INFO 250 - Information Visualization

—— Project 2 ——

Team 27

December 19, 2025

Members:

Name	Student ID	Email	Role
Weijie Ge	320230941861	gewj2023@lzu.edu.cn	Leader
Zhengyi Li	320230942111	lzhengyi2023@lzu.edu.cn	
Hongye Li	320230942041	lhongye2023@lzu.edu.cn	
Yuxiang Liu	320230942201	liuyuxiang2023@lzu.edu.cn	

Abstract

Have you ever looked at a chart and felt more confused than before? Choosing the right AI model is hard work, and it becomes even harder when the data comparing them is confusing. We looked at a key chart from the recent "Spider 2.0" paper, which tries to explain why AI agents fail at complex database tasks. The original chart was a fancy, multi-layered circle that looked nice but was very difficult to read. It hid the most important information behind confusing angles and tilted text. To fix this, we turned the complex circle into a clear, sorted bar chart. Our improved version tells a simple, instant story: it shows exactly where the AI makes mistakes, without forcing the reader to solve a visual puzzle first.

Keywords: Error Analysis, Spider 2.0, Cognitive Load, Hierarchical Visualization, Text-to-SQL.

1. Introduction

Everyone is talking about how amazing Generative AI is at writing code. But when we look at the actual numbers, there is a different story. We encountered a fascinating study, "Spider 2.0: Evaluating Language Models on Real-World Enterprise Text-to-SQL Workflows," published at ICLR 2025 (https://spider2-sql.github.io). It shows that even the smartest AI models often struggle when faced with messy, real-world business data. Naturally, we were curious: **Why are these powerful models failing?**

To find the answer, we looked at the paper's main error analysis chart. At first glance, it was colorful and impressive. But as we tried to understand which errors were the biggest problems, we got stuck. The chart was a complicated "sunburst" design—a circle within a circle. We had to twist our heads to read the text and squint our eyes to guess which slice was bigger.

A good chart should answer questions, not create more work.

We realized that this chart was a perfect candidate for improvement. It prioritized looking "cool" over being clear. In this project, we stripped away the complexity. We moved from a circular layout to a linear one, sorted the data by importance, and made the labels easy to read. Our goal was simple: to let the data speak for itself so that anyone can see the story at a glance.

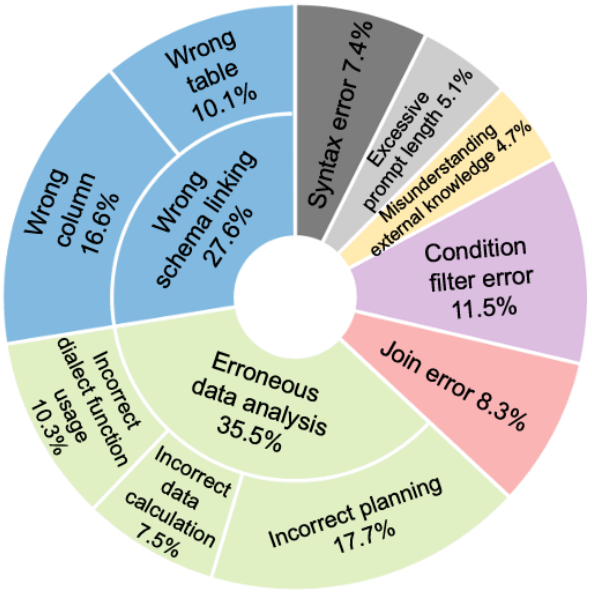

Figure 1: The original visualization from the Spider 2.0 paper (ICLR 2025)

2. Background and Challenges

To understand the stakes, think of **Spider 2.0** as a "final exam" for AI. Unlike previous simple tests, Spider 2.0 throws agents into the deep end: massive cloud databases like **Snowflake and BigQuery** with complex structures. The results were shocking—even the best models passed only about **20% of the time**. This leaves researchers with an urgent question: **Why are the models failing the other 80% of the time?**

The authors attempted to answer this with **Figure 4: Statistics of errors** (shown here as Figure 1), categorizing 300 failure cases to guide future fixes. However, the visualization acts as a barrier. The authors chose a **nested doughnut chart**, prioritizing art over clarity. This design forces our brains to do hard geometry—comparing curved slices accurately is nearly impossible, and the rotated text makes reading a physical pain.

To prove why the chart failed, we programmatically replicated it using Python to extract the raw numbers (Figure 2). This confirmed our suspicion: the data is actually a simple list of percentages. By forcing this linear data into a radial shape, the authors turned a clear story about AI limitations into a confusing visual puzzle, obscuring the very insights they meant to share.

