Computer Science and Application 计算机科学与应用, 2025, 15(7), 1-8
Published Online July 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.157174

Hans汉斯

基于双向增强和多阶监督的Text2SQL训练语料生成

黄　浩

中国电子科技集团公司第十研究所，四川 成都

收稿日期：2025年6月4日；录用日期：2025年7月4日；发布日期：2025年7月14日

摘　要

针对Text2SQL任务中训练语料人工标注成本高、场景覆盖有限的问题，本文提出一种基于双向增强与多阶监督的语料生成框架。该方法通过问题到SQL的正向增强与SQL到问题的逆向增强构建双向数据流，结合大语言模型的上下文理解与代码生成能力，创新性地引入四阶段监督审查机制(提问多样性扩充、提问质量审查、SQL自动生成、生成质量审查)，极大地提高了低资源条件下训练语料生成的效率与质量。实验表明，该方法生成的语料所训练出来的模型执行准确率相较于传统人工标注语料微调模型提升了16.3%，相较于少样本提示学习方法提升了35.7%。其次，在语料的泛化迁移性方面，本文方法生成的语料对模型尺寸大小和提问难易程度的适应性都高于人工少量标注方式。

关键词

双向增强，多阶监督，Text2SQL，训练语料生成，低语言学习

Text2SQL Training Corpus Generation Based on Bidirectional Enhancement and Multi-Stage Supervision

Hao Huang

The 10th Research Institute of China Electronics Technology Group Corporation, Chengdu Sichuan

Received: Jun. 4th, 2025; accepted: Jul. 4th, 2025; published: Jul. 14th, 2025

Abstract

To address the challenges of high annotation costs and limited scenario coverage in Text2SQL training corpus construction, this paper proposes a corpus generation framework based on bidirectional enhancement and multi-stage supervision. The method constructs a bidirectional data flow through

文章引用: 黄浩. 基于双向增强和多阶监督的 Text2SQL 训练语料生成[J]. 计算机科学与应用, 2025, 15(7): 1-8.
DOI: 10.12677/csa.2025.157174

question-to-SQL forward enhancement and SQL-to-question reverse enhancement, combines the contextual understanding and code generation capabilities of large language models (LLMs), and innovatively introduces a novel four-stage supervision and verification mechanism (question diversity expansion, question quality verification, SQL auto-generation, and generation quality verification), significantly improving the efficiency and quality of corpus generation under low-resource conditions. Experiments demonstrate that models trained with this generated corpus achieve a 16.3% improvement in execution accuracy compared to models fine-tuned with traditional human-annotated corpora and a 35.7% improvement over few-shot prompt learning methods. Furthermore, in terms of the generalization and transferability of the corpus, the corpus generated by this paper's method is more adaptable to both model size and question difficulty levels than the manually annotated small-scale approach.

Keywords

Bidirectional Enhancement, Multi-Stage Supervision, Text2SQL, Training Corpus Generation, Low-Resource Learning

1. 引言

Text2SQL 技术作为自然语言处理与数据库系统的核心接口，在面向企业级数据库应用的实证研究中，已展现出显著的规模化应用潜力[1]。行业基准测试[2]表明，典型企业数据库交互场景下，约 35.2%的查询请求可通过 Text2SQL 技术实现端到端自动化处理。特别地，该技术在高复杂度查询场景(如多表 JOIN、嵌套子查询)中的经济价值尤为突出：在金融风控领域，基于大语言模型的 SQL 生成系统可减少 78%的人工 SQL 编写耗时；在医疗数据分析场景中，Text2SQL 技术使非专业人员的复杂数据检索准确率提升至 89.2% [3]。然而，现有技术在实际落地中面临两大瓶颈：数据稀缺性与领域泛化性。

主流数据集如 WikiSQL [4]仅覆盖单表简单查询，而 Spider [2]虽包含 10,181 个跨领域样本，但其复杂嵌套查询(含 GROUP BY/HAVING/JOIN)占比不足 18%。人工标注成本方面，测算显示，专业标注员构建高质量“问题-SQL”对的边际成本高达$7.2/对，且错误率随 SQL 复杂度呈指数增长(如 5 层嵌套查询的标注错误率达 41.3%)。尽管近期 LLM 驱动的方法(如 C3 [5]、DAIL-SQL [6])在零样本场景取得突破，但其性能仍受限于训练语料的领域狭窄性——例如在 BIRD 基准[7]中，DAIL-SQL 的执行准确率(EX)较 Spider 下降 19.8 个百分点，暴露显著的领域迁移瓶颈。

