

الجمهورية الشعبية الديمقراطية الجزائرية
People's Democratic Republic of Algeria
وزارة التعليم العالي و البحث العلمي
Ministry of Higher Education and Scientific Research
المدرسة العليا للإعلام الآلي 8 ماي 1945 - سيدي بلعباس
Higher School of Computer Science
8 Mai 1945 - Sidi Bel Abbès



Thesis

To obtain the diploma of **Master's Degree**
Field of Study: **Computer Science**
Specialization: **Artificial Intelligence and Data Science**

Theme

A Comparative Study of Text-to-SQL Systems on the BIRD Benchmark

Presented by
Zahra Boucheta

Defended on: **June 29, 2026**
In front of the jury composed of

Dr. TAOULI Amina
Dr. NAIB Rabiaa
Dr. CHAIB Souleyman
Dr. ABDELHAK Soumia

President of the Jury
Examiner
Supervisor
Incubator Representative

Academic Year: 2025/2026

Abstract

Large language models (LLMs) have increasingly been integrated into Text-to-SQL systems, enabling the transformation of natural language specifications into executable SQL queries over structured data sources. Despite notable advances in this direction, a persistent gap remains between the performance of current automated approaches and the capabilities exhibited by expert users in realistic database interaction scenarios.

This study undertakes a systematic comparative analysis of eight state-of-the-art Text-to-SQL frameworks, evaluated under the BIRD benchmark, which is designed to reflect complex, real-world database reasoning and query generation tasks. The examined systems are analyzed along multiple dimensions, including architectural design paradigms, benchmark performance, computational efficiency, and training methodologies.

The empirical findings suggest that performance gains are more strongly associated with system-level design choices and orchestration strategies than with increases in model scale alone. In addition, approaches incorporating reinforcement learning exhibit improved compositional generalization when compared to conventional supervised fine-tuning paradigms.

Furthermore, schema grounding emerges as a central unresolved challenge across all evaluated systems, with preliminary estimates indicating potential improvements of approximately five percentage points in execution accuracy if adequately addressed [1]. Although the strongest-performing system achieves an execution accuracy of 81.67%, a measurable gap remains relative to estimated human performance, on the order of approximately eleven percentage points. These observations suggest that the development of robust and generalizable natural language interfaces for complex database systems continues to constitute an open research problem.

Keywords: Text-to-SQL, Large Language Models, BIRD Benchmark, Schema Linking, Reinforcement Learning, Test-Time Scaling, Execution Accuracy, Natural Language Interfaces to Databases

ملخص

شهدت النماذج اللغوية الكبيرة تطوراً ملحوظاً في مجال أنظمة تحويل اللغة الطبيعية إلى SQL، مما أتاح تحويل الاستفسارات المكتوبة باللغة الطبيعية إلى استعلامات قابلة للتنفيذ على قواعد البيانات المهيكلية. وعلى الرغم من التقدم الكبير المحقق في هذا المجال، لا تزال هناك فجوة واضحة بين أداء الأنظمة الآلية الحالية والأداء الذي يحققه المستخدمون الخبراء عند التفاعل مع قواعد البيانات في البيئات الواقعية. تقدم هذه الدراسة تحليلاً مقارناً منهجياً ثمانية من أحدث أنظمة تحويل اللغة الطبيعية إلى SQL، والتي تم تقييمها باستخدام معيار BIRD المصمم لمحاكاة مهام الاستدلال المعقدة وتوليد الاستعلامات على قواعد البيانات الواقعية. وقد شمل التحليل عدة جوانب، من بينها البنى المعمارية للأنظمة، والأداء على المعيار المرجعي، والكفاءة الحاسوبية، ومنهجيات التدريب المعتمدة. تشير النتائج التجريبية إلى أن التحسينات في الأداء ترتبط بدرجة أكبر بالاختيارات المعمارية على مستوى النظام واستراتيجيات التوجيه والتنسيق، مقارنةً بالاعتماد على زيادة حجم النموذج فقط. كما أظهرت الأنظمة المعتمدة على التعلم المعزز قدرة أفضل على التعميم في المهام المعقدة مقارنةً بالأنظمة المدربة باستخدام أساليب الضبط الدقيق التقليدية المعتمدة على التعلم الخاضع للإشراف. علاوةً على ذلك، يبرز ربط مخطط قاعدة البيانات باعتباره التحدي الرئيسي غير المحلول في جميع الأنظمة المدروسة، حيث تشير التقديرات الأولية إلى إمكانية تحقيق تحسن يقارب خمس نقاط مئوية في دقة التنفيذ عند معالجة هذه المشكلة بفعالية [1]. وعلى الرغم من أن أفضل الأنظمة أداءً حقق دقة تنفيذ بلغت حوالي 82%، فإن فجوة ملحوظة لا تزال قائمة مقارنةً بالأداء البشري المقدّر، وتبلغ نحو إحدى عشرة نقطة مئوية. وتشير هذه النتائج إلى أن تطوير واجهات لغة طبيعية قوية وقابلة للتعميم للتفاعل مع قواعد البيانات المعقدة لا يزال يمثل تحدياً بحثياً مفتوحاً.