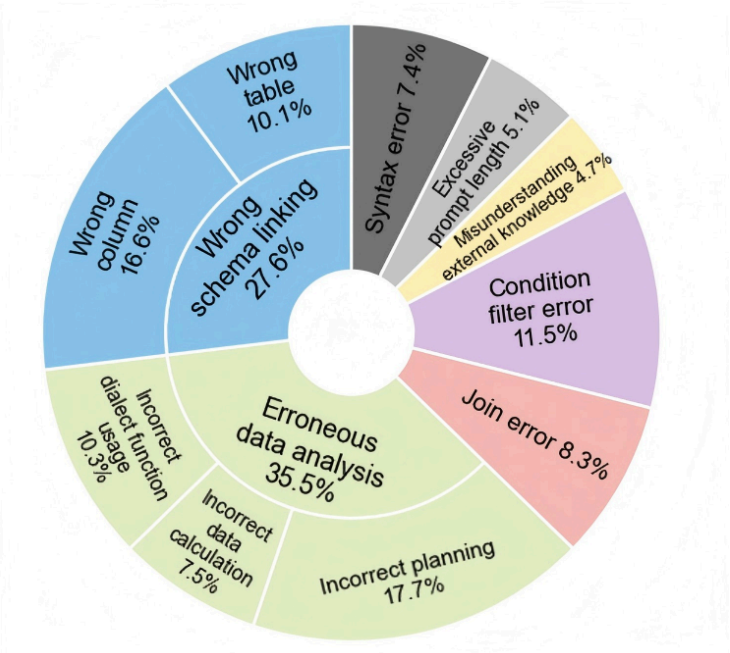

Figure 2: Our replication of the original chart using Python

Despite the valuable data hidden inside, the original visualization felt like a barrier to understanding. The authors used a **nested doughnut chart**, which creates four specific headaches for the reader:

a. It makes comparison a guessing game. Can you tell the difference between 26% and 35% just by looking at a curve? Human brains are bad at comparing the size of curved slices, especially when they aren't lined up straight. In the original chart, trying to figure out if "Wrong Schema Linking" (26.7%) is bigger or smaller than "Erroneous Data Analysis" (35.5%) takes way too much mental effort. We shouldn't have to squint to compare numbers.

b. The "Head-Tilt" Problem. The text labels are rotated around the circle. Forcing us to physically twist our necks to read the text isn't just annoying—it breaks our focus. A good visualization should be readable instantly without gymnastics.

c. The connections are lost. The chart tries to show two layers: the big categories (inner ring) and the specific errors (outer ring). But with the circular shape, it's difficult to visually grasp how much the specific sub-errors contribute to the total weight of the major errors.

d. It's not sorted. The slices are arranged randomly, not by size. **Why make the reader hunt for the answer?** Instead of immediately seeing the biggest problem, our eyes have to wander around the circle, searching for the most critical error types.

We chose to fix this chart because it prioritizes "looking cool" over readability. Our goal is to untangle this data so the critical error patterns stand out instantly.

3. Improvement Process

According to the above problems, we have made the following improvements.

Step One: Optimizing Chart Representation

The original multi-level doughnut chart relied on reading angles and areas to determine percentages. This is a common flaw: our brains are poor at comparing slices and circles, which requires a lot of us to expend mental effort to decode. To fix this, we have converted the chart to a vertical bar chart. This shifts the measurement from confusing areas to simple lengths. By doing this, we have instantly improved the chart's accuracy and ensured that the visual height perfectly reflects the actual data value, making comparisons effortless.

