

# AGENTSM: Semantic Memory for Agentic Text-to-SQL

Asim Biswal<sup>△○</sup>, Chuan Lei<sup>\*○</sup>, Xiao Qin<sup>\*□</sup>, Aodong Li<sup>○</sup>,  
Balakrishnan Narayanaswamy<sup>○</sup>, Tim Kraska<sup>○</sup>

<sup>○</sup>Amazon Web Services    <sup>△</sup>University of California, Berkeley    <sup>○</sup>Oracle Corporation    <sup>□</sup>Snowflake Inc.  
abiswal@berkeley.edu, chuan.lei@oracle.com, xiao.qin@snowflake.com  
{aodongli, muralibn, timkrask}@amazon.com

## ABSTRACT

Recent advances in LLM-based Text-to-SQL have achieved remarkable gains on public benchmarks such as BIRD and Spider. Yet, these systems struggle to scale in realistic enterprise settings with large, complex schemas, diverse SQL dialects, and expensive multi-step reasoning. Emerging agentic approaches show potential for adaptive reasoning but often suffer from inefficiency and instability—repeating interactions with databases, producing inconsistent outputs, and occasionally failing to generate valid answers. To address these challenges, we introduce Agent Semantic Memory (AGENTSM), an agentic framework for Text-to-SQL that builds and leverages interpretable semantic memory. Instead of relying on raw scratchpads or vector retrieval, AGENTSM captures prior execution traces—or synthesizes curated ones—as structured programs that directly guide future reasoning. This design enables systematic reuse of reasoning paths, which allows agents to scale to larger schemas, more complex questions, and longer trajectories efficiently and reliably. Compared to state-of-the-art systems, AGENTSM achieves higher efficiency by reducing average token usage and trajectory length by 25% and 35%, respectively, on the Spider 2.0 benchmark. It also improves execution accuracy, reaching a state-of-the-art accuracy of 44.8% on the Spider 2.0 Lite benchmark.

## 1 INTRODUCTION

Text-to-SQL seeks to enable interaction with structured data by translating natural language tasks into executable SQL queries. This capability is particularly valuable in enterprise data analytics and business intelligence, where non-technical users need to extract insights from large, complex databases without mastering SQL or detailed schema knowledge.

Recent advances in large language models (LLMs) [37], prompting strategies [21], and post-training techniques [34] have driven notable progress in Text-to-SQL performance across benchmarks such as BIRD [16] and Spider [35]. However, most existing Text-to-SQL systems remain difficult to scale in realistic enterprise settings, where challenges such as deep nested schemas, diverse SQL dialects, and domain-specific business logic lead to degraded accuracy and efficiency.

To evaluate Text-to-SQL systems under more realistic conditions, the Spider 2.0 benchmark [14] was recently introduced, featuring complex, multi-dialect databases and extremely long contexts. Traditional Text-to-SQL approaches—including vector-based schema retrieval [15], candidate generation with majority voting [21], and self-consistency decoding [13, 27]—struggle when applied on their own to this benchmark, revealing fundamental limitations in their ability to generalize and *scale*. These shortcomings have led to

growing interest in agentic Text-to-SQL methods [7, 26, 30], where agents iteratively interact with databases to inspect schemas, validate partial queries, and adapt to dialectal differences.

While agentic systems demonstrate improved adaptability to large and complex schemas, there is a challenging tradeoff between computational cost and accuracy that makes these systems difficult to scale. Agents often incur substantial **redundancy**, repeatedly retracing identical exploration steps across queries on the same database, which inflates execution cost and latency. They are also prone to **planning variance**, where suboptimal initial reasoning paths lead to inconsistent or failed outcomes. Moreover, the inherently iterative nature of these systems results in **high computational cost and latency**, as each step consumes additional tokens and time. Under practical constraints such as step limits, token limits, or latency budgets, these inefficiencies significantly reduce both overall accuracy and efficiency [17].

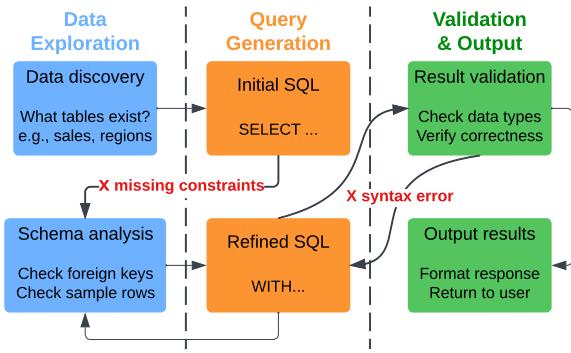
To address these limitations, we present Agent Semantic Memory (AGENTSM), a scalable, stable, and efficient agentic framework for Text-to-SQL. Instead of treating each query as an isolated interaction, AGENTSM introduces a structured semantic memory that captures and reuses trajectories from prior executions. Each trajectory is enriched and stored with semantic annotations to support future retrieval and reasoning. When a new query arrives on a known database, AGENTSM reuses relevant portions of prior trajectories, eliminating redundant exploration and ensuring more consistent behavior. In addition, frequently co-occurring tool sequences are automatically coupled into composite tools, shortening trajectories and improving execution efficiency.

Beyond Text-to-SQL, AGENTSM provides a generalizable foundation for other agentic data tasks—including information extraction, data cleaning, and data transformation—where structured trajectory reuse is critical for scalable and high-quality performance.

In summary, our contributions are as follows:

- (1) We present AGENTSM, an agentic framework that captures and leverages structured semantic memory for enterprise-level Text-to-SQL.
- (2) We design a structured semantic memory that encodes prior trajectories in an interpretable and retrievable format. This memory enables agents to retrieve relevant past experiences, improving efficiency and reasoning consistency.
- (3) We introduce composite tools to AGENTSM that streamline decision-making, reduce latency, and alleviate hallucination in agent planning and tool usage during complex multi-step query generation. These tools reduce both agent turns and token usage.
- (4) We conduct extensive experiments demonstrate that AGENTSM achieves state-of-the-art accuracy of 44.8% on the Spider 2.0 Lite

\*Work done at Amazon Web Services.



**Figure 1:** The standard workflow for Text-to-SQL agents is a series of steps alternating between data exploration, query generation, and answer validation.

benchmark. Ablation studies further show that AGENTSM effectively shortens average trajectory length by 25% and improves execution accuracy by 35%.

