

Text-to-SQL as Dual-State Reasoning: Integrating Adaptive Context and Progressive Generation

Zhifeng Hao^{1,2} Qibin Song¹ Ruichu Cai^{1,3} Boyan Xu^{1*}

¹School of Computer Science, Guangdong University of Technology

²College of Science, Shantou University

³Peng Cheng Laboratory

haozhifeng@stu.edu.cn

{gdutsongqibin, cairuichu, hpakyim}@gmail.com

Abstract

Recent divide-and-conquer reasoning approaches, particularly those based on Chain-of-Thought (CoT), have substantially improved the Text-to-SQL capabilities of Large Language Models (LLMs). However, when applied to complex enterprise databases, such methods struggle to maintain coherent reasoning due to limited context capacity, unreliable schema linking, and weak grounding in database semantics. To overcome these issues, we introduce DSR-SQL, a **Dual-State Reasoning** framework that models Text-to-SQL as an interaction between an adaptive context state and a progressive generation state. The first constructs a compact, semantically faithful environment by refining large schemas and selecting relevant structures, while the second formalizes SQL synthesis as feedback-guided state transitions, enabling the model to self-correct and align with user intent. Without any post-training or in-context examples, DSR-SQL achieves competitive performance, reaching 35.28% execution accuracy on Spider 2.0-Snow and 68.32% on BIRD development set. Our implementation will be open-sourced at: <https://github.com/DMIRLAB-Group/DSR-SQL>.

1 Introduction

Text-to-SQL enables non-technical users to access and analyze structured data through natural language, bridging the gap between human intent and relational databases (Qin et al., 2022; Sui et al., 2023; Chen et al., 2025; Liu et al., 2025a). In enterprise environments, this capability forms the core of many business intelligence systems—powering ad-hoc querying, dashboard generation, and multi-table analytics that support operational and strategic decision-making.

Existing Text-to-SQL benchmarks such as Spider capture only moderate schema complexity and thus fail to reflect the realities of enterprise-scale

databases, which often span hundreds of interdependent tables and evolve continuously over time. More recent corpora like Spider 2.0 (Lei et al., 2024) and BIRD (Li et al., 2023) reveal the key bottleneck in such settings: as schema size and domain heterogeneity grow, LLMs struggle to maintain accurate *query-schema alignment* within a limited context window. Without sufficient grounding in database semantics, a single-pass generation process tends to produce SQL that is syntactically correct but semantically misaligned with user intent. Overcoming this misalignment in a zero-shot setting—without labeled data or task-specific tuning—remains an open challenge for both research and practical deployment (Chen et al., 2021; Liu et al., 2023; Dong et al., 2023; Li et al., 2025a).

Prior work has explored two main strategies to handle large and complex schemas. The first focuses on *schema compression and retrieval*, where methods such as retrieval-augmented generation or locality-sensitive hashing prune the schema to fit within model context limits (Talaei et al., 2024; Liu et al., 2025b). While these techniques improve efficiency, they introduce irreversible information loss: once a relevant table or column is omitted, the model can no longer recover the missing linkage during generation. A complementary line of research attempts to *decompose* complex questions into sub-problems through static CoT prompting (Pourreza and Rafiei, 2023; Xie et al., 2024). As illustrated in Figure 1, however, such decompositions remain detached from the true database content—each reasoning step proceeds without feedback, resulting in schema mismatches and execution failures.

Both directions reveal a deeper limitation: existing approaches treat reasoning as a one-way process. The model’s *context state* (what it sees) and *generation state* (how it reasons) evolve independently, preventing effective correction or adaptation. Bridging this divide requires a framework

* Corresponding author, hpakyim@gmail.com

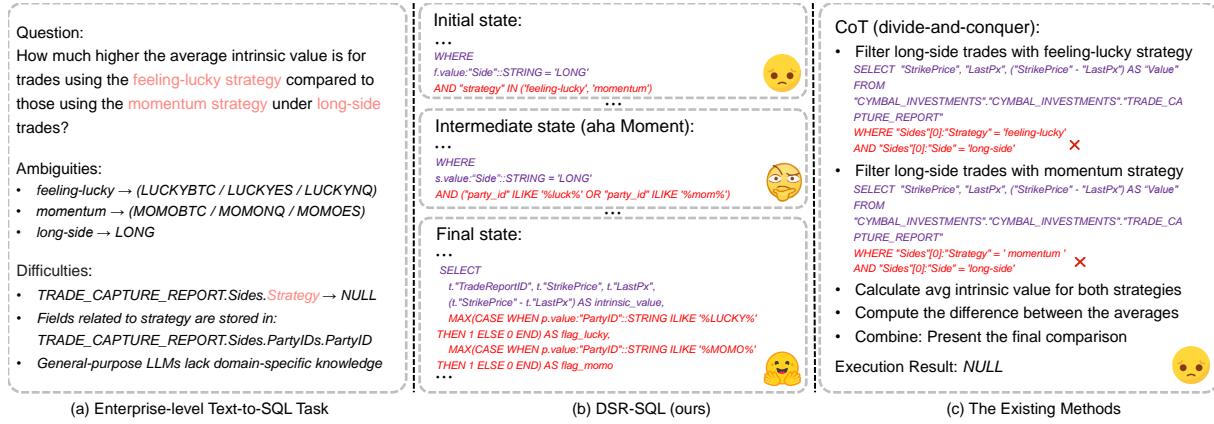


Figure 1: A motivating example illustrating the need for dual-state reasoning. Panel (a) shows an enterprise-level Text-to-SQL query containing domain-specific terminology from quantitative trading, such as *feeling-lucky strategy*, *momentum strategy*, and *long-side trades*. The database, however, stores these entities in entirely different formats. Panel (c) demonstrates the failure of static, CoT-based divide-and-conquer reasoning, which splits the problem linguistically but remains unaware of the true schema, resulting in execution failure. In contrast, panel (b) shows DSR-SQL, which leverages execution feedback to revise intermediate states, probe the database when mismatches occur, and ultimately reach an “aha moment” that correctly aligns user intent with database structure.

where these two states interact dynamically through feedback—motivating the design of our dual-state reasoning paradigm.

Building on this insight, we introduce **DSR-SQL**, a *dual-state reasoning* framework that reconceptualizes Text-to-SQL as the interaction between an *adaptive context state* (what the model can see) and a *progressive generation state* (how it reasons). Rather than aligning the full schema in a single pass, DSR-SQL incrementally refines both states through execution signals: it condenses and selects schema fragments to maintain a compact, task-relevant environment, and it evolves partial SQL hypotheses based on their results. Complementing the failure case already shown in Figure 1 (panel c), Figure 1 (panel b) illustrates how DSR-SQL uses execution feedback to revise intermediate states, actively probe the database when mismatches are detected, and converge to the correct query. This dual-state formulation turns Text-to-SQL from static mapping into grounded, feedback-conditioned reasoning. In contrast to the failure in Figure 1 (panel c), panel (b) shows DSR-SQL revising states via execution feedback and converging to the correct query. Our main contributions can be summarized as follows:

- We propose **DSR-SQL**, a novel *dual-state reasoning* framework that reformulates Text-to-SQL as an interactive process between adaptive context management and feedback-guided generation. This formulation enables LLMs to dynami-

cally couple perception and reasoning rather than relying on a static schema view.

- DSR-SQL introduces two complementary reasoning loops: a context-state loop that incrementally refines database visibility through schema condensation and adaptive selection, and a generation-state loop that evolves partial SQLs under execution feedback. Their co-evolution allows robust query–schema alignment even under enterprise-scale complexity.
- In a fully *zero-shot* setting—without Post-training or in-context exemplars—DSR-SQL achieves strong execution accuracy on Spider 2.0-Snow and BIRD, demonstrating that test-time dual-state reasoning can match or surpass methods requiring additional supervision.

2 Related Work

The Text-to-SQL task, aiming to convert natural language into executable SQL statements, has evolved from early sequence-to-sequence models (Cai et al., 2018; Qi et al., 2022) and pre-trained architectures (Wang et al., 2021; Cai et al., 2021) to the current paradigm dominated by LLMs (Pourreza et al., 2024). Despite these advantages, LLMs, particularly in zero-shot scenarios, often struggle with complex queries due to schema misinterpretation and logical reasoning errors (Li et al., 2023). To mitigate this, recent zero-shot methods have moved beyond the limitations of static Chain-of-Thought prompting, instead leveraging execution

feedback for post-hoc correction and validation of generated SQL candidates (Pourreza and Rafiei, 2023; Xie et al., 2024; Lee et al., 2024). However, these methods share a fundamental limitation: they primarily treat feedback as an external signal for post-hoc optimization or validation of a complete SQL candidate (Qu et al., 2025). In contrast, DSR-SQL deeply integrates feedback into the iterative generation process itself, using it not just to correct errors, but to actively explore and understand the database’s structure, thereby achieving a more fundamental alignment between user intent and database semantics.