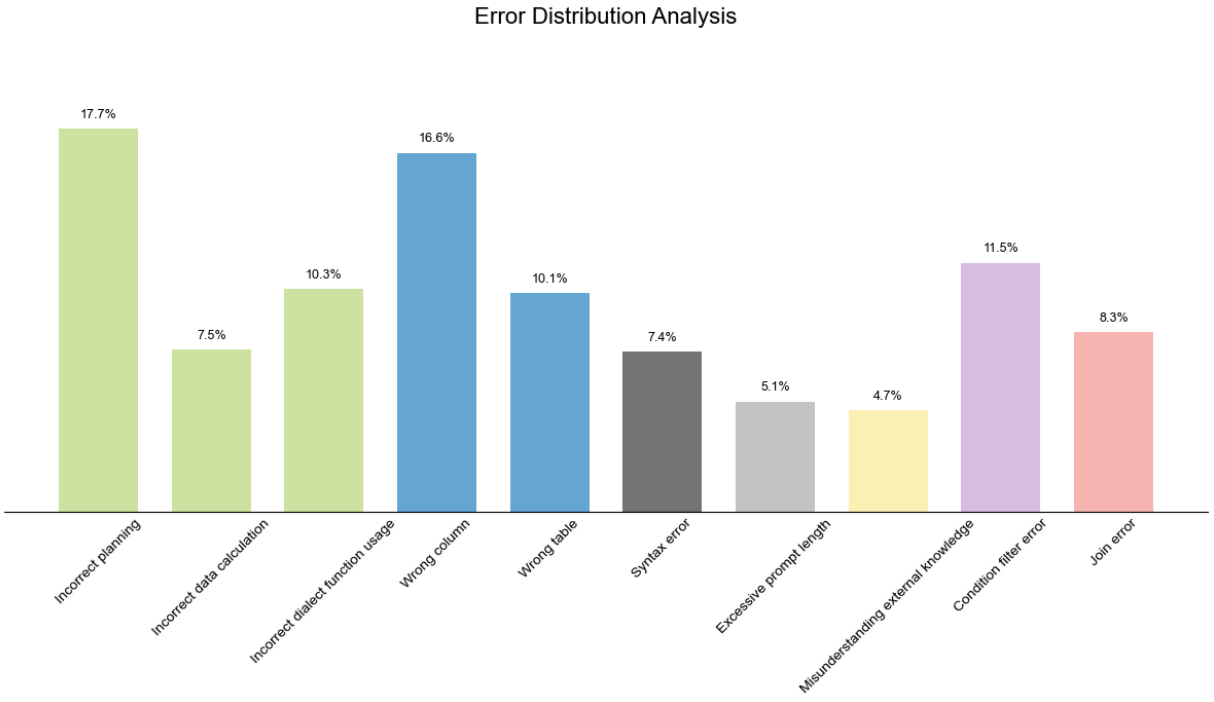

Figure 3: Step One

Step Two: Sort by Value in Descending Order

The presentation above was unsorted, and this forces the reader to search randomly for the most important data points, wasting time and mental energy. We have solved this by sorting the bars by value, from largest to smallest. This prioritized arrangement immediately directs their attention to the biggest error sources first, allowing the chart to clearly tell its story about what matters most. Crucially, we maintained the structure of the original data groups while applying this sorting.