## 2 BACKGROUND AND PROBLEM FORMULATION

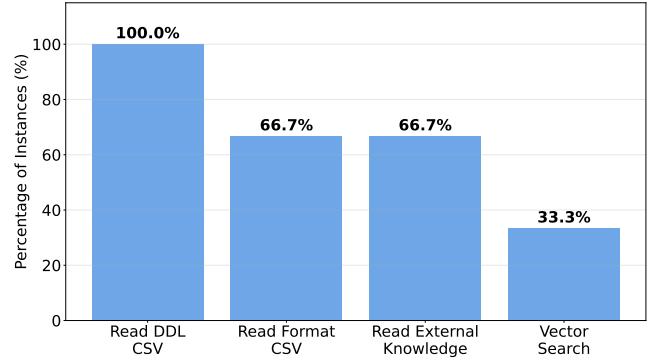
### 2.1 Agentic Text-to-SQL

Modern ReAct-based agents [33] follow a standard pattern of reasoning and action steps. At each step, the agent observes the current state of the task, reasons about the next appropriate action, and executes an action that often involves tool use. For example, a code agent observes intermediate program states or execution outputs, reasons about modifications or next steps, and acts by generating and executing code to advance toward the final solution.

Figure 1 shows the standard workflow of an agentic Text-to-SQL system, which generally follows three phases: data exploration, SQL query generation/execution, and response validation. During the data exploration phase, the agent explores and understands the database schema, identifying relevant tables, columns, and values needed for the task. In enterprise environments, schema information is often parsed and stored offline in files, which the agent can reference alongside simple exploratory SQL queries to understand the data. The agent then attempts to generate SQL queries to answer the given task, often in parts, before synthesizing a final query. Without encountering any errors, the agent can choose to validate its response or directly output it. When errors occur, such as syntax violations or unexpected query results, the agent can revise the SQL or return to data exploration for additional context, depending on the nature of the error.

We identify three key opportunities within these standard workflows that motivate our proposed solution.

**2.1.1 Repeated exploration.** Analysis of agent reasoning patterns reveals that data exploration is *inherently repetitive* across queries on the same database. In fact, Liu et al. [17] highlight the repetitive nature of agent trajectories on the BIRD benchmark [16], reporting that fewer than 10–20% of the trajectories are distinct.



**Figure 2:** Distribution of the first seven steps in agent trajectories on the firebase from Spider 2 [14], reporting the percentage of instances in which each specific action is taken.

We observe the same behavior for agents operating in general Text-to-SQL settings. In the initial steps, the agent consistently performs a set of basic exploration actions. It begins by reading schema files or performing PRAGMA queries to obtain understanding of the database schema, including table structures, column names, and data types. If external knowledge files are available, the agent reads them for additional context about the database or task. If the agent requires additional help dealing with particularly large schemas, then it consults the help of tools, such as vector search, or issues additional SQL queries to understand the contents of candidate tables.

**Opportunity 1.** *Reuse prior trajectories* to eliminate redundant exploration and improve efficiency on new questions over the same database.

**2.1.2 Strategy selection.** Traditional Text-to-SQL systems have converged toward highly precise, pipeline-style solutions that consist of a fixed sequence of steps to every question [21, 27]. Similarly, agentic Text-to-SQL approaches remain heavily reliant on prompt engineering [7], where carefully designed instructions define the agent’s workflow and tool use to ensure consistency.

However, enforcing a single dominant strategy across all queries can be suboptimal in terms of both accuracy and efficiency. While standardized routines can be effective for common cases, they often fail on out-of-distribution queries. For example, Figure 2 reveals that the agent chose to perform vector search in only 30% of questions on the ‘firebase’ database in Spider 2.0, likely due to the nested schema and sufficient prior knowledge gained from prior exploration steps. In such cases, adhering to a rigid strategy can introduce irrelevant context, unnecessary computation, and wasted steps that otherwise contribute to more targeted actions to solve the given query.

**Opportunity 2.** *Dynamically adapt trajectories* according to the characteristics of each task and database, rather than relying on a fixed, one-size-fits-all strategy.

**2.1.3 Reducing variance.** Agent behavior in Text-to-SQL generation inherently exhibits variance across runs. Small deviations, including trivial syntax errors, in intermediate steps can derail the

reasoning process, causing the agent to abandon a correct reasoning path and pursue an alternative one. This instability leads to inconsistent query generation and, consequently, variation in final results [31].

Deterministic behavior cannot be enforced reliably even by setting the model temperature to zero for agents. Moreover, even if a strictly deterministic setting was possible, it prevents the agent from exploring alternative reasoning paths, limiting its ability to discover solutions to previously unseen or complex tasks [36]. Common Text-to-SQL strategies, such as candidate generation and query refinement, explicitly rely on controlled model variance to improve the likelihood of a correct SQL being generated. However, once new contexts are introduced into the reasoning workflow, subsequent steps can diverge unpredictably, even under deterministic settings.

**Opportunity 3.** Optimize the *trade-off between tool complexity and trajectory length* to minimize variance and enhance accuracy and robustness.

## 2.2 Problem Statement

In this paper, we aim to develop an agentic Text-to-SQL system that, given a natural language query  $q$ , a database  $D$  and a set of tools  $\mathcal{U}$ , produces an optimized reasoning trajectory

$$\tau(q, D, u) = \arg \max_{\tau \in T(q, D, \mathcal{U})} \text{Acc}(\tau)$$

where  $T(q, D, u)$  denotes the space of executable reasoning trajectories composed of tool usages and intermediate reasoning steps, and  $\text{Acc}(\tau)$  measures the accuracy of the generated SQL query.

## 3 METHODOLOGY

Figure 3 depicts a full overview of the AGENTSM framework, which consists of two agents with a trajectory-reading method and careful tool design to improve agent performance on Text-to-SQL tasks.

### 3.1 AGENTSM Architecture

**Agents.** We introduce two main agents in our framework, the planner agent and the schema linking agent. Inspired by the ReAct framework, agents alternate between observation, reasoning, and action cycles. In our implementation, both agents are coding agents that take actions by writing Python code. Unlike prior methods, however [14], we do not restrict the functions or code that the agent can write to a limited, predefined action space. Instead, we provide a broad range of authorized libraries which the agent may use to help with data exploration and analysis, including Pandas and NumPy.

The **planner agent** serves as the system’s core reasoning component. It constructs a high-level execution plan, iteratively issues SQL queries, performs reasoning and data wrangling, and applies query refinements before producing the final answer. Unlike multi-agent pipelines that delegate SQL generation to a dedicated agent[29], we intentionally integrate SQL generation within the planner’s reasoning loop. This design choice ensures that the planner agent directly leverages the schema and data understanding gathered through exploration, avoiding the context loss commonly observed when control is handed off to specialized sub-agents. Moreover, by

letting the planner generate and execute SQL, AGENTSM avoids unnecessary inter-agent communication overhead and latency.