3 Method

In this section, we introduce **DSR-SQL**, a structured framework designed to enhance the Text-to-SQL capabilities of general-purpose LLMs on complex enterprise-level databases. DSR-SQL reformulates the task as a two-state process that alternates between *context construction* and *generation evolution*, enabling the model to reason effectively under limited context windows. We begin with a formal problem formulation and then detail the function of each component within the DSR-SQL pipeline, as illustrated in Figure 2.

3.1 Problem Formulation

Given a natural-language question \mathcal{Q} , a database schema $\mathcal{S} = (\mathcal{T}, \mathcal{C})$, and optional external knowledge \mathcal{K} , the goal of Text-to-SQL is to generate a SQL query \mathcal{Y} whose execution result is equivalent to that of the reference query on the same database.

To handle large-scale databases under context-length constraints, DSR-SQL divides the process into two interacting components.

(1) Context-State Construction. Before generation, the model operates within a manageable *context state* $\mathcal{E} = (\mathcal{S}', \mathcal{K}')$ that defines the visible schema and knowledge. DSR-SQL constructs this state by refining lengthy inputs, selecting question-relevant schema elements, and aligning linguistic mentions with database semantics. This compact environment retains the essential information needed for accurate generation.

(2) Generation-State Evolution. Given \mathcal{E} , the model generates SQL through an iterative reasoning process that maintains a *generation state* $s_t = (\mathcal{Y}_t, \mathcal{R}_t)$, where \mathcal{Y}_t is a partial query and \mathcal{R}_t is its execution feedback. The state is updated step

by step until a valid and executable query is obtained, progressively aligning the predicted SQL with user intent and database semantics.

3.2 Context-State Construction

3.2.1 Schema and Knowledge Refinement

As the first stage of **Context-State Construction**, DSR-SQL refines large-scale database schemas and verbose external knowledge to build a compact yet semantically faithful environment state $(\mathcal{S}', \mathcal{K}')$. This step ensures that subsequent reasoning operates within the context capacity of LLMs while preserving essential information.

Schema Refinement. We reorganize the schema by identifying and consolidating **structurally equivalent tables**—those that share nearly identical column layouts but differ only in versioning or temporal suffixes (e.g., GA_SESSIONS_20160801–GA_SESSIONS_20170801). Each series is represented by a single canonical table, accompanied by a concise annotation automatically generated by an LLM to describe the merged pattern (see Appendix A). Columns that contain consistently null or uninformative values are pruned, yielding a reduced schema $\mathcal{S}' = (\mathcal{T}', \mathcal{C}')$ with $|\mathcal{T}'| \ll |\mathcal{T}|$.

Knowledge Refinement. Similarly, verbose external knowledge \mathcal{K} is summarized into \mathcal{K}' through LLM-based extraction and synthesis, retaining only information directly relevant to the question—such as units, enumerations, or temporal ranges.

As shown in Figure 3, most refined M-schemas after this reduction fall below 128K tokens, enabling efficient context utilization for later stages of Context-State Construction.

3.2.2 Adaptive Schema Selection

As the second stage of **Context-State Construction**, DSR-SQL dynamically selects a subset of the refined schema that is most relevant to the input question \mathcal{Q} . Given the compressed schema \mathcal{S}' , the model identifies a task-specific subset $\mathcal{S}_{\text{sub}} \subseteq \mathcal{S}'$ through an adaptive exploration strategy that balances coverage and efficiency.

Global Exploration. When the schema size remains within the context limit, the model performs global exploration: an LLM infers potential SQL structures over the full schema and collects tables that appear across high-confidence candidates as the relevant subset \mathcal{S}_{sub} .

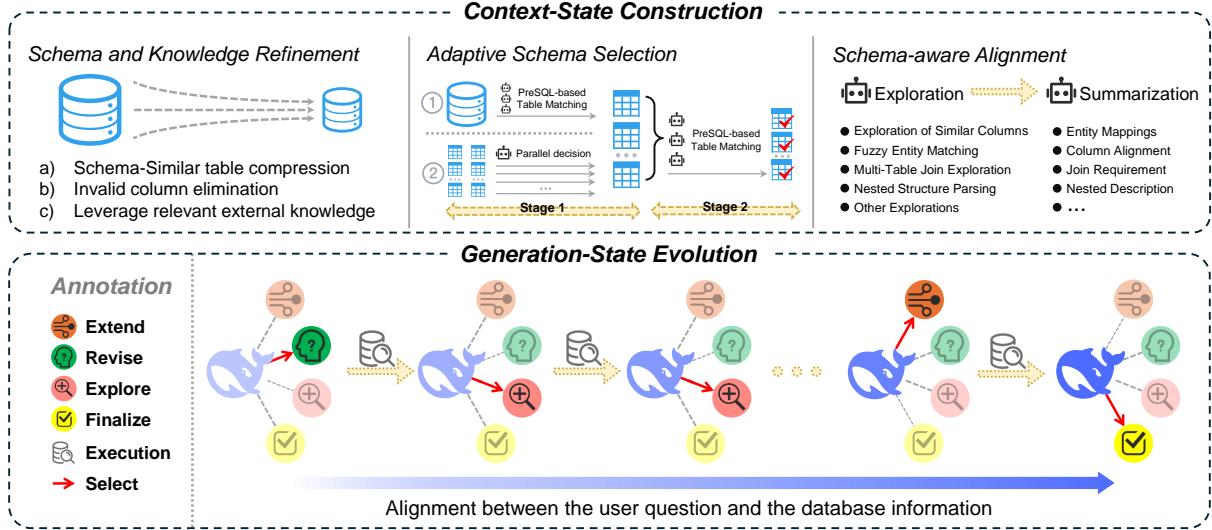


Figure 2: Overview of the DSR-SQL framework for enterprise-level Text-to-SQL. DSR-SQL consists of four components: Schema and Knowledge Refinement to handle large-scale inputs, Adaptive Schema Selection to identify relevant tables, Schema-aware Alignment to bridge linguistic and database semantics, and Generation-State Evolution guided by execution feedback to iteratively align user intent with query logic.



Figure 3: Schema token counts (M-Schema format) for selected databases in Spider 2.0-Snow. The comparison shows the original schema size (orange), the size after *Schema and Knowledge Refinement* (green), and the average per-query size after *Adaptive Schema Selection* (blue-and-white striped).

Partitioned Exploration. For larger databases that exceed the context window, DSR-SQL adopts a partitioned exploration strategy to ensure robust table selection under varying schema scales. It evaluates each table independently through query-table similarity prompts and retains only those predicted to be relevant.

Through this adaptive selection process, most per-query database schemas are thus reduced to under 32K tokens (see Figure 3), greatly alleviating the context burden. The detailed procedure is provided in Appendix E as pseudocode.

3.2.3 Schema-aware Alignment

The final stage of **Context-State Construction** establishes an initial semantic grounding between the natural-language question and the selected schema \mathcal{S}_{sub} . To obtain a content-aware alignment prior, DSR-SQL performs a two-phase procedure of lightweight exploration and summarization.

Exploration. An LLM first generates a small set of simple *probing queries* (e.g., `SELECT COUNT(*)`, `SELECT DISTINCT`) to inspect key value distributions and entity relationships within \mathcal{S}_{sub} . The execution results of these queries reveal frequent values, typical ranges, and common join patterns.

Summarization. The obtained results are then synthesized by another LLM prompt into a concise description $\mathcal{K}_{\text{align}}$ that captures salient domain information (e.g., enumerations, units, synonyms). This alignment prior provides the generator with explicit semantic cues, helping it avoid redundant or inefficient reasoning steps in the subsequent generation stage (see Appendix C for an illustration).

3.3 Generation-State Evolution

After constructing the contextual environment, DSR-SQL performs **generation-state evolution**, an iterative reasoning process guided by execution feedback. Rather than generating the entire SQL in a single pass, the model maintains a generation state $s_t = (\mathcal{Y}_t, \mathcal{R}_t)$, where \mathcal{Y}_t is a partial query and \mathcal{R}_t is its execution result. At each step, this state

evolves based on feedback signals and the prepared context $\mathcal{X}' = (\mathcal{Q}, \mathcal{S}_{\text{sub}}, \mathcal{K}', \mathcal{K}_{\text{align}})$.

This process can be conceptualized as a state-driven reasoning procedure governed by four recurrent actions, each triggered by distinct feedback conditions:

Extend. When the current execution \mathcal{R}_t yields valid and informative results, the model extends the partial query to handle the next aspect of the user question, progressively composing a more complete SQL structure.

Revise. If execution feedback reveals a semantic or logical inconsistency, the model revises the existing partial query, correcting the reasoning path based on error signals from the database.