针对上述挑战，本文提出双向增强范式与多阶监督机制的创新融合。与传统的单方向数据增强不同，本方法通过问题→SQL 的正向生成与 SQL→问题的逆向补全构建数据闭环，并引入四阶段质量审查管道。实验表明，该方法生成的语料所训练出来的模型执行准确率相较于传统人工标注语料微调模型提升了 16.3%，相较于少样本提示学习方法提升了 35.7%。

2. 相关工作

2.1. Text2SQL 数据增强技术

现有数据增强方法可分为三类：1) 基于规则模板的方法：WikiSQL [4]采用固定句式替换实体，但受限于模板覆盖度(仅能生成 17%的嵌套查询)；2) 弱监督生成方法：IRNet 通过 SQL 骨架解析生成候选问

题，但依赖精确的语法树对齐，错误传播率达 28%；3) LLM 驱动方法：C3 [5]利用 ChatGPT 生成候选 SQL，但缺乏逆向验证机制，导致语义一致性不足(SBERT 相似度均值仅 0.63)。相比而言，本文方法通过双向数据流设计，在增大标注语料数量的同时，也提高了其多样性。

2.2. Text2SQL 模型架构演进

Text2SQL 模型发展可分为三代：

1、Seq2Seq 基础架构：早期工作如 SQLNet [8]采用注意力机制编码数据库 schema，但在复杂 JOIN 预测上准确率不足 50%；

2、结构感知模型：RAT-SQL [9]引入关系感知 Transformer，将 Spider 开发集 EX 提升至 65.6%，但其需要全量标注数据支持；

3、LLM 时代方法：1) 模型微调：DAIL-SQL [6]通过动态示例选择实现 86.6%的 EX，但每个查询需消耗 12.7 k tokens；2) 任务分解：DIN-SQL [10]采用任务分解策略，将准确率提升至 85.3%，但对 BIRD 等工业基准的适应性差(EX 仅 55.9%)；3) 零样本/少样本：C3 [5]通过提示工程实现零样本学习，但其在低频率查询类型(如 WITH 子句)上失败率达 73%。

本工作的核心价值在于为上述模型提供高质量、跨领域的训练语料。例如，使用本文生成语料所训练出来的模型执行准确率相较于传统人工标注语料微调模型提升了 16.3%，相较于少样本提示学习方法提升了 35.7%。其次，在语料的泛化迁移性方面，本文方法生成的语料对模型尺寸大小和提问难易程度的适应性都高于人工少量标注方式。

3. 本文方法

针对 Text2SQL 任务中训练语料人工标注成本高、场景覆盖有限的问题，本文提出了基于双向增强和多阶监督的 Text2SQL 训练语料生成方法，从问题到 SQL 和 SQL 到文本两个方向，结合大模型在数据准备、数据扩充、质量审核等多个阶段的监督审查机制，以较小的注释成本自动生成大量可靠且通用的“问题-SQL”对，显著提高了语料生成的效率和质量，技术方案如下。

