld, Li Zhang, Karthik Ramanathan, Sesh Sadasivam, Rui Zhang, and Dragomir R. Radev. 2018. Improving text-to-sql evaluation methodology. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 351–360.

Bryan A. Garner. 2013. *HBR Guide to Better Business Writing*. Harvard Business Review Press.

Mehul Gathani and at et al. 2022. What-if analysis for business users: Current practices and future opportunities. *arXiv preprint arXiv:2212.13643*.

Gemini Team, Google DeepMind. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next-generation agentic capabilities. Technical Report.

Google DeepMind. 2024. Introducing gemini 2.0: our new ai model for the agentic era. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december\protect\penalty\z@-2024/. Accessed: 2025-10-07.

Larry E Greiner and Flemming Poulfelt. 2005. *Consulting to Management*. Springer.

Jiayu Guo, Yu Guo, Martha Li, and Songtao Tan. 2025. Flame: Financial large-language model assessment and metrics evaluation. *Preprint*, arXiv:2501.06211.

Nan Huo, Xiaohan Xu, Jinyang Li, Per Jacobsson, Shipei Lin, Bowen Qin, Binyuan Hui, Xiaolong Li, Ge Qu, Shuzheng Si, and 1 others. 2025. Bird-interact: Re-imagining text-to-sql evaluation for large language models via lens of dynamic interactions. *arXiv preprint arXiv:2510.05318*.

Souratn Jain. 2024. Revolutionizing business consulting with generative ai: Exploring transformative models for strategic decision-making, innovation, and operational excellence. *World Journal of Advanced Research and Reviews*, 22:2273–2284.

Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, Xuanhe Zhou, Chenhao Ma, Guoliang Li, Kevin C.C. Chang, Fei Huang, Reynold Cheng, and Yongbin Li. 2023a. Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Jinyang Li, Xiaolong Li, Ge Qu, Per Jacobsson, Bowen Qin, Binyuan Hui, Shuzheng Si, Nan Huo, Xiaohan Xu, Yue Zhang, and 1 others. 2025. Swe-sql: Illuminating llm pathways to solve user sql issues in real-world applications. *arXiv preprint arXiv:2506.18951*.

Zhenyu Li, Xiaoguang Zhang, Jing Liu, Chenguang Sun, Lei Zhang, and Ming Zhou. 2023b. Bird: A large-scale dataset for cross-domain text-to-sql. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Peixian Ma, Xialie Zhuang, Chengjin Xu, Xuhui Jiang, Ran Chen, and Jian Guo. 2025. Sql-r1: Training natural language to sql reasoning model by reinforcement learning. *arXiv preprint*, arXiv:2504.08600.

Simon Malberg, Nikita Grigorev, Victoria Lea Rein, Barbara Borne Bass, Maria Alejandra Gelvez Alvarez, and Georg Groh. 2025. Bridging ai and business: Are large language models good management consultants? In *ECIS 2025 Proceedings*, Artificial Intelligence in Organizations and Society. Paper No. ECIS2025-1108, CRP.

Meta AI. 2025. The llama 4 herd: The beginning of a new era of natively multimodal intelligence. https://ai.meta.com/blog/llama-4-multimodal-intelligence. Official model release. No peer-reviewed technical report available at the time of writing.

Barbara Minto. 2009. *The Pyramid Principle: Logic in Writing, Thinking, and Problem Solving*, 3rd edition. Financial Times Prentice Hall, Harlow, UK.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025. Introducing gpt-5. https://openai.com/research/introducing-gpt-5. Official model release. No public technical report available at the time of writing.

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.

Iurii Perga. 2024. The role of ai (large language models) in managerial decision-making: Benefits and challenges. *Naukovi Perspektyvy (Naukovì perspektivi)*.

Michael E. Porter and Mark R. Kramer. 2011. Creating shared value. *Harvard Business Review*, 89(1-2):62–77.

Foster Provost and Tom Fawcett. 2013. *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. O'Reilly Media, Inc.

Zeeshan Rasheed, Muhammad Waseem, Kari Systä, and Pekka Abrahamsson. 2024. Large language model evaluation via multi ai agents: Preliminary results. *Preprint*, arXiv:2404.01023.

Galit Shmueli. 2010. To explain or to predict? *Statistical Science*, 25(3):289–310.

Edward H. Shortliffe and Martin J. Sepúlveda. 2021. Clinical decision support in the era of artificial intelligence. *JAMA*, 325(1):9–10.

Ruihao Shui, Yixin Cao, Xiang Wang, and Tat-Seng Chua. 2023. A comprehensive evaluation of large language models on legal judgment prediction. *arXiv preprint arXiv:2310.11761*.

Spider Team. 2023. Spider2.0: Towards comprehensive evaluation of text-to-sql systems. https://yale-lily.github.io/spider2/.

Eric J. Topol. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR*.

Pengcheng Xu, Tao Yu, Tianze Shi, Rui Jiang, Chenglong Li, Zhiwei Dong, Tao Zhao, Ruoxi Cao, Jun Yan, Thomas Lin, and et al. 2021. Sparc and cosql: Cross-domain context-dependent text-to-sql semantic parsing. In *ACL*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *NeurIPS*.

Zhewei Yao, Guoheng Sun, Lukasz Borchmann, Zheyu Shen, Minghang Deng, Bohan Zhai, Hao Zhang, Ang Li, and Yuxiong He. 2025. Arctic-text2sql-r1: Simple rewards, strong reasoning in text-to-sql. *arXiv preprint*, arXiv:2505.20315.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018a. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3911–3921.

Tao Yu, Rui Zhang, Michihiro Yasunaga, Yi Chern Tan, Xi Victoria Lin, Suyi Li, Akiko Eriguchi, and Danqi Chen. 2018b. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, and 1 others. 2023. Judging llms by llms: Benchmarking large language models with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *CoRR*, abs/1709.00103.

Xinyi Zhou, Frank F. Xu, and et al. 2023. Limitations of current text-to-sql evaluation frameworks. In *Workshop on Data-Centric AI, NeurIPS*.

## A Detailed Descriptions of CORGI Business Scenarios

We provide the following additional details for the 10 business scenarios covered in CORGI.