Explore. When execution produces unexpected or empty outputs, the model issues short exploratory queries to probe relevant value ranges or relational connections before resuming the main reasoning path.

Finalize. Once all sub-goals have been resolved and execution outcomes align with the user’s intent, the model synthesizes a coherent and executable SQL query, terminating the evolution process.

Through these feedback-driven transitions, DSR-SQL progressively aligns natural-language intent with database semantics, forming a directed, acyclic reasoning trajectory that converges to the final SQL query. The detailed prompting design and pseudocode are provided in Appendix F.

4 Experiments

4.1 Benchmarks

To comprehensively evaluate the performance of our proposed DSR-SQL framework, we conducted extensive experiments on two widely-used, cross-domain Text-to-SQL benchmarks: Spider 2.0(Lei et al., 2024) and BIRD(Li et al., 2023). Spider 2.0 is an emerging and highly challenging benchmark designed to simulate real-world, enterprise-level application scenarios. Its database schemas are exceptionally complex, averaging approximately 800 columns per database, and the required SQL queries present a formidable challenge in terms of complexity. For this evaluation, we focus on its sub-task, Spider 2.0-Snow, which requires generating SQL in the Snowflake dialect. In contrast, the BIRD benchmark places a greater emphasis on understanding fine-grained schema details, such as

distinguishing between similarly named columns and handling minor typographical errors in questions, making it equally challenging.

4.2 Evaluation Metrics

Following the standard in the field, we adopt the widely recognized Execution Accuracy (EX) as our core performance metric. It is important to note that the definition of EX actually differs between the two benchmarks. BIRD employs a strict evaluation standard, requiring the execution results of a generated query to be an exact match with the gold standard in both content and column order. Spider 2.0-Snow, however, utilizes a more lenient evaluation script. Acknowledging the potential ambiguity in user intent, its evaluation only requires that the returned results contain all the core information of the gold answer, without constraining the order of columns or the presence of any additional ones, thereby minimizing false negatives in the evaluation process.

4.3 Implementation Details

Our experiments employ several LLMs from the DeepSeek and Qwen series, including DeepSeek-R1-0528 (DS-R1), DeepSeek-V3-0324 (DS-V3), DeepSeek-V3.1 (DS-V3.1), and Qwen3-30B-A3B-Instruct-2507 (Qwen3-30B). Unless otherwise specified, the decoding temperature was set to 0.2. For the Generation-State Evolution, the maximum number of iterations is set to 20 for Snow and 10 for BIRD development datasets, reflecting the differing query complexities of the datasets. In addition, a query correction loop (up to 5 iterations) following Pourreza et al. (2024) was applied to all methods to fix minor syntactic errors. For more details on the comparison with baselines, see Appendix G.

4.4 Comparison Results

Results on Spider2-Snow Table 1 shows the performance comparison on the Snow benchmark. With a single generation path, DSR-SQL paired with DS-R1 achieves an EX of 35.28%. This result represents a 6.03% absolute improvement over ReFORCE, which relies on an 8-path voting mechanism. Notably, our framework with the DS-V3 model also surpasses ReFORCE paired with the much stronger GPT-4o model (20.84%). These findings demonstrate the effectiveness of the DSR-SQL framework in handling complex databases without requiring multi-path generation. We con-

Method	Voting	EX (%)
Spider-Agent + GPT-4o	✗	10.05
Spider-Agent + DS-R1	✗	10.79
ReFORCE + GPT-4o	✓	20.84
ReFORCE + DS-R1	✓	29.25
DSR-SQL + DS-V3	✗	22.67
DSR-SQL + DS-R1	✗	35.28

Table 1: Comparison on Spider 2.0-Snow. The "Voting" column indicates whether multi-path voting was used: a ✓ denotes that voting was applied, while a ✗ denotes that it was not.

duct a detailed error analysis on the Snow benchmark; see Appendix H for details.

Results on BIRD In Table 2, we provide a detailed performance comparison of DSR-SQL against mainstream Text-to-SQL frameworks on the BIRD development set. To ensure a fair assessment of each method’s foundational performance, the number of samples was uniformly set to one for all baselines.

The results demonstrate that, under a zero-shot setting, DSR-SQL achieves strong execution accuracies of 68.32% and 65.58% when applied to two leading LLMs (DS-V3.1 and Qwen3-30B). This performance surpasses all existing zero-shot methods and is comparable to—or even exceeds—most approaches based on Post-training and in-context examples. While a gap remains compared to certain hybrid methods, it is important to note that those approaches often depend on closed-source models and proprietary Post-training datasets, which substantially increase computational costs and raise potential data privacy concerns.

4.5 Ablation and Module Analysis

4.5.1 Overall Ablation Study

We performed an ablation study on the Spider 2.0-Snow benchmark based on the DS-R1 model (see Table 3) to measure each module’s contribution to DSR-SQL. The results show that every module is essential. The Schema and Knowledge Refinement (SKR) module is foundational, tackling the core challenge of limited context capacity by condensing oversized schemas. Building on this, the Adaptive Schema Selection (ASS) and Schema-aware Alignment (SAA) components are crucial for con-

structing a compact, semantically faithful environment by selecting relevant structures and grounding linguistic terms in database semantics. Omission of any module produces measurable performance drops of varying magnitudes. Notably, the Generation-State Evolution (GSE) module leverages this carefully constructed context state, using feedback-guided state transitions to incrementally align user intent with the database. This approach attains a higher performance ceiling than conventional divide-and-conquer methods, corroborating the effectiveness of our dual-state design.

4.5.2 Analysis of Adaptive Schema Selection

We evaluate the efficiency of our Adaptive Schema Selection approach across multiple dimensions, with results presented in Table 4. On the challenging Snow dataset, our ASS module achieves a recall of 91.13%, a significant improvement over the 62.59% achieved by the matching method in ReFoRCE. This higher accuracy comes at the cost of increased token consumption (34.35K vs. 23.23K). However, our approach reduces the average number of LLM calls from 13.29 to 6.06, indicating a more efficient reasoning process. On the BIRD dataset, ASS achieves over 98% recall with both tested models, demonstrating its robustness.

4.5.3 Analysis of Schema-aware Alignment

We conducted an ablation study on the Schema-aware Alignment (SAA) module to assess its contribution to the overall framework. To ensure a fair and consistent comparison, we reused the intermediate results from the Exploration phase across all relevant experimental setups. The results are presented in Table 5.

The analysis reveals several key findings. First, removing the SAA module entirely results in a substantial performance drop of approximately 3.5%. Second, providing the model with only the raw (`SQL`, `execution result`) pairs from the Exploration phase yields performance considerably lower than that of the full framework. This suggests that without the Summarization step, the LLM struggles to extract deep insights from the long-context, unprocessed data. Third, using Summarization alone, without the data-grounded context from Exploration, leads to a significant performance drop and can even be detrimental, yielding results comparable to a simple divide-and-conquer approach (see Appendix D). This is likely due to model hallucinations caused by the lack of concrete, exploratory

Method	EX (%)			
	Simple	Moderate	Challenging	Total
<i>Hybrid</i>				
CHASE-SQL + Gemini	—	—	—	74.90
XiYan-SQL + Multi-model	—	—	—	73.34
AskData + GPT4o	—	—	—	75.36
Contextual-SQL	—	—	—	73.50
<i>Post-training</i>				
XiYanSQL-QwenCoder-32B	73.62	60.34	51.72	67.54
Arctic-Text2SQL-R1-7B	73.62	62.50	57.24	68.71
Databricks-RLVR-32B	—	—	—	70.80
OmniSQL-32B	71.78	54.31	50.34	64.47
<i>In-context Learning</i>				
DAIL-SQL + GPT4	62.49	43.44	38.19	54.43
MCS-SQL + GPT4	70.40	53.10	51.40	63.40
RSL-SQL + DS-V3.1	72.11	59.27	53.79	66.49
OpenSearch-SQL + DS-V3.1	74.05	62.07	57.24	68.84
<i>Zero-shot</i>				
DIN-SQL + GPT4	—	—	—	50.72
MAC-SQL + GPT4	—	—	—	55.54
MAG-SQL + DS-V3.1	71.24	59.27	55.17	66.10
DS-V3.1	64.54	50.43	49.66	58.87
Qwen3-30B	64.00	48.06	38.62	56.78
DSR-SQL + Qwen3-30B	70.70	59.27	53.10	65.58
DSR-SQL + DS-V3.1	72.65	61.21	63.45	68.32

Table 2: Comparison of Methods on the BIRD Development dataset.

Method	EX (%)
DSR-SQL (Full)	35.28
DSR-SQL w/o SKR	22.48 (\downarrow 12.80)
DSR-SQL w/o ASS	24.49 (\downarrow 10.79)
DSR-SQL w/o SAA	31.80 (\downarrow 3.48)
DSR-SQL w/o GSE	27.79 (\downarrow 7.49)

Table 3: Ablation study of different components in DSR-SQL on the Spider 2.0-Snow dataset.

information.