The **schema linking agent** is managed by the planner agent, and its role is to perform fine-grained data exploration for the task, when the planner agent requires deeper inspection of candidate tables. It has access to specialized tools including a vector search tool, which performs a similarity search on a precomputed index of schema information to narrow a list of tables and columns relevant for a given question. The schema linking agent operates within a small budget (e.g., 5 steps) to perform basic queries to interpret nested structures, probe candidate columns, and validate potential join paths, before returning its findings (i.e., mapping question tokens to schema elements) to the planner agent. Delegating deeper exploration to a separate agent allows the planner agent to not drift in its reasoning and keep only relevant information in context.

Although prior work has proposed complex multi-agent systems (e.g., introducing query fixers, decomposers, or evaluators [7]), we deliberately keep AGENTSM’s architecture simple with two tightly coupled agents. This decision is guided by empirical findings [9, 28] that excessive agent decomposition leads to communication overhead, fragmented memory, and increased behavioral variance. In Text-to-SQL tasks, contextual knowledge is often highly localized: understanding a schema’s nested hierarchy requires maintaining persistent state over multiple exploration and reasoning cycles. Therefore, introducing additional agents would necessitate a shared memory [4, 24] to prevent context fragmentation. However, shared memory introduces nontrivial challenges in retrieval efficiency, consistency management, and error propagation, which we leave as future work.

**Tools.** Each agent in AGENTSM is equipped with a carefully selected set of tools designed to balance flexibility, robustness, and efficiency. The planner agent is provided with tools for trajectory retrieval, output validation, and result saving. The trajectory retrieval tool allows the planner to review reasoning traces from semantically similar questions before starting a new task, helping it avoid redundant exploration steps. The validation and saving tools ensure that the final SQL and results follow consistent formatting and can be correctly persisted. The schema linking agent, on the other hand, is equipped with tools specialized for fine-grained schema exploration, particularly the vector search tool that retrieves semantically relevant tables and columns from a precomputed index. Restricting this capability to the schema linking agent prevents redundant searches between the two agents.

Both agents share a set of general-purpose tools for local directory exploration and SQL execution on supported backends such as BigQuery, Snowflake, and SQLite. To improve execution robustness, we incorporate self-refinement into the SQL execution tools. When a query fails or returns an empty result, the SQL execution tools automatically attempt a correction based on the received feedback. This mechanism reduces the likelihood that a single execution error derails the reasoning trajectory.

In addition to these basic tools, we introduce *composite tools*, special types of tool that combine the logic of single-purpose tools that frequently co-occur in consecutive steps. These composite tools, described in Section 3.3, simplify planning, reduce unnecessary tool usage, and improve consistency across runs.

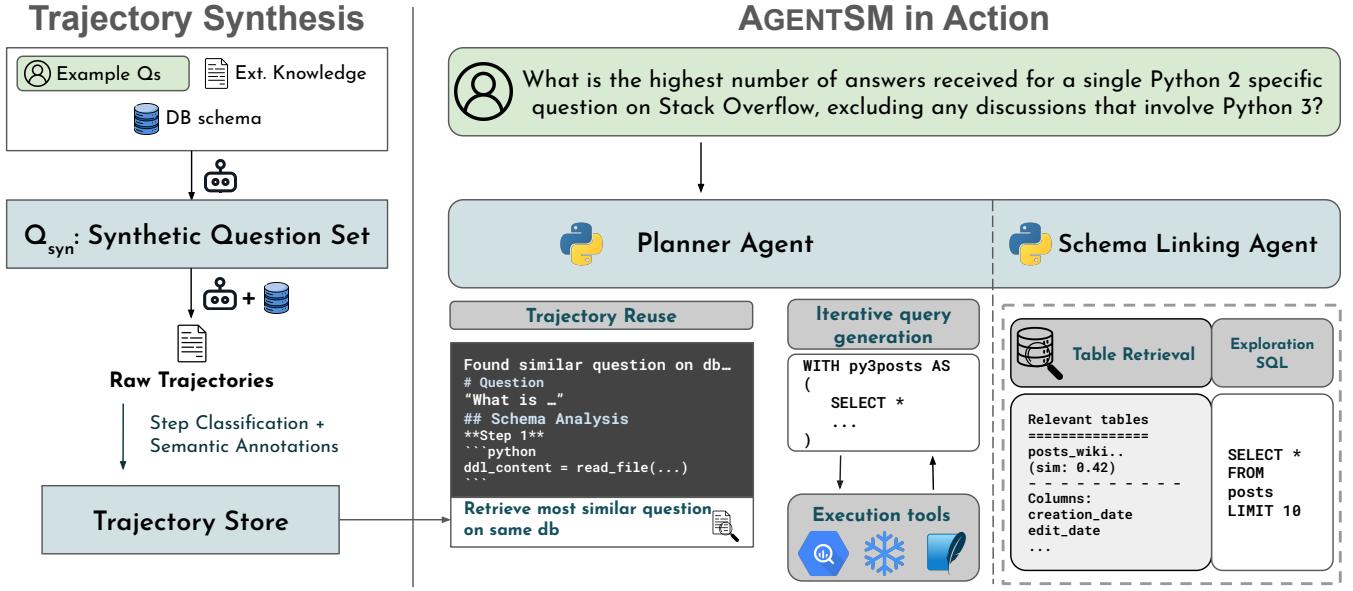


Figure 3: An overview of the AGENTSM architecture that leverages trajectories in structured semantic memory.

### 3.2 Trajectory Synthesis and Retrieval

**Synthetic question exploration.** We make a key insight that data exploration is inherently reusable: once an agent has examined the schema, inspected columns, and issued exploratory queries, those interaction traces remain valuable for subsequent questions on the same database. Crucially, this property extends to *synthetic questions*, questions that can be formulated and executed offline that are deliberately designed to elicit data exploration. By enriching the trajectory store with such exploration-rich traces, AGENTSM effectively reduces redundant probing and enhances agent performance on real queries.

Algorithm 1 outlines the process for generating the full synthetic question set  $Q_{\text{syn}}$ . First, we ensure that each database present in our query distribution is covered by at least one synthetic question (Lines 2-4). We allocate additional synthetic questions across databases proportionally to the distribution these databases on the query set, up to a fixed total question budget  $n$  (Line 6). This allocation strategy naturally emphasizes large, complex schemas, where additional exploration traces are most beneficial (Lines 7-11).