**1. Online food delivery platform (e.g. Doordash)** This type of business connects users with a variety of local restaurants, and connects delivery personnel with different orders. The platform differentiates itself through personalized features including allergen tracking, prime club membership, customizable favorite lists, and family account management. Dynamic business features such as time-bound promotions and surge pricing allow restaurants to optimize engagement and profitability. Users can browse restaurants by categories, view detailed menus organized into sections and dishes, and place order. Orders progress through detailed status stages, and are fulfilled by a fleet of delivery personnel tracked by availability and vehicle type, with their performance recorded in a delivery history log. Users can rate not only restaurants but also individual dishes and delivery service, while restaurants can respond directly to reviews. The system supports secure payments, real-time tracking, promotional codes, and supports reviews and replies to ensure feedback loops.

**2. App store (e.g. Google Play)** This type of business connects developers and users through app discovery, download, and monetization. Users can log in socially, maintain balances, and use multiple payment methods; developers can publish apps with version control, pricing models, and platform support; apps are categorized and tagged

for enhanced discoverability; and usage is tracked through downloads, installations, rankings, and reviews.

**3. Online clothing store (e.g. Lululemon)** This type of business usually operates an online and offline fashion retail platform that offers a diverse range of products categorized by department, category, and collection, and is tailored to various customer needs through uses, features, and fit types. The schema is centered on the Products table, which connects to departments, categories, collections, and is extended via junction tables for product uses and features. Each product links to one or more SKUs, which represent sellable items with inventory, pricing, and shipment details. Customer data is also extensively modeled, including multiple addresses, payment methods, loyalty membership, and order history, while promotional codes, returns, cancellations, and fulfillment options are designed to support scalable and efficient operations across both digital and physical retail channels.

**4. Freelancing platform (e.g. Upwork)** This type of business operates a freelancing platform that connects employers and freelancers globally. Users can be differentiated between employers and freelancers. Employers can post jobs, each categorized and requiring specific skills. Freelancers can maintain a detailed profile with portfolio, education, experience and certifications. They can apply for the jobs that employers posted, which may lead to contracts. The wage is handled by different payment methods. After the completion of contracts, employers and freelancers can give reviews to each other.

**5. Online room booking platform (e.g. Airbnb)** This type of business operates in the short-term rental and vacation housing market, connecting hosts who want to rent out properties with guests seeking accommodation. Users can act as either Hosts or Guests. Listings are tied to specific neighborhoods, cities, and countries through hierarchical foreign keys. Each listing includes details such as RoomTypes, pricing, amenities, availability, and policies. Booking activities are linked to users and listings, and supported by payment infrastructure (Payments, PaymentMethods, UserPaymentMethods). The system also tracks user engagement through Reviews and UserFavorites.

**6. Personalized products (e.g. Persona Nutrition)** This type of business operates a platform that offers customized products tailored to individual customer needs. Users typically begin by completing a detailed assessment, which is then analyzed to provide personalized recommendations. For example, in the case of nutrition, the platform might offer custom vitamin packs, but this model can extend to other product categories such as skincare, beauty, and wellness products. Each product is linked to specific SKUs, representing items tailored to the customer's preferences. The business model is often subscription-based, ensuring regular delivery of the customized products. Additionally, the platform may offer reordering features, customer support for further customization, and a feedback loop to refine future product recommendations based on user reviews and changing needs.

**7. Car Rental services (e.g.Turo)** This type of business operates a peer-to-peer rental platform that connects vehicle owners with individuals seeking short-term car rentals. Vehicle owners, known as hosts, can list their cars on the platform, specifying the car's availability, location, and rental price. Renters, known as guests, can browse these listings, select a vehicle, and book it for the desired period. The platform charges a commission on each transaction, with the commission percentage varying based on factors like the type of vehicle and rental duration. Each rental is managed through the system, tracking reservations, pricing, insurance, and payment processing. Additionally, features like user verification, vehicle delivery, and 24/7 support ensure a seamless and secure experience for both parties. The platform also allows renters to rate hosts and vice versa after each transaction.

**8. Second-hand luxury goods marketplace (e.g. TheRealReal)** This type of business operates a consignment platform specializing in the sale of pre-owned luxury goods, including clothing, handbags, jewelry, and art. Sellers consign their items to the platform, where experts authenticate each item before listing it for sale. The platform earns revenue by taking a commission from each sale, with rates varying based on the category and sale price of the item. The items are categorized and linked to detailed product pages, showcasing the brand, condition, and price. Buyers can browse and purchase authenticated luxury items from a wide range of high-end brands. The platform also offers additional services such as home pickup, personal-

ized styling, and premium membership, enhancing the overall customer experience. Additionally, customers can sell their own pre-owned items through the platform, fostering a circular economy for luxury goods.

**9. UX-driven advertising analytics platform (e.g. Mixpanel + Amplitude + AppsFlyer hybrid)** This type of business provides a full-stack data platform focused on optimizing digital advertising and customer experience through behavioral analytics, A/B testing, and multi-channel attribution. The platform integrates data from advertisement impressions, clicks, user sessions, content views, and CRM records to offer a unified view of the customer journey. Advertisers and product teams can launch campaigns, define audience segments, and test multiple ad creatives or landing pages using built-in experimentation tools. The platform supports real-time event tracking and visual funnel analysis, allowing businesses to identify drop-offs, optimize conversion paths, and personalize experiences at scale. The company also enables dynamic cohort building, personalized campaign delivery, and feedback-loop integration with external ad platforms (e.g., Meta, Google), helping brands make more informed, data-driven decisions to maximize ROI and user engagement.

**10. E-commerce enablement platform (e.g. Shopify)** This type of business provides an end-to-end platform that enables entrepreneurs, SMBs, and large enterprises to launch, operate, and scale online storefronts with minimal technical complexity. The platform offers drag-and-drop website builders, integrated payment processing, inventory and order management, and multi-channel sales capabilities across marketplaces, social media, and in-person POS systems. It integrates data from product catalogs, customer interactions, and transaction histories to support a seamless retail experience. Merchants can run marketing campaigns, personalize product recommendations, and automate workflows using built-in analytics and third-party apps. The platform supports real-time inventory syncing, flexible fulfillment options, and automated tax and shipping calculations. Businesses can enhance customer engagement through CRM tools, loyalty programs, and abandoned cart recovery.

