Novelty Assessment Report

Paper: BIRD-INTERACT: Re-imagining Text-to-SQL Evaluation via Lens of Dynamic Interactions
PDF URL: https://openreview.net/pdf?id=nHrYBGujps
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-27

Abstract

Large language models (LLMs) have demonstrated remarkable performance on single-turn text-to-SQL tasks, but real-world database applications predominantly require multi-turn interactions to handle ambiguous queries, execution errors, and evolving user requirements. Existing multi-turn benchmarks fall short of capturing this complexity, either by treating conversation histories as static context or by limiting evaluation to narrow, read-only (SELECT-ONLY) operations, thereby failing to reflect the challenges encountered in production-grade database assistant. In this work, we introduce BIRD-INTERACT, a benchmark that restores this missing realism through: (1) a **comprehensive interaction environment** that couples each database with a hierarchical knowledge base, metadata files, and a function-driven user simulator, enabling models to solicit clarifications, retrieve knowledge, and recover from execution errors without human supervision; (2) two **evaluation settings** reflecting real-world interaction settings which contain a pre-defined conversational protocol (c-Interact) and a more open-ended agentic setting (a-Interact) in which the model autonomously decides when to query the user simulator or explore the DB environment; (3) a **challenging task suite** that covers the full CRUD spectrum for both business-intelligence and operational use cases, guarded by executable test cases. Each task features ambiguous and follow-up sub-tasks, requiring LLMs to engage in dynamic interaction. The suite is organized into two sets: a full set (**BIRD-INTERACT-FULL**) of 600 tasks which unfold up to **11,796** dynamic interactions for a comprehensive overview of performance and a lite set (**BIRD-INTERACT-LITE**) of 300 tasks, with simplified databases for detailed behavioral analysis of interactions, and fast development of methods. Our empirical results highlight the difficulty of BIRD-INTERACT: the most recent flagship model GPT-5 completes only **8.67%** of tasks in the c-Interact setting and **17.00%** in the a-Interact setting on the full task suite. Further analysis via memory grafting and Interaction Test-time Scaling (ITS), validate the importance of effective interaction for achieving success in complex, dynamic text-to-SQL tasks.

Core Task Landscape

This paper addresses: **Multi-Turn Interactive Text-to-SQL with Dynamic User Clarification**

A total of **46 papers** were analyzed and organized into a taxonomy with **18 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Interactive Clarification and Ambiguity Resolution**
- **Multi-Turn Conversational Text-to-SQL**
- **User Feedback-Based Query Correction**
- **Execution-Guided Refinement and Validation**
- **Query Generation and Refinement Techniques**
- **Domain-Specific and Application-Oriented Systems**
- **Surveys and Comprehensive Reviews**

Complete Taxonomy Tree

- Multi-Turn Interactive Text-to-SQL with Dynamic User Clarification Survey Taxonomy
- Interactive Clarification and Ambiguity Resolution
  - Proactive Ambiguity Detection and Question Generation (4 papers)
  - [1] An integrated interactive framework for natural language to SQL translation (Yuankai Fan, 2023) View paper
  - [7] AmbiSQL: Interactive Ambiguity Detection and Resolution for Text-to-SQL (Ding Zhongjun, 2025) View paper
  - [19] Interactive Text-to-SQL via Expected Information Gain for Disambiguation (Qiu, 2025) View paper
  - [32] Sphinteract: Resolving Ambiguities in NL2SQL through User Interaction (Fuheng Zhao, 2024) View paper
  - Parser-Independent Interactive Frameworks (2 papers)
  - [20] Model-based Interactive Semantic Parsing: A Unified Framework and A Text-to-SQL Case Study (Yao, 2019) View paper
  - [25] â□□What Do You Mean by That?â□□ - a Parser-Independent Interactive Approach for Enhancing Text-to-SQL (Yuntao Li, 2020) View paper
- Multi-Turn Conversational Text-to-SQL
  - Conversational Benchmark and Evaluation Frameworks ★ (4 papers)
  - [0] BIRD-INTERACT: Re-imagining Text-to-SQL Evaluation via Lens of Dynamic Interactions (Anon et al., 2026) View paper
  - [14] Cosql: A conversational text-to-sql challenge towards cross-domain natural language interfaces to databases (Yu Tao, 2019) View paper
  - [17] Rethinking Text-to-SQL: Dynamic Multi-turn SQL Interaction for Real-world Database Exploration (Guo Tianyu, 2025) View paper
  - [35] BIRD-INTERACT: Re-imagining Text-to-SQL Evaluation for Large Language Models via Lens of Dynamic Interactions (Huo, 2025) View paper
  - Conversational Dialogue Systems and Interfaces (4 papers)
  - [11] Conversational access to structured knowledge exploiting large models (Manto, 2022) View paper