Notably, even a baseline divide-and-conquer method improves when augmented with our Schema-aware Alignment module, further underscoring the positive guiding role the module plays in SQL generation.

4.5.4 Analysis of Generation-State Evolution

The Generation-State Evolution module plays a pivotal role in DSR-SQL, enabling dynamic alignment with and understanding of complex databases to facilitate accurate SQL generation. We conducted

a corresponding ablation study on both the Spider 2.0-Snow and BIRD development datasets to quantify its impact.

The results demonstrate that, as shown in Table 6, compared to a static divide-and-conquer approach, our dynamic method holds a clear advantage on complex, large-scale databases like Spider 2-Snow, with a performance improvement of approximately 7.49%. On the BIRD development dataset, our method also exhibits superior performance in resolving ambiguous and challenging user queries, highlighting the generalizability of our approach.

Furthermore, we analyzed the accuracy rates across different generation paths within the SQL Generation module. We categorize the generation process into three distinct path types (note: this classification is independent of the dataset's official difficulty ratings):

- **Straightforward Path:** Paths that construct the final SQL using only **Extend** and **Finalize** actions. This indicates that the LLM has high confidence in its solution and requires no additional information.

Dataset	Method	Model	Precision (%)	Recall (%)	Avg. Tokens	Avg. LLM Calls
Snow	ReFORCE	DS-V3	58.09	62.59	23.23K	13.29
	DSR-SQL	DS-V3	75.62	91.13	34.35K	6.06
BIRD	DSR-SQL	DS-V3.1	63.07	98.41	3.67K	1.00
	DSR-SQL	Qwen3-30B	64.59	98.83	3.80K	1.00

Table 4: Performance Analysis of Adaptive Schema Selection. We set the model temperature to 1.2 and generate three samples to improve recall. ReFoRCE defaults to using the full schema, invoking its table matching method only when the context exceeds 200K tokens. Thus, its matching cost is low but has minimal impact on its final SQL accuracy. Token counts may vary slightly across models.

Method	EX (%)
DSR-SQL (Full)	35.28
DSR-SQL w/o SAA	31.8 (\downarrow 3.48)
DSR-SQL w/o Summarization	32.35 (\downarrow 2.93)
DSR-SQL w/o Exploration	28.70 (\downarrow 6.58)
divide-and-conquer w SAA	27.79 (\downarrow 7.49)
divide-and-conquer w/o SAA	25.05 (\downarrow 10.23)

Table 5: Evaluating the performance impact of the Schema-aware Alignment module on SQL generation on the Spider2.0-Snow dataset.

Dataset	Method	Model	EX (%)
Snow	divide-and-conquer	DS-R1	27.79
	DSR-SQL		35.28
BIRD	divide-and-conquer	DS-V3.1	63.17
	DSR-SQL		68.32

Table 6: Performance Analysis of the Generation-State Evolution module, where divide-and-conquer utilizes the Schema-aware Alignment module.

- **Refinement Path:** Paths composed of a combination of **Extend**, **Revise**, and **Finalize** actions. This signifies that intermediate execution results may have partially deviated from the user’s requirements.
- **Exploratory Path:** Paths that utilize the **Explore** action. This suggests that the user’s question is particularly difficult or ambiguous, necessitating in-depth database exploration to formulate the correct SQL.

The detailed results are shown in Figure 4. DSR-SQL consistently outperforms across all path types, with the largest gains on the **Exploratory Path**, improving relative accuracy by 14.95% and 20.64% on Spider 2-Snow and BIRD, respectively.

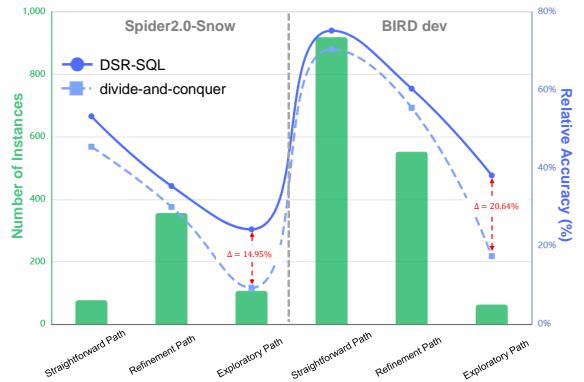


Figure 4: Accuracy comparison of the SQL Generation module by generation path type on the Spider 2.0-Snow and BIRD dev datasets, with divide-and-conquer making use of the Schema-aware Alignment module.

5 Conclusions

In this work, we introduced DSR-SQL, a dual-state reasoning framework that models Text-to-SQL as the interaction between an adaptive context state and a progressive generation state. This formulation enables LLMs to maintain coherent reasoning over complex enterprise databases, addressing the challenges of context limitation and schema grounding. Experiments on Spider 2.0-Snow and BIRD demonstrate that DSR-SQL achieves strong zero-shot performance without Post-training or in-context examples. Beyond these results, our findings suggest that balancing how models manage external context and evolve internal reasoning states can be an effective path toward more reliable structured reasoning. We hope this perspective encourages further exploration into state-aware reasoning frameworks across other data-intensive tasks.

Limitations

While our approach demonstrates superior performance over current SOTA methods on Spider 2.0-

Snow and BIRD in a zero-shot scenario, it still lags behind certain techniques that leverage Post-training, in-context examples, or multi-path generation. Addressing this performance disparity will be a key focus of future work.

References

- Sheshansh Agrawal and Thien Nguyen. 2025. [Open-sourcing the best local text-to-sql system](#).
- Alnur Ali, Ashutosh Baheti, Jonathan Chang, Ta-Chung Chi, Brandon Cui, Andrew Drozdov, Jonathan Franklin, Abhay Gupta, Pallavi Koppol, Sean Kulinski, Jonathan Li, Dipendra Misra, Krista Opsahl-Ong, Jose Javier Gonzalez Ortiz, Matei Zaharia, and Yue Zhang. 2025. [A state-of-the-art sql reasoning model using rlrvr](#). *Preprint*, arXiv:2509.21459.
- Ruichu Cai, Boyan Xu, Xiaoyan Yang, Zhenjie Zhang, Zijian Li, and Zhihao Liang. 2018. [An encoder-decoder framework translating natural language to database queries](#). *Preprint*, arXiv:1711.06061.
- Ruichu Cai, Jinjie Yuan, Boyan Xu, and Zhifeng Hao. 2021. Sadga: Structure-aware dual graph aggregation network for text-to-sql. *Advances in Neural Information Processing Systems*, 34.
- Zhenbiao Cao, Yuanlei Zheng, Zhihao Fan, Xiaojin Zhang, and Wei Chen. 2024. Rsl-sql: Robust schema linking in text-to-sql generation. *arXiv preprint arXiv:2411.00073*.
- Albert Chen, Manas Bunde, Gaurav Ahlawat, Patrick Stetz, Zhitao Wang, Qiang Fei, Donghoon Jung, Audrey Chu, Bharadwaj Jayaraman, Ayushi Panth, Yatin Arora, Sourav Jain, Renjith Varma, Alexey Ilin, Iuliia Melnychuk, Chelsea Chueh, Joyan Sil, and Xiaofeng Wang. 2025. [Text-to-sql for enterprise data analytics](#). *Preprint*, arXiv:2507.14372.
- Yongrui Chen, Xinnan Guo, Chaojie Wang, Jian Qiu, Guilin Qi, Meng Wang, and Huiying Li. 2021. [Leveraging table content for zero-shot text-to-sql with meta-learning](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5):3992–4000.
- Minghang Deng, Ashwin Ramachandran, Canwen Xu, Lanxiang Hu, Zhewei Yao, Anupam Datta, and Hao Zhang. 2025. [Reforce: A text-to-sql agent with self-refinement, consensus enforcement, and column exploration](#). *Preprint*, arXiv:2502.00675.
- Xuemei Dong, Chao Zhang, Yuhang Ge, Yuren Mao, Yunjun Gao, lu Chen, Jinshu Lin, and Dongfang Lou. 2023. [C3: Zero-shot text-to-sql with chatgpt](#). *Preprint*, arXiv:2307.07306.
- Dawei Gao, Haibin Wang, Yaliang Li, Xiuyu Sun, Yichen Qian, Bolin Ding, and Jingren Zhou. 2023. [Text-to-sql empowered by large language models: A benchmark evaluation](#). *CoRR*, abs/2308.15363.
- Dongjun Lee, Choongwon Park, Jaehyuk Kim, and Heesoo Park. 2024. [Mcs-sql: Leveraging multiple prompts and multiple-choice selection for text-to-sql generation](#). *Preprint*, arXiv:2405.07467.
- Fangyu Lei, Jixuan Chen, Yuxiao Ye, Ruisheng Cao, Dongchan Shin, Hongjin Su, Zhaoqing Suo, Hongcheng Gao, Wenjing Hu, Pengcheng Yin, Victor Zhong, Caiming Xiong, Ruoxi Sun, Qian Liu, Sida Wang, and Tao Yu. 2024. [Spider 2.0: Evaluating language models on real-world enterprise text-to-sql workflows](#). *Preprint*, arXiv:2411.07763.
- Boyan Li, Jiayi Zhang, Ju Fan, Yanwei Xu, Chong Chen, Nan Tang, and Yuyu Luo. 2025a. [Alpha-sql: Zero-shot text-to-sql using monte carlo tree search](#). *Preprint*, arXiv:2502.17248.
- Haoyang Li, Shang Wu, Xiaokang Zhang, Xinmei Huang, Jing Zhang, Fuxin Jiang, Shuai Wang, Tieying Zhang, Jianjun Chen, Rui Shi, Hong Chen, and Cuiping Li. 2025b. [Omnisql: Synthesizing high-quality text-to-sql data at scale](#). *Preprint*, arXiv:2503.02240.
- Jinyang Li, Binyuan Hui, Ge Qu, Binhu Li, Jiaxi Yang, 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. 2023. [Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls](#). *Preprint*, arXiv:2305.03111.
- Aiwei Liu, Xuming Hu, Lijie Wen, and Philip S. Yu. 2023. [A comprehensive evaluation of chatgpt’s zero-shot text-to-sql capability](#). *Preprint*, arXiv:2303.13547.
- Xinyu Liu, Shuyu Shen, Boyan Li, Peixian Ma, Runzhi Jiang, Yuxin Zhang, Ju Fan, Guoliang Li, Nan Tang, and Yuyu Luo. 2025a. [A survey of text-to-sql in the era of llms: Where are we, and where are we going?](#) *IEEE Transactions on Knowledge and Data Engineering*.
- Yifu Liu, Yin Zhu, Yingqi Gao, Zhiling Luo, Xiaoxia Li, Xiaorong Shi, Yuntao Hong, Jinyang Gao, Yu Li, Bolin Ding, and Jingren Zhou. 2025b. [Xiyuan-sql: A novel multi-generator framework for text-to-sql](#). *Preprint*, arXiv:2507.04701.
- Mohammadreza Pourreza, Hailong Li, Ruoxi Sun, Yeounoh Chung, Shayan Talaei, Gaurav Tarlok Kakkar, Yu Gan, Amin Saberi, Fatma Ozcan, and Sercan O. Arik. 2024. [Chase-sql: Multi-path reasoning and preference optimized candidate selection in text-to-sql](#). *Preprint*, arXiv:2410.01943.
- Mohammadreza Pourreza and Davood Rafiei. 2023. [Din-sql: Decomposed in-context learning of text-to-sql with self-correction](#). *Preprint*, arXiv:2304.11015.
- Jiexing Qi, Jingyao Tang, Ziwei He, Xiangpeng Wan, Yu Cheng, Chenghu Zhou, Xinbing Wang, Quanshi Zhang, and Zhouhan Lin. 2022. [Rasat: Integrating](#)