## B Database Population Details

**Step 1: Construct a schema.** We collect information from publicly available websites, applications, and technical documents to extract domain-specific workflows and business logic across diverse verticals. Based on this information, we construct entity–relationship diagrams and schemas that approximate the operational structures of real-world business.

**Step 2: Simulate data** To capture complexities beyond static schemas and to approximate realistic commercial scenarios, we introduce three types of simulation rules to guide data generation. These rules are modeled in isolation as a simplifying assumption; no compounding effects are considered. This serves as the baseline for our study.[6] For example, payment method distributions are not linked to seasonal trends: during Black Friday simulations, the proportions of payment methods are generated randomly rather than conditioned on sales surge. To illustrate the rules **a,b,c**, we simulate *Persona Nutrition*, a personalized health and wellness subscription service, as a case study.

**(a) Business Operation Rules.** These are deterministic rules that directly reflect the company's operational logic. At *Persona Nutrition*, new customers must complete (i) an *Initial Assessment*. If the health assessment score falls below 3.0, (ii) a *Low Rating Trigger* automatically generates consultation offers. Finally, (iii) *Follow-up Assessments* are conducted every six months or upon user request. These operational constraints are kept consistent with the underlying schemas.

**(b) Latent Feature Distribution.** Demographic and behavioral patterns are instantiated through predefined distributions. For example, figure 2a illustrates that at *Persona Nutrition*, the customer base consists of 35% trial users (0–3 months, price-sensitive), 44% regular subscribers (3–12 months, consistent users), 17% loyal customers (1–3 years, high engagement), and 4% VIP customers (3+ years, premium services).

**(c) Latent Seasonal Trends.** Our data also mimic realistic fluctuations. Figure 2b illustrates these seasonal dynamics at *Persona Nutrition*: subscriptions increase during events such as New Year's resolutions (+45%) and Black Friday promotions (+27%), while decreases are observed during periods such as post-New Year resolution (–33%)

[6] Future work will model interactions across rules.

and summer vacation (–31%).

We employ Gemini-2.5-flash-lite[7] (DeepMind, 2025) to synthesize data consistent with defined schemas and simulation rules. To ensure fidelity, all generated data is manually reviewed and randomly tested. Appendix C provides an example prompt used for data generation.

# C Prompts

## C.1 Data Simulation Rules Generation

The following is the prompt for generating data simulation rules, using Shopify as an example.

**Prompt for Generating Data Simulation Rules**

**INSTRUCTIONS:** Assume you are a data scientist working on a text-to-SQL benchmark project. You now need to generate data simulation rules for Shopify database. To better approximate real business scenario of Shopify, you have: (1) Shopify description; (2) Link to Shopify official website; (3) Schema (***shopify.sql***); (4) Real-world user experience report. You can retrieve information from reliable online sources.

**OBJECTS:** Your need to design simulation rules three categories of data simulation rules: (1) Business Operation Rules: Deterministic rules that directly encode Shopify's operational logic; (2) Latent Feature Distributions: Probabilistic rules capturing demographic and behavioral regularities; (3) Latent Trends: Temporal dynamics that reflect realistic fluctuations (e.g., seasonal demand shifts).

**REQUIREMENTS:** All rules should be documented in ***rules.md***, clearly delineated by category. When constructing the rules, adhere strictly to the tables, columns, and relationships explicitly defined in ***shopify.sql***. Do not introduce assumptions about missing or inferred schema elements.

## C.2 Database Population

The following is the prompt for populating database, using Shopify as an example.

[7]Version date: Sept 2025

**Prompt for Populating Database**

**INSTRUCTIONS:** Assume you are a data scientist working on a text-to-SQL benchmark project. You now need to populate the Shopify database. You are provded with: ((1) Schema (***shopify.sql***); (2) Data simulation rules (***rules.md***).

**OBJECTS:** You should populate the Shopify database by strictly following ***shopify.sql*** and ***rules.md***. You should populate the database as realistic as possible, simulating the real e-commerce business enablement platform scenario.

**REQUIREMENTS:** The database should be stored in ***shopify.db***.

## C.3 Answer Generation

The following are the prompts for answer generation.

**Prompt for SQL Queries Generation**

**INSTRUCTIONS:** You are an expert SQL analyst. Given a database schema and a business question, generate multiple comprehensive SQL queries (up to {max_queries}) that will help answer the question,

**GUIDELINES:** (1) Generate at most {max_queries} different SQL queries that explore the question comprehensively; (2) Return each query in a separate code block with a brief description.

Format your response as:
Query 1: [Brief description]

```
SELECT ...
FROM ...
WHERE ...
```

Query 2: [Brief description]

```
SELECT ...
FROM ...
WHERE ...
```

And so on.

**User Prompt:**
Database Schema: {schema}

Question: {question}
Please generate multiple SQL queries to comprehensively answer this question:

**Prompt for Final Answer Generation**

**INSTRUCTIONS:** You are a senior data analyst. Given a business question and multiple SQL queries with their results, provide a comprehensive analysis that synthesizes all the information.

Original Question: {question}
All Queries and Results: {query_list}, {result_list}

### C.4 Evaluation

The following is the prompt for metric evaluation.

**Prompt for Metric Evaluation**

**INSTRUCTIONS:** You are an expert evaluator. Please evaluate the following answer based on the given question using the {metric} metric.

Question: {question} Answer: {answer} Metric Definition: {metric}: {metric definition} Evaluation Criteria: {criteria}

**REQUIREMENTS:** Please provide your evaluation in the following JSON format:
{{
"Score": <score 0-5>,
"Reasoning": "<brief explanation of the score>"
}}
Please be precise and provide clear reasoning for your score. The output format should be strictly as above. Do not include any other text other than the JSON format.

## D Complete Strategic Management Consulting-based Evaluation Framework

Tables 6 and 7 present the complete version of strategic management consulting–inspired evaluation framework.

We use Labubu as an example and use the following three concrete examples to illustrate the how we evaluate the answers using the framework. The three concrete examples are:

- Type 2: Why Q2 revenue declined by 12% compared to Q1?
- Type 3: Given the past 12 months of sales and campaign performance, forecast Q3 revenue and justify your approach.
- Type 4: Revenue growth has slowed for two consecutive quarters. What strategic actions should Labubu take to restore growth?