    - [27] Marrying Dialogue Systems with Data Visualization: Interactive Data Visualization Generation from Natural Language Conversations (Yuanfeng Song, 2024) View paper
    - [28] QueryGenie: Making LLM-Based Database Querying Transparent and Controllable (Chen LongFei, 2025) View paper
    - [40] Conversational interfaces for unconventional access to business relational data structures (Pavel KostelnÃk, 2021) View paper
    - Multi-Turn Agentic and Reinforcement Learning Approaches (3 papers)
    - [9] MARS-SQL: A multi-agent reinforcement learning framework for Text-to-SQL (Yang HaoLin, 2025) View paper
    - [31] CIRCLE: Multi-Turn Query Clarifications with Reinforcement Learning (Erbacher, 2023) View paper
    - [34] MTSQL-R1: Towards Long-Horizon Multi-Turn Text-to-SQL via Agentic Training (Guo, 2025) View paper
- User Feedback-Based Query Correction
    - Post-Generation Natural Language Correction (3 papers)
    - [3] Fisql: Enhancing text-to-sql systems with rich interactive feedback (Qian, 2025) View paper
    - [16] NL-EDIT: Correcting Semantic Parse Errors through Natural Language Interaction (Awadallah, 2021) View paper
    - [41] Speak to your Parser: Interactive Text-to-SQL with Natural Language Feedback (Elgohary, 2020) View paper
    - Continual Learning from User Feedback (2 papers)
    - [21] Utilizing Past User Feedback for More Accurate Text-to-SQL (Matthias Urban, 2025) View paper
    - [26] Continual Learning of Domain Knowledge from Human Feedback in Text-to-SQL (Thomas Cook, 2025) View paper
    - Interactive Explanation and Visualization Interfaces (2 papers)
    - [4] SQLucid: Grounding Natural Language Database Queries with Interactive Explanations (Yuan Tian, 2024) View paper
    - [15] IntelliExplain: Enhancing Conversational Code Generation for Non-Professional Programmers (Yan Hao, 2024) View paper
- Execution-Guided Refinement and Validation
    - Iterative Execution Feedback Loops (4 papers)
    - [6] Enhancing text-to-SQL parsing through question rewriting and execution-guided refinement (Wenxin Mao, 2024) View paper
    - [10] Retrieval-augmented gpt-3.5-based text-to-sql framework with sample-aware prompting and dynamic revision chain (Chunxi Guo, 2023) View paper
    - [23] Optimizing Reasoning for Text-to-SQL with Execution Feedback (Bohan Zhai, 2025) View paper
    - [24] ExCoT: Optimizing Reasoning for Text-to-SQL with Execution Feedback (Zhai, 2025) View paper
    - Execution-Based Optimization and Reasoning (2 papers)
    - [12] Review, Refine, Repeat: Understanding Iterative Decoding of AI Agents with Dynamic Evaluation and Selection (S Chakraborty, 2025) View paper