For each selected schema  $S$ , we prompt an LLM with the schema, external knowledge file, and a set of existing questions to generate diverse candidate questions that involves different operators, tables, and columns. The number of questions generated per schema follows the same distribution as the full query set. The prompts used to generate candidate questions can be found in the repository<sup>1</sup>.

Each synthetic question  $q \in Q_{\text{syn}}$  is then consumed by an agent equipped with SQL and file-reading tools only to produce a trajectory  $T(q)$ , which is stored for reuse. Although the answers to these synthetic questions are not evaluated, the resulting trajectories provide dense and diverse exploration traces. When the agent

#### Algorithm 1: Synthetic Question Generation.

GenQ( $S, K, Q$ ) generates a new question using an LLM over the database schema  $S$ , external knowledge  $K$ , and existing questions  $Q$  for  $S \in \mathcal{D}$ .  
Allocation( $S, n, p$ ) returns an integer budget  $k$  for the max number of questions on  $S$  based on the query distribution  $p(Q)$  and the remaining global budget  $n$ .

---

**Input:** Databases  $\mathcal{D}$ , target count  $n$ , query distribution  $p(Q)$   
**Output:** Synthetic question set  $Q_{\text{syn}}$

```

1  $Q_{\text{syn}} \leftarrow \emptyset;$ 
2 foreach schema  $S \in \mathcal{D}$  do
3      $q \leftarrow \text{GenQ}(S, K, Q);$ 
4      $Q_{\text{syn}} \leftarrow Q_{\text{syn}} \cup \{q\};$ 
5 while  $|Q_{\text{syn}}| < n$  do
6      $k \leftarrow \text{Allocation}(S, n, p(Q));$ 
7     for  $i \leftarrow 1$  to  $k$  do
8         if  $|Q_{\text{syn}}| = n$  then
9             break
10             $q \leftarrow \text{GenQ}(S, K, Q);$ 
11             $Q_{\text{syn}} \leftarrow Q_{\text{syn}} \cup \{q\};$ 
12 return  $Q_{\text{syn}};$ 

```

---

encounters new queries in future, these traces provide rich context, allowing it to leverage past exploration rather than repeating it.

**Step classification.** Agent workflows are non-linear, consisting of sequence of distinct states that the agent enters and exits during execution. In practice, a Text-to-SQL agent alternates among three primary phases depicted in Figure 1: (1) exploring the database to understand the data, (2) formulating and executing candidate

<sup>1</sup><https://tinyurl.com/xcey6h33>

Method	Avg Steps	Accuracy (%)
No trajectory	22.62	25
Naive	22.62	25
Markdown	16.50	50
JSON	15.12	50

**Table 1: Average trajectory length (in steps) and execution accuracy on a sample of Spider 2.0 Lite questions, under different trajectory formats. Reading raw, unstructured trajectories yield no improvement compared to without using any trajectory. In contrast, structured representations, in either Markdown or JSON format, significantly reduce the average number of steps and improve the overall execution accuracy.**

queries, and (3) validating results or producing an answer. Similar stage decompositions have been observed in prior work [17].

Grounding trajectory reuse in these phases is essential: exploration steps are often generalizable across questions, execution steps encode database-specific reasoning, and validation steps highlight patterns of error correction and output formatting. Step classification helps avoid presenting the agent with a noisy sequence of mixed tool invocations and rendering reuse ineffective.

To classify steps, we apply a lightweight text-pattern matching using regular expressions rather than relying on step position in the trajectory. While exploration often occurs early, agents may re-enter this exploration phase later when encountering execution errors or missing context. We therefore identify intent based on tool usage and query patterns. For example, file listing and reading operations are matched as exploration, while complex SQL queries beginning with ‘WITH’ (common table expressions) are almost always recognized as main query execution. We also explored LLM-based step classification and observed comparable accuracy but substantially higher cost and latency.

Each classified step is parsed and saved in a separate trajectory file. During trajectory reading, the agent selectively loads only the trajectory segment corresponding to the relevant phase for reference. For example, during data exploration phase, the agent retrieves and reads the relevant exploration trajectories that include basic schema analysis and external knowledge reading actions, enabling efficient reuse of prior context without redundant steps.

**Trajectory structure.** Agent trajectories are typically long, complex sequences that interleave text, code, reasoning traces, and execution outputs. Reading raw trajectories introduces significant noise and increases the *lost-in-the-middle* effect, where valuable information is buried amid verbose agent debugging logs and intermediate outputs.

To help the agent extract useful knowledge more effectively, we impose a structured representation on each trajectory. In particular, we store agent trajectories in markdown format, with semantically meaningful headers automatically generated by a lightweight LLM (e.g., Claude Haiku 4.5). This structure segments the trajectory into interpretable sections, improving both retrieval and readability.

As shown in Table 1, structured trajectories, whether represented in markdown or JSON, yield substantial improvements in

both execution accuracy and average step reduction compared to unstructured logs, based on a sample of 20 benchmark questions. We adopt markdown as the format for its consistency with external knowledge files and human readability. The implementation details of step classification and trajectory structure generation can be found in the repository<sup>2</sup>.

**Trajectory selection.** Having established a structured representation for storing agent trajectories, we next describe how these stored trajectories are retrieved and reused during inference. Among the synthetic questions  $Q_{syn}$  for which trajectories are synthesized and saved, we first restrict retrieval to those associated with the same database as the given question  $q$ . This filtering narrows the search space to relevant trajectories:

$$Q_d = \{q' \in Q_{syn} \mid db(q') = db(q)\} \quad (1)$$

From this filtered subset  $Q_d$ , we then select the most relevant trajectory  $T$  based on its associated question  $q'$  that has the highest semantic similarity to the given question  $q$ :