3.1. 问题到 SQL 的多阶段增强

图 1 展示了问题到 SQL 的多阶段增强流程，其从查询问题端出发，经过问题扩充、质量审查、SQL 自动生成、SQL 错误纠正四个阶段的增强处理，得到最终结果。

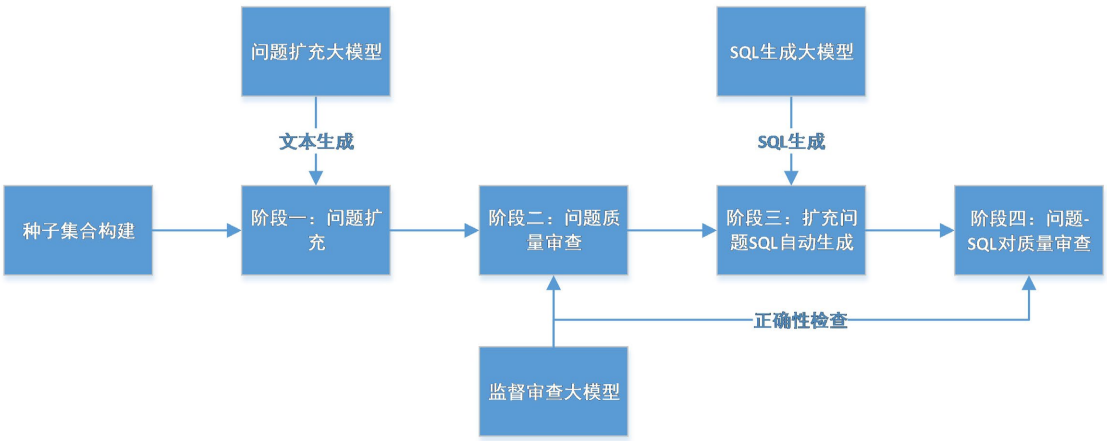

Figure 1. Question-to-SQL multi-stage enhancement process
图 1. 问题到 SQL 的多阶段增强过程

1) 种子集合构建

首先，从用户处收集真实的自然语言问题，并人工标注对应的 SQL 语句，形成高质量的种子集合。

2) 多阶段监督审查增强

第一阶段：利用大语言模型的上下文理解能力，以种子集合为正向示例，结合数据库表结构和字段信息，生成多样化的问题集合。

第二阶段：大模型担任监督角色，对生成的问题进行质量审查，过滤低质量和无法回答的问题，修正表达不清的问题。

第三阶段：基于大模型的代码生成能力，结合种子集合中的“问题-SQL”对，自动为问题清单中的每个问题生成 SQL 语句。

第四阶段：利用大模型的 SQL 语法知识和数据库表设计信息，对生成的“问题-SQL”对进行正确性审查，确保语料的准确性。

3.2. SQL 到问题的多阶段增强

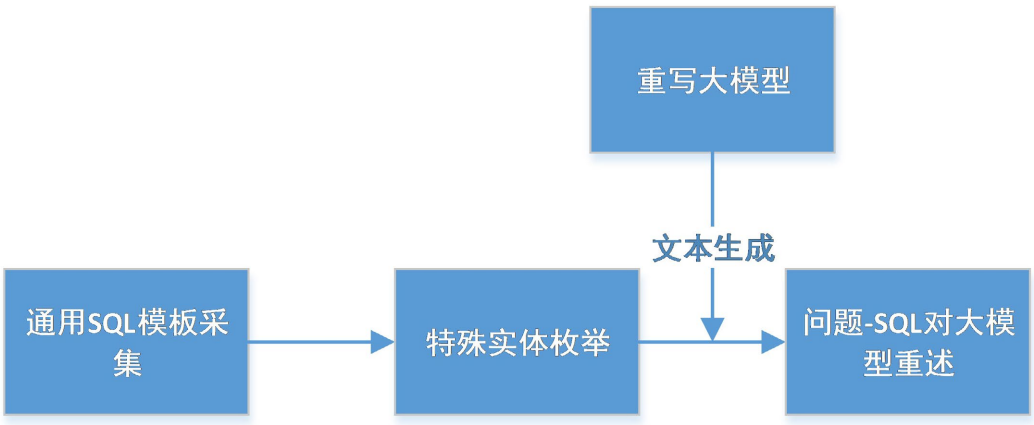

Figure 2. SQL-to-question multi-stage enhancement process
图 2. SQL 到问题的多阶段增强过程

如图 2 所示，SQL 到问题的多阶段增强包含通用模板生成、特殊实体枚举和大模型重述三个步骤：

1) 种子集合构建

采用公开数据集 Spider 中的常见 SQL 模板，结合数据库表结构和字段，生成通用的“问题-SQL”对。

2) 多阶段监督审查增强

针对库表中的特殊实体，枚举其可能的取值，并反向代入到问题模板中，生成多样化的问题。

3) 大模型重述

使用大语言模型对生成的问题进行自然语言重述，确保问题符合中文语法习惯，同时保持原意不变。

4. 实验

4.1. 实验方法

为了测试本文方法的有效性，选取了公认数据集 Spider。Spider 是一个大规模的跨域数据集。由 10,181 个问题-SQL 对组成，涉及 200 多个数据库，共分为四个难度级别，即简单、中等、困难和超困难。实验方法如图 3 所示。