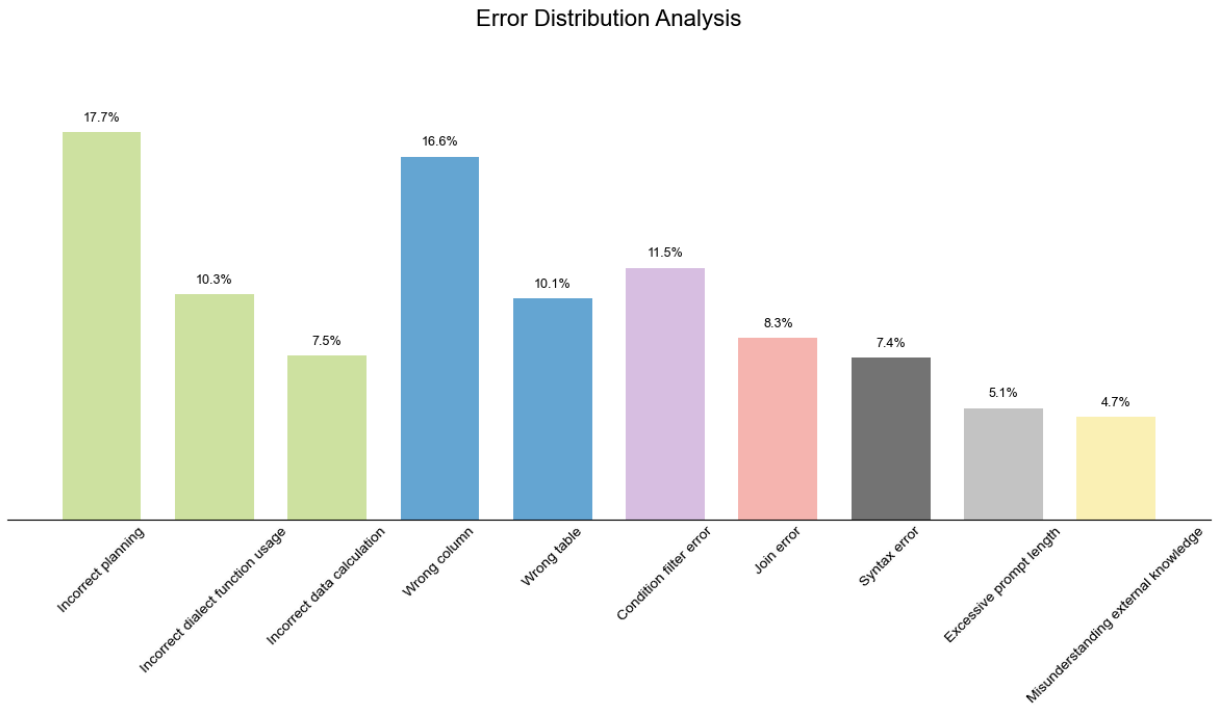

Figure 4: Step Two

Step Three: Convert to a Horizontal Chart

While our new vertical bar chart improved data encoding, the long category labels presented a practical barrier. In visualization design, we must avoid any element that forces us to physically adjust or guess information—such as tilting one's head to read an axis. To solve this, we have transitioned the entire structure into a **Horizontal Bar Chart**. This structural adjustment is the textbook solution for long label management, allowing verbose names to be displayed perfectly flat. This key change has optimized the Text Flow and eliminated reader friction, making the chart instantly accessible and scannable.

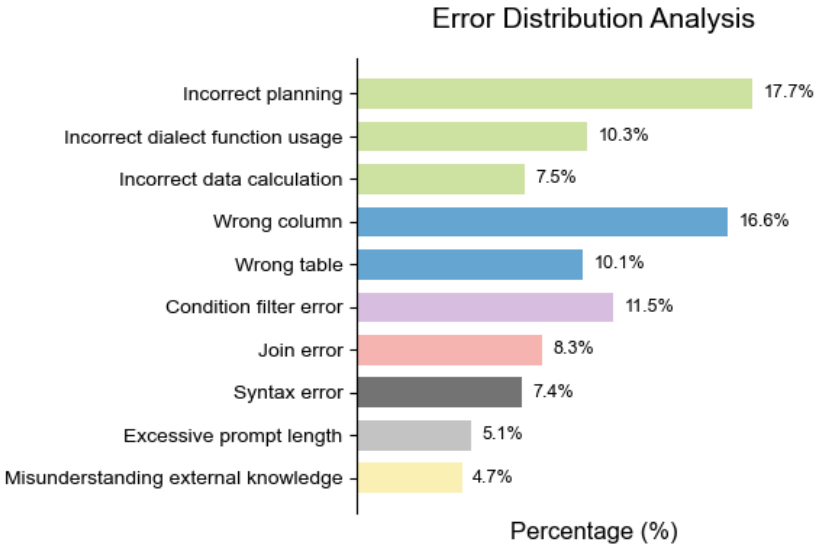

Figure 5: Step Three

Step Four: Add Hierarchical Total Display:

The original chart attempted to show the overall structure of errors, but its nested, multi-level design made it nearly impossible for us to accurately compare the total weight of one major category against another. To solve this, we have introduced a **Hierarchical Total Display**. We have used light-colored background bars whose length precisely represents the sum of the sub-category percentages. This feature uses the simple principle of visual grouping—that objects in a common colored area belong together—to instantly show us which smaller errors contribute to which larger error category. Furthermore, by removing redundant lines and borders, we followed the **Data-Ink principle**, ensuring every remaining element on the page contributes essential information, not clutter.