Table 6: Detailed Evaluation Rubric (Part 1)

| Category | Metric | Scope | Definition | Scores and Criteria | | | Sources |
|---|---|---|---|---|---|---|---|
| | | | | 1wo (Fails) | 3 (Meets) | 5 (Exceeds) | |
| **Structure** | Structural Organization | Type 2, 3, 4 | Whether the response presents a clear central thesis and organizes its subpoints in a mutually exclusive and collectively exhaustive (MECE) manner, with a logically hierarchical structure that enables easy comprehension and traceability of arguments. | The response is disorganized, lacks a main idea, and has redundant or illogical subpoints, making it not MECE. *Example*: "Our sales are down. We need to focus on marketing and then also on our customer base, especially the new ones. Also, products." | The response has a central idea but is only partially MECE; subpoints may overlap or not be fully exhaustive. The logic requires some effort to follow. *Example*: "Sales are down because of marketing and sales team performance. We should also look at our promotional strategies." | The response is a perfectly coherent pyramid, with a single main point and subpoints that are perfectly MECE and easy to grasp. *Example*: "Sales are down due to Price and Volume. We will analyze each of these areas separately to find the root cause." | Minto, B. (2009). The Minto Pyramid Principle: Logic in Writing, Thinking, and Problem Solving. |
| | Logical Consistency | Type 2, 3, 4 | Whether the response is internally consistent and contextually aligned, with a complete and non-contradictory chain of reasoning that connects evidence, intermediate claims, and conclusions. | The response contradicts the context or is internally contradictory, with fundamentally flawed or incomplete reasoning. *Example*: "Our Q3 sales were 100M dollar, which is a 10% decrease from last quarter's 80M dollar." | The response is mostly consistent but has minor logical gaps, making it difficult to follow. *Example*: "Our Q3 sales grew significantly because we launched a new product, which increased our market share." | The response is perfectly aligned with the context and has a complete, sound, and seamless chain of reasoning. *Example*: "Our Q3 sales grew by 10% from Q2, from 80M to 88M dollar, which was a direct result of our new marketing campaign targeting urban areas." | Cosentino, M.P., 2024. Case in Point 12: Complete Case Interview. Burgee Press. |
| | Conciseness | Type 2, 3, 4 | Whether the response communicates necessary information efficiently, minimizing redundancy and unnecessary elaboration while preserving analytical clarity. | The response is overly verbose, repetitive, and uninformative. *Example*: "In order to improve our sales performance, it is imperative that we strategically and proactively implement a new initiative designed to enhance overall revenue generation and expand our presence in the market." | The response is of reasonable length but could be more concise, containing some redundant details. *Example*: "To improve our sales, we should put a new initiative in place to increase revenue and market presence." | The response is concise and to the point, using the minimum words to convey maximum information efficiently. *Example*: "To increase sales, we will launch a new initiative to expand our market presence." | Garner, B.A., 2013. HBR guide to better business writing (HBR Guide Series). Harvard Business Review Press. |
| **Data Sense** | Information Relevance and Adequacy | Type 2, 3, 4 | Whether the response selects and integrates relevant internal or external information that is sufficient and appropriate to support the analysis, reasoning, or decision being made. | The response is irrelevant or inadequate, missing critical information or misinterpreting the user's intent. *Example*: "The Q3 revenue decline was caused by the cancellation of an insignificant user's account." | The response is functional but misses critical details or context, leaving significant follow-up questions. *Example*: "The Q3 revenue decline was caused by decreased sales in the US market." | The response is relevant and comprehensive, providing all necessary information to fully address the query and anticipate user needs. *Example*: "The Q3 revenue decline was due to a 10% drop in unit volume, driven by increased competition and a new competitor's product launch that targeted our key demographic." | Voorhees, E. M. (2001). The TREC question answering track. Natural Language Engineering, 7(4), 361-378. |
| | Trend Awareness | Type 2, 3, 4 | Whether relevant temporal patterns, external market signals, or contextual dynamics are identified and meaningfully incorporated into the analysis, rather than treating the problem as static. | The response is static, providing data without considering trends. *Example*: "The sales in next November will be the same as October." (Ignoring Black Friday trend) | The response acknowledges trends but fails to integrate them meaningfully into the analysis. *Example*: "Next month is holiday season, but we do not need to take this into account when managing inventories." | XXX | Binder, J. and Watkins, M.D., 2024. To solve a tough problem, reframe it. Harvard Business Review, 102(1-2), pp.80-89. |
| | Model Selection Rationality | Type 3 | Whether the chosen predictive or analytical model is appropriate for the problem context, justified with clear reasoning, and accompanied by an explicit acknowledgment of its assumptions and limitations. | An inappropriate or unsuited model is used for the prediction, showing a lack of fundamental data science understanding. *Example*: "We used a simple linear regression model to predict next year's sales, ignoring seasonality and market shifts." | A reasonable model is selected, but the choice is not justified, and its limitations are not acknowledged. *Example*: "We used an ARIMA model to predict sales." | A rational and well-justified model is selected for the prediction, with a clear explanation of its reasoning and limitations. *Example*: "We chose a Prophet model to forecast sales due to its strong performance on time-series data with clear seasonal and holiday effects, which are present in our dataset." | Brea, C., Anderson, J., Bartling, R., Mayo, R., & Mueller, F. (2018). Demand forecasting with advanced analytics. Bain & Company. |
| **Insightfulness** | Out-of-Box Thinking | Type 2, 3, 4 | Whether the response introduces non-obvious yet justified perspectives, heuristics, or reframing strategies that extend beyond routine or formulaic approaches, while remaining logically grounded. | The reasoning is conventional, relying only on standard formulas or predetermined steps. *Example*: "Used the standard 5 Forces analysis without adjustment." | The reasoning adapts a standard framework to the problem context but does not introduce a novel concept. *Example*: "Used a modified Profitability Tree to incorporate a new variable." | The response develops a novel analytical approach or heuristic to solve the problem, going beyond expected logical steps. *Example*: "Developed a new heuristic to model customer retention based on unstructured social data, resulting in a novel segmentation." | Backler, W., Iny, A., D'Intino, N., Parker, E., & Hirashita, S. (2025, January 14). Navigating the future with strategic foresight. Boston Consulting Group. |
| | Root Cause Depth | Type 2 | Whether the analysis moves beyond surface-level symptoms to identify well-supported underlying causes, demonstrating causal depth and explanatory rigor. | The response only describes symptoms and offers no root cause analysis. *Example*: "Sales are down because customers are not buying our products." | The analysis is shallow or based on unsupported assumptions, providing weak conclusions. *Example*: "Sales are down due to a lack of marketing and high prices." | The analysis is thorough and multi-faceted, identifying plausible, well-supported root causes and proposing mitigations. *Example*: "Sales are down due to a new competitor that entered the market with lower prices and a superior product. This has made our existing product non-competitive, leading to a 20% drop in volume." | inder, J. and Watkins, M.D., 2024. To solve a tough problem, reframe it. Harvard Business Review, 102(1-2), pp.80-89. |
| | Assumption Appropriate-ness | Type 3 | Whether assumptions underlying predictions or projections are explicitly stated, realistic, logically coherent, and aligned with available evidence. | The analysis is based on unrealistic or unstated assumptions, invalidating the conclusion. *Example*: "Our new product will be a success." (Based on no data) | Some assumptions are appropriate but are not clearly stated or justified, leading to a confusing analysis. *Example*: "We assume that Gen Z will respond positively to our new product due to its viral potential." | All assumptions are clearly stated, justified, and realistic, with conclusions that logically follow and are fully consistent. *Example*: "We assume a 5% increase in market share in Q4, based on our successful pilot program and customer feedback showing a high intent to purchase." | Mulligan, C., Northcote, N., Röder, T., & Vesuvala, S. (2021, September 9). The strategy-analytics revolution. McKinsey & Company. |