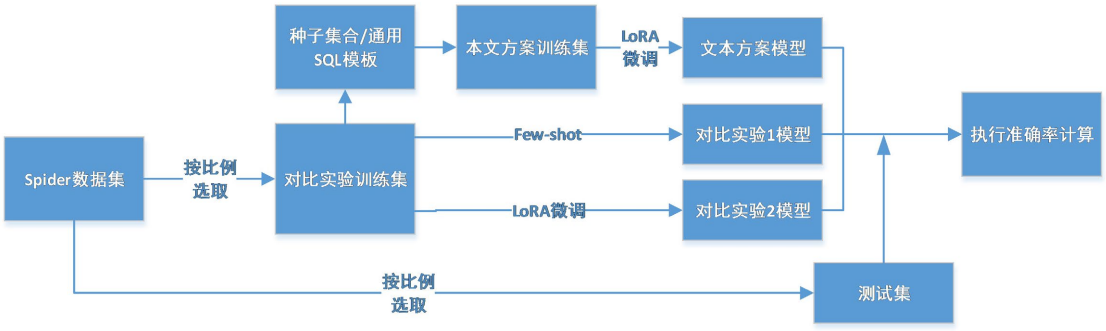

Figure 3. Experimental method process
图 3. 实验方法过程

第一步：训练集和测试集选取。我们按照一定比例从四个难度级别中各随机选取一些问答对，作为大模型训练的训练数据集。同样的方法，获取另一组问答对，作为测试数据集。

第二步：对比实验 1 设置，基于 few-shot 的大模型 SQL 生成。当预测问题答案时，将训练集作为示例样本动态加载于其提示词上下文中。动态样本选取方式采用文本相似度召回算法，即选择与问题文本语义最相似的训练样本。

第三步：对比实验 2 设置，基于 Lora 微调后大模型的 SQL 生成。基于 Lora 算法，使用训练数据集，微调得到基线模型。

第四步：本文技术方案。选取训练集中的自然语言问题作为“问题到 SQL 的多阶段增强”中的种子集合，选取训练集中 SQL 语句，将其提炼为“SQL 到问题的多阶段增强”中的通用 SQL 模板。然后完成训练集合的生成和扩充。同样地，基于 Lora 算法微调得到基线模型。

第五步：评价指标计算。本文选取执行精度(Execution Accuracy, EX)作为评价模型生成 SQL 准确率的方式。执行精度指的是生成的 SQL 查询在数据库上运行的结果与标准答案运行结果相一致。根据生成的 SQL 查询能否正确执行且返回预期结果的比例来计算执行准确率，见计算公式(1)：

$$LA = \text{正确执行的查询数量/总测试查询数量} \tag{1}$$

4.2. 实验设置

4.2.1. 共性实验设置

训练集数量：100，测试集数量：100，Few-shot 示例数量：3 个。

微调大模型：为评测技术方案的泛化性，采用三组模型进行微调，包括 qwen3-8b、qwen3-4b 和 qwen3-1.7b。

LoRA 微调参数配置如表 1 所示。

Table 1. LoRA parameter setup
表 1. LoRA 参数设置

参数名称	描述	设置值
lora_rank	低秩矩阵的秩	16
lora_alpha	缩放因子	32
学习率	模型每次更新的幅度	$3e^{-6}$
批次大小	每次训练中处理的样本数量	8
优化器	用于模型参数更新的算法	AdamW

4.2.2. 本文方案设置

1) 问题扩充大模型

模型选取：qwen3-8b。

提示词设置：你是一个 Text2SQL 专家，你的任务是参考“问题种子集合”，结合“数据库表设计信息”，请按下面的步骤一步步思考：1、从 when、where、which、where、how、why 等角度丰富提问的角度；2、仔细阅读数据库表设计，从单表、多表联查的角度给出可能的问题；3、模拟种子清单中的提问方式或习惯，使问题更加贴合中文表达习惯。尽可能从多个维度提出更多用户可能的问题。返回格式要求：要求以字符串列表返回，每个元素为一个问题字符串。问题种子集合为：xxx。数据库表设计信息为：xxxx。

2) 监督审查大模型

模型选取：qwen3-32b。

提示词设置：你是一个提问质量审核专家，你的任务是参考“数据库表设计信息”，对“问题清单”中的问题进行质量审查。请按下面的步骤一步步思考：1、现有数据库无法回答的问题，标记为过滤；2、问题表述不符合中文习惯，标记为修正，并给出修正后问题；3、无任何意义的问题标记为过滤。返回格式要求：要求以 json 格式返回，key 为问题序号，问题类型，改写后问题。问题清单为：xxx。数据库表设计信息为：xxxx。

3) SQL 生成大模型

模型选取：qwen3-32b。

提示词设置：你是一个 Text2SQL 专家，你的任务是参考“问题种子集合”，结合“数据库表设计信息”，给出“问题清单”中每个问题对应的 SQL 语句。返回格式要求：要求以 json 格式返回，key 为问题序号，问题对应 SQL。问题种子集合为：xxx。数据库表设计信息为：xxxx。问题清单为：xxx。