    - [38] TS-SQL: Test-driven Self-refinement for Text-to-SQL (Wenbo Xu, 2025) View paper
    - Validation and Accuracy Assurance Frameworks (3 papers)
    - [5] What Do You Mean? Using Large Language Models for Semantic Evaluation of NL2SQL Queries (Noor, 2025) View paper
    - [43] Ensuring Data Accuracy in Text-to-SQL Systems: A Comprehensive Validation Framework (Piyush Pandey, n.d.) View paper
    - [45] Automated Evaluation of Database Conversational Agents (Matheus de Souza Silva, n.d.) View paper
- Query Generation and Refinement Techniques
    - Retrieval-Augmented and In-Context Learning (2 papers)
    - [2] In-context reinforcement learning with retrieval-augmented generation for Text-to-SQL (R Toteja, 2025) View paper
    - [22] Reliable Answers for Recurring Questions: Boosting Text-to-SQL Accuracy with Template Constrained Decoding (S Jivani, 2025) View paper
    - Multi-Agent and Decomposition Approaches (2 papers)
    - [33] SQLfuse: Enhancing Text-to-SQL Performance through Comprehensive LLM Synergy (Zhang Tingkai, 2024) View paper
    - [37] MAG-SQL: Multi-Agent Generative Approach with Soft Schema Linking and Iterative Sub-SQL Refinement for Text-to-SQL (Xie Wenxuan, 2024) View paper
    - Schema Linking and Refinement (1 papers)
    - [46] Schema-Refiner: Synergizing Knowledge Graphs and LLMs for Proactive Schema Refinement in Text-to-SQL (J Wang, n.d.) View paper
    - Rationalization and Chain-of-Thought Methods (1 papers)
    - [39] Rationalization Models for Text-to-SQL (Rossiello, 2025) View paper
- Domain-Specific and Application-Oriented Systems
    - Business Intelligence and Requirements Automation (2 papers)
    - [29] Based on BERT-GPT-GNN converged architecture: intelligent generation engine for complex SQL queries in business intelligence (Shiwei Chu, 2025) View paper
    - [36] Automating Business Intelligence Requirements with Generative AI and Semantic Search (Nimrod Busany, 2024) View paper
    - Non-English and Multilingual Text-to-SQL (1 papers)
    - [30] Large language model based automated translation of natural language to SQL (KanburoÄlu, 2025) View paper
- Surveys and Comprehensive Reviews (5 papers)
    - [8] Exploring Large Language Models in Information Systems: A Survey (Filippo Bianchini, 2025) View paper
    - [13] Chat-SQL: Natural Language text to SQL Queries based on Deep Learning Techniques (K MAJHADI, 2024) View paper
    - [18] Conversational Text-to-SQL: A Comprehensive Survey of Paradigms, Challenges, and Future Directions (Yu-Fei Yang, 2025) View paper
    - [42] Interactive Intelligent Systems (M Zhou, 2024) View paper
    - [44] Synthesizing SQL from Natural Language (A Doshi, 2020) View paper