Table 7: Detailed Evaluation Rubric (Part 2)

| Category | Metric | Scope | Definition | Scores and Criteria | | | Sources |
|---|---|---|---|---|---|---|---|
| | | | | 1 (Fails) | 3 (Meets) | 5 (Exceeds) | |
| **Operation Implementability** | Actionability | Type 4 | Whether recommendations are executable, concrete, and tied to measurable steps or KPIs. | The recommendation is too vague or abstract to be executed, with no concrete steps. *Example*: "The company should improve its online presence." | The recommendation is high-level and lacks specific, measurable steps for implementation. *Example*: "The company should launch a social media campaign and update its website." | The recommendation is highly actionable, concrete, and includes specific steps tied to business KPIs. *Example*: "Launch a TikTok campaign with five influencer partners, with a budget of 50,000 dollar, and redesign the homepage with A/B testing on two new layouts." | Garner, B.A., 2013. HBR guide to better business writing (HBR Guide Series). Harvard Business Review Press. |
| | Time-Based Planning | Type 4 | Whether recommendations include a realistic implementation timeline with clear milestones. | The recommendation lacks any timeline or sense of urgency. *Example*: "We recommend a new marketing campaign." | The recommendation includes a vague, general timeline that does not allow for effective planning. *Example*: "We will launch the new campaign in the near future." | The recommendation has a specific, realistic, and well-justified timeline with clear milestones for effective planning. *Example*: "We will launch the new marketing campaign in Q4, with a goal of completion by December 15th." | Brea, C., Anderson, J., Bartling, R., Mayo, R., & Mueller, F. (2018). Demand forecasting with advanced analytics. Bain & Company. |
| **Purpose Alignment** | Goal Orientation | Type 4 | Whether the recommendations are explicitly linked to clearly defined strategic objectives or measurable business goals (e.g., SMART criteria), with a coherent value proposition. | The recommendation is not linked to any clear goals or business value. *Example*: "We suggest you hire a new VP of Marketing." | The recommendation is linked to a general goal that is not specific or measurable. *Example*: "We suggest you hire a new VP of Marketing to increase brand awareness." | The recommendation is directly tied to a specific, measurable, and achievable (SMART) business goal with a clear value proposition. *Example*: "We suggest you hire a new VP of Marketing with the goal of increasing our market share by 5% and our social media engagement by 20% in the next fiscal year." | Harvard Business Review Analytic Services. (2022). Implementing strategic goals for organizational success. Harvard Business School Publishing. Sponsored by Lattice |
| | Stakeholder Orientation | Type 4 | Whether the recommendation considers and balances impacts on key non-shareholder stakeholders. | The recommendation focuses exclusively on financial metrics (e.g., ROI) and ignores all non-shareholder impacts. *Example*: "Recommend mass layoffs to cut costs by 20% without considering employee morale or retention." | Mentions key stakeholders (e.g., customers) but does not integrate their perspectives into the final recommendation or mitigation plan. *Example*: "The plan may impact employees, but the financial return justifies it." | The recommendation explicitly balances financial gains with stakeholder well-being, including measures to mitigate negative effects on non-shareholder groups. *Example*: "Recommend a phased automation rollout, offsetting job losses with a robust retraining program for high-value roles." | Bain & Company. (2023, January). Stakeholder strategies – Management tools. ) |
| **Compliance** | Risk Management | Type 4 | Whether risks are identified, warned, and accompanied by concrete mitigation strategies. | The response fails to identify any risks or contains content that exacerbates risk. *Example*: "The new product launch will be a guaranteed success." | Some risks are identified, but warnings or mitigation suggestions are inadequate. *Example*: "A new product launch could fail, so we should be careful." | The response provides comprehensive risk warnings, identifies potential risk points, and offers concrete mitigation suggestions. *Example*: "We recognize the risk of market saturation and mitigate this by conducting a thorough competitive analysis, allocating a contingency budget for unexpected costs, and a phased rollout to a test market." | Guo, J., Guo, Y., Li, M., & Tan, S. (2025). Flame: Financial large-language model assessment and metrics evaluation. arXiv preprint arXiv:2501.06211. |
| | Regulatory Compliance | Type 4 | Whether recommendations adhere to applicable laws, regulations, and industry standards, including disclaimers. | The response is non-compliant with relevant laws and regulations, posing a severe risk. *Example*: "To bypass data privacy regulations, you should use customer data from our partners without their explicit consent." | The response is generally compliant but lacks specific disclaimers or fails to fully address all regulatory requirements. *Example*: "Our new data collection process will adhere to all regulations, but we will not be adding a legal disclaimer to our website." | The response is fully compliant with all applicable laws and regulations, including adequate disclaimers and privacy protocols. *Example*: "All of our new data collection processes are fully compliant with GDPR and CCPA, and we have implemented a clear, user-facing privacy policy and consent form to ensure full legal adherence." | Guo, J., Guo, Y., Li, M., & Tan, S. (2025). Flame: Financial large-language model assessment and metrics evaluation. arXiv preprint arXiv:2501.06211. |
| | Ethical Responsibility | Type 4 | Whether the recommendations uphold principles of fairness, non-discrimination, and social responsibility, supported by specific safeguards or mitigation mechanisms where risks of harm exist. | The recommendation is demonstrably biased or perpetuates known social inequalities, or ignores explicit ethical consequences. *Example*: "The loan risk model shows bias against applicants from specific zip codes; recommend proceeding without mitigation." | Includes a general statement about "fairness" but lacks concrete, measurable steps to mitigate identified bias or ensure equitable outcomes. *Example*: "We should try to be fair, but the model output is final." | The recommendation includes specific, actionable steps (e.g., bias monitoring, disparate impact analysis) to ensure equitable outcomes and promote the well-being of the affected demographic. *Example*: "Implement a specific re-weighting algorithm to address observed demographic bias in the output and monitor fairness metrics in real time." | Cowgill, B., Dell'Acqua, F., Deng, S., Hsu, D., Verma, N. and Chaintreau, A., 2020, July. Biased programmers? Or biased data? A field experiment in operationalizing AI ethics. In Proceedings of the 21st ACM Conference on Economics and Computation (pp. 679-681). |