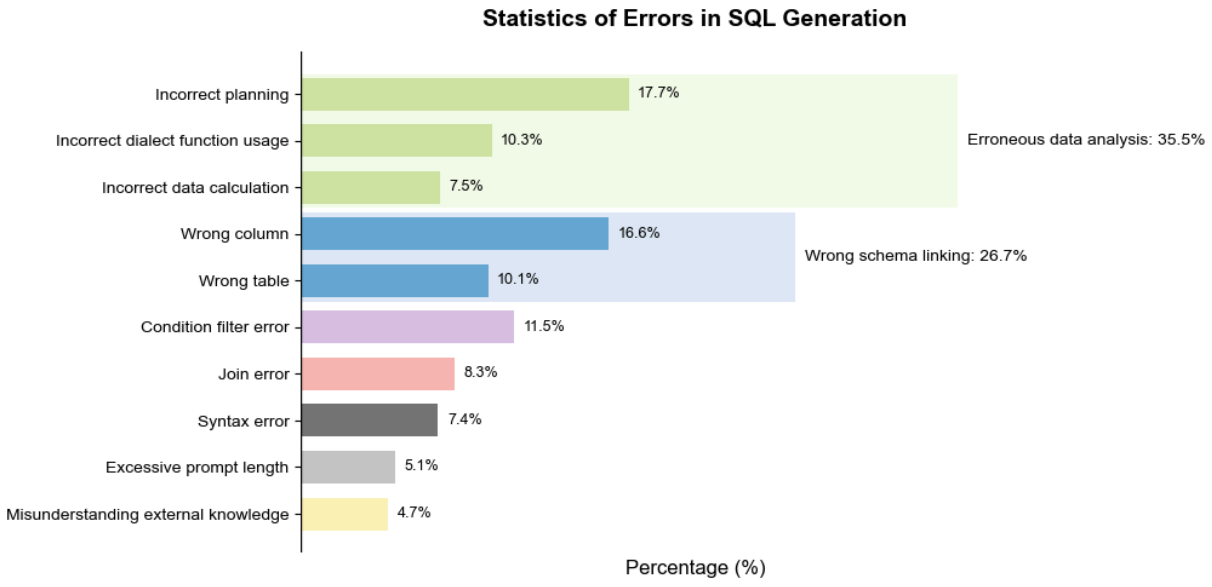

Figure 6: Step Four

Collectively, these modifications allow us to present the content concisely and, crucially, to extract new, actionable insights that were previously obscured by the complexity of the original chart.

4. Improved Visualization Exploration

The final improved visualization, which consolidates sub-category percentages with their corresponding major group totals, now powerfully clarifies the **Error Distribution in SQL Generation**. This clarity allows us to move beyond mere recognition of flaws and extract critical, actionable insights regarding where model failures are most concentrated.

(1) Major Category Dominance: Where is the Primary Failure Point?

The Observation: With the Hierarchical Total Display in place, we immediately see that **Erroneous data analysis (35.5%)** is the single largest source of failure, clearly dominating **Wrong schema linking (26.7%)** and the un-grouped errors.

The Insight and Action: This structure reveals that models are struggling most significantly with interpreting the user's intent and performing the correct calculations or planning required for complex SQL. To achieve the largest immediate gain in model reliability, future research efforts must prioritize interventions that enhance the model's internal data flow and reasoning engine, specifically targeting the causes of **Erroneous data analysis**.

(2) Sub-Category Severity: What Are the Top Three Errors?

The Observation: By sorting all sub-categories by value, we can precisely identify the three most severe individual errors: **Incorrect planning (17.7%)**, **Wrong column (16.6%)**, and **Condition filter error (11.5%)**. Crucially, these top three errors belong to three different major categories (Erroneous data analysis, Wrong schema linking, and an un-grouped error, respectively).

The Insight and Action: This finding challenges the notion that fixing one major group will solve the problem. The simultaneous prominence of errors across different types (planning, column selection, and filtering) implies that the model suffers from **multiple, independent systemic weaknesses**. Development efforts should be segmented to address these three distinct failure modes in parallel, rather than focusing solely on the largest major category.

(3) The Unseen Problem: What's Masking the Join and Syntax Errors?

The Observation: Errors like **Join error (8.3%)** and **Syntax error (7.4%)** now appear much smaller when compared to the top errors. However, under the original angular visualization, these errors might have been visually distorted or given undue weight due to their position or color.

The Insight and Question: While the improved visualization accurately minimizes the perceived severity of these two errors, we must still ask: Why do basic structural errors like Syntax and Join failures still persist, even if at lower rates? These are foundational SQL elements. The continued presence of these errors suggests a **lack of fundamental language constraint mastery**, which must be addressed to ensure robust, production-ready SQL generation, regardless of the percentage.

5. Conclusion and Future Direction

This paper demonstrates a systematic methodology for optimizing a problematic, multi-level visualization to better facilitate data comprehension and analysis. By providing a detailed, step-by-step explanation of the improvement process—including the shift to length encoding, prioritized sorting, and the use of hierarchical total display—this work guides the reader through the key cognitive and design considerations essential for creating effective visualizations. The comparison between the initial nested doughnut chart and the final refined bar chart clearly highlights how these changes drastically reduce reader friction and allow for the extraction of specific, actionable insights that were previously obscured by the original chart's complexity. Through this process, this paper strongly emphasizes the **importance of scientifically designed visualizations in data science**, underscoring the value of effective visual communication for accurate insight extraction and deeper analysis.

6. Reference

- https://arxiv.org/abs/2411.07763
- https://onlinelibrary.wiley.com/doi/10.1207/s15516709cog1202_4
- https://dl.acm.org/doi/10.1145/102377.115768
- https://aclanthology.org/D18-1425/