Narrative

Core task: multi-turn interactive text-to-SQL with dynamic user clarification. The field addresses scenarios where natural language queries are inherently ambiguous or incomplete, requiring systems to engage users in clarifying dialogues before generating accurate SQL. The taxonomy organizes research into several main branches: Interactive Clarification and Ambiguity Resolution focuses on detecting and resolving unclear user intent through targeted questions; Multi-Turn Conversational Text-to-SQL examines how systems maintain context across dialogue turns, with benchmarks like CoSQL[14] and Dynamic Multi-turn SQL[17] establishing evaluation frameworks; User Feedback-Based Query Correction explores mechanisms for incorporating explicit corrections; Execution-Guided Refinement and Validation leverages query results to iteratively improve outputs; Query Generation and Refinement Techniques develops core algorithmic advances; Domain-Specific and Application-Oriented Systems tailors solutions to particular use cases; and Surveys and Comprehensive Reviews synthesize broader trends, as seen in Conversational Text-to-SQL Survey[18] and LLM Information Systems Survey[8].

Recent work reveals contrasting strategies for handling ambiguity and multi-turn interaction. Some approaches emphasize proactive clarification, such as AmbiSQL[7] detecting ambiguous queries and Expected Information Gain[19] optimizing question selection, while others like Fisql Interactive Feedback[3] and SQLucid[4] focus on refining queries through iterative user feedback loops. Execution-guided methods, including Execution Feedback Reasoning[23] and ExCoT[24], validate generated SQL against database results to trigger refinement. BIRD-INTERACT[0] situates itself within the conversational benchmark branch alongside CoSQL[14] and Dynamic Multi-turn SQL[17], providing a framework for evaluating how well systems handle dynamic clarification across multiple turns. Compared to BIRD-INTERACT LLM[35], which explores LLM-specific strategies on the same benchmark, BIRD-INTERACT[0] emphasizes the broader evaluation infrastructure. The interplay between proactive ambiguity detection, feedback incorporation, and execution validation remains an active area, with open questions around balancing user burden against query accuracy.

Related Works in Same Category

The following **3 sibling papers** share the same taxonomy leaf node with the original paper:

1. Cosql: A conversational text-to-sql challenge towards cross-domain natural language interfaces to databases

Authors: Yu Tao, Zhang Rui, Er, He Yang, Li Suyi, et al. (39 authors total) | **Year/Venue**: 2019 | **URL**: View paper

Abstract

We present CoSQL, a corpus for building cross-domain, general-purpose database (DB) querying dialogue systems. It consists of 30k+ turns plus 10k+ annotated SQL queries, obtained from a Wizard-of-Oz (WOZ) collection of 3k dialogues querying 200 complex DBs spanning 138 domains. Each dialogue simulates a real-world DB query scenario with a crowd worker as a user exploring the DB and a SQL expert retrieving answers with SQL, clarifying ambiguous questions, or otherwise informing of unanswerable qu...

Relationship Analysis

Both papers belong to the Conversational Benchmark and Evaluation Frameworks category, focusing on multi-turn text-to-SQL evaluation with dynamic interactions. CoSQL provides a Wizard-of-Oz collected dataset with static conversation transcripts and focuses primarily on SELECT-only queries, while BIRD-INTERACT introduces a dynamic interactive environment with a function-driven user simulator, executable test cases, and full CRUD spectrum coverage including ambiguity resolution and follow-up tasks. The key distinction is that BIRD-INTERACT enables dynamic interaction evaluation where models autonomously navigate conversations, whereas CoSQL evaluates against predetermined dialogue trajectories.

---

2. Rethinking Text-to-SQL: Dynamic Multi-turn SQL Interaction for Real-world Database Exploration

Authors: Guo Tianyu, Liang Hao, Li Yuying, Cai Qi-feng, Wei Jingxuan, et al. (8 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Recent advances in Text-to-SQL have achieved strong results in static, single-turn tasks, where models generate SQL queries from natural language questions. However, these systems fall short in real-world interactive scenarios, where user intents evolve and queries must be refined over multiple turns. In applications such as finance and business analytics, users iteratively adjust query constraints or dimensions based on intermediate results. To evaluate such dynamic capabilities, we introduce D...

Relationship Analysis

Both papers belong to the Conversational Benchmark and Evaluation Frameworks category, focusing on multi-turn interactive text-to-SQL evaluation with dynamic user interactions. They overlap in addressing the limitations of static single-turn benchmarks by introducing interactive environments with user simulators, executable databases, and multi-turn task sequences covering CRUD operations. However, BIRD-INTERACT emphasizes ambiguity resolution through a two-stage function-driven user simulator with hierarchical knowledge bases and distinguishes between protocol-guided (c-Interact) and agentic (a-Interact) evaluation modes, while DySQL-Bench focuses on automated task synthesis via tree-structured database representations and evaluates models through a triadic user-agent-database interaction framework with hash-based state verification.