## E Evaluation Examples

**Type 2: Why did download numbers in March 2025 much higher than other months? (Appstore db)**

**Responder: Gemini-2.5-flash-lite**
The data reveals a significant spike in app downloads in March 2025, with **3006 downloads** compared to an average of approximately **1800–1900 downloads** in surrounding months (Query 1). This analysis investigates the contributing factors behind this surge.

**Key Drivers of the Download Increase:**

- **Dominance of "BlackMonkey" App:** The primary driver of the download surge in March 2025 was overwhelmingly the app **"BlackMonkey"**, which accounted for **1222 downloads** (Query 2). This single app represents a substantial portion of the total downloads for that month, far exceeding any other individual app.
- **Developer "KidsCloud" Success:** The developer **"KidsCloud"** appears to be closely linked to this surge, reporting **1208 downloads** in March 2025 (Query 7). Given that "BlackMonkey" is the top-downloaded app and "KidsCloud" is the top-downloaded developer, it is highly probable that "BlackMonkey" is developed by "KidsCloud" and is the reason for their exceptional download numbers.
- **Category Performance:** The surge is concentrated within the **"Sports"** (1364 downloads) and **"Food & Drink"** (1251 downloads) categories (Query 4). This suggests that "BlackMonkey" (or any other significant contributors from KidsCloud) likely falls into one or both of these categories.
- **Age Group Engagement:** The "Teen" (**1824 downloads**) and "Kids" (**1820 downloads**) age groups were the most active in March 2025 (Query 5). This aligns with the hypothesis that "BlackMonkey" is an app popular with younger demographics, potentially a game or an entertainment app.
- **Platform Distribution:** While iOS and Android show high download numbers (**2307** and **2299** respectively), macOS and Windows also saw a considerable increase in downloads compared to other months (Query 3). This indicates that the surge in downloads for "BlackMonkey" (and potentially other apps from KidsCloud) was not limited to a single platform.
- **Pricing Model:** Both **"paid"** (**1284 downloads**) and **"free"** (**1200 downloads**) apps saw a significant increase in downloads, with subscription models also showing a moderate rise (**522 downloads**) (Query 6). This suggests that the increased downloads might be a mix of paid app popularity and free app adoption, or that "BlackMonkey" itself might have a tiered pricing model or in-app purchases.

**App Releases and Updates in March 2025:** Query 8 lists several app releases and updates in March 2025. While many popular apps like PicsArt, Nike Training Club, and Clash of Clans had updates, none of these individually appear to have the download volume to explain the overall surge. The app **"BlackMonkey"** does not appear in the list of updated or released apps in March 2025. This suggests that either "BlackMonkey" had no significant update or release in March, or that the surge is driven by an app not captured in this specific query (e.g., an app with a very long name not easily identifiable, or an app released prior to March but gaining significant traction due to external factors).

**Conclusion:** The significant increase in app downloads in March 2025 was primarily driven by the app **"BlackMonkey"**, likely developed by **"KidsCloud"**. This app appears to be particularly popular among **teen and kid** demographics and has likely

contributed heavily to the **Sports** and **Food & Drink** categories. The download surge was observed across major platforms (iOS, Android, macOS, Windows) and was influenced by both paid and free app downloads. Further investigation could focus on understanding the specific marketing efforts or viral nature of "BlackMonkey" that led to this exceptional performance.
**Evaluator: GPT-4o**
Scores:

- Structure
  - Argument Soundness: 5
  - Logical Coherence: 5
  - Verbosity: 3
- Data Sense
  - Information Adequacy: 5
  - Trend Awareness: 3
- Insightfulness
  - Out-of-box Thinking: 3
  - Root Cause Depth: 5

**Type 3: What will our return rate be during the next holiday season? (Clothing E-commerce DB)**

**Responder: GPT-4o**

**Comprehensive Analysis.** The original question seeks to estimate the return rate for the next holiday season based on historical data. By synthesizing the results of the provided queries, we can derive insights into the overall return rate, the primary reasons for returns, the categories with the highest return rates, and the impact of customer loyalty tiers on returns. These insights will help us predict and potentially mitigate returns during the next holiday season.

**1. Overall Return Rate (Query 1)**

The overall return rate during the last holiday season (November 1, 2022, to December 31, 2022) was **11.52%**. Out of 382 total orders, 44 were returned. This provides a baseline for estimating the return rate for the upcoming holiday season. If similar conditions persist (e.g., product offerings, customer base, and operational factors), we can expect a return rate in the range of **11–12%**.