$$q^* = \arg \max_{q' \in Q_d} \text{sim}(q', q) \quad (2)$$

$$T = \text{traj}(q^*) \quad (3)$$

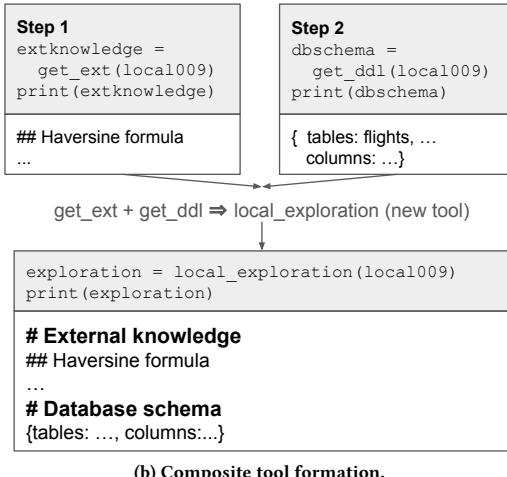
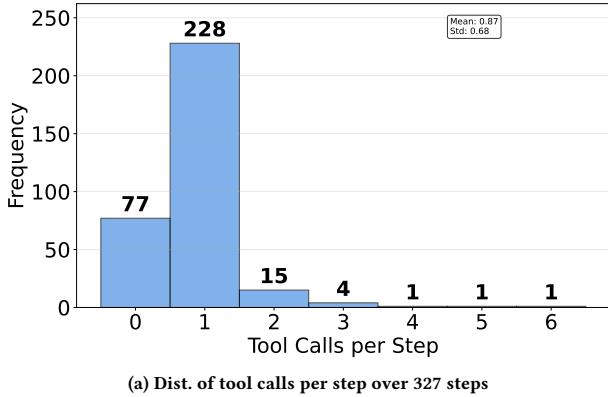
This strategy is effective for data exploration reuse, as semantically similar questions tend to probe the same set of tables during their initial exploration steps. The agent naturally prioritizes tables with their schema semantically similar to the question before performing deeper searches. Consequently, selecting trajectories based on question similarity leads to effective trajectory reuse. Further improvements in retrieval quality could be achieved through finer-grained alignment between questions and stored trajectories. For example, schema-level retrieval could identify which tables are relevant to a question and restrict trajectory retrieval to those involving the same tables. Alternatively, trajectory-level alignment could leverage the agent’s initial plan to retrieve similar prior executions and adjust the reasoning strategy before execution. We leave these directions to future work, as fine-grained retrieval risks missing relevant context essential for accurate reasoning.

### 3.3 Composite Tools

**Composite tool construction.** Across trajectories, we observe that certain tool pairs consistently co-occur in sequence. Figure 4b illustrates how the `get_ext` → `get_ddl` pattern frequently appears when the agent needs to access an external knowledge file and the database schema at the beginning of the exploration phase. As such, these tools can be combined into a single composite tool `local_exploration`.

As shown in Figure 4a, agents rarely chain multiple tool invocations within a single reasoning step, instead issuing one call per step, even when such sequences are highly regular and deterministic. This pattern leads to inefficiency: the agent spends valuable steps on executing repetitive exploration routines rather than focusing on higher-level reasoning and problem solving.

<sup>2</sup><https://tinyurl.com/492ymhc>



**Figure 4: Analysis of tool usage and composite tool construction.** (a) shows that most steps use zero or one tool, motivating composition. (b) illustrates how frequently co-occurring tools are merged into a higher-level composite tool.

To mitigate this overhead, we introduce *composite tools*, which combine commonly co-occurring tool sequences into single higher-level operations. Formally, we construct a composite tool:

$$C = \langle t_1, t_2, \dots, t_k \rangle \quad \text{if} \quad \frac{f(t_1, t_2, \dots, t_k)}{N} \geq \tau$$

where  $t_i \in \mathcal{T}$ ,  $\mathcal{T}$  is the tool set,  $f(\cdot)$  counts the trajectories containing the subsequence  $(t_1, t_2, \dots, t_k)$ ,  $N$  is the total number of trajectories, and  $\tau$  is a chosen support threshold.

In practice, AGENTSM follows several heuristics to construct effective composite tools while preserving modularity. First, tools are combined only when they co-occur within the same reasoning phase (e.g., data exploration or validation), ensuring that composite tools remain semantically coherent. Second, tools that appear across multiple phases are excluded from composition to preserve their agility for reuse in different contexts. Finally, to avoid excessive aggregation, we constrain the maximum size of composite tools, preventing all tools collapse into a single monolithic operator.

Each composite tool is then assigned a descriptive name and natural language summary by an LLM, which the agent can reference

during prompting. At inference time, the agent uses these composite tools directly, effectively replacing repeated multi-step exploration with single high-level actions and freeing more trajectory budget for reasoning.

## 4 EVALUATION

In this section, we evaluate the effectiveness of AGENTSM on Spider 2, a widely recognized benchmark for agentic Text-to-SQL systems. Our evaluation is designed to answer the following key questions:

- (1) How do trajectory synthesis and retrieval influence the agent's efficiency and accuracy?
- (2) How does the optimized tool design improve agent's efficiency and accuracy?

### 4.1 Evaluation Setup

**Dataset.** Spider 2.0 [14] is a large-scale benchmark of 547 Text-to-SQL workflow problems that emulate the challenges of enterprise-scale data analysis. Unlike traditional Text-to-SQL benchmarks, Spider 2.0 requires solving tasks that involve extremely long contexts, nested schemas, and diverse SQL dialects. The databases in Spider 2.0 often span thousands of columns and are deployed in production systems such as BigQuery and Snowflake.

This makes Spider 2.0 a natural workload for agentic approaches, where agents must iteratively plan and issue queries to reach a final answer. We select this benchmark to evaluate our method, as our framework is specifically designed to operate in such complex environments that reward iterative reasoning and tool use.

**Metrics.** We measure execution accuracy as the percentage of execution output matches to the gold output using the evaluation function provided by the authors of the original benchmark. We adopt this metric as per the original benchmark. We report accuracy for BigQuery, Snowflake, and SQLite questions, along with an overall result.

We also report the average agent trajectory length in both reasoning steps, as well as the number of input and output tokens. We also report average end-to-end latency for a single execution.

**Experimental setup.** We evaluate our method using the Claude 3-7 [1] and Claude 4 Sonnet [2] models. Several agent frameworks have risen in popularity, including LangChain [12], smolagents [8], etc. While our approach is framework-agnostic, for evaluation we implement it using smolagents as the foundation of our agentic Text-to-SQL system.

For all evaluated systems, we employ the MiniLM-L6-v2 model [23] from SentenceTransformers for generating embeddings and use the FAISS [10] library for vector similarity search. We run ArcticSQL-7B with vLLM [11] on a cluster of 8×A100 80GB GPUs.