---

3. BIRD-INTERACT: Re-imagining Text-to-SQL Evaluation for Large Language Models via Lens of Dynamic Interactions

Authors: Huo, Nan, Xu Xiaohan, Li Jinyang, Qin Bo-wen, et al. (26 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Large language models (LLMs) have demonstrated remarkable performance on single-turn text-to-SQL tasks, but real-world database applications predominantly require multi-turn interactions to handle ambiguous queries, execution errors, and evolving user requirements. Existing multi-turn benchmarks fall short by treating conversation histories as static context or limiting evaluation to read-only operations, failing to reflect production-grade database assistant challenges. We introduce BIRD-INTERA...

⚠ Similarity Notice

These papers share nearly identical titles, abstracts, and technical content, describing the same BIRD-INTERACT benchmark with the same evaluation settings (c-Interact and a-Interact), user simulator design, and experimental results. The only notable difference is the author attribution (anonymous vs. named authors), suggesting these are different submission versions of the same work rather than distinct papers.

Contributions Analysis

Overall novelty summary. The paper introduces BIRD-INTERACT, a benchmark for multi-turn interactive text-to-SQL that couples databases with hierarchical knowledge bases, metadata, and a function-driven user simulator. It resides in the 'Conversational Benchmark and Evaluation Frameworks' leaf alongside three sibling papers: CoSQL, Dynamic Multi-turn SQL, and another conversational evaluation framework. This leaf contains four papers total within a taxonomy of 46 papers across 18 leaf nodes, indicating a moderately populated but not overcrowded research direction focused specifically on multi-turn conversational evaluation rather than single-turn or execution-only methods.

The taxonomy reveals neighboring branches addressing related but distinct challenges. 'Proactive Ambiguity Detection and Question Generation' (four papers) focuses on pre-generation clarification, while 'Multi-Turn Agentic and Reinforcement Learning Approaches' (three papers) explores agent-based long-horizon tasks. 'Conversational Dialogue Systems and Interfaces' (four papers) emphasizes end-to-end dialogue management rather than benchmark design. BIRD-INTERACT bridges evaluation infrastructure with interaction modeling, distinguishing itself from execution-guided refinement branches that automate correction without user involvement and from post-generation feedback systems that rely on explicit user corrections after SQL generation.

Among 15 candidates examined through limited semantic search, none clearly refute the three identified contributions. The comprehensive interaction environment (nine candidates examined, zero refutable) and dual evaluation settings of c-Interact and a-Interact (six candidates examined, zero refutable) show no substantial prior overlap within this search scope. The function-driven user simulator received no candidate examination, suggesting either novelty or insufficient search coverage in that specific dimension. These

statistics reflect a constrained literature search rather than exhaustive field coverage, indicating that within the examined top-15 semantically similar papers, no direct precedents emerged.

Based on the limited search scope of 15 candidates, the work appears to occupy a distinct position within conversational text-to-SQL benchmarking, particularly through its integration of knowledge bases, metadata, and autonomous user simulation. The taxonomy structure confirms this sits in a moderately active but not saturated research direction. However, the analysis does not cover broader benchmark literature outside the top-15 semantic matches, and the zero-refutation finding reflects search limitations rather than definitive novelty claims across the entire field.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: BIRD-INTERACT benchmark with comprehensive interaction environment

Description: The authors develop a new benchmark featuring an interactive environment that includes databases, hierarchical knowledge bases, metadata files, and a function-driven user simulator. This environment enables models to solicit clarifications, retrieve knowledge, and recover from execution errors without human supervision, addressing limitations of static conversation transcripts in existing benchmarks.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. A requirements driven framework for benchmarking semantic web knowledge base systems

URL: View paper

Brief Assessment

Requirements Driven Benchmark[56] focuses on developing a framework for benchmarking semantic web knowledge base query systems, not interactive text-to-SQL environments with user simulators and dynamic database interactions.

---

2. Clear-kgqa: Clarification-enhanced ambiguity resolution for knowledge graph question answering

URL: View paper

Brief Assessment

Clear-kgqa[51] focuses on knowledge graph question answering with entity and intent disambiguation, not text-to-SQL tasks with databases, hierarchical knowledge bases, and user simulators as described in the original contribution.