**2. Primary Reasons for Returns (Query 2)**

The analysis of return reasons reveals the following key insights:

- The top reasons for returns were **"Didn't fit" (26.47%)**, **"Wrong size" (23.53%)**, and **"Quality issue" (23.53%)**. Together, these three reasons accounted for over 73% of all returns.
- Other reasons, such as **"Defective item" (14.71%)** and **"Changed mind" (11.76%)**, were less frequent but still notable.

This suggests that sizing and quality are the primary drivers of returns. To reduce the return rate, the company could:

- Improve size guides and provide more detailed product descriptions.
- Enhance quality control processes to minimize defective items.
- Offer virtual fitting tools or customer reviews to help customers make more informed decisions.

**3. Return Rates by Product Category (Query 3)**

The return rates varied significantly across product categories:

- **Dresses (16.96%)**, **Swimwear (15.70%)**, and **Shoes (13.97%)** had the highest return rates. These categories are likely more prone to sizing and fit issues, which aligns with the reasons for returns identified in Query 2.
- Categories like **Bottoms (8.45%)** and **Formal Wear (10.00%)** had lower return rates.

This indicates that the company should focus on addressing fit-related concerns in high-return categories like Dresses, Swimwear, and Shoes. For example:

- Offering free exchanges for different sizes in these categories could reduce the likelihood of outright returns.
- Providing detailed size charts and customer reviews specific to these categories could also help.

**4. Impact of Customer Loyalty Tiers on Returns (Query 4)**

The return rates by loyalty tier were as follows:

- **Silver (14.47%)** and **Gold (14.12%)** customers had the highest return rates.
- **Platinum (7.50%)** customers had the lowest return rate, followed by **Bronze (10.39%)** and customers with no loyalty tier (10.94%).

This suggests that higher-tier loyalty customers (Silver and Gold) are more likely to return items, possibly because they purchase more frequently or feel more confident in the return process. Platinum customers, on the other hand, may be more selective in their purchases, leading to fewer returns.
To address this:

- The company could analyze the purchasing behavior of Silver and Gold customers to identify patterns leading to higher returns.
- Offering personalized recommendations or targeted size/fit guidance for these tiers could help reduce returns.

**5. Key Takeaways and Recommendations**

Based on the analysis, the following insights and actions can help predict and mitigate the return rate for the next holiday season.

**Predicted Return Rate.** The return rate for the next holiday season is likely to be in the range of **11–12%**, consistent with the last holiday season.

**Recommendations to Reduce Returns.**

1. **Address Sizing and Fit Issues:**
   - Focus on high-return categories like Dresses, Swimwear, and Shoes by improving size guides, offering virtual fitting tools, and providing detailed product descriptions.
   - Consider offering free exchanges for size-related issues to reduce outright returns.
2. **Enhance Quality Control:**
   - Since "Quality issue" accounted for 23.53% of returns, improving quality checks and addressing common defects could significantly reduce returns.
3. **Targeted Interventions for Loyalty Tiers:**
   - Analyze the purchasing behavior of Silver and Gold customers to identify patterns leading to higher returns.
   - Provide personalized recommendations or targeted size/fit guidance for these tiers.
4. **Monitor and Address Seasonal Trends:**
   - Since the holiday season often involves gifting, returns due to "Changed mind" or "Wrong size" may increase. Offering gift receipts or extended return windows could improve customer satisfaction while managing returns.

**Evaluator: Gemini-2.5-flash-lite**
Scores:

- Structure
  - Argument Soundness: 5
  - Logical Coherence: 5
  - Verbosity: 0
- Data Sense
  - Information Adequacy: 3

- Trend Awareness: 3
- Model Selection Rationale: 0

- Insightfulness
  - Out-of-box Thinking: 0
  - Assumption Appropriateness: 3
  - Root Cause Depth: 5

**Type 4: How can I increase the repeat purchase rate? (Shopify DB)**

**Responder: GPT-4o**

**Comprehensive Analysis.** Based on the query results, there is no data for customers with repeat purchases (i.e., customers with more than one order). This indicates that all existing customers have made only a single purchase. The absence of repeat orders may result from several factors, including limited post-purchase engagement, lack of incentives to return, or misaligned marketing strategies. Below is an analysis of the findings and actionable recommendations to improve the repeat purchase rate.

**1. Key Observations**

**(a) No Repeat Purchases.** Query 1 confirms that there are no customers with more than one order (`total_orders = 1`). Subsequent queries (2–8) analyzing patterns such as discounts, campaigns, delivery, or refunds also return no results, since no repeat customers exist.
**(b) Marketing Campaigns.** Despite substantial ad spending (e.g., up to $24,156.55), Query 5 shows that marketing campaigns have not generated repeat customers, suggesting a weak alignment between marketing efforts and retention goals.
**(c) Fulfillment and Delivery.** Query 6 reports no delivery delays for repeat customers; however, this finding is irrelevant given the absence of such customers.
**(d) Discounts and Promotions.** Query 4 reveals no data on discounts driving repeat purchases, implying that either discount-based campaigns are lacking or ineffective.
**(e) Product Offerings.** Query 3 indicates that no specific product variant or collection has encouraged repeat purchases.
**(f) Refunds.** Query 8 shows no refund data related to repeat customers—unsurprising given their absence—suggesting that refund issues are not the primary barrier to repeat purchases.

**2. Recommendations to Increase Repeat Purchases**

**1. Understand Customer Behavior (Root-Cause Analysis).**

- Conduct post-purchase surveys or interviews to identify reasons customers do not return.
- Evaluate satisfaction with product quality, pricing, and service experience.
- Verify that product value propositions align with customer expectations.

**2. Enhance Customer Experience.**

- **Personalization:** Use purchase history and preferences to recommend relevant products and send follow-up emails.
- **Onboarding:** Deliver a seamless first-purchase experience that builds trust and recall.
- **Fulfillment:** Offer fast and reliable delivery to reinforce satisfaction.
- **Quality Assurance:** Maintain consistent product quality to encourage repeat buying.