relational structures into pretrained seq2seq model for text-to-sql. *Preprint*, arXiv:2205.06983.

Bowen Qin, Binyuan Hui, Lihan Wang, Min Yang, Jinyang Li, Binhu Li, Ruiying Geng, Rongyu Cao, Jian Sun, Luo Si, Fei Huang, and Yongbin Li. 2022. A survey on text-to-sql parsing: Concepts, methods, and future directions. *Preprint*, arXiv:2208.13629.

Ge Qu, Jinyang Li, Bowen Qin, Xiaolong Li, Nan Huo, Chenhao Ma, and Reynold Cheng. 2025. Share: An slm-based hierarchical action correction assistant for text-to-sql. *arXiv preprint arXiv:2506.00391*.

Vladislav Shkapenyuk, Divesh Srivastava, Theodore Johnson, and Parisa Ghane. 2025. Automatic metadata extraction for text-to-sql. *Preprint*, arXiv:2505.19988.

Guanghu Sui, Zhishuai Li, Ziyue Li, Sun Yang, Jingqing Ruan, Hangyu Mao, and Rui Zhao. 2023. Re-boost large language model-based text-to-sql, text-to-python, and text-to-function – with real applications in traffic domain. *Preprint*, arXiv:2310.18752.

Shayan Talaei, Mohammadreza Pourreza, Yu-Chen Chang, Azalia Mirhoseini, and Amin Saberi. 2024. Chess: Contextual harnessing for efficient sql synthesis. *Preprint*, arXiv:2405.16755.

Bailin Wang, Richard Shin, Xiaodong Liu, Oleksandr Polozov, and Matthew Richardson. 2021. Rat-sql: Relation-aware schema encoding and linking for text-to-sql parsers. *Preprint*, arXiv:1911.04942.

Bing Wang, Changyu Ren, Jian Yang, Xinnian Liang, Jiaqi Bai, LinZheng Chai, Zhao Yan, Qian-Wen Zhang, Di Yin, Xing Sun, and Zhoujun Li. 2025. Mac-sql: A multi-agent collaborative framework for text-to-sql. *Preprint*, arXiv:2312.11242.

Wenxuan Xie, Gaochen Wu, and Bowen Zhou. 2024. Mag-sql: Multi-agent generative approach with soft schema linking and iterative sub-sql refinement for text-to-sql. *Preprint*, arXiv:2408.07930.

Xiangjin Xie, Guangwei Xu, Lingyan Zhao, and Ruijie Guo. 2025. Opensearch-sql: Enhancing text-to-sql with dynamic few-shot and consistency alignment. *Preprint*, arXiv:2502.14913.

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. *Preprint*, arXiv:2505.20315.

A Table description generation

For the tables that underwent Schema Refinement, we used an LLM to generate a simple table description. Specifically, we fed all table names and their corresponding column descriptions for a set of structurally equivalent tables into a prompt, instructing the LLM to generate a concise description

that summarizes them. An example is provided in Figure 5.

Database Name: GA360

The naming convention for these tables with the same structure is:

1. `GA_SESSIONS_[DATE]`, where [DATE] represents the date of the session data in ‘YYYYMM-MDD’ format, **ranging from 20160801 to 20170801**, covering a full 13-month period from August 1, 2016, to August 1, 2017.
2. These tables represent daily partitions of Google Analytics session-level data, capturing metrics such as pageviews, sessions, bounces, hits, and session duration for each day.
3. All tables share the same core schema, but an important schema change occurs starting July 1, 2017: **tables from 20170701 through 20170801 include an additional column clientId, which is absent in earlier tables**. This suggests an enhancement to track client-level identifiers for user journey analysis starting in July 2017.

Figure 5: Example of Table description

B M-Schema

M-Schema, as proposed by Liu et al. (2025b), is a semi-structured representation of a database schema. Compared to simple database creation statements, it encapsulates richer information, such as value examples and column descriptions. In this work, our configuration largely follows the original implementation, with two key modifications. First, for the value examples of each column, we first deduplicate them and then randomly select up to three distinct values, with each example truncated to a maximum length of 50 characters. Second, for tables that have been compressed (as described in our preprocessing step), we introduce a new `# Table Description` field to articulate their specific characteristics. An illustrative example of our modified M-Schema format is shown in Figure 6.

C Example of Different Generation-State Evolution Paths

This section presents an example of a single Text-to-SQL task being solved via two distinct evolution paths, as shown in Figure 7. Although both paths ultimately yield the correct result, the “Circuitous Path” requires multiple steps to arrive at the solution. In contrast, the “Optimal Path” gen-

M-Schema: GA360

```
[DB_ID] GA360
# Table: GA360.GOOGLE_ANALYTICS_SAMPLE.GA_SESSIONS_20160801
[
  (visitNumber: NUMBER, The session number for this user. If this is the first session, then this is set to 1., Examples: [1, 3, 30]),
  (visitId: NUMBER, An identifier for this session. This is part of the value usually stored as the _utmb cookie. This is only unique to the user. For a completely unique ID, you should use a combination of `fullVisitorId` and `visitId`., Examples: [1470046245, 1470072494, 1470078988]),
  (visitStartTime: NUMBER, The timestamp (expressed as POSIX time)., Examples: [1470046245, 1470072494, 1470078988]), (date: TEXT, The date of the session in YYYYMMDD format., Examples: [20160801]),
  ...
]
# Table Description: The naming convention for these tables with the same structure is: GOOGLE_ANALYTICS_SAMPLE.GA_SESSIONS_[DATE], where [DATE] represents the date of the session data in `YYYYMMDD` format, ranging from 20160801 to 20170801, covering a full 13-month period from August 1, 2016, to August 1, 2017.
...
```

Figure 6: Example of M-Schema

erates the answer in a single step by leveraging the implicit null-value filtering of an INNER JOIN. This comparison highlights the importance of our Schema-aware Alignment module in guiding the model toward more efficient and direct reasoning.