---

3. Simulated User Behavior for Recommender Systems Applied to the MIND Dataset

URL: View paper

Brief Assessment

MIND User Simulation[55] focuses on simulating user behavior for news recommender systems using the MIND dataset, not on interactive database benchmarks with knowledge bases and user simulators for text-to-SQL tasks.

---

4. SAGE: A Top-Down Bottom-Up Knowledge-Grounded User Simulator for Multi-turn AGent Evaluation

URL: View paper

Brief Assessment

SAGE User Simulator[57] focuses on user simulation for multi-turn agent evaluation in business contexts (customer support, sales), not on interactive database benchmarks with SQL generation and execution environments.

---

5. From Conversation to Query Execution: Benchmarking User and Tool Interactions for EHR Database Agents

URL: View paper

Brief Assessment

EHR Database Agents[54] focuses on interactive question answering over electronic health records with simulated users and tools for clinical data access, while the original paper addresses text-to-SQL with hierarchical knowledge bases and user simulators for general database interactions. The domains and specific interaction mechanisms differ substantially.

---

6. BIRD-INTERACT: Re-imagining Text-to-SQL Evaluation for Large Language Models via Lens of Dynamic Interactions

URL: View paper

Brief Assessment

BIRD-INTERACT LLM[35] focuses on text-to-SQL evaluation with databases, knowledge bases, and user simulators, while the original paper addresses general RL frameworks for agent training. These are distinct application domains with different technical objectives.

---

7. Simulating Users in Interactive Web Table Retrieval

URL: View paper

Brief Assessment

Web Table Retrieval Simulation[59] focuses on simulated interactive web table retrieval with query reformulation strategies, not text-to-SQL tasks with databases, knowledge bases, and user simulators for resolving SQL ambiguities.

---

8. BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models

URL: View paper

Brief Assessment

BioModels Database[52] is a repository for quantitative kinetic models in systems biology, not an interactive database benchmark for text-to-SQL evaluation. The candidate focuses on storing and distributing biochemical simulation models with metadata annotations, while the original paper presents an interactive environment for evaluating LLMs on dynamic text-to-SQL tasks with user simulators and knowledge bases.

---

9. Studying the effectiveness of conversational search refinement through user simulation

URL: View paper

Brief Assessment

Conversational Search Refinement[53] focuses on conversational search query clarification with a user simulator for search intent refinement, not on text-to-SQL tasks with databases, knowledge bases, and SQL execution environments as in the original paper.

Contribution 2: Two evaluation settings: c-Interact and a-Interact

Description: The authors propose two distinct evaluation modes: c-Interact tests models' ability to follow structured conversational protocols, while a-Interact evaluates autonomous planning where models decide when to query users or explore the database environment. These settings reflect different real-world interaction scenarios for database assistants.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. ChatbotSQL: Conversational agent to support relational database query language learning

URL: View paper

Brief Assessment

ChatbotSQL[47] is an educational conversational agent for teaching SQL to students, not an evaluation framework for database query systems. It does not propose evaluation settings for assessing text-to-SQL models.

---

2. Conversational Text-to-SQL: A Comprehensive Survey of Paradigms, Challenges, and Future Directions

URL: View paper

Brief Assessment

Conversational Text-to-SQL Survey[18] is a survey paper that reviews existing paradigms in conversational text-to-SQL systems. It does not propose specific evaluation settings but rather categorizes existing approaches. The original paper introduces novel evaluation frameworks (c-Interact and a-Interact) for dynamic text-to-SQL interaction, which is a distinct methodological contribution not addressed by a survey of the field.

---

3. BIRD-INTERACT: Re-imagining Text-to-SQL Evaluation for Large Language Models via Lens of Dynamic Interactions

URL: View paper

Brief Assessment

BIRD-INTERACT LLM[35] proposes c-Interact and a-Interact settings specifically for database query systems, whereas the original paper's evaluation settings target general reinforcement learning agent training. The technical contexts differ fundamentally.