**3. Introduce Loyalty Programs.**

- Implement a points-based system rewarding repeat purchases.
- Provide tiered benefits (e.g., Bronze, Silver, Gold) to incentivize engagement.

**4. Run Retargeting and Re-engagement Campaigns.**

- Re-engage one-time customers via email, SMS, or social media.

- Highlight new arrivals, exclusive offers, and limited-time discounts.
- Use abandoned-cart reminders to recover potential repeat buyers.

**5. Leverage Discounts and Incentives.**

- Offer second-purchase discounts (e.g., "10% off your next order").
- Provide free shipping or gifts for repeat customers.
- Introduce bundles or subscriptions to foster habitual purchases.

**6. Improve Marketing Effectiveness.**

- Shift focus from acquisition to retention-oriented marketing.
- Target existing customers with retargeting ads and personalized messaging.
- Evaluate the ROI of current channels (affiliate, social) and expand to email or partnership marketing.

**7. Expand Product Offerings.**

- Introduce complementary or add-on products to encourage follow-up purchases.
- Bundle products to increase perceived value and convenience.

**8. Build Trust and Brand Loyalty.**

- Communicate brand values and share authentic customer stories.
- Feature reviews or user-generated content to build credibility.
- Engage actively with customers on social platforms.

**9. Monitor and Optimize.**

- Track metrics such as repeat customer rate, average order value, and customer lifetime value.
- Continuously experiment with retention tactics and optimize based on results.

**3. Immediate Next Steps**

1. **Survey existing customers** to collect feedback on satisfaction and barriers to repeat purchases.
2. **Launch a "Welcome Back" campaign** offering targeted discounts to first-time buyers.
3. **Analyze marketing ROI** and reallocate budget toward retention activities.
4. **Design a loyalty program** to reward repeat behavior.

**Evaluator: GPT-4o**
**Scores:**

- Structure
  - Argument Soundness: 5
  - Logical Coherence: 5
  - Verbosity: 3
- Data Sense
  - Information Adequacy: 5
  - Trend Awareness: 3
  - Model Selection Rationale: 3
- Insightfulness
  - Out-of-box Thinking: 5
  - Assumption Appropriateness: 3
- Operation Implementability
  - Actionability: 5
  - Time-based Planning: 0
- Purpose Alignment
  - Goal Orientation: 5
  - Stakeholder Orientation: 5
- Compliance
  - Risk Management: 5
  - Regulatory Compliance: 3
  - Ethical Responsibility: 5

## F Implementation Details of Bradley–Terry Framework

We use the Bradley–Terry framework to model pairwise preferences and implement it with the `choix`

| Model | T2 | T3 | T4 | T4_A | T4_B |
|---|---|---|---|---|---|
| Gemini-2.0-flash-lite | 3.06 | 3.05 | 3.05 | 3.012 | 2.993 |
| Gemini-2.5-flash-lite | 3.08 | 3.10 | 3.08 | 3.042 | 3.026 |
| GPT-4o | 3.38 | 3.32 | 3.27 | 3.244 | 3.231 |

Table 8: Impact of different aggregation methods. T denotes question type.

PyPI package (v0.3.6). We assign each item $i$ a latent strength parameter $\theta_i > 0$. For $n$ comparisons, we uniformly sample a pair $(i, j)$ from all $\binom{n}{2}$ possible pairs and determine the outcome through a Bernoulli trial with $P(i \succ j) = \frac{\theta_i}{\theta_i+\theta_j}$. We record $(a, b)$ when $a$ is preferred over $b$, producing a consistent dataset for ranking evaluation under stochastic preferences.

## G Discussion of Aggregation Methods

We examine whether increasing the influence of the three dimensions specific to type 4 tasks, namely Operational Implementability (OI), Purpose Alignment (PA), and Compliance (C), alters the overall performance ranking for type 4 tasks. We consider two alternative weighting schemes: (i) *Method A*, which doubles the weights of OI, PA, and C while keeping the weights of Structure, Data Sense, and Insightfulness unchanged; and (ii) *Method B*, which assigns a weight of 30% to each of OI, PA, and C, and 10% to each of the remaining three dimensions.

Table 8 shows that: (i) all models achieve substantially lower scores on type 4 tasks than on type 2 and type 3 tasks, and this trend persists under both weighting schemes A and B, with the performance drop being especially pronounced for GPT-4o (OpenAI, 2024) on type 4; (ii) the relative ranking of models remains stable across all task types (GPT-4o > Gemini-2.5-flash-lite > Gemini-2.0-flash-lite (Google DeepMind, 2024)), and this ordering is fully consistent across type 2–4 tasks, including their weighted variants; and (iii) increasing the weights of the three type 4 dimensions does not alter the overall ranking.

## H Kendall's Tau ($\tau$)

We use Kendall's tau ($\tau$) to assess the consistency between automatic evaluation scores and human expert judgments. Kendall's tau is a nonparametric statistic that measures the ordinal association between two ranked variables. It quantifies ranking consistency by comparing the relative ordering of all pairs of observations.

Given a set of paired observations $(x_i, y_i)$, Kendall's tau is defined as:

$$\tau = \frac{P - Q}{P + Q}, \tag{1}$$

where $P$ denotes the number of *concordant* pairs and $Q$ denotes the number of *discordant* pairs. A pair $(i, j)$ is concordant if the relative ordering of $x_i$ and $x_j$ agrees with that of $y_i$ and $y_j$, and discordant otherwise. The value of $\tau$ lies in the interval $[-1, 1]$, where $\tau = 1$ indicates perfect agreement, $\tau = -1$ indicates perfect disagreement, and $\tau = 0$ indicates no ordinal association.

Table 9: CORGI vs. other benchmarks. #D and #T/DB denote the number of databases and the average number of tables per database, respectively.

| Dataset | #D | #T/DB | Domain | BI Question |
|---|---|---|---|---|
| Spider | 138 | 5.1 | Cross | × |
| WikiSQL | – | 1 | Cross | × |
| BIRD | 37 | 7.3 | Cross | × |
| Advising | 1 | 10 | Single | × |
| **CORGI** | **10** | **26** | **Cross** | ✓ |

## I Additional Benchmark Statistics

Table 9 presents additional statistics of the CORGI benchmark.