D divide-and-conquer Baseline

This section details the divide-and-conquer baseline method used for comparison in our main experiments. For the BIRD dataset, we adopted a state-of-the-art Chain-of-Thought (CoT) prompting strategy similar to that of Pourreza et al. (2024), which includes a fixed set of few-shot examples to guide the LLM in generating the target SQL. For the Spider 2.0-Snow dataset, our baseline model was the DeepSeek R1. In accordance with the official guidelines for this model, we did not use few-shot examples. we prompted the LLM to directly follow a sub-question → sub-SQL → final SQL decomposition format to generate the answer. An example of the CoT decomposition used for a sample from the BIRD development set is provided in the Figure 8 .

E Adaptive Schema Selection Algorithm

In this section, we present the pseudocode of the Adaptive Schema Selection algorithm; see Algorithm E for details.

Algorithm 1 Adaptive Schema Selection

Require: Question \mathcal{Q} , refined schema \mathcal{S}' , refined knowledge \mathcal{K}' , LLM \mathcal{M} , samples k , threshold θ_{\max}

Ensure: Question-relevant schema subset \mathcal{S}_{sub}

```

1:  $\mathcal{T}' \leftarrow \text{GetTab}(\mathcal{S}')$ 
2: if  $|\mathcal{T}'| = 1$  then
3:    $\mathcal{T}_{\text{final}} \leftarrow \mathcal{T}'$ 
4: else
5:    $\mathcal{S}_w \leftarrow \text{ReprSch}(\mathcal{T}', 'M')$ 
6:   if  $\text{TokCnt}(\mathcal{S}_w) > \theta_{\max}$  then
7:      $\mathcal{S}_w \leftarrow \text{ReprSch}(\mathcal{T}', 'DDL')$ 
8:     if  $\text{TokCnt}(\mathcal{S}_w) > \theta_{\max}$  then
9:        $\mathcal{T}_c \leftarrow \emptyset$ 
10:      for each table  $\mathcal{T}_j \in \mathcal{T}'$  in parallel
11:         $\mathcal{Y}_j \sim P_{\mathcal{M}}(\cdot | \mathcal{Q}, \mathcal{K}', \text{SchOf}(\mathcal{T}_j))$ 
12:         $\mathcal{T}_c \leftarrow \mathcal{T}_c \cup \text{TabFromSQL}(\mathcal{Y}_j)$ 
13:      end for
14:    else
15:       $\mathcal{T}_c \leftarrow \text{SampTab}(\mathcal{Q}, \mathcal{K}', \mathcal{S}_w, \mathcal{M}, k)$ 
16:    end if
17:  else
18:     $\mathcal{T}_c \leftarrow \text{SampTab}(\mathcal{Q}, \mathcal{K}', \mathcal{S}_w, \mathcal{M}, k)$ 
19:  end if
20:   $\mathcal{S}_c \leftarrow \text{ReprSch}(\mathcal{T}_c)$ 
21:   $\mathcal{T}_{\text{final}} \leftarrow \text{SampTab}(\mathcal{Q}, \mathcal{K}', \mathcal{S}_c, \mathcal{M}, k)$ 
22: end if
23:  $\mathcal{S}_{\text{sub}} \leftarrow \text{ReprSchCol}(\mathcal{T}_{\text{final}})$ 
24: return  $\mathcal{S}_{\text{sub}}$ 

```

Question: What is the phone number of the school that has the highest average score in Math?

Optimal Path (with Schema-aware Alignment)

Optimal Sub-question: Find the phone number of the school with the highest average math score

Optimal Sub-SQL: SELECT schools.Phone FROM satscores INNER JOIN schools ON satscores.cds = schools.CDSCode WHERE satscores.AvgScrMath IS NOT NULL AND schools.Phone IS NOT NULL ORDER BY satscores.AvgScrMath DESC LIMIT 1

Phone
(408) 366-7700



Circuitous Path (without Schema-aware Alignment)

Sub-question1: Find the `cds` code and the score of the school with the absolute highest average math score.

Sub-SQL1: SELECT cds, AvgScrMath FROM satscores ORDER BY AvgScrMath DESC LIMIT 1

cds	AvgScrMath
1611760135244	699

Sub-question2: Using the `cds` code of the top-scoring school, find its phone number in the `schools` table.

Sub-SQL2: SELECT Phone FROM schools WHERE CDSCode = (SELECT cds FROM satscores ORDER BY AvgScrMath DESC LIMIT 1)

Phone
None

Sub-question3: Query top 3 students' scores and phone numbers to investigate missing contacts.

Sub-SQL3: SELECT T2.AvgScrMath, T1.Phone FROM satscores AS T2 LEFT JOIN schools AS T1 ON T1.CDSCode = T2.cds ORDER BY T2.AvgScrMath DESC LIMIT 3

AvgScrMath	Phone
699	None
698	(408) 366-7700
691	(408) 366-7600

Sub-question4: Retrieve the information of the top-ranked student who has a phone number.

Sub-SQL4: SELECT T1.Phone FROM schools AS T1 INNER JOIN satscores AS T2 ON T1.CDSCode = T2.cds ORDER BY T2.AvgScrMath DESC LIMIT 1

Phone
(408) 366-7700



Figure 7: Example of Decomposition Paths: With vs. Without the Schema-aware Alignment module

F Generation-State Evolution

This section provides the pseudocode of Generation-State Evolution; see Algorithm 2 for details.

Algorithm 2 Generation-State Evolution

```

Require: Context  $\mathcal{X}' = (\mathcal{Q}, \mathcal{S}_{\text{sub}}, \mathcal{K}', \mathcal{K}_{\text{align}})$ ,  

LLM  $\mathcal{M}$   

Ensure: The final SQL query  $\mathcal{Y}$   

1:  $\mathcal{H} \leftarrow \emptyset$   

2:  $t \leftarrow 0$   

3:  $\mathcal{Y}_1 \leftarrow \text{GenInitialQuery}(\mathcal{X}')$   

4: loop  

5:    $t \leftarrow t + 1$   

6:    $\mathcal{R}_t \leftarrow \text{Execute}(\mathcal{Y}_t)$   

7:    $s_t \leftarrow (\mathcal{Y}_t, \mathcal{R}_t)$   

8:    $\mathcal{H} \leftarrow \mathcal{H} \cup \{s_t\}$   

9:    $\alpha \leftarrow \text{SelectAction}(\mathcal{X}', \mathcal{H})$   

10:  if  $\alpha = \text{Finalize}$  then  

11:     $\mathcal{Y} \leftarrow \text{GenFinalQuery}(\mathcal{X}', \mathcal{H})$   

12:    break  

13:  else  

14:     $\mathcal{Y}_{t+1} \leftarrow \text{GenNextQuery}(\mathcal{H}, \alpha)$   

15:  end if  

16: end loop  

17: return  $\mathcal{Y}$ 

```

G Compared Baselines

For a fair and comprehensive comparison, we selected appropriate state-of-the-art baseline methods for each benchmark. On the **Spider 2.0-Snow dataset** (Lei et al., 2024), we chose the representative agentic framework Spider-Agent (Lei et al., 2024) and ReFoRCE (Deng et al., 2025), which is based on multi-path generation and voting, as our primary baselines. For the **BIRD dataset** (Li et al., 2023), our comparison encompasses four mainstream categories of methods: (1) **Hybrid methods**, including CHASE-SQL (Pourreza et al., 2024), XiYan-SQL (Liu et al., 2025b), the method from (Shkapenyuk et al., 2025), and Contextual-SQL (Agrawal and Nguyen, 2025); (2) **Post-training methods**, such as XiYanSQL-QwenCoder-32B (Liu et al., 2025b), Arctic-Text2SQL-R1-7B (Yao et al., 2025), Databricks-RLVR-32B (Ali et al., 2025), and OmniSQL-32B (Li et al., 2025b); (3) **In-Context Learning**, including DAIL-SQL (Gao et al., 2023), MCS-SQL (Lee et al., 2024), RSL-SQL (Cao et al., 2024), and OpenSearch-SQL (Xie et al., 2025); and (4) **Zero-shot methods**, such as DIN-SQL (Pourreza and Rafiei, 2023), MAC-SQL (Wang et al., 2025), and MAG-SQL (Xie et al., 2024). Furthermore, to ensure a fair comparison, we re-evaluated the performance of several baselines us-

divide-and-conquer Example

Question: List the city and state of members enrolled under electrical and computer engineering department.

Evidence: "Electrical and Computer Engineering Department" is the department; members enrolled refers to position = 'Member'.

1. divide-and-conquer:

Analysis: The question asks for city and state from the zip_code table, filtered to members who are enrolled (position = 'Member') and whose major's department is "Electrical and Computer Engineering Department". This requires joining member, major, and zip_code tables.