---

4. Target: Benchmarking table retrieval for generative tasks

URL: View paper

Brief Assessment

Target Table Retrieval[48] focuses on table retrieval for generative tasks (question answering, fact verification, text-to-sql) rather than conversational versus agentic evaluation modes for database query systems. The candidate does not address multi-turn interaction protocols or autonomous planning settings.

---

5. Toward A Self-Evolving Agent In Multi-Turn Dialogue Question-Answering Systems

URL: View paper

Brief Assessment

Self-Evolving Agent[50] focuses on multi-turn dialogue QA systems with domain-specific learning and modular agent design, not on database query evaluation settings like c-Interact and a-Interact.

---

6. Conversational vs Traditional: Comparing Search Behavior and Outcome in Legal Case Retrieval

URL: View paper

Brief Assessment

Conversational Legal Retrieval[49] focuses on comparing conversational versus traditional search paradigms in legal case retrieval, not on evaluating different interaction modes (protocol-guided vs. agentic) for database query systems. The candidate does not address evaluation settings that distinguish between structured conversational protocols and autonomous planning capabilities.

Contribution 3: Function-driven user simulator with two-stage approach

Description: The authors introduce a two-stage user simulator design where an LLM first parses clarification requests into predefined symbolic actions (AMB, LOC, UNA), then generates responses based on these actions and annotated ground-truth SQL. This approach prevents ground-truth leakage and ensures predictable, controllable simulator behavior while maintaining context-aware interactions.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Appendix: Text Similarity Detection

Textual similarity detection checked 17 papers and found 4 similarity segment(s) across 2 paper(s).

The following **2 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. BIRD-INTERACT: Re-imagining Text-to-SQL Evaluation for Large Language Models via Lens of Dynamic Interactions

Detected in: Core Task (sibling), Contribution: contribution_1, Contribution: contribution_2

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

---

2. Cosql: A conversational text-to-sql challenge towards cross-domain natural language interfaces to databases

Detected in: Core Task (sibling)

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] BIRD-INTERACT: Re-imagining Text-to-SQL Evaluation via Lens of Dynamic Interactions View paper
- [1] An integrated interactive framework for natural language to SQL translation View paper