**Baselines.** For our main experiments on Spider 2.0 Lite, we compare AGENTSM against two baselines: SpiderAgent [14] and a standard coding agent (CodingAgent) implemented using the smolagents [8] framework equipped with basic SQL execution tools described in Section 2. We evaluate both baselines with Claude 3-7 [1] and Claude 4 Sonnet [2] as the underlying LLMs. The performance of

Method	EX (%)				Avg. Length			Avg Latency (s)
	BigQuery	Snowflake	SQLite	Overall	Steps	Input Toks	Output Toks	
SpiderAgent (claude-3-7-sonnet)	37.6	12.6	40.7	28.7	18.9	200K	4K	363.2
SpiderAgent (claude-4-sonnet)	-	-	-	27.8*	-	-	-	-
CodingAgent (claude-4-sonnet)	27.3	11.1	42.2	24.7	18.1	200K	4K	325.5
AgentSM (claude-3-7-sonnet)	40.5	33.3	42.2	38.4	16.8	299K	5K	226.4
AgentSM (claude-4-sonnet)	<b>52.2</b>	<b>35.0</b>	<b>51.9</b>	<b>44.8</b>	16.4	300K	5K	247.1
AgentSM (claude-4-sonnet, gold tables)	62.0	44.0	79.2	57.6	15.8	268K	5K	194.3

Table 2: Comparison of execution accuracy, step/token length, and latency for methods on the Spider 2.0 Lite dataset. \*This result is quoted directly from the original SpiderAgent paper.

other state-of-the-art systems is available on the official Spider 2.0 leaderboard<sup>3</sup>.

## 4.2 Main Results

Table 2 contains the execution accuracy for each method on the full Spider 2.0 Lite benchmark with 547 questions. We report execution accuracy for BigQuery instances, Snowflake instances, and SQLite instances along with overall accuracy.

**Accuracy.** AGENTSM substantially outperforms the prior agentic baseline SpiderAgent [14], achieving higher accuracies across each SQL dialect and an overall 14.1% improvement in execution accuracy. AGENTSM also outperforms a standard coding agent implemented using the same agent framework[8] and model [2] by over 20%. Based on our evaluation, AGENTSM would rank No. 1 on the Spider 2.0 Lite leaderboard, with an overall accuracy of 44.8% at the time of paper submission. Notably, we achieve this result without using a powerful reasoning model such as Qwen3 [3] or OpenAI’s o3 [19], which other top systems rely on.

Across all SQL dialects, the questions on the Snowflake database are the most challenging for the agents. The platform has more questions on databases with larger and nested schemas. For this platform alone, our method performs better with an older model Claude 3-7 Sonnet than Claude-4 Sonnet.

**Efficiency.** To understand how AGENTSM improves agent efficiency, we revisit the standard agentic workflow (i.e., CodingAgent) and analyze how our method reshapes the agent’s progression through three key stages: database exploration, query execution, and answer validation. We randomly sample 75 questions from the Spider 2.0 Lite dataset and classify each step in the agent’s trajectory as one of the three stages, based on the reasoning trace and executed queries. We aggregate these stage annotations across instances to measure the proportion of each trajectory devoted to exploration versus execution.

As shown in Figure 5, AGENTSM transitions from exploration to execution significantly earlier, completing tasks with fewer overall steps. In contrast, a standard coding agent not utilizing synthesized trajectories repeatedly issues exploratory queries and spends most of its steps analyzing schemas rather than composing or validating SQL queries. By leveraging structured semantic memory, AGENTSM alleviates redundant data exploration, maintains consistent reasoning efficiency, and scales gracefully to larger databases and more complex questions.

<sup>3</sup><https://spider2-sql.github.io/>

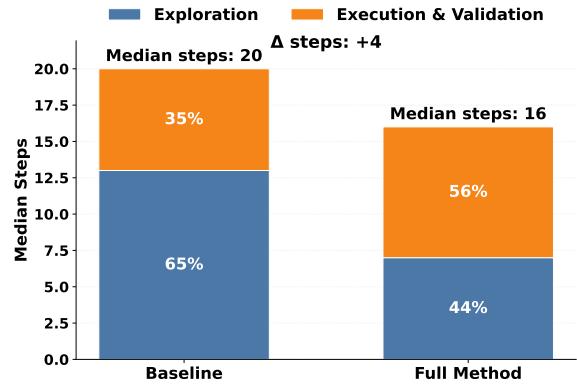


Figure 5: Trajectory composition: median steps split into exploration vs. execution and validation.

## 4.3 Ablation Studies

**4.3.1 Effectiveness of trajectory reading and composite tools.** We conduct ablation studies on two key components of AGENTSM, the synthesized trajectory and the composite tools, to report their respective effects on accuracy and efficiency.

Table 3 reports the average trajectory length and accuracy for a sample of 75 examples from Spider 2.0 Lite with and without trajectory reading. We observe that enabling trajectory reading reduces the average trajectory length of the agent by 25% while improving accuracy by 35%. This is consistent with the results reported in Section 4.2.

Composite tools have a similarly pronounced impact on agent trajectory length and accuracy. Without composite tools, the agent requires the largest number of steps across all sample instances, yet achieving the lowest overall execution accuracy.

The observed accuracy gains naturally stem from the reduction in average trajectory length, which can be attributed to two factors: (1) longer trajectories often correspond to inherently more complex questions, and (2) as trajectory length increases, the agent becomes more likely to reach the maximum step threshold before solving the question. By reducing the number of steps spent on data exploration, AGENTSM effectively mitigates the issues, allowing the agent to allocate more steps toward reasoning and to iteratively refine its final queries before reaching the maximum steps.

**4.3.2 Impact of gold tables.** We further evaluate AGENTSM on Spider 2.0 Lite with gold tables (as provided by Spider 2.0). Using the gold tables and columns, our method achieves **55.3%** execution

Method	Execution Accuracy		Steps			Avg. Length			Latency	
	EX (%)	$\Delta_{\text{EX}}$	Total	Avg	$\Delta_{\text{Steps}}$	Input Toks	Output Toks	$\Delta_{\text{Input}}$	Time (s)	$\Delta_{\text{Time}}$
No trajectory reading	17.3	-34.7	1527	20.36	+4.37	360K	5.6K	+8K	348.6	+96.2
No composite tools	16.2	-35.8	1557	20.76	+4.77	398K	5.8K	+46K	523.7	+271.3
<b>Full method</b>	<b>52.0</b>	—	<b>1167</b>	<b>15.99</b>	—	<b>352K</b>	<b>5.2K</b>	—	<b>252.4</b>	—

**Table 3: Effects of synthesized trajectory and composite tools on execution accuracy, step count, token usage, and latency over 75 sampled questions. Deltas are computed relative to the full method. Removing either component increases token usage and latency while sharply decreasing execution accuracy.**

accuracy on the question subset (listed in Table 2). The results indicate a substantial improvement in execution accuracy compared to the base configuration of our method. This highlights that schema linking is still a significant hurdle in these complex database scenarios, as agents often struggle to identify the right tables and columns when data and the given question are ambiguous.