Pseudo SQL:

```
SELECT T3.city, T3.state FROM member AS T1
INNER JOIN major AS T2 ON T1.link_to_major
= T2.major_id INNER JOIN zip_code AS T3 ON
T1.zip = T3.zip_code WHERE [enrolled member]
AND [electrical and computer engineering department]
```

Sub-question 1: enrolled member

Analysis: Filter on position column in member table.

Pseudo SQL: T1.position = 'Member'

Sub-question 2: electrical and computer engineering department

Analysis: Filter on department column in major table.

Pseudo SQL: T2.department = 'Electrical and Computer Engineering Department'

2. Assembling SQL:

Sub-question 1 (enrolled member): T1.position = 'Member'

Sub-question 2 (electrical and computer engineering department): T2.department = 'Electrical and Computer Engineering Department'

Main Question (city and state of members):

```
SELECT T3.city, T3.state FROM member AS T1
INNER JOIN major AS T2 ON T1.link_to_major
= T2.major_id INNER JOIN zip_code AS T3 ON
T1.zip = T3.zip_code WHERE T1.position = 'Member'
AND T2.department = 'Electrical and Computer
Engineering Department'
```

3. Simplification and Optimization:

The query uses efficient INNER JOINS and straightforward WHERE conditions. No further optimization is needed.

```
<sql>
SELECT T3.city, T3.state FROM member AS
T1 JOIN major AS T2 ON T1.link_to_major =
T2.major_id JOIN zip_code AS T3 ON T1.zip =
T3.zip_code WHERE T1.position = 'Member' AND
T2.department = 'Electrical and Computer Engineering
Department'
</sql>
```

ing their open-source implementations under a unified model and experimental environment.

H Error Analyses

In our evaluation on the Spider 2-Snow dataset, we analyzed 121 samples that have corresponding gold SQL queries. Of these, the 70 cases that DSR-SQL failed to predict correctly were selected for our manual analysis. Despite the limited sample size, we believe this analysis reveals generalizable error patterns. Unlike previous studies, our analysis introduces the concept of DSR-SQL's sub-states, enabling a more precise localization of error origins.

Our findings reveal that Text-to-SQL errors are often not attributable to a single cause but rather exhibit a cascading effect. For instance, the root cause of some Schema Linking errors can be traced back to an initial Intent Misunderstanding. Therefore, in analyzing each failure case, we documented and categorized all error types involved, rather than assigning a single, primary cause.

As shown in Figure 9, we classify the errors into five major categories: **Logical Construction Error**, **Intent Misunderstanding**, **Complex Scenario Failure**, **Schema Linking Error**, and **Data Processing Error**. We elaborate on each category and its specific sub-types below:

Logical Construction Error This was the most prevalent category, accounting for approximately 27.11% of errors. It refers to cases where the model partially understood the user's intent but deviated in constructing the SQL logic.

- Aggregation logic error:** The generated SQL failed to accurately reflect the user's aggregation intent when summarizing, counting, or grouping data. The core issue is a deviation in the data summarization method, leading to results with incorrect values, structure, or granularity.

- Condition filter error:** The model correctly identified the filtering intent and attempted to construct conditions using WHERE, HAVING, or JOIN ON, but the resulting logic did not accurately or completely capture the user's constraints. This often stems from the model's unfamiliarity with database-specific vocabulary combined with imprecise user questions.

- Order of operations error:** The model correctly identified most of the necessary

Figure 8: An example of the divide and conquer method.

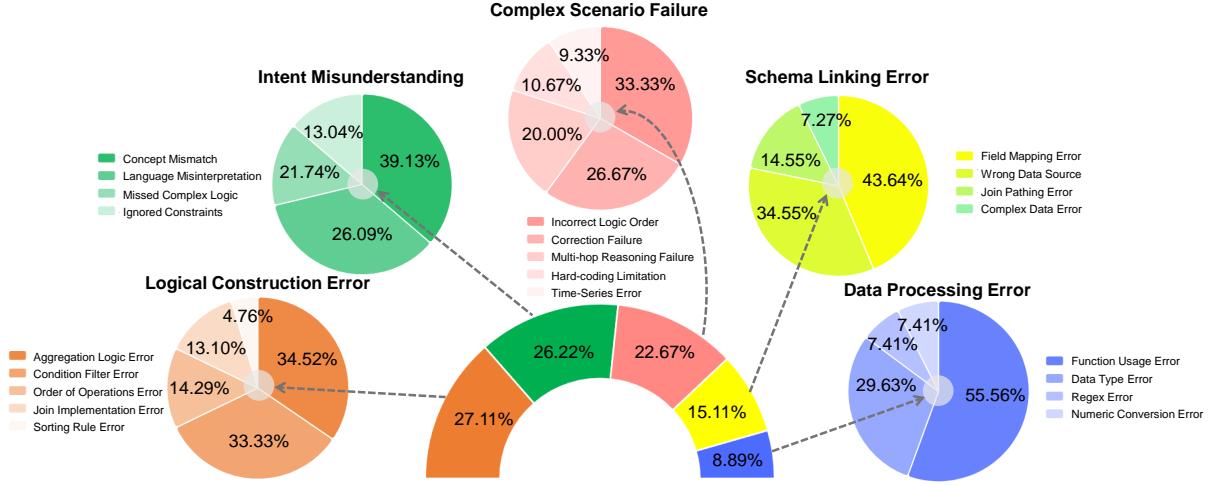


Figure 9: Error Statistics of Spider2.0-Snow

columns and tables but erred in translating the mathematical logic or computational steps from natural language into SQL. The error lies not in selecting the wrong data, but in the incorrect method of calculation. For example, in a gene expression analysis, the predicted SQL calculated $\text{LOG}(\text{AVG}(\text{value}) + 1)$, which aligns with common intuition, whereas the domain convention is $\text{AVG}(\text{LOG}(\text{value} + 1))$. Due to the non-linearity of the logarithm function, these two orders yield vastly different results.

- **Join Implementation Error:** The generated SQL failed to accurately reflect the implicit or explicit data relationships required to combine data from multiple tables. This fundamentally corrupts the data foundation of the query, guaranteeing an incorrect final result even if subsequent logic is sound. Examples include selecting the wrong JOIN keyword (e.g., INNER JOIN vs. LEFT JOIN), leading to improper inclusion or exclusion of rows.
- **Sorting rule error:** The model recognized the need for sorting or ranking but failed to implement the logic according to the sorting criteria, direction, or priority explicitly or implicitly stated by the user.

Intent Misunderstanding This category was the root cause of many failures, accounting for approximately 26.22% of errors. It typically manifests as the model’s failure to correctly grasp the core intent, business logic, or key constraints of the user’s question.

- **Concept Mismatch:** The model identified a key business term or concept, but the **operational definition** or **computational scope** it used to translate it into SQL logic was inconsistent with the gold query’s intent. For example, for the term “frequency,” the LLM interpreted it as the physical count of a text’s occurrences in a file, whereas the gold standard defined it as the total number of unique (repository, language) pairs associated with that text.
- **Language Misinterpretation:** Although the model parsed the keywords and syntax of the question, it failed to capture the deeper semantics, contextual constraints, or implicit logical relationships. This deviation is not a simple mapping error but arises during the translation from “intent” to “logic.” For instance, translating “lacking a ‘wikidata’ tag” into a logically impossible condition such as `key = ‘wikidata’ AND key IS NULL`.
- **Missed Complex Logic:** When a user’s question required multiple sequential and dependent computational steps, the model incorrectly simplified, merged, or flattened this complex process into a single step.
- **Ignored Constraints:** The model understood the core query objective but generated SQL that failed to adhere to all specified conditions, resulting in a dataset that was too broad, improperly filtered, or logically incomplete.

Complex Scenario Failure This category refers to failures in scenarios requiring multi-step reason-

ing or complex business logic.

- **Incorrect Logic Order:** When a question required a specific sequence of logical steps (e.g., filtering, then aggregating, then calculating), the model failed to arrange these steps correctly. The error lies in the overall data flow and processing pipeline of the query, even if individual SQL clauses are syntactically correct.
- **Correction Failure:** When attempting to recover from an intermediate error (such as a query returning an empty set), the model not only failed to correct the logical flaw but also made a mistaken causal attribution (often concluding "the data does not exist"). Based on this false conclusion, it completely distorted the user's original intent, leading to an invalid or irrelevant final SQL query.
- **Multi-hop Reasoning Failure:** When addressing complex questions that require traversing one or more intermediate entities or logical steps to establish a data relationship, the model failed to construct the complete query logic. This reflects an inability to discover or correctly utilize non-direct, implicit connection paths between entities.
- **Hard-coding Limitation:** Within the DSR-SQL framework, the LLM, upon observing intermediate execution results, may opportunistically hard-code partial results from a previous step into subsequent queries. This practice undermines the query's generalizability.
- **Time-Series Error:** When processing time-dimensional data, the model failed to correctly understand and implement complex time-series analysis logic. This involved a fundamental misunderstanding of temporal continuity, dependencies, and window calculations, leading to incorrect time-series metrics.