الكلمات المفتاحية: تحويل اللغة الطبيعية إلى SQL، النماذج اللغوية الكبيرة، معيار BIRD، ربط مخطط قاعدة البيانات، التعلم المعزز، التوسع وقت الاستدلال، دقة التنفيذ، واجهات اللغة الطبيعية لقواعد البيانات.

Résumé

Les grands modèles de langage (Large Language Models, LLMs) sont désormais largement intégrés aux systèmes Text-to-SQL, permettant la traduction de requêtes exprimées en langage naturel en requêtes SQL exécutables sur des bases de données relationnelles. Malgré les progrès significatifs réalisés dans ce domaine, un écart de performance persistant subsiste entre les approches automatisées actuelles et les capacités démontrées par des experts humains dans des environnements réalistes d’interrogation de bases de données.

Cette étude propose une analyse comparative systématique de huit systèmes Text-to-SQL de l’état de l’art, évalués sur le benchmark BIRD, conçu pour représenter des scénarios complexes et réalistes de raisonnement sur bases de données et de génération de requêtes SQL. Les systèmes étudiés sont analysés selon plusieurs dimensions, notamment leurs paradigmes architecturaux, leurs performances sur benchmark, leur efficacité computationnelle, ainsi que leurs stratégies d’entraînement et d’inférence.

Les résultats expérimentaux montrent que les gains de performance sont principalement attribuables aux choix d’architecture système et aux stratégies d’orchestration, plutôt qu’à l’augmentation de la taille des modèles à elle seule. Par ailleurs, les approches basées sur l’apprentissage par renforcement présentent une meilleure capacité de généralisation compositionnelle que les méthodes classiques de fine-tuning supervisé.

En outre, l’ancrage au schéma (schema linking) apparaît comme un défi fondamental non résolu dans l’ensemble des systèmes étudiés. Des estimations préliminaires suggèrent que son amélioration pourrait conduire à un gain d’environ cinq points de pourcentage en exact match execution accuracy [1]. Bien que le système le plus performant atteigne une exactitude d’exécution de 81,67 %, un écart significatif subsiste par rapport à la performance humaine estimée, de l’ordre d’environ onze points de pourcentage. Ces résultats soulignent que la conception d’interfaces en langage naturel robustes et généralisables pour des systèmes de bases de données complexes demeure un problème de recherche ouvert.

Mots-clés : Text-to-SQL, grands modèles de langage, benchmark BIRD, ancrage de schéma, apprentissage par renforcement, passage à l’échelle lors de l’inférence, exactitude d’exécution, interfaces en langage naturel pour bases de données.