- [2] In-context reinforcement learning with retrieval-augmented generation for Text-to-SQL View paper
- [3] Fisql: Enhancing text-to-sql systems with rich interactive feedback View paper
- [4] SQLucid: Grounding Natural Language Database Queries with Interactive Explanations View paper
- [5] What Do You Mean? Using Large Language Models for Semantic Evaluation of NL2SQL Queries View paper
- [6] Enhancing text-to-SQL parsing through question rewriting and execution-guided refinement View paper
- [7] AmbiSQL: Interactive Ambiguity Detection and Resolution for Text-to-SQL View paper
- [8] Exploring Large Language Models in Information Systems: A Survey View paper
- [9] MARS-SQL: A multi-agent reinforcement learning framework for Text-to-SQL View paper
- [10] Retrieval-augmented gpt-3.5-based text-to-sql framework with sample-aware prompting and dynamic revision chain View paper
- [11] Conversational access to structured knowledge exploiting large models View paper
- [12] Review, Refine, Repeat: Understanding Iterative Decoding of AI Agents with Dynamic Evaluation and Selection View paper
- [13] Chat-SQL: Natural Language text to SQL Queries based on Deep Learning Techniques View paper
- [14] Cosql: A conversational text-to-sql challenge towards cross-domain natural language interfaces to databases View paper
- [15] IntelliExplain: Enhancing Conversational Code Generation for Non-Professional Programmers View paper
- [16] NL-EDIT: Correcting Semantic Parse Errors through Natural Language Interaction View paper
- [17] Rethinking Text-to-SQL: Dynamic Multi-turn SQL Interaction for Real-world Database Exploration View paper
- [18] Conversational Text-to-SQL: A Comprehensive Survey of Paradigms, Challenges, and Future Directions View paper
- [19] Interactive Text-to-SQL via Expected Information Gain for Disambiguation View paper
- [20] Model-based Interactive Semantic Parsing: A Unified Framework and A Text-to-SQL Case Study View paper
- [21] Utilizing Past User Feedback for More Accurate Text-to-SQL View paper
- [22] Reliable Answers for Recurring Questions: Boosting Text-to-SQL Accuracy with Template Constrained Decoding View paper
- [23] Optimizing Reasoning for Text-to-SQL with Execution Feedback View paper
- [24] ExCoT: Optimizing Reasoning for Text-to-SQL with Execution Feedback View paper
- [25] â□□What Do You Mean by That?â□□ - a Parser-Independent Interactive Approach for Enhancing Text-to-SQL View paper
- [26] Continual Learning of Domain Knowledge from Human Feedback in Text-to-SQL View paper
- [27] Marrying Dialogue Systems with Data Visualization: Interactive Data Visualization Generation from Natural Language Conversations View paper
- [28] QueryGenie: Making LLM-Based Database Querying Transparent and Controllable View paper
- [29] Based on BERT-GPT-GNN converged architecture: intelligent generation engine for complex SQL queries in business intelligence View paper
- [30] Large language model based automated translation of natural language to SQL View paper
- [31] CIRCLE: Multi-Turn Query Clarifications with Reinforcement Learning View paper
- [32] Sphinteract: Resolving Ambiguities in NL2SQL through User Interaction View paper
- [33] SQLfuse: Enhancing Text-to-SQL Performance through Comprehensive LLM Synergy View paper
- [34] MTSQL-R1: Towards Long-Horizon Multi-Turn Text-to-SQL via Agentic Training View paper
- [35] BIRD-INTERACT: Re-imagining Text-to-SQL Evaluation for Large Language Models via Lens of Dynamic Interactions View paper
- [36] Automating Business Intelligence Requirements with Generative AI and Semantic Search View paper
- [37] MAG-SQL: Multi-Agent Generative Approach with Soft Schema Linking and Iterative Sub-SQL Refinement for Text-to-SQL View paper
- [38] TS-SQL: Test-driven Self-refinement for Text-to-SQL View paper
- [39] Rationalization Models for Text-to-SQL View paper
- [40] Conversational interfaces for unconventional access to business relational data structures View paper
- [41] Speak to your Parser: Interactive Text-to-SQL with Natural Language Feedback View paper
- [42] Interactive Intelligent Systems View paper
- [43] Ensuring Data Accuracy in Text-to-SQL Systems: A Comprehensive Validation Framework View paper
- [44] Synthesizing SQL from Natural Language View paper
- [45] Automated Evaluation of Database Conversational Agents View paper
- [46] Schema-Refiner: Synergizing Knowledge Graphs and LLMs for Proactive Schema Refinement in Text-to-SQL View paper
- [47] ChatbotSQL: Conversational agent to support relational database query language learning View paper
- [48] Target: Benchmarking table retrieval for generative tasks View paper
- [49] Conversational vs Traditional: Comparing Search Behavior and Outcome in Legal Case Retrieval View paper
- [50] Toward A Self-Evolving Agent In Multi-Turn Dialogue Question-Answering Systems View paper
- [51] Clear-kgqa: Clarification-enhanced ambiguity resolution for knowledge graph question answering View paper
- [52] BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models View paper
- [53] Studying the effectiveness of conversational search refinement through user simulation View paper
- [54] From Conversation to Query Execution: Benchmarking User and Tool Interactions for EHR Database Agents View paper
- [55] Simulated User Behavior for Recommender Systems Applied to the MIND Dataset View paper
- [56] A requirements driven framework for benchmarking semantic web knowledge base systems View paper
- [57] SAGE: A Top-Down Bottom-Up Knowledge-Grounded User Simulator for Multi-turn AGent Evaluation View paper
- [58] LUBM: A benchmark for OWL knowledge base systems View paper
- [59] Simulating Users in Interactive Web Table Retrieval View paper