#### 4.4 Error Analysis

We provide a detailed error analysis on the full Spider 2.0 Lite evaluation to better understand the remaining failure cases of AGENTSM. Among the 245 incorrect instances, 34% occur on BigQuery, 22% on SQLite, and 44% on Snowflake, which shows the highest relative error rate. Within Snowflake, 5% of failures stem from exceeding the step budget, 30% from schema-linking errors caused by nested schemas, and the remainder from logical or dialect-specific syntax errors during query generation.

We further examine how structured semantic memory influences the agent’s success rate. When relevant trajectories are available, AGENTSM can reuse past reasoning steps to provide useful context for new queries. This is particularly effective for schema-linking, where prior trajectories often highlight useful table and join patterns, helping the agent quickly identify the relevant portions of a large schema. In contrast, the benefit is limited for queries that demand complex mathematical operations or intricate CTE reasoning. These tasks typically require deeper reasoning or problem decomposition that cannot be easily inferred from previous trajectories.

Finally, we observe that AGENTSM’s performance varies across domains. Commonly used databases (e.g., city, weather, census) achieve 60–78% accuracy, whereas heterogeneous databases (e.g., github\_repos, idc) lag behind at 14–40%. These results suggest that AGENTSM excels in domains with consistent schema patterns and recurring reasoning structures, but remains challenged by highly domain-specialized data sources.

## 5 RELATED WORK

**Text-to-SQL.** Recent Text-to-SQL work has spanned supervised fine-tuning, prompt engineering, and reinforcement learning based solutions. Supervised systems such as DIN-SQL [22] and MAC-SQL [29] utilize a pipeline with structured decoding, schema linking, and decomposition components to generate SQL queries. Prompting based methods [7, 21] leverage LLMs in few-shot or zero-shot settings, often augmented with modular pipelines that include candidate generation, majority voting, and execution verification. More recently, reinforcement techniques using group relative policy optimization (GRPO) engineer rewards to improve model performance on Text-to-SQL tasks [5]. Multi-agent systems, such as Agentic-Data, explore utilizing a combination of planning, exploration, and validation agents to iteratively work towards an answer [26]. These

techniques have steadily improved overall performance on the Spider [35] and BIRD [16] benchmarks, however they often assume clean, well-aligned schemas and struggle with ambiguity, nested queries, and relational complexity found in enterprise settings.

**Coding Agents.** Recent work on coding agents explore how LLMs can use external tools, such as web-search and code, to solve tasks through iterative reasoning. Approaches like ReACT [32] and Reflexion [25] highlight the effectiveness of planning, execution, and feedback stages for complex tasks. In line with these approaches, agentic Text-to-SQL systems decompose query generation into steps of interleaved tool executions and reasoning, allowing models to inspect schemas, validate partial queries, and iteratively refine the final answer. SpiderAgent [14] formalizes this trend as an initial solution for Spider2.0. Recent works [18] have also extended this broad strategy with memory-guided refinement. While effective, most of these solutions rely on a fresh state for the agent or maintain some intra-task memory, unlike our solution which leverages inter-task memory to re-use database exploration reasoning across examples.

**Agent Memory.** Research on agentic memory can be broadly categorized into action-oriented memory, focused on persisting agent state, and knowledge-oriented memory, focused on persisting the knowledge agents gain from interaction. The most basic approach treats the language model’s context window as a scratchpad-like working memory, where agents maintain notes. MemGPT [20] addresses the challenge of limited context windows by implementing a memory hierarchy that mimics operating system memory management, effectively bridging the previous two approaches. More recently, Mem0[6] utilized graph representations of memory, maintaining memory with standard graph knowledge base structure: a collection of nodes (entities) and edges (relationships between entities). These forms of memory are limited in the amount of structure they can represent in their information for the agent.

## 6 CONCLUSION

In this paper, we introduce AGENTSM, a framework that enables agents to reuse reasoning steps across related Text-to-SQL tasks within the same database. By exploiting the inherent repetition in data exploration, AGENTSM synthesizes, stores and retrieves structured trajectories to substantially reduce redundant exploration and improve execution efficiency. Moreover, our solution, equipped with semantic memory, enhances scalability, allowing agents to handle larger schemas and more complex queries. As a result, AGENTSM delivers significant gains in efficiency across agent turns, token usage, and latency. On the Spider 2.0 Lite benchmark, AGENTSM achieves an execution accuracy of 44.8%.