Schema Linking Error This category refers to the model's failure to correctly understand the database structure and establish the correct mapping from natural language to database elements.

- **Field Mapping Error:** When a table contained columns with highly similar names or functions, the model struggled to make the correct choice, especially when deeper semantic

analysis was required for differentiation. For example, distinguishing between assignee and assignee_harmonized, where the latter is a standardized field that provides more complete results.

- **Wrong Data Source:** When the database contained multiple tables with overlapping functions, similar names, but different data granularities or versions, the model failed to identify the most authoritative or best-matching source. This is one of the most common and impactful linking errors.
- **Join Pathing Error:** The model failed to correctly identify, select, or construct the inter-table join paths necessary to answer the user's question. This is a fundamental logical error, such as failing to find the correct "bridge table" or choosing an incorrect join path.
- **Complex Data Type Error:** The model was unable to generate correct SQL to query data stored in semi-structured or non-atomic columns (e.g., JSON, ARRAY). It tended to apply query logic suited for flat tables, failing when it needed to parse, un-nest, or navigate within complex structures.

Data Processing Error This category includes technical errors made by the model during data processing and transformation.

- **Function Usage Error:** The generated SQL might be syntactically valid but failed to accurately implement the user's business logic due to a misunderstanding of the selection, parameters, or application context of an SQL function. For example, while the TO_TIMESTAMP_* family of functions all convert timestamps, the model might confuse their subtle differences in time zone handling.
- **Data Type Error:** Although the model identified a column's basic type (e.g., number, string), it erred in handling its specific format, intrinsic semantics, or operational requirements, leading to format misunderstandings or improper type conversions.
- **Regex Error:** The regular expression used in the SQL was unable to accurately, completely, or robustly match or extract the required information from the target text field. This usually

stemmed from an insufficient understanding of the target text’s structure, diversity, or edge cases.

- **Numeric Conversion Error:** When performing mathematical operations or type conversions on numerical values (to change their unit, scale, or precision), the process did not align with the user’s requirements or the data’s intrinsic logic, leading to skewed numerical results.

Overall, our analysis demonstrates that even powerful reasoning models still face numerous challenges when tackling highly difficult Text-to-SQL benchmarks like Spider 2-Snow. Looking ahead, future research in the Text-to-SQL domain should continue to focus on several key areas: more refined task context management, more accurate alignment of domain-specific database knowledge with user intent, and enhancing model adaptability to different SQL dialects.

I Case Study: Leveraging ‘Revise’ Actions for Feedback-Driven Correction

We present a case study that demonstrates the LLM’s ability to successfully leverage execution feedback to correct its SQL generation trajectory, a process we term the **Refinement Path**. As shown in Figure 10, upon observing the execution feedback from the preceding sub-SQL, the LLM re-evaluated the current business logic and implemented a corresponding revision step. It expanded the calculation scope from ‘Active Accounts’ to ‘All Accounts,’ ensuring that single-transaction surges (e.g., spiking from 0 to 10 million) were correctly identified as the ‘largest differences.’

J Case Study: How Execution Feedback Drives Problem Solving

The utilization of execution feedback is a core tenet of the DSR-SQL framework, and we present a successful case study in Figure 11. As illustrated, in the initial stage, the LLM relies on its general knowledge to analyze the user’s question and generate a corresponding SQL query, which results in an empty execution result. In Stage 2, the LLM initiates an **Explore** action to identify the column where the ‘strategy’ entity is likely stored. Stage 4 is the most pivotal phase: through a greedy search, the LLM discovers the association between the

user-expressed entity and its corresponding representation in the database, thereby achieving **Semantic Alignment**. At this point, having gathered all necessary context to answer the question, the model proceeds to the final generation stage. This case clearly demonstrates how DSR-SQL leverages execution feedback to familiarize itself with the structure of an unknown database, progressively aligning user intent. It showcases the practical utility of DSR-SQL in navigating unfamiliar and complex scenarios.

K Other results

Subsequent to the submission of our manuscript, the official evaluation script for Spider 2.0 underwent a major revision, which may lead to inconsistencies in reported scores. Although all experiments in our paper were conducted using the original version of the script, we report comparison results on both the old and new versions in Table 7 to ensure transparency and consistency. Furthermore, we also report the performance of DSR-SQL on the multi-dialect Spider 2.0-Lite benchmark, on which we achieved a third-place ranking at the time of our submission.

Benchmark	Script Version	EX (%)
Spider 2.0-Snow	Old	35.28
	New	52.83
Spider 2.0-Lite	Old	-
	New	46.80

Table 7: Performance of DSR-SQL on the Spider 2.0 benchmark, along with other results.

L Disclosure of Large Language Model Usage

During the development phase, a LLM was utilized to assist in the generation of code for several utility functions. Additionally, an LLM was employed during the manuscript preparation phase for grammatical polishing and refinement. All authors have reviewed and verified the correctness of the LLM-generated content and unanimously consent to its inclusion in this work.

Question: Which six Ethereum addresses, excluding '0x000', have the largest absolute differences between their previous and current balances from the tokens at addresses '0x0d8775f648430679a709e98d2b0cb6250d2887ef0' and '0x1e15c05cbad367f044cbfbafda3d9a1510db5513'?

Sub-question1: Gather all transfers for the two tokens, breaking each transfer into incoming and outgoing transactions for non-zero addresses.

Sub-question2: Compute per-address absolute balance differences for the specified tokens by analyzing last transaction amounts and ensuring sufficient transaction history.

NOTE: The LLM inferred that calculating a "balance difference" necessitated a >comparison between a historical state (Previous) and current state (Current). >Consequently, it constructed a filtering condition that excluded all addresses with >only a single transaction.

Generated SQL Fragment:

```
WHERE "rn" = 1
AND "tx_count" >= 2 -- Enforced requirement for at least 2 transactions
```

LLM Reasoning Process (Revise):

<think>...But this data looks anomalous: the top change values are suspiciously clustered in the 9xxx range, and the Tx_Count is stuck exactly at the filter threshold of 2... The prompt never explicitly required a minimum of two transactions. Why did I impose this constraint? In reality, for an address with a single transaction, the 'previous balance' is inherently zero, and the 'current balance' is the transaction amount itself—this constitutes a massive, valid 'difference' in its own right... </think>

Sub-question3: Compute per-address absolute balance differences for the specified tokens by analyzing last transaction amounts, including addresses with only one transaction.

Generated SQL Fragment:

```
-- Removed the constraint: AND "tx_count" >= 1
WHERE "rn" = 1
```

Figure 10: Leveraging 'Revise' Actions for Feedback-Driven Correction

Question: How much higher the average intrinsic value is for trades using the feeling-lucky strategy compared to those using the momentum strategy under long-side trades?

Phase 1

Sub-question: Extract deduplicated intrinsic values and strategies for trades with long-side positions using either 'feeling-lucky' or 'momentum' strategies

Sub-SQL: `SELECT f.value:"Strategy":STRING ... FROM ... LATERAL FLATTEN(input => t."Sides") f ...`

Comment: Normal assumption for generating SQL

Phase 2 (Structure Exploration)

Sub-question: Identify all existing keys in Sides array for LONG-side trades

Sub-SQL: `SELECT DISTINCT fk.value AS "key_name" FROM ... LATERAL FLATTEN(input => OBJECT_KEYS(f.value)) fk ...`

Comment: Explore columns that may store policy-related fields.

Phase 3

Sub-question: Explore PartyID values in PartyIDs array for strategy indicators

Sub-SQL: `SELECT DISTINCT party_val.value:"PartyID":STRING ... WHERE "party_id" ILIKE ANY ('%feeling-lucky%', '%momentum%')`

Comment: Based on the process of elimination, the field is highly likely to be stored within the nested column "PartyIDs."

Phase 4 (Semantic Alignment)

Sub-question: Scan PartyID values for partial strategy indicators

Sub-SQL: `... WHERE ("party_id" ILIKE '%luck%' OR "party_id" ILIKE '%mom%') ...`

Comment: Greedy search for potentially relevant fields

Phase 5

Sub-question: Compute average intrinsic value difference... using PartyID patterns

Sub-SQL: `: ...AVG(CASE WHEN p.value:"PartyID":STRING ILIKE '%LUCKY%' THEN ... END) - AVG(CASE WHEN p.value:"PartyID":STRING ILIKE '%MOMO%' THEN ... END)`

Comment: Collect sufficient context and directly generate the final SQL.

Figure 11: How Execution Feedback Drives Problem Solving