4) 问题重写大模型

模型选取：qwen3-8b。

提示词设置：你是一个 Text2SQL 专家，你的任务是参考“数据库表设计信息”，对“问题-SQL 对”清单进行质量审查。请按下面的步骤一步步思考：1、如果 SQL 语句正确，则跳过；2、如果 SQL 语句错误，标记为错误，并给出修正后的 SQL；返回格式要求：要求以 json 格式返回，key 为“问题-SQL 对”序号，改写后 SQL。“问题-SQL 对”清单为：xxx。数据库表设计信息为：xxxx。

4.3. 实验结果与分析

本文设置了两组对比实验，对比实验 1 为基于提示工程的 few-shot 方式，作为基线，对比实验 2 为基于原始语料集数据进行 LoRA 微调。三组实验分别在 qwen3-8b、qwen3-4b 和 qwen3-1.7b 三组不同尺寸的基座模型上运行。实验测试数据集由简单、中等、困难、超困难 4 个等级组成。

4.3.1. 总体实验结果分析

从表 2 可以看出，相比于少样本的提示工程方法，大模型微调方案准确率有所提升。而本文基于双向增强和多阶段监督方法生成的语料，其微调后模型的 SQL 生成准确率要比仅采用少量人工标注数据进行高效参数微调的大模型表现提高 16.3%。

Table 2. Overall experimental results
表 2. 总体实验结果

实验方法	Few-shot (基线)	原始人工语料 LoRA	本文方法
准确率	42%	49%	57%
相比基线提升百分比	/	16.7%	35.7%

4.3.2. 不同基础模型参数量对实验结果的影响

从图 4 可以看出，模型微调后的整体准确率随着基座模型的增大而提高。另外，对于三种不同尺寸的模型基座，本文方法生成的语料微调效果均高于少量人工标注语料，一定程度反映本文方法的泛化性较好。

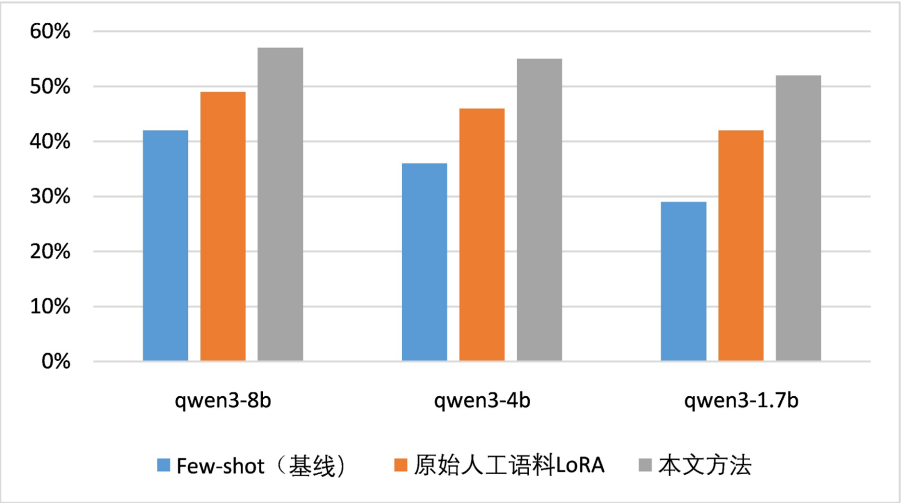

Figure 4. Performance of different size base model after finetune
图 4. 不同尺寸基座模型微调后表现

4.3.3. 不同困难程度的测试集对实验结果的影响

从图 5 可以看出，由本文方法合成语料微调得到的模型在困难及超困难测试集上的表现均好于对比试验，进一步说明针对复杂场景的 SQL 生成，训练语料的数量、多样性很重要。

