# From Text-to-SQL Benchmarks to LLM-Powered Data Analytics: An Overview of Recent Advances

Marczis Bálint

## Abstract

The long-standing challenge of Text-to-SQL (translating human language into structured SQL queries) addresses the needs of users across many fields, such as business analytics, research, and public data management, where individuals must work with large datasets but lack SQL expertise. Early deep neural network approaches struggled with schema understanding and generalization, leading to the development of more advanced LLM-based solutions. This paper summarizes four representative works in this area: the Spider and BIRD benchmarks, Zhu et al.'s comprehensive survey of LLM-based Text-to-SQL methods, and CoddLLM, a recent LLM model explicitly trained for data analytics. These studies illustrate the progression from traditional neural models toward domain-specialized LLMs capable of making data analysis more intuitive, accurate, and widely accessible.

## 1. INTRODUCTION

Natural language interfaces to databases (NLIDB) have long represented a key challenge in the intersection of data management and natural language processing. The goal is to allow users to express analytical intents in natural language and automatically translate them into executable structured queries such as SQL. Early work on semantic parsing provided foundational methods, but research in this field accelerated dramatically with the introduction of large-scale, standardized benchmarks and the recent emergence of large language models (LLMs). This paper outlines the evolution of this research direction through four works: Spider [1], BIRD [2], Zhu et al.'s survey [3], and CoddLLM [4].

The **Spider dataset** [1], introduced in 2018, represented a turning point for Text-to-SQL research by providing the first large-scale, human-labeled benchmark designed to test a model's ability to generalize across different database schemas. Unlike earlier datasets, Spider separated training and test databases, ensuring that models could not simply memorize query patterns. Containing over 10,000 natural language questions and 200 databases spanning diverse domains, Spider became the standard testbed for evaluating semantic parsing models. Its results demonstrated that most systems at the time such as Seq2SQL, SQLNet and other deep neural network–based models performed poorly when faced with unseen database schemas. This revealed the field's key challenge: enabling robust schema understanding and logical reasoning across domains rather than relying on pattern memorization.

Building on this foundation, the **BIRD benchmark** [2] expanded the Text-to-SQL task into real-world and large-scale contexts, introducing the first benchmark explicitly tailored to LLMs. Whereas Spider focused on synthetic, well-structured data, BIRD simulated the messiness of enterprise databases, featuring incomplete column names, denormalized schemas, and inconsistent metadata. With over 98,000 natural language questions and 12,000 databases, BIRD provided a more realistic environment to assess whether large language models could serve as practical database interfaces. Experimental results revealed that even the most capable models, such as *GPT-4*, achieved only around **54.9%** execution accuracy, while human performance reached over **92%**. These findings highlighted that despite their strong language understanding, LLMs still struggle with noisy, large-scale data and schema inconsistencies, exposing a persistent gap between linguistic fluency and true database reasoning. This transition marked a shift from academic datasets to industrial-scale evaluation.

As research attention shifted toward LLM-based Text-to-SQL systems, comprehensive analyses of their strengths and weaknesses emerged. The survey by **Zhu et al. (2024)** [3] synthesized this progress, offering a structured taxonomy and evaluation of LLM-enhanced Text-to-SQL approaches. The authors found that prompt-based LLM systems, while impressive in linguistic flexibility, often fail at key analytical reasoning tasks. Common issues include incorrect schema linking, flawed joins and aggregations, and hallucinated columns or tables. The survey emphasized that such failures stem from the lack of explicit data-structure understanding in current models and identified the need for domain specific fine-tuning and structured data representation highlighting the importance of moving beyond pure prompting strategies.

Beyond identifying failure patterns, the survey categorized recent LLM-based approaches into four main paradigms: *prompt engineering*, *fine-tuning*, *task-specific training*, and *LLM agent methods*. Prompt engineering leverages large pretrained models through well-crafted instructions (e.g., zero-shot, few-shot, or chain-of-thought prompting), requiring no additional training but remaining highly sensitive to prompt phrasing. Fine-tuning techniques, by contrast, adapt model parameters on domain specific Text-to-SQL datasets, either fully or with efficient low-rank adaptation methods such as LoRA and QLoRA, improving schema reasoning but demanding significant computational resources. Task specific training from scratch, using transformer or mixture of experts architectures, enables direct supervision of SQL generation but sacrifices gen-

eral language capabilities. Finally, emerging LLM-agent frameworks employ multiple interacting components such as query planners, schema interpreters, and error-correction modules to achieve self reflective reasoning. Zhu et al. concluded that while current LLMs show remarkable promise, achieving robust Text-to-SQL performance will require hybrid approaches that integrate the structured reasoning of symbolic systems with the contextual understanding of language models.

This motivation directly inspired the development of **CoddLLM** [4], a 2025 work proposing an LLM explicitly post-trained for data analytics. Rather than relying solely on prompts, CoddLLM introduces a new data-centric "recipe" for fine-tuning LLMs using synthetic data generation focused on database manipulation and representation. The model, based on Mistral-NeMo-12B, is trained on tasks that bridge tables and text, such as schema creation and table-to-text translation, allowing it to internalize database reasoning and schema comprehension. To evaluate these capabilities, the authors released *AnalyticsMMLU*, a new benchmark containing analytical reasoning questions and database-centric challenges. The results showed that CoddLLM achieved a **24.9% average improvement** in Text-to-SQL performance compared to its base model, and outperformed **GPT-4o by 12.1%** in table selection accuracy. Across eight datasets, CoddLLM obtained the highest overall accuracy and demonstrated superior schema comprehension and reasoning capabilities, setting a new benchmark for LLMs designed for structured data understanding.

Beyond these quantitative gains, the authors emphasized that true database reasoning still remains an open challenge for large language models. They identified several key directions for future research: improving LLMs' ability to reason over complex joins and nested queries, integrating symbolic reasoning modules for precise schema grounding, and enhancing the interpretability of generated SQL statements. Furthermore, they proposed extending the model's training to encompass multimodal data sources such as text, tables, and charts, reflecting real-world analytical workflows. Ultimately, the authors argue that bridging natural language understanding with formal data logic will be crucial for the next generation of foundation models capable of autonomous, trustworthy, and context-aware data analytics.

Taken together, these four works illustrate the evolution of the Text-to-SQL and LLM for analytics landscape: from standardized benchmarking **Spider**, through real-world scalability **BIRD**, to diagnostic understanding of model limitations **Zhu et al.**, and finally to purpose-built solutions **CoddLLM**. The progression reflects a broader shift in data management research from viewing LLMs as mere text generators toward recognizing them as emerging analytical agents capable of understanding, reasoning about, and interacting with structured data.

## References


[1] Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Bo Tan, Xi Victoria Li, et al. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3911–3921, 2018.

[2] Bo Li, Ruoxi Ding, Yeyun Fang, Weize Lin, Weijia Zhao, Dayiheng Liu, Li Chen, Zhifang Liu, Tiejun Zhao, and Rui Zhang. Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. *arXiv:2305.03111*, 2023.

[3] Xiaohu Zhu, Qian Li, Lizhen Cui, and Yongkang Liu. Large language model enhanced text-to-sql generation: A survey. *arXiv:2410.06011*, 2024.

[4] Zexuan Wang, Qiang Zhang, Yifan Wu, Yujia Li, Yichi Yang, and Qi Luo. Coddllm: Empowering large language models for data analytics. *arXiv:2502.00329*, 2025.