## REFERENCES

- [1] Anthropic. 2025. Claude 3.7 Sonnet. <https://www.anthropic.com/clause>. Large Language Model.
- [2] Anthropic. 2025. Claude 4 Sonnet. <https://www.anthropic.com/clause>. Large Language Model.
- [3] Jiawei Bai, Wei Zhang, Yifei Li, et al. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2503.06749* (2025). <https://arxiv.org/abs/2503.06749>
- [4] Xiaohe Bo, Zeyu Zhang, Quanyu Dai, Xueyang Feng, Lei Wang, Rui Li, Xu Chen, and Ji-Rong Wen. 2024. Reflective Multi-Agent Collaboration based on Large Language Models. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 138595–138631. [https://proceedings.neurips.cc/paper\\_files/paper/2024/file/fa54b0edce5eef0bb07654e8ee800cb4-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2024/file/fa54b0edce5eef0bb07654e8ee800cb4-Paper-Conference.pdf)
- [5] Kai Cheng, Xiaojun Zhou, and Weidong Chen. 2024. Text-to-SQL with Reinforcement Learning via Execution-Guided Reward Modeling. *arXiv preprint arXiv:2403.12345* (2024).
- [6] Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. Mem0: Building Production-Ready AI Agents with Scalable Long-Term Memory. *arXiv:2504.19413 [cs.CL]* <https://arxiv.org/abs/2504.19413>
- [7] 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. *arXiv:2502.00675 [cs.CL]* <https://arxiv.org/abs/2502.00675>
- [8] Hugging Face. 2025. *SmolAgents*. <https://github.com/huggingface/smolagents> Lightweight framework for building language model agents.
- [9] Mourad Gridach, Jay Nanavati, Khalidoun Zine El Abidine, Lenon Mendes, and Christina Mack. 2025. Agentic AI for Scientific Discovery: A Survey of Progress, Challenges, and Future Directions. *arXiv:2503.08979 [cs.CL]* <https://arxiv.org/abs/2503.08979>
- [10] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-Scale Similarity Search with FAISS. In *IEEE Transactions on Big Data*. IEEE. <https://doi.org/10.1109/TBDA.2019.2921572>
- [11] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- [12] LangChain. 2025. *LangChain*. <https://github.com/langchain-ai/langchain> Large Language Model application framework.
- [13] Dongjun Lee, Choongwon Park, Jaehyun Kim, and Heesoo Park. 2024. MCS-SQL: Leveraging Multiple Prompts and Multiple-Choice Selection For Text-to-SQL Generation. *arXiv:2405.07467 [cs.CL]* <https://arxiv.org/abs/2405.07467>
- [14] Fangyu Lei, Jixuan Chen, Yuxiao Ye, Ruisheng Cao, Dongchan Shin, Hongjin Su, Zhaoqing Suo, Hongcheng Gao, Wenjing Hu, Pengcheng Yin, et al. 2024. Spider 2.0: Evaluating language models on real-world enterprise text-to-sql workflows. *arXiv preprint arXiv:2411.07763* (2024).
- [15] 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. 2025. OmniSQL: Synthesizing High-Quality Text-to-SQL Data at Scale. *Proc. VLDB Endow.* 18, 11 (Sept. 2025), 4695–4709. <https://doi.org/10.14778/3749646.3749723>
- [16] Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhu Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, et al. 2024. Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. *Advances in Neural Information Processing Systems* 36 (2024).
- [17] Shu Liu, Soujanya Ponnappalli, Shreya Shankar, Sepanta Zeigham, Alan Zhu, Shubham Agarwal, Ruiqi Chen, Samion Suwito, Shuo Yuan, Ion Stoica, Matei Zaharia, Alvin Cheung, Natacha Crooks, Joseph E. Gonzalez, and Aditya G. Parameswaran. 2025. Supporting Our AI Overlords: Redesigning Data Systems to be Agent-First. *arXiv:2509.00997 [cs.AI]* <https://arxiv.org/abs/2509.00997>
- [18] Zihan Liu, Junhao Xu, Dongji Guo, et al. 2024. CodeAct: LLM-Based Autonomous Programming via Tool-Use and Reflection. *arXiv preprint arXiv:2402.16055* (2024). <https://arxiv.org/abs/2402.16055>
- [19] OpenAI. 2025. OpenAI o3 Reasoning Model. <https://platform.openai.com/docs/models>. Large Language Model.
- [20] Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2024. MemGPT: Towards LLMs as Operating Systems. *arXiv:2310.08560 [cs.AI]* <https://arxiv.org/abs/2310.08560>
- [21] 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. *arXiv:2410.01943 [cs.LG]* <https://arxiv.org/abs/2410.01943>
- [22] Mohammadreza Pourreza and Davood Rafiei. 2023. DIN-SQL: Decomposed In-Context Learning of Text-to-SQL with Self-Correction. *arXiv:2304.11015 [cs.CL]* <https://arxiv.org/abs/2304.11015>
- [23] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. <https://www.sbert.net/>
- [24] Alireza Rezazadeh, Zichao Li, Ange Lou, Yuying Zhao, Wei Wei, and Yujia Bao. 2025. Collaborative Memory: Multi-User Memory Sharing in LLM Agents with Dynamic Access Control. *arXiv:2505.18279 [cs.MA]* <https://arxiv.org/abs/2505.18279>
- [25] Noah Shinn, Joseph Labash, and Ashwin Gopinath. 2023. Reflexion: Language Agents with Verbal Reinforcement Learning. In *Advances in Neural Information Processing Systems (NeurIPS) Workshop on LLMs*. <https://arxiv.org/abs/2303.11366>
- [26] Ji Sun, Guoliang Li, Peiyao Zhou, Yihui Ma, Jingzhe Xu, and Yuan Li. 2025. AgenticData: An Agentic Data Analytics System for Heterogeneous Data. *arXiv:2508.05002 [cs.DB]* <https://arxiv.org/abs/2508.05002>
- [27] Shayan Talaei, Mohammadreza Pourreza, Yu-Chen Chang, Azalia Mirhoseini, and Amin Saberi. 2024. CHESS: Contextual Harnessing for Efficient SQL Synthesis. *arXiv:2405.16755 [cs.LG]* <https://arxiv.org/abs/2405.16755>
- [28] Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O’Sullivan, and Hoang D. Nguyen. 2025. Multi-Agent Collaboration Mechanisms: A Survey of LLMs. *arXiv:2501.06322 [cs.AI]* <https://arxiv.org/abs/2501.06322>
- [29] Bing Wang, Changyu Ren, Jian Yang, Ximian 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. *arXiv:2312.11242 [cs.CL]* <https://arxiv.org/abs/2312.11242>
- [30] Yihan Wang, Peiyu Liu, and Xin Yang. 2025. LinkAlign: Scalable Schema Linking for Real-World Large-Scale Multi-Database Text-to-SQL. *arXiv:2503.18596 [cs.CL]* <https://arxiv.org/abs/2503.18596>
- [31] Junde Wu, Jiayuan Zhu, Yuyuan Liu, Min Xu, and Yueming Jin. 2025. Agentic Reasoning: A Streamlined Framework for Enhancing LLM Reasoning with Agentic Tools. *arXiv:2502.04644 [cs.AI]* <https://arxiv.org/abs/2502.04644>
- [32] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2022. ReAct: Synergizing Reasoning and Acting in Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*. <https://arxiv.org/abs/2210.03629>
- [33] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. *arXiv:2210.03629 [cs.CL]* <https://arxiv.org/abs/2210.03629>
- [34] 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:2505.20315 [cs.CL]* <https://arxiv.org/abs/2505.20315>
- [35] 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. 2019. Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task. *arXiv:1809.08887 [cs.CL]* <https://arxiv.org/abs/1809.08887>
- [36] Bingxi Zhao, Lin Geng Foo, Ping Hu, Christian Theobalt, Hossein Rahmani, and Jun Liu. 2025. LLM-based Agentic Reasoning Frameworks: A Survey from Methods to Scenarios. *arXiv:2508.17692 [cs.AI]* <https://arxiv.org/abs/2508.17692>
- [37] Glenn Zorpette. 2025. Large language models are improving exponentially. <https://spectrum.ieee.org/large-language-model-performance>