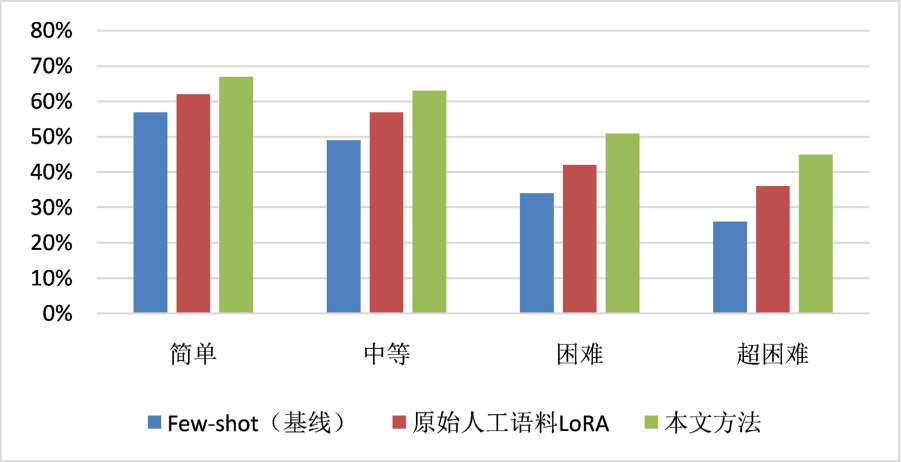

Figure 5. Performance of test sets with different levels of difficulty
图 5. 不同困难程度测试集上的表现

5. 结论与展望

本文针对 Text2SQL 任务中训练语料人工标注成本高、场景覆盖不足的问题，提出基于双向增强与多阶监督的语料生成框架。通过问题→SQL 的正向增强与 SQL→问题的逆向增强构建双向数据流，结合大

语言模型的上下文理解与代码生成能力，实现了高效、高质量的语料生成。实验表明，基于本文方法生成的语料训练的模型，其执行准确率(EX)达到 57%，较传统人工标注语料微调模型提升 16.3%，较少样本提示学习方法提升 35.7%。且在复杂查询场景(困难/超困难级别)中，准确率优势进一步扩大，验证了生成语料对高阶逻辑的覆盖能力。泛化迁移性突出：实验显示，在三种不同参数量级(8B/4B/1.7B)的基座模型上，本文方法均保持稳定的性能增益。这表明生成语料对模型尺寸和领域分布具有强适应性，突破了传统语料对专业标注的强依赖。

尽管本文方法在语料生成效率与模型性能上取得显著进展，仍存在以下改进方向：1) 跨模型协同验证：现有方法基于单一 LLM 进行双向增强，后续可通过多模型协同生成与验证，降低模型特异性偏差风险。2) 领域自适应增强：针对垂直领域(如金融、医疗)的复杂查询需求，需探索基于模式图嵌入(Schema Graph Embedding)的领域知识注入方法，提升嵌套子句(如 WITH、HAVING)的生成准确率。

参考文献

[1] Deng, N., Chen, Y. and Zhang, Y. (2022) Recent Advances in Text-to-SQL: A Survey of What We Have and What We Expect. arXiv: 2208.10099.

[2] Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task. https://github.com/taoyds/spider

[3] Marshan, A., Almutairi, A.N., Ioannou, A., Bell, D., Monaghan, A. and Arzoky, M. (2024) Medt5sql: A Transformers-Based Large Language Model for Text-to-SQL Conversion in the Healthcare Domain. *Frontiers in Big Data*, **7**, Article 1371680. https://doi.org/10.3389/fdata.2024.1371680

[4] Zhong, V., Xiong, C. and Socher, R. (2017) Seq2SQL: Generating Structured Queries from Natural Language Using Reinforcement Learning. arXiv: 1709.00103.

[5] Dong, X., Zhang, C., Ge, Y., Mao, Y., Gao, Y., Lin, J., Lou, D., *et al*. (2023) C3: Zero-Shot Text-to-SQL with ChatGPT. arXiv: 2307.07306.

[6] Gao, D., Wang, H., Li, Y., Sun, X., Qian, Y., Ding, B. and Zhou, J. (2023) Text-to-SQL Empowered by Large Language Models: A Benchmark Evaluation. arXiv: 2308.15363.

[7] Li, Y., Guo, J., Yu, W., *et al*. (2023) BIRD: A New Benchmark for Cross-Domain Text-to-SQL Generation. ACL.

[8] Xu, X., Liu, C., Song, D., Zhang, Y., Shah, A., Tian, Y. and Salakhutdinov, R. (2017) SQLNet: Generating Structured Queries from Natural Language Without Reinforcement Learning. ACL.

[9] Wang, B., Shin, R., Liu, X., Polozov, O. and Richardson, M. (2020) RAT-SQL: Relation-Aware Schema Encoding and Linking for Text-To-SQL Parsers. In: Jurafsky, D., Chai, J., Schluter, N. and Tetreault, J., Eds., *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 7567-7578. https://doi.org/10.18653/v1/2020.acl-main.677

[10] Pourreza, M. and Rafiei, D. (2024) Din-SQL: Decomposed in-Context Learning of Text-to-SQL with Self-Correction. 37*th Conference on Neural Information Processing Systems* (*NeurIPS* 2023), New Orleans, 10-16 December 2